WHO Indoor Air Quality Guidelines: household fuel combustion

Methods used for evidence assessment

Prepared by: Nigel Bruce[1,2], Annette Pruss-Ustun[1], Dan Pope[2], Heather Adair-Rohani[1], Eva Rehfuess[3]

Affiliations:
[1]World Health Organization, Public Health, Social and Environment Determinants of Health, Geneva, Switzerland
[2]Department of Public Health and Policy, University of Liverpool, UK
[3]Institute for Medical Informatics, Biometry and Epidemiology, University of Munich, Germany

Disclaimer:

The work presented in this technical paper for the WHO indoor air quality guidelines: household fuel combustion has been carried out by the listed authors, in accordance with the procedures for evidence review meeting the requirements of the Guidelines Review Committee of the World Health Organization.

Full details of these procedures are described in the Guidelines, available at: http://www.who.int/indoorair/guidelines/hhfc ; these include declarations by the authors that they have no actual or potential competing financial interests. The review was conducted in order to inform the development of recommendations by the Guidelines Development Group. Some of the authors are staff members of, or consultants to, the WHO. The authors alone are responsible for the views expressed in this publication, which do not necessarily represent the views, decisions, or policies of the WHO.

This article should not be reproduced for use in association with the promotion of commercial products, services or any legal entity. The WHO does not endorse any specific organization or products. Any reproduction of this article cannot include the use of the WHO logo.

Contents

Summary

Background

The interventions that are required to mitigate health and other adverse consequences of current global patterns of household fuel combustion are 'complex', that is, they require not only effective technologies and cleaner fuels, but also action by multiple stakeholders across society in order to ensure equitable and lasting adoption. The development of recommendations to address these issues therefore requires drawing on a wide range of evidence including population studies of fuel use and exposure, laboratory emissions data, epidemiological studies of exposure and health outcomes risk, intervention impacts studies, qualitative evidence on user perceptions about change, and policy analysis. These sources of evidence use very disparate methods and research paradigms, and randomised trials – the gold standard of evidence of effectiveness – are few due partly to the practical difficulties of conducting these, but also as their relevance in evaluating complex interventions can be limited.

Objectives

The objectives of this paper are to:

1. Document the types of evidence required for the guidelines;
2. Discuss the strengths and limitations of GRADE for evaluating the quality of this evidence and determining the strength of recommendations;
3. Describe the rationale and approach of the revised methodology used to address such limitations as were identified, and;
4. Describe how the revised methods were applied to the current guidelines.

Methods and findings

The standard method of assessing quality and strength of evidence for the purposes of WHO guideline recommendations known as GRADE provides a valuable framework for assessing quality of bodies of evidence, and for the step of moving from evidence to recommendations. This system does not easily allow for the assessment of all of the evidence sources relevant to this topic, and tends to assess much of what is available as low or very low quality. In the judgment of the Guidelines Development Group, this was felt to undervalue the contribution of this evidence to formulating recommendations. In addition, GRADE does not include assessment of the degree of consistency between the various components of evidence making up the 'causal chain' that relates interventions ultimately to health outcomes.

Modifications to GRADE have been developed in order to address these issues. This revised approach, termed 'grading of evidence for public health interventions' (GEPHI), is described with an explanation of how each stage in the process has been applied to the systematic reviews, and other sources of evidence including summaries of laboratory emission studies and models. In summary, the revisions include:

- Entering non-randomized experimental studies into the grading table as moderate quality;
- Allowing upgrading for (i) consistency of findings across different settings and/or study designs, and (ii) analogous evidence that supports the findings, for example from other sources of combustion pollution;
- Using GRADE domains to guide quality assessment of evidence not amenable to systematic review and meta-analysis;
- Assessing consistency of findings of bodies of evidence relating to different links in the causal chain model.

The final stage of the process, that is, using GRADE decision tables to determine the strength of each recommendation, remains relevant and applicable to this topic and has been carried out in the standard manner, albeit termed ‘decision table for strength of recommendations’ to acknowledge that different methods (GEPHI) have been used for assessing evidence quality.

Conclusions

The GEPHI methodology accommodated the assessment of quality for all types of evidence contributing to recommendations. Most of the evidence was rated as being of moderate quality, although some was rated as low or high quality; these assessments were in line with the generally good level of consistency between components of the evidence contributing to the causal chain. It will be useful to assess the usefulness and validity of these methods when applied to other environmental and public health interventions.

1. Introduction

1.1 Evidence review for guidelines development

This paper describes the methods used in reviewing evidence for the development of recommendations which follow guidance provided by the WHO Guidelines Review Committee *(1)*, with some modifications and additional perspectives appropriate to the topic and the nature of evidence available.

As with other areas of health policy, choosing interventions that modify the environment in order to protect and promote health must be supported by systematically collected and synthesized evidence and an appreciation of the confidence in the relevant body of evidence. No scheme for grading the strength of evidence for the purposes of making recommendations on interventions is universally agreed upon, but a scheme called GRADE (Grading of Recommendations Assessment, Development and Evaluation)*(2)* is becoming increasingly popular and the preferred approach recommended for the development of WHO guidelines *(1)*.

The GRADE system provides a valuable and systematic means of assessing both the quality of bodies of evidence, and the strength of recommendations based on that evidence. Key questions and the importance of outcomes are specified at the outset. Quality of evidence relating to each key question is determined through GRADE profile tables, whereby study design, methodological strengths and weaknesses, consistency, publication bias and other issues are used to derive an overall score on a 4-point scale from high to very low quality. The strength of recommendations informed by the evidence is assessed using GRADE decision tables, which consider not only the quality of evidence, but also the balance of benefits and harms, values and preferences, and resource implications. Thus, high quality evidence does not guarantee that a recommendation should be strong, if for example, there are concerns about intervention harms, or costs are prohibitive. Conversely, weaker evidence does not necessarily preclude a strong recommendation if there are other very compelling reasons for this. Generally, however, a strong recommendation should be supported by high to moderate quality evidence.

GRADE was developed primarily for application in the field of clinical medicine, where a substantial proportion of studies are randomized trials, amenable to systematic review and meta-analysis. Recommendations in the field of public health (and including environmental health) will normally draw on an evidence base dominated by other types of study design and evidence, as randomized controlled trials may not be feasible and/or are difficult to conduct, and the limitations to the application of GRADE in this respect have been discussed in the literature *(3-8)*.

These considerations are very relevant to the current topic of household fuel combustion and the harms resulting from air pollution and related issues, for which the quality of evidence assessment methods used by GRADE are not ideally suited, and more specifically not to the range and nature of evidence that needs to be compiled for developing effective recommendations. The key issues are that:

1. GRADE does not discriminate between non-randomized experimental studies and observational designs (with no investigator-led intervention) such as cross sectional, cohort and case-control, whereas in the context of household energy interventions (at least), the former group can provide higher quality evidence;
2. The criteria within GRADE to upgrade observational studies do not allow the expression of increased confidence in the evidence where (i) consistency in findings across settings, study designs and research groups is observed; and (ii) analogous evidence is available. An example of the second points is the contribution that

evidence from smoking (both active and second hand) can make to the understanding of health risk of using wood and other biomass fuels in the home, based on the fact that all three of these sources expose individuals to similar combustion mixtures.

3. GRADE does not accommodate the potential contributions of alternative sources of evidence such as laboratory, mechanistic or animal studies, the principles of other disciplines (e.g. physiology, engineering, toxicology, chemistry, physics), and other research paradigms, for example qualitative research.

Environmental health interventions, together with other complex public health interventions such as in the area of nutrition, typically also draw heavily on non-epidemiological evidence for their effectiveness to be judged. The imperative of thorough, transparent assessment of evidence however remains, if recommendations are to have validity. The objectives of this overview of the evidence review methods used for the guidelines are therefore to describe and discuss:

1. The types of evidence available for developing recommendations for improved air quality in respect of household fuel combustion.
2. The rationale and application of the concept of a causal chain, as a framework for linking these various types of evidence.
3. The applicability of GRADE methods and the rationale for the modifications used.
4. The specific revisions made, and how these have been applied in developing the guidelines.

1.2 Types of evidence contributing to the development of recommendations

Interventions in the field of household energy, as well as those for environmental and public health more generally, are characterized as ‘complex interventions’, where multiple components (e.g. technology, behaviour) implemented at multiple levels (e.g. community, household, individual level) and by multiple sectors (e.g. energy, environment, health, education) interact to bring about a range of short-term and long-term health and non-health benefits *(9, 10)*.

As a consequence, assessment of the impact of interventions for reducing the adverse health impacts of household fuel combustion, together with an understanding of the most effective ways to deliver them, can draw on a wide range of evidence, including:

- **Population-based surveys** and **cross-sectional studies** describing the extent and distribution (e.g. in relation to poverty) of types of household fuels and associated technologies, and levels of household air pollution (HAP) and personal exposure.
- **Laboratory-based testing** of the performance of combustion technologies, including rates of emission of health-damaging pollutants, and safety (i.e. from burns and scalds).
- **Modelling,** including (i) linking emission rates with area pollution concentrations, and (ii) exposure-response functions combining exposure and risk data from multiple sources of combustion pollution.
- **Epidemiological studies** of the links between fuel use, HAP and a range of disease and safety outcomes, the majority of which to date have been observational.
- Experimental studies, including **randomized controlled trials**, **controlled and uncontrolled before-and-after** and designs, used to measure the impact of improved stoves and cleaner fuels on HAP and exposure, and – if available – on health and safety outcomes.
- A range of study designs, including **qualitative, quantitative and policy/case studies** which provide evidence on factors enabling and limiting adoption of interventions.

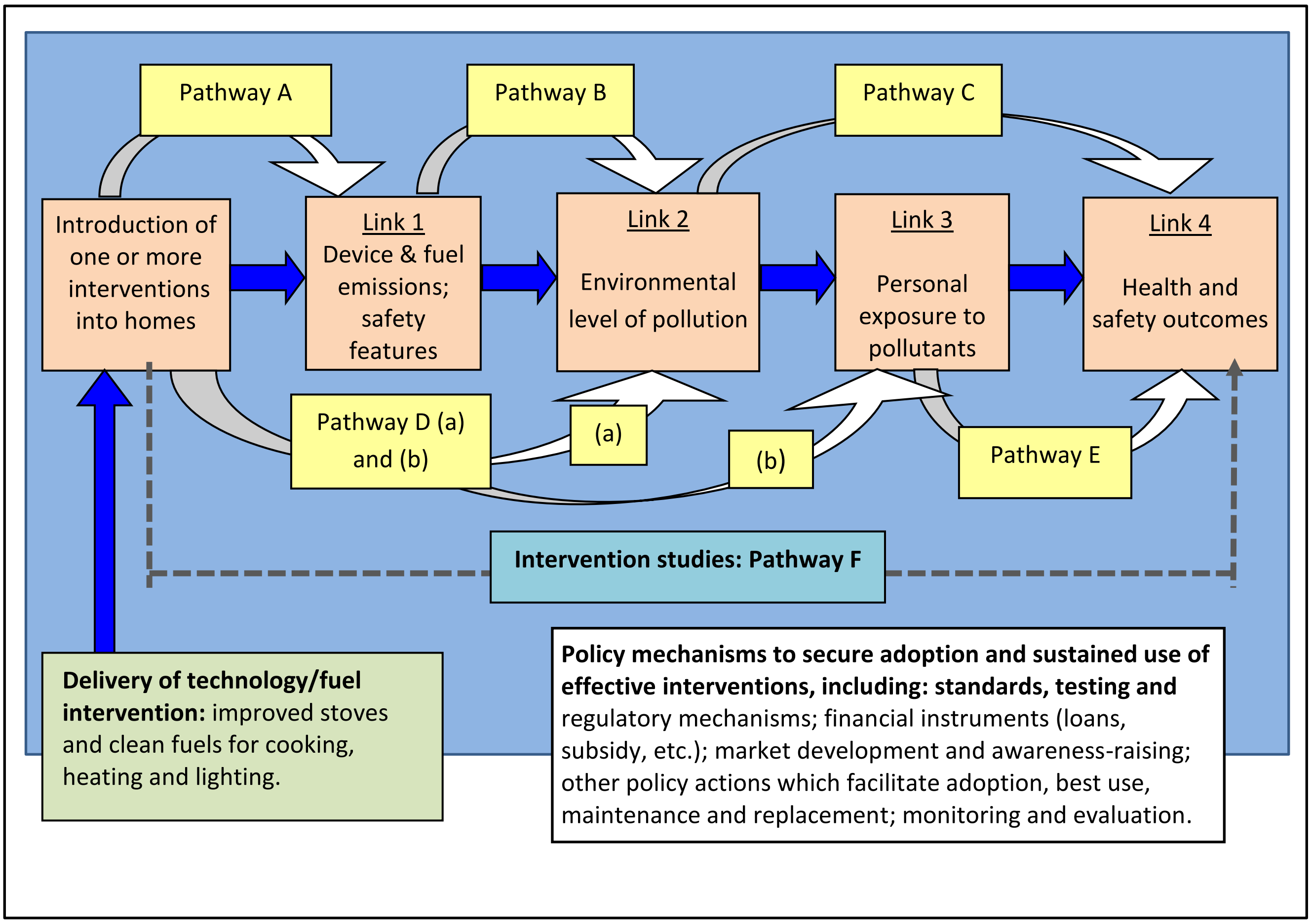
- **Economic evaluation studies** providing evidence on the cost-effectiveness and benefit to cost ratios for interventions.

The evidence review methods used for the guidelines have been developed to allow the evaluation and incorporation of all of the above types of evidence into the recommendations.

1.3 The causal chain: rationale and value for defining key questions for evidence review

In order to address this complexity, both in respect of the nature of the interventions, and well as the range of evidence informing recommendations, it is useful to understand the effects of these interventions through a "causal chain" (see Figure 1). Using this approach, evidence informing sequential and multiple links in the chain can be evaluated, and the overall consistency of evidence relating interventions to health outcomes can be assessed.

Figure 1: Causal chain relating household energy technology, fuel and other interventions to health and safety outcomes via intermediate links


Pathway A
Pathway B
Pathway C
Introduction of one or more interventions into homes
Link 1
Device & fuel emissions; safety features
Link 2
Environmental level of pollution
Link 3
Personal exposure to pollutants
Link 4
Health and safety outcomes
Pathway D (a) and (b)
(a)
(b)
Pathway E
Intervention studies: Pathway F
Delivery of technology/fuel intervention: improved stoves and clean fuels for cooking, heating and lighting.
Policy mechanisms to secure adoption and sustained use of effective interventions, including: standards, testing and regulatory mechanisms; financial instruments (loans, subsidy, etc.); market development and awareness-raising; other policy actions which facilitate adoption, best use, maintenance and replacement; monitoring and evaluation.

The focus of this causal pathway is the source of the combustion emissions (for cooking, heating, lighting and other purposes in the home), since reduction of emissions is the most critical underlying factor for measures aimed at achieving the air quality guidelines.

It is recognized, however, that other aspects of the home environment (for example, ventilation through windows and eaves) and behaviour (in how the stove is used, time spent by individuals in various micro-environments in and around the home) also play a part on total dose of air pollutants and hence health effects, and these impact on the causal chain at varying points. Insofar as available evidence allows, these are also considered. Examples of the factors that can be assessed at each stage of the causal chain are shown in Table 1, and the ways in which the different types of evidence described in Section 2 above provide information on different components of the causal chain, illustrated by the 'pathways' shown in Figure 1, and elaborated in Table 2.

Table 1: Examples of factors that may be assessed at each link:

Interventions	Emissions and safety features	Environmental level	Personal exposure	Health and safety outcomes
• Improved solid fuel stoves • Clean fuels and associated technologies for cooking, heating and lighting	• Emission rates of toxic pollutants directly into homes • Emissions (e.g. via flue) to ambient air • Inherent safety (e.g. stability, enclosed flame, raised surface)	• Concentrations of pollutants in kitchen and other areas of the home. • Concentrations of pollutants in ambient air that can enter the home.	• Exposure of children, women and men to pollutants • Function of time spent in various micro-environments • Exposure to open flame or risk of falling pots with hot liquids	• Range of child and adult health outcomes from exposure to pollutants (ALRI, COPD, lung cancer, CVD, etc.) • Safety outcomes (e.g. burns, scalds) • Other health and socio-economic impacts

Table 2: Examples of the types of evidence providing information on pathways in the causal chain

Pathway	Type of evidence	Explanation
A	Laboratory emissions testing	Provides information on the rates of emissions of toxic pollutants, for example in relation to a unit of energy delivered.
B	Emissions model	Relates emission rates to predicted concentrations in the home, based on assumptions about duration of use, air exchange rates and kitchen volume.
C	Epidemiological studies	Investigate the risk of a range of disease outcomes among those using more polluting fuels compared to groups with lower exposure, for example using clean fuels; such studies may or may not include measurement of HAP and/or exposure.
D (a,b)	Experimental studies	Randomized and non-randomized experimental studies that measure the impact of introducing an in improved stove or clean fuel on pollutant concentrations or personal exposure.
E	Epidemiological studies	Studies which include exposure assessment may allow investigation of the relationship between exposure and disease risk.
F	Experimental and observational studies	Randomized and non-randomized experimental, and some observational, studies that investigate the impact of an improved stove or clean fuel directly on risk of health outcomes

In addition, and not included in the illustration of pathways in Figure 1 is evidence on factors influencing effective and equitable adoption of improved technologies, cleaner fuels and other interventions, as well as maintenance and replacement; these are indicated in the box in Figure 1, and reported in full in Review 7 (Factors Influencing Adoption).

Reducing the complexity of environmental health interventions by picking out a single link in the causal chain (e.g. by only considering an impact of exposure reductions on a health outcome under highly controlled circumstances) may under- or overestimate actual effectiveness in the field *(10)*.

Also, it is now recognized by organizations such as the Public Health group of the UK National Institute for Health and Care Excellence that understanding and making recommendations about public health interventions should not only seek to answer the question “what works”, but also “how it works, and for whom, in which circumstances”*(11, 12)*. Not least, this is critical in efforts to distinguish between the failure of a concept (e.g. an improved cooking stove) and implementation failure (e.g. inappropriate manufacture and maintenance, failure to fully address user needs, or inappropriate use of a subsidy that discourages a sense of ownership).

The foregoing issues are taken into consideration in formulating the evidence review scoping questions for the new guidelines. These questions are elaborated in Section 4.1.

1.4 Applicability of GRADE to evaluation of evidence for IAQ guidelines

In light of the importance of contributions from the different types of evidence discussed in Sections 1.2 and 1.3 above, four main issues arose in considering the applications of GRADE methods to interventions for reducing the adverse health and other impacts of household energy use:

a) **Assessment of the value of non-randomized experimental studies*:*** Actions in the field of environmental health – for practical, ethical, political and cost reasons – are rarely amenable to randomized controlled trials that directly assess the health impacts of interventions. Following the current GRADE approach, a majority of studies providing information on environmental health interventions would therefore start off as low-quality with limited opportunity for upgrading the quality of evidence, and may be further down-graded for indirectness (e.g. examining concentrations, personal exposure or short-term health impacts rather than long-term impacts such as mortality due to cancer or cardiovascular disease). The method does not distinguish between true observational designs (e.g. cohort studies, case-control studies, cross-sectional studies) and non-randomized experimental designs (before-and-after studies and quasi-experimental study designs) which involve investigator-led changes to the stove and can provide higher quality evidence (see rationale in Section 3.1).

b) **Assessment of the quality of non-epidemiological evidence*:*** As noted, effectiveness in the field of environmental health is often characterized by causal chains and consequently draws on a combination of epidemiological studies and physical or engineering principles, animal or in vitro laboratory studies and qualitative evidence. In GRADE, all studies other than standard epidemiological study designs are rated as of very low quality, or may not be included at all.

c) **Taking account of analogous evidence*:*** GRADE assessment does not provide for recognition of evidence from similar types of exposure (“analogies”), such as evidence form active and second-hand smoking in the case of biomass fuel combustion.

d) **Integration of different types of evidence that can contribute to assessing effectiveness:**

> Environmental health studies often cover only part of the causal chain. GRADE would treat each piece of the evidence relating to effectiveness on its own (which may result in multiple pieces rated low or very low), rather than allowing for an overall assessment of the insights provided through different sources of evidence, and the extent to which these are consistent.

1.5 Revised methodology: grading of evidence for public health (GEPHI)

In order to address these issues, a number of modifications to GRADE were developed, focusing on assessment of the quality of evidence. These modifications are described in detail below in Section 3, but in summary include:

- Entering non-randomized experimental studies into the grading table as moderate quality;
- Allowing upgrading for (i) consistency of findings across different settings and/or study designs, and (ii) analogous evidence that supports the findings, for example from other sources of combustion pollution;
- Using GRADE domains to guide quality assessment of evidence not amenable to systematic review and meta-analysis;
- Assessing consistency of findings of bodies of evidence relating to different links in the causal chain model.

For the assessment of the strength of a recommendation, the standard approach using the GRADE decision table (involving an assessment of the findings on overall quality of evidence, values and preferences, balance of costs and benefits, and resource implications) was applied, as this remains entirely relevant to the issues that need to be considered.

In Section 2 we present a brief overview of the GRADE classification for reference, and then describe the modifications employed in developing these guidelines in Section 3. The actual steps taken in applying these methods to the current guidelines are described in Section 4. Collectively, this modified approach is referred to in the current volume as **Grading of Evidence for Public Health Interventions**, or GEPHI.

2. Overview of GRADE evidence review classification

2.1 Quality assessment

GRADE categorizes evidence into four levels of quality: high, moderate, low and very low, defined as shown in Table 3 and determined according to the criteria in the standard GRADE Table, details of which are included in Table 4.

Table 3: GRADE: Four levels of evidence and their significance

Quality level	Definition
High (++++)	We are very confident that the true effect lies close to that of the estimate of the effect
Moderate (+++)	We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different
Low (++)	Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect
Very low (+)	We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect

Source: Balshem et al 2011 *(13)*

When entering a set of studies into the GRADE profile table, randomized trials start as 'high' quality, observational evidence (including experimental studies that are not randomized) starts as 'low' quality, while other (non-epidemiological, e.g. laboratory-based testing) would be entered as ‘very low’ – if included at all. A body of evidence can be downgraded or upgraded according to the criteria in Table 4. Further detail on the assessment of evidence can be found in the publications of the GRADE working group *(14)*.

2.2 Strength of evidence for causal inference

One critical aspect of the definitions on the strength of evidence in Table 3 is confidence in the intervention effect estimate. In this regard, it is useful to make a distinction between:

a) Strength of evidence for causal inference, for which Bradford Hill viewpoints for distinguishing causation from association in environmental epidemiology are often referred to (see Box A1) *(15)*, and;
b) The quality of evidence for the intervention effect size (confidence in the estimate), for which GRADE may be used.

Box 1: Bradford-Hill viewpoints

1. Strength of association
2. Consistency across populations, study designs, etc.
3. Specificity
4. Temporality (exposure precedes outcome)
5. Biological gradient (dose-response)
6. Biological plausibility
7. Coherence with natural history, animal studies, etc.
8. Experiment
9. Analogy

While these assessments have much in common, it is quite possible to have good evidence for causal inference in respect of an association between HAP exposure and one or more of the disease outcomes (and by implication that reducing exposure will reduce the risk of that disease), but rather lower confidence about the size of the intervention effect.

Accordingly, in reviewing evidence, separate assessments are made of (i) the extent to which causality can be inferred (by reference to the Bradford-Hill viewpoints), and (ii) confidence about estimates of the impacts that interventions can be expected to have on the various health outcomes (by application of GEPHI).

In respect of assessing the strength of evidence for causal inference in the absence of a strong body of experimental evidence, we have drawn on the discussion by Bradford Hill that none of the viewpoints he discusses, including experimental evidence, is required for concluding that an association is causal, albeit the potential importance and strength of experimental evidence in this regard is clearly recognized:

“Here then are nine different viewpoints from all of which we should study association before we cry causation. What I do not believe – and this has been suggested – that we can usefully lay down some hard-and-fast rules of evidence that must be obeyed before we can accept cause and effect. None of my nine viewpoints can bring indisputable evidence for or against the cause-and-effect hypothesis and none can be required as a sine qua non.” (15)

We refer several times to the work of Bradford Hill as this framework is useful and has stood the test of time (see for example Howick et al. 2009 *(16)*), albeit with some qualifications such as the viewpoint on specificity of effect, which has been recognized in applying these viewpoints.

The evidence base is not wholly devoid of experimental evidence, although this is in the minority. The synthesis of evidence contributing to the causal chain in Annex 4.5 includes an assessment of consistency across evidence types, and particularly, consistency with the limited experimental and other intervention-based evidence that is available.

3. Revisions made to GRADE methods

3.1 Quality of non-randomized experimental evidence

An important study design issue is the handling of evidence from non-randomised experimental studies. Although these may take different forms, in the context of evidence available to these guidelines, two main types were available:

- **Before and after** designs with no comparison group, commonly used for testing the short-term impact of a new stove or fuel on HAP, exposure or fuel consumption.
- More complex **quasi-experimental** designs with one or more comparison groups; these have been rather less commonly used but application has included testing of new stoves and fuels as above, and other outcomes such as incidence of common symptoms (cough, phlegm production, sore eyes, headaches, etc.) and burns.

GRADE treats all non-randomised experimental studies as observational study designs, and hence these enter the table as low quality evidence. This is important because these study designs do provide key evidence for the new guidelines, providing the most extensive evidence base for the effectiveness of so-called improved stoves in everyday use and a more limited assessment for clean fuels including liquefied petroleum gas (LPG), electricity and ethanol. The contribution of these studies involves treating HAP and personal exposure levels as a proxy for health risk, which has become more valid (less indirect) now that integrated exposure-response functions are available (see Review 4). Given the important contribution of this type of evidence, a clear rationale for assigning a level of quality when entered into the table was needed.

The approach taken was informed by the recognition that, while not of the same quality as randomised controlled trials (RCTs), these designs do provide evidence which is of a different nature and quality than do observational comparisons of homes that have adopted different types of cooking stoves or fuels usually of their own volition and at their own expense. The main reason for the difference in strength of evidence is that the new stove has been introduced into the same home in which the baseline assessment is made, hence avoiding much of the confounding that arises in between-home observational comparisons. Such studies are experimental in the sense that an implementing agency introduces the new technology and/or fuel, and evaluation of the impacts on HAP and/or exposure is carried out involving baseline and post-intervention measurements.

The before and after design in which, for example, $PM_{2.5}$ (a key measure of particulate air pollution) is measured in a group of homes first with the traditional stove and then with the new stove, is a weaker design than that in one or more comparison groups are used with no change to the stove. Indeed, if in the latter design, efforts are made to match comparison homes closely with those that will receive interventions, and follow-up is contemporaneous, the design is quite robust.

The revisions to GRADE take account of these non-randomised experimental study designs, given their important contribution to the available evidence; provision has been made by adding these designs into the modified table at the level of moderate evidence (between RCTs and observational evidence), Table 4. In order to allow for the potentially lower strength of evidence from the uncontrolled ‘before and after’ type of study, these studies were more likely (than controlled designs) to be downgraded if quality assessment suggested there was good reason to do so.

In addition, factors may change over time during the follow-up period, for example seasonal practices, numbers of family members being cooked for. In the better quality non-randomized experimental studies, these factors have been recorded and controlled for if needed. In applying GEPHI assessment, if a substantial proportion of the non-randomized intervention studies have not examined or addressed these issues and are judged to be subject to bias, these have been downgraded.

3.2 Additional criteria for GEPHI assessment

When basing recommendations on observational studies that start off as "low" quality evidence, the two following criteria have been proposed to contribute to a better assessment of confidence in the effect estimate (5).

a) **Consistency:** Consistent evidence that is found across multiple settings, geographical locations and/or over time, and across diverse epidemiological study designs and/or gathered by different researchers suggests that the effect can be reproduced under highly variable conditions that are unlikely to be subject to the same sources of confounding and bias. This draws on the Bradford Hill viewpoints: specifically, the original paper by Bradford Hill refers to ‘a variety of situations and techniques’.*(6, 15)* In the present context, ‘techniques’ have been interpreted as ‘study design’, as this seems the most appropriate term when dealing with a varied set of epidemiological studies.

b) **Analogy:** coherent evidence on the effect of similar environmental health interventions or exposures that operate through the same or a similar mechanism. This is, for example, seen with health impacts of exposure to air pollution generated through similar combustion processes, including household air quality from solid fuel combustion, outdoor air quality from vehicles and fossil fuel power generation, second-hand and active smoking. In this regard, it is noted that tobacco is a form of biomass, in common with the most widely used solid household fuel. Further confidence in effect estimates can be derived from consistency in levels of risk and exposure. For example, second hand smoke (SHS) and HAP from solid fuel use have both been linked with an increased risk of low birth weight (LBW) *(17)*; SHS typically results in lower levels of exposure to combustion pollutants than does solid fuel use, and is associated with a correspondingly lower risk (and a smaller effect on mean birth weight).

As these criteria are not currently among those which would permit upgrading of the evidence in GRADE, these have been added to the GEPHI assessment, as shown in Table 4. These criteria have not been applied at the level of a body of evidence based on one type of study design (e.g. controlled before-and-after studies) but are applied when examining a body of evidence based on multiple study designs (e.g. RCTs and uncontrolled before-and-after studies combined). Upgrading for these two criteria was carried out even after downgrading for lower quality in the initial stage of assessment.

3.3 Revised (GEPHI) assessment table

In light of the foregoing discussion, the following specific modifications have been incorporated into the GEPHI assessment, as shown in Table 4.

- Non-randomized experimental studies, including uncontrolled ‘before and after’ studies and 'quasi-experimental’ designs with comparison groups would enter the table as 'moderate evidence'.
- In all other respects, existing GRADE scoring rules have been applied in the initial assessment. Evidence was only upgraded in the presence of risk of bias where there was a statistically significant large (or very large) effect.

- Non-randomized experimental studies were more likely to be downgraded if uncontrolled, and/or there was evidence that time-varying factors has not been adequately considered or adjusted for.
- Where there was consistent evidence from multiple studies in different settings for example countries and regions of the world, and population groups, and/or across differing study designs, upgrading of +1 has been applied.
- Where consistent supporting evidence was available for analogous exposures (e.g. outdoor air pollution, smoking) known to operate through the same or a similar mechanism, upgrading of +1 has been applied.
- When one or both of these latter two criteria were met, upgrading was carried out even in the presence of downgrading in the initial stage of assessment.

Table 4: Modified (GEPHI) assessment table: modifications are highlighted in blue

Quality of evidence	Study design	Lower the quality in presence of	Raise the quality in presence of
High	Randomized trial	**Study limitations:** -1 Serious limitations -2 Very serious limitations -1 Important **inconsistency** **Directness:** -1 Some uncertainty -2 Major uncertainty -1 **Imprecise data** -1 High probability of **reporting bias**	**Strong association:** +1 Strong, no plausible confounders, consistent and direct evidence +2 Very strong, no major threats to validity and direct evidence +1 Evidence of a **dose-response** gradient +1 All **plausible confounders** would have reduced effect **Additional criteria (applied across a body of evidence based on multiple study designs)[1]:** +1 **Consistency** across multiple studies in different settings +1 **Analogy** across other exposure sources
Moderate	**Quasi-experimental (with controls) and before and after (uncontrolled) studies[2]**		
Low	Observational study		
Very low	Any other evidence		

3.4 Comparing revised (GEPHI) and standard GRADE assessment

In order to allow an appreciation of the impact of these modifications, the assessment of evidence using the standard GRADE profile table has been preserved in the revised approach by providing both an 'intermediate' and a 'final' assessment.

For health outcome evidence (summarised in Review 4) for which no before and after or quasi-experimental intervention studies were available, the standard GRADE assessments can be directly compared with the 'intermediate' assessment in the GEPHI table, with accompanying explanations for application of the new criteria in the table and relevant text.

[1] Upgrading for these criteria was carried out even if the evidence had been downgraded in the initial stage of assessment.

[2] To distinguish between controlled and uncontrolled non-randomized designs in this category, the latter were more likely to be downgraded for lower quality.

For evaluation of the impacts of improved stoves and clean fuels on HAP and exposure (see Review 6), for which non-randomized experimental studies were the most important design (in fact, virtually all were uncontrolled before-and-after designs), comparison can be made with the standard GRADE method by reducing the 'intermediate' assessment level by 1 as a result of this evidence being entered as 'moderate' rather than 'low'.

3.5 Methods for quality assessment of other evidence contributing to recommendations

In section 1.2, the wide range of types of evidence contributing to the recommendations was described. Even with the modifications to GRADE described above, not all of this evidence is amenable to this method of evidence review. The approaches adopted are specific to the topics and evidence reviewed for these guidelines, and a generic description is neither possible nor useful.

The GRADE domains (see Box 2) were found useful and have therefore been used as a guide, for example in assessing the risk of bias, or whether or not publication bias may be present (in circumstances where funnel plot asymmetry cannot be employed).

In using these domains, judgements have been made based on the available evidence, but in no case has any overall numerical scoring has been attempted. A final quality assessment has been expressed using the same terms and definitions as set out in Table 3, namely: high, moderate, low and very low. Further details of which reviews have been assessed in this way are provided in Table 7, and in the respective reviews.

Box 2: GRADE domains used to guide evidence quality assessment where GEPHI table not applicable:

- Number of studies
- Study design
- Risk of bias
- Indirectness
- Inconsistency (heterogeneity)
- Imprecision
- Publication bias

4. Application of evidence review methods to Guidelines

As noted in the introduction, the methods used for evidence review for the purposes of developing recommendations and guidance follow GRC guidance. These steps, together with the modification described in Section 3, have been applied as follows:

4.1 Scoping questions for evidence review

Based on the policy objectives for the guidelines, the following four main scoping questions setting out the issues to be addressed by the guideline recommendations were developed:

1. What device and fuel emission rates are required to meet WHO (annual average) air quality guideline and interim target-1 for $PM_{2.5}$, and the (24-hr average) air quality guideline for CO?
2. In light of the acknowledged challenges in securing rapid adoption and sustained use of very low emission household energy devices and fuels, particularly in low income settings, what approach should be taken during this transition?
3. Should coal be used as a household fuel?
4. Should kerosene be used as a household fuel?

4.2 Evidence required to address the scoping questions

The first step in evidence search and retrieval was to identify and define the evidence required to address the scoping questions. Due to the nature of the policy challenges being addressed and also the fact that very few experimental studies have directly assessed the impact of alternative interventions on health, several distinct areas of evidence were required for each scoping question. These 'areas of evidence' are summarized in Table 5. Those amenable to the PICO format are indicated, and elaborated further below.

Table 5: Areas of evidence sought for each scoping question

Scoping question	Evidence required for scoping question	Framing of evidence requirement (topic)
1. Emission rates to meet AQGs	a. Published WHO air quality guidelines	a. Reference to AQGs for selected pollutants ($PM_{2.5}$, CO).
	b. Emission rates of key pollutants from traditional devices/fuels and intervention options	b. Summary of laboratory and field test results for $PM_{2.5}$, CO (and other important pollutants).
	c. Relationships between emission rates and indoor air quality	c. Model relating emission rates with predicted concentrations for $PM_{2.5}$ and CO.
2. Policy during transition	a. Disease risks from HAP and estimated effective sizes for impacts of interventions	a. Summary of evidence relating HAP to specific disease outcomes, causal evidence and effect sizes: defined by **PICO-1** (below).
	b. Relationships between level of exposure and level of (major) disease risk across the full range of exposure seen with intervention options	b. Summary of evidence on exposure-response relationships for (major) disease outcomes.
	c. Levels of HAP and exposure experienced by populations using traditional stoves/fuels, and intervention options	c. Summary of observational population-based studies with measured average $PM_{2.5}$ and CO.
	d. Impacts of interventions on HAP levels achieved with stoves/fuels in everyday use.	d. Summary of observational (where relevant) and experimental studies (randomized and non-randomized) with measured average $PM_{2.5}$, and CO: defined by **PICO-2** (below.)
	e. Nature and extent of barriers to transition to improved solid fuel stoves and clean fuels.	e. Summary of evidence on factors influencing the adoption and sustained used of interventions.
3. Coal use	a. Health impacts of solid fuel use	a. As for 2(a) and (b) above.
	b. Health risks specific to household use of coal	b. Summary of evidence on carcinogenicity, toxic contaminants, and constraints on clean combustion of coal.
4. Kerosene use	a. Health risks specific to household use of kerosene	a. Summary of evidence on kerosene use, levels of pollutants, and health impacts

Reflecting the varied and broad nature of the evidence, reviews of differing types were judged to best fulfil these requirements. The following types of review have been prepared:

- A **systematic review** (with meta-analysis if included).
- A **summary of a systematic review** (with meta-analysis if included), where the review summarizes a recently conducted or published systematic review.
- A **summary and synthesis of systematic reviews and other evidence**, where the review brings together summarizes of completed SRs (with meta-analyses if included), and other evidence, and includes some synthesis of this evidence.
- A **model**, which here is used to describe the emissions rate model in Review 3, and the integrated exposure-response functions (IERs) in Review 4.
- A **narrative review**, where an overview of a set of issues is provided that has not been the subject of systematic, defined literature search.

Further detail on which type of review was used for each area of evidence, and the nature of evidence included, is provided in Table 7.

4.3 Framing of questions amenable to PICO format

As shown in Table 5, two of the areas of evidence (topics) were amenable to framing using the PICO format, namely those addressing (i) impacts of interventions on health outcomes [PICO-1, 2(a) in Table 5], and (ii) impacts of interventions on household levels of $PM_{2.5}$ and CO [PICO-2, 2(d) in Table 5]. These are presented below, with additional explanation of the rationale for the outcomes included.

Impacts of interventions on health outcomes (PICO-1)

Although it was judged important for the purposes of the full guidelines project to review evidence for all child and adult health outcomes linked to HAP exposure, the GDG determined that the focus should be on important outcomes, that is, those which impact on child survival and development (e.g. ALRI, low birth weight, stillbirth) and those otherwise responsible for large burden of disease in the 2010 GBD study (e.g. COPD, CVD): these outcomes are indicated in the PICO table below with an asterisk (*).

While these studies provide the largest and most robust source of evidence on the impacts of interventions on health outcomes (in terms of estimated risk reduction), the lack of HAP and/or personal exposure measurement in most mean that the exposure levels associated with these 'impact effect' findings can only be estimated. Furthermore, this leaves the question of risk levels with 'intermediate' exposure reductions essentially unanswered.

This latter (and critical) area of evidence is provided by the exposure-response evidence, and in particular by the integrated exposure response functions (IER) which are covered by topic 2(b) in Table 5. This important IER evidence is not available for all of the 'important' disease outcomes listed in the PICO-1 table: where it is available this is indicated in the table below by inclusion of [IER].

PICO-1: Impact of interventions reducing HAP exposure on health outcomes

PICO-1	Description	
Population	The 2.8 billion people*(18)* using solid fuels, that is biomass (wood, animal dung, crop wastes, charcoal) or coal as their primary cooking and heating fuel, with open fires or traditional stoves.	
Intervention	Clean fuels (LPG, electricity), and/or a range of 'improved' solid fuel stoves, delivering substantial reductions in HAP exposure. Exposure levels have mostly not been measured in the relevant studies, but have been estimated to generally lie between the WHO annual IT-1 of 35 µg/m^3 $PM_{2.5}$ at the lower end of the range and 75 µg/m^3 $PM_{2.5}$ at the upper end.	
Comparison	Households using solid fuels with traditional stoves	
Outcome	**Child (under 5 years)**	**Adult**
	Acute lower respiratory infections (ALRI)* [IER]	Chronic Obstructive Pulmonary Disease (COPD)* [IER]
	Low birth weight*	Lung cancer with coal exposure* [IER]
	Stillbirth*	Lung cancer with biomass exposure* [IER]
	Stunting*	Cardiovascular disease*[1] [IER]
	All-cause mortality under 5 years*	Cataract
	Cognitive development	Other cancers
		Asthma (adult and child)

[1] Very few primary studies are available on the risk of cardiovascular outcomes with exposure to solid fuels or associated HAP levels, and health risk assessment has relied on interpolation from risk functions for other combustion sources; as a consequence, the GEPHI assessment table has not been used for this outcome.

Impact of interventions in everyday use on household levels of $PM_{2.5}$ and CO (PICO-2)
The second area of evidence amenable to the PICO format deals with the impacts of alternative solid fuel and clean fuel intervention options on kitchen levels of $PM_{2.5}$ and CO, when these devices and fuels are in everyday use, although eligible studies were not found for all of the interventions listed. The main outcomes are average kitchen levels of the two key pollutants selected as the focus for the recommendations, that is, $PM_{2.5}$ and CO.

PICO-2: Impact of interventions on average levels of household air pollution

PICO-2	Description	
Population	The 2.8 billion people *(18)* using solid fuels, that is biomass (wood, animal dung, crop wastes, charcoal) or coal as their primary cooking and heating fuel, with open fires or tradition stoves.	
Intervention	**Improved solid fuel stoves**	**Clean fuels**
	With chimney	Liquefied petroleum gas (LPG)/ natural gas
	Without chimney	Ethanol
	Mixed (stove plus other improvements to kitchen and cooking arrangements)	Biogas
		Solar cookers
		Electricity
Comparison	Households using solid fuels with traditional stoves	
Outcome	Average 24-hr (or 48-hr) concentrations of: • Kitchen $PM_{2.5}$ • Kitchen carbon monoxide (CO)	

4.4 Defining other questions and topics

Three other important issues for which evidence reviews were conducted are as follows:

- **Safety*:*** although not the result of poor air quality, the safety risks (burns, scalds, poisoning from ingestion of liquid fuel) associated with household energy use were identified as important since interventions that reduce emissions of health damaging pollutants cannot be assumed also to be safer. The findings of the systematic review on this topic (Review 10) have informed the 'General Considerations' which apply to all of the specific recommendations, in addition to contributing directly to the evidence used for the recommendation on the household use of kerosene.
- **Adoption*:*** as noted in the introduction, there are significant challenges for policy in achieving rapid and sustained adoption of much cleaner household energy interventions, particularly in low-income settings. The systematic review of factors influencing adoption and sustained use of improved stoves and clean fuels (Review 7) informs plans for the development and testing of guidance and tools to support implementation.
- **Synergies between health and climate impacts*:*** household fuel combustion can have significant impacts on climate through both efficiency of combustion and the nature of the emissions, in addition to the health impacts. A review of evidence on the net climate impacts (warming) from inefficient use of non-sustainable biomass and emissions from incomplete fuel combustion is reported in Review 11. This informs a best practice recommendation on maximizing health 'co-benefits' in climate change mitigation policy that addresses household energy.

4.5 Example of revised (GEPHI) assessment table

The revised assessment table used for evidence informing the PICO-defined questions is shown (in blank form) in Table 6. For each body of evidence, profiles are summarised separately by major study design category, as follows:

a) Randomized controlled trials, including cluster-randomized trials
b) Non-randomized experimental studies
c) Observational studies (e.g. cohort studies, case-control studies, cross-sectional studies)

In line with GRADE, if there are sufficient higher-quality study designs [e.g. groups (a) and (b)] and these are considered to provide sufficient evidence overall, lower-quality study designs [e.g. group (c)] have not contributed to the final assessment.

4.6 Summary of assessment of quality of evidence

The methods used for the assessment of (i) individual studies contributing to systematic reviews and (ii) quality of overall evidence available from each review, is summaries in Table 7. The methods used for evaluation of the quality and strength of evidence described in Table 7 were judged to be the best available given the nature of the questions and the evidence available. Other work underway, for example in a project funded by the European Centre for Disease Prevention and Control (ECDC), a group of researchers is attempting to modify GRADE for some of these broader types of questions and evidence, including qualitative evidence[3].

[3] See PRECEPT project: http://www.rki.de/EN/Content/Prevention/PRECEPT/PRECEPT_node_en.html).

Table 6: Format for summarizing assessment of evidence reviews for PICO-defined questions

Study design	No of studies	Risk of bias	Inconsistency [statistical heterogeneity	Indirect-ness	Precision [power]	Publication bias	Other[1] [e.g. large effect]	No. of subjects[2]		Effect and 95% CI	Quality
								Int	Con		
RCT (individual and cluster)											
Uncontrolled before-and-after, Quasi-experimental with comparison groups											
Cross-sectional, C/control, cohort											
	Intermediate assessment										
Final score	Additional criteria[3]:				Explanation						
	1. Consistency across studies of differing designs, different setting, etc (+1) 2. Analogy of evidence from other combustion exposure sources (+1)				Final assessment						

[1] Only large effect (statistically significant RR>2) has been used for upgrading if the group of studies was downgraded for any reason

[2] As intervention and control groups do not apply to observational designs, and the designation of subjects varies between study designs (i.e. case/control, subjects in a cohort study, etc.), the total number of subjects is recorded

[3] Upgrading for additional criteria was carried out even if there had been downgrading for in the initial stage of assessment.

Table 7: Summary of systematic and other evidence reviews, methods for individual study assessment, and overall evidence quality assessment.

Further detail of the assessment of quality, including upgrading and downgrading for specific outcomes, is available in the assessments of quality of evidence for each of the main recommendations (Annexes 4-7).

Review number and (short) title and area of evidence (topic) addressed	Type of review and evidence included	Methods for assessment of individual study quality	Methods for assessment of quality of the set of evidence compiled in the review
Review 1: Fuel use Global and regional summary of household fuel and technology use for cooking, heating and lighting	**Synthesis and analysis:** Includes a synthesis and analysis of survey data and nationals statistics reports from low, middle and high income countries on the fuels and technologies used by households for cooking, heating and lighting, as well as a summary of trends in solid cooking fuel use based on modelled estimates.	All surveys or reports included in the analysis had to meet minimum criteria for date, population representation, and methods of data collection (e.g. survey questions used).	A global dataset was available for cooking fuel use. Data related to heating, lighting and cooking devices were less comprehensive and in some cases no aggregated figures by region could be calculated and only national level statistics are presented. Inconsistencies in data collection or reporting methods were reconciled through grouping of response data or exclusion from analysis.
Review 2: Pollutant emissions Summary of laboratory and field test results for $PM_{2.5}$, CO (and other important pollutants).	**Systematic review:** Includes mainly laboratory and a smaller number of field-based measurements of emissions of major pollutants from a representative range of stove and fuel types.	Suitability of the emission test protocol used, adherence to protocol and quality assurance of methods used.	Suitability of test protocols used for emissions testing, and extent of adherence to protocols and quality control/assurance by laboratories. Issues for interpretation of results from laboratory and field testing.
Review 3: Emissions model Model relating emission rates to predicted concentrations for $PM_{2.5}$ and CO.	**Model:** Includes a review of alternative modelling approaches, and a description of the single zone box model using Monte Carlo simulation to link emission rates (inputs) with distributions of average concentrations $PM_{2.5}$ and CO in the home (outputs).	Not applicable, although the limitations of the model and currently available data inputs are discussed.	Validation of predicted kitchen concentrations of $PM_{2.5}$ and CO against empirical data from homes in India and countries from other regions; interpretation of findings in light of evidence of emissions from multiple sources in real life settings, including from neighbouring homes and other sources of combustion-derived air pollution.
Review 4: Health impacts of household air pollution (HAP) Summaries of evidence relating HAP to specific disease outcomes, causal evidence and effect sizes: defined by PICO-1, exposure-response evidence; risk from use of gas, and impacts of smoke reduction on vector-borne disease.	**Summary and synthesis of systematic reviews and other evidence:** 1. Summary and synthesis of completed systematic reviews and meta-analyses (SRMA) of epidemiological studies linking HAP exposure (from solid fuels and gas) to a range of health outcomes.	1. Intervention impact estimates: Evaluation of study quality was carried out using versions of Newcastle-Ottawa scale, adapted to each type of study design.	1. Intervention impact estimates: Bradford Hill viewpoints to assess strength of causal evidence; assessment using GEPHI[4] for quality and precision of intervention effect estimates (details of grading by disease outcome are provided in Annex 5).
	2. Summary of all available exposure-response evidence including newly developed integrated	2. Integrated exposure-response functions. All studies of risk associated with	2. IER functions: GRADE domains (number of studies, study design, risk of bias, indirectness, imprecision,

[4] GEPHI: Grading of evidence for public health intervention (see Section 2.3.3)

Review number and (short) title and area of evidence (topic) addressed	**Type of review and evidence included**	**Methods for assessment of individual study quality**	**Methods for assessment of quality of the set of evidence compiled in the review**
	exposure-response models for several disease outcomes.	household air pollution which contributed to these integrated models were assessed individually as described above.	inconsistency, and publication bias) were used as a guide, and applied in the most appropriate way given the nature of the available evidence. Assessment was made (i) generically for the IER approach and model assumptions, and (ii) in respect of specific issues for each disease outcome.
	3. Summary of published SRMA of health risks from household use of gas.	3. Individual study quality assessment was not available, but general quality issues for studies (especially for exposure assessment) were identified and discussed.	3. Assessment focussed on the inconsistency across findings of the SRMA between the two measures of exposure (gas and NO_2) and the two outcomes (wheeze and asthma), and potential explanation for this.
	4. Summary of published systematic review of risk of vector-borne disease (VBD) from potential interventions to reduce HAP exposure.	4. Individual study quality assessment was not available, but general quality issues including confounding are discussed.	4. Meta-analysis was not attempted, so formal assessment with GEPHI not conducted. Assessment focuses on lack of experimental studies, and potential for confounding.
Review 5: Population levels of HAP and exposure Summary of observational population-based studies with measured average $PM_{2.5}$ and CO.	**Systematic review:** Includes studies which have measured 24-hour or 48-hour concentrations of PM and CO in kitchens, other rooms within homes, in the local ambient air, and personal exposure of these same pollutants for men, women and children.	Methods used for selecting samples of homes and individuals, protocols for measurements of $PM_{2.5}$ and CO, and evidence of quality assurance procedures.	GRADE domains (number of studies, study design, risk of bias, indirectness, imprecision, inconsistency, and publication bias) were used as a guide. All eligible studies provided measures of long-term average (e.g. 24-hr or 48-hr) levels of pollutants, which increased consistency of the findings.
Review 6: Intervention impacts Summary of observational (where relevant) and experimental (randomized and non-randomized) studies with measured average $PM_{2.5}$ and CO defined by PICO-2.	**Systematic review and meta-analysis:** Includes studies which provide data on average home-based 24-hr or 48-hr kitchen and/or personal $PM_{2.5}$, PM_4 or CO using either experimental designs, or observational studies of intervention programmes.	Evaluation of study quality using versions of Newcastle-Ottawa scale, adapted to each type of study design, including adequacy of description and application of standardized HAP measurement.	Assessment using GEPHI to assess the quality and precision of estimates for each pollutant ($PM_{2.5}$, CO), and for each group of stove or clean fuel intervention (details of grading by intervention type and pollutant are provided in Annex 5).
Review 7: Factors influencing adoption Summary of evidence on factors influencing the adoption and	**Systematic review:** Includes quantitative, qualitative and policy/case studies in low and middle income countries, reporting on factors influencing adoption and/or	*Quantitative studies*: Used versions of Newcastle-Ottawa scale, adapted to each type of study design. *Qualitative studies*: Used methods	GRADE domains (number of studies, study design, risk of bias, indirectness, inconsistency, imprecision for quantitative evidence, publication bias) were used as a guide, and applied in the

Review number and (short) title and area of evidence (topic) addressed	Type of review and evidence included	Methods for assessment of individual study quality	Methods for assessment of quality of the set of evidence compiled in the review
sustained use of interventions.	sustained use of improved solid fuel stoves, and four types of clean fuel (LPG, biogas, alcohol, solar cookers)	described by Harden et al 2009 *(18)* *Policy and case studies:* Used methods for case studies described by Atkins and Sampson 2002*(19)*	most appropriate way given the nature of the available evidence; consistency of findings across different studies designs and settings was important.
Review 8: Coal Summary of evidence on carcinogenicity, toxic contaminants, and interventions to reduce adverse health effects including bans and other restrictions on household use of coal.	**Summary and synthesis of systematic reviews and other evidence:** 1. Synthesis of studies on health risks from coal arising from products of incomplete combustion, drawing on published systematic review, and other studies identified through systematic search. 2. Summary of evidence from IARC monograph on carcinogenicity of emissions from household coal use. 3. Systematic review of studies (intervention and observational) relating to health risks from toxins in coal.	For carcinogenicity, methods used are those described by IARC. Quality assessment of studies of health risks from coal use (from products of incomplete combustion, and from toxic contaminants) was based on evaluation of methods used for sampling, exposure and outcome assessment, and analysis including adjustment for confounding. This information was used to provide an overall quality assessment.	Three distinct components contributed to the overall evidence available for coal: carcinogenicity, health effects from products of incomplete combustion (PIC), and toxic contaminants. *Carcinogenicity:* IARC methods are based on assessment of human epidemiology, animal evidence and mechanistic evidence. *PIC effects:* For lung cancer, the GEPHI assessment from Review 4 (Health impacts of household air pollution) was used. For other (non-cancer) outcomes, due to small numbers of studies and heterogeneity of outcome definitions, meta-analysis was not conducted and GEPHI was not applied. GRADE domains (number of studies, study design, risk of bias, indirectness, inconsistency, imprecision for quantitative evidence, publication bias) were used as a guide. *Toxic contaminants:* Assessment of quality was based on the combination of studies (including experimental studies) reporting coal toxin content, emission and area concentrations of toxins, and population studies of specific outcomes (e.g. fluorosis, arsenicosis) in areas where households burn coal.
Review 9: Kerosene Summary of evidence on kerosene use, levels of pollutants, and health impacts.	**Summary of systematic review:** Includes studies reporting on kerosene use (fuel type/grade and devices) for cooking, heating and lighting; emissions of major pollutants and areas concentrations; epidemiological studies on health risks with kerosene	Evaluation of study quality was based on methods for exposure and outcome assessment, and analysis. A formal quality scoring tool was not used.	Due to the substantial heterogeneity in study methods, quality and findings, meta-analysis was not attempted for any of the health outcomes, and the GEPHI assessment table not applied. GRADE domains (number of studies, study design, risk of bias, indirectness, inconsistency, imprecision for

Review number and (short) title and area of evidence (topic) addressed	Type of review and evidence included	Methods for assessment of individual study quality	Methods for assessment of quality of the set of evidence compiled in the review
	use in the home.		quantitative evidence, publication bias) were used as a guide.
Review 10: Safety Summary of evidence on burns, scalds and poisoning.	**Systematic review:** Includes studies reporting on rates of burns and poisoning from household energy use, risk factors, impact of interventions. Also includes a description of a newly developed safety testing protocol for solid fuel stoves.	Evaluation of intervention study quality using versions of the Newcastle-Ottawa scale, adapted to each type of study design.	For studies of risk factors, methodological issues such as case selection - most were drawn from facilities with few population studies - were a quality concern. Experimental studies (both randomized and non-randomized designs available) were generally of good quality, but too variable in terms of interventions and outcomes to carry out meta-analysis.
Review 11: Climate impacts Summary of evidence on climate impacts (warming) from inefficient use of non-sustainable biomass, and emissions from incomplete fuel combustion.	**Narrative review:** In view of the complexity of climate science (and this not being the main focus of the guidelines), a narrative review providing an overview of the impacts of household fuel combustion on climate was provided. This draws on a recent comprehensive UNEP report on the effects on climate of short-acting pollutants and other published studies.	Individual climate science studies on the impacts of household fuel combustion pollutants on climate warming were not assessed separately.	The overall evidence provided by the climate science studies on the impacts of household fuel combustion pollutants on climate warming was not assessed. The consistency of evidence indicating substantive net warming effects draws strongly on the conclusions of the UNEP report.

4.7 Assessment of coherence of evidence contributing to the causal chain

An assessment of the coherence of evidence forms part of the overall evidence evaluation for Recommendation 2, and is presented in Section A4.5 of the Evidence Profile for this recommendation (Annex 4). This uses the causal chain model (Figure 1) as a unifying framework.

This first examines how consistent the emission rates from various stoves and fuel options (Review 2) are when compared against observed levels of HAP and exposure in populations (Reviews 5 and 6), and with those predicted by the emissions model (Review 3). This assessment takes into account what is known about use of these various stoves and fuels in practice and the factors which influence such behaviour (Review 7).

This is followed by an assessment of the consistency of observed HAP and exposure reductions with the findings of epidemiological studies of health risks (Reviews 4 and 8), both in respect of binary exposure classification (e.g. use of solid fuel vs. clean fuels) and also with reference to the exposure-response functions (Review 4) which are especially important in this regard. The degree of coherence is discussed, and explanations advanced for any evidence which does not ‘fit’ the pathways proposed in the causal chain.

This overall assessment of the degree of coherence provides an important check on the wide range of evidence informing recommendations. Thus, each component of the evidence contributes to our understanding of the links between, for example, new technologies/fuels and health impacts, taking into account how these perform in everyday use, user's perspectives and behaviour, and the influences of various aspects of policy. The more coherent the findings of the various components, the greater can be our confidence in recommendations based on this evidence. On the other hand, aspects which are less coherent can point towards areas where, for example, anticipated impacts are not being realised and where further technical, research and policy attention are required.

4.8 Decision tables for strength of recommendations

As noted in the introduction, GRADE provides for an assessment of strength of recommendations through consideration of the quality of evidence and three other sets of issues (values and preferences, balance of costs and benefits, and resource implications) that influence whether a recommendation should be 'strong' or 'conditional'. The definitions used in the guidelines for strong and conditional recommendations are as follows:

- A **strong** recommendation is one that the guideline development group agrees that the quality of the evidence combined with certainty about the values, preferences, benefits and feasibility of this recommendation means it should be implemented in most circumstances;

- A **conditional** recommendation is one for which there was less certainty about the combined quality of evidence and values, preferences, benefits and feasibility of this recommendation meaning there may be circumstances in which it will not apply.

This assessment has been made using the GRADE decision table, as shown below (Table 8), in blank form. Although, as has been described in the foregoing sections of this overview, the topic of the guidelines has required a modified approach to assessment of evidence quality, the principles set out in the GRADE decision table remain relevant and appropriate, and have therefore not been modified. Since the methods for assessment quality (the first consideration in the decision table) uses GEPHI rather than GRADE, however, the tables have been termed ***Decision tables for strength of recommendations***.

For each recommendation, summaries of the evidence quality, values and preferences, benefits vs. harms, and resource implications, have been prepared based on the respective evidence reviews, and are presented in Annexes 3-6.

Table 8: The Decision table used for assessing strength of recommendations

Recommendation: [Number and text]		
Factors influencing strength of recommendations	**Decision (all response options shown)**	**Explanation**
Quality of evidence (based on GEPHI methods) (The higher the quality of the evidence, the more likely a strong recommendation is warranted.)	• High • Moderate • Low • Very low	
Balance of benefits versus harms and burdens (The larger the difference between the benefits and harms, the more likely a	• Benefits clearly outweigh harms • Benefits and harms are • balanced • Potential harms clearly	

Recommendation: [Number and text]		
Factors influencing strength of recommendations	**Decision (all response options shown)**	**Explanation**
strong recommendation is warranted. The smaller the net benefit and the lower the certainty for that benefit, the more likely a conditional recommendation is warranted.)	• outweigh potential benefits	
Values and preferences (The greater the variability or uncertainty in values and preferences, the more likely a conditional recommendation is warranted.)	• No major variability • Major variability	
Resource use (The higher the costs of an intervention, that is, the more resources consumed, the more likely a conditional recommendation is warranted.)	• Less resource intensive • More resource intensive	

5. Conclusions

A broad range and type of evidence was needed to address the guideline scoping questions, including laboratory test studies, models, qualitative and policy studies, in addition to standard epidemiological designs. In addition, for the latter, few randomised trials were available, but non-randomized experimental studies were an important resource.

Evaluation of the quality and consistency of this evidence required a modified approach to GRADE. This involved development of a causal chain as a framework to link the different sets of evidence, additional criteria for assessing quality, and an evaluation of the degree of consistency across the casual chain. This revised methodology was termed ‘grading of evidence for public health’, or GEPHI.

For systematic reviews with defined PICO questions and quantitative summaries using meta-analysis, modified GEPHI assessment tables incorporating additional criteria were applied, and were set out to allow comparison between a standard GRADE rating and the GEPHI alternative. The latter (GEPHI) method usually, but not always, rated evidence as being of higher quality than the GRADE method. For other bodies of evidence not amenable to quantitative summary (that is, meta-analysis was not possible or deemed suitable), GRADE domains were found to be a useful means of guiding the evaluation of quality; this resulted in standard descriptions of quality (high, moderate, low and very low), but no attempt was made to derive a numerical quality score.

Most of the evidence used to inform recommendations was rated as of moderate quality, although some was rated low and some high. Some confidence in the validity of these ratings was provided by a generally good level of consistency between the findings of sets of evidence representing the various parts of the causal chain.

The decision table for strength of recommendations employed the GEPHI quality assessments as described above, but otherwise used unmodified GRADE methodology for values and preferences, balance of benefits and costs, and resource implications.

Overall, the methods met the evidence assessment requirements for these guidelines adequately, but would benefit from further application and evaluation for other types of environmental and public health intervention.

References

1. WHO handbook for guideline development. Geneva: World Health Organisation; 2012.
2. GRADE Working Group: Grading quality of evidence and strength of recommendations. British Medical Journal. 2004;328:1490. doi: 10.1136/bmj.328.7454.1490.
3. Durrheim DN, Rheingold A. Modifying the GRADE framework could benefit public health Journal of Epidemiology and Community Health. 2010;64(5):387.
4. Barbui C, Dua T, van Ommeren M, Yasamy MT, Fleischmann A, Clark N, et al. Challenges in developing evidence-based recommendations using the GRADE approach: The case of mental, neurological and substance use disorders. PLOS Medicine. 2010;7(8):e1000322. doi: doi:10.1371/journal.pmed.1000322.
5. Rehfuess EA, Akl EA. Current experience with applying the GRADE approach to public health interventions: an empirical study. BMC public health. 2013;13(9).
6. Rehfuess E, Bruce N, Pruss-Ustin A. GRADE for the advancement of public health. Journal of Epidemiology and Community Health. 2011;65(6):559. doi: 10.1136/jech.2010.130013.
7. Duclos P, Durrheim DN, Reingold A, Bhutta Z, Vannice K,Rees H. Developing evidence-based immunization recommendations and GRADE. Vaccine 2012;31(1):12-9. doi: 10.1016/j.vaccine.2012.02.041.
8. ECDC. Evidence-based methodologies for public health – How to assess the best available evidence when time is limited and there is lack of sound evidence. Stockholm: European Centre for Disease Prevention and Control, 2011 Contract No.: ISBN 978-92-9193-311-2.
9. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. Developing and evaluating complex interventions: the new Medical Research Council guidance. Br Med J 2008;337:a1655. doi: 10.1136/bmj.a1655.
10. Rehfuess EA, Bartram J. Beyond direct impact: evidence synthesis towards a better understanding of effectiveness of environmental health interventions. International Journal of Hygiene and Environmental Health, 2014;217(2-3):155-9 doi: 10.1016/j.ijheh.2013.07.011.
11. Pawson R. Evidence based policy: a realist perspective. London: Sage; 2006.
12. NICE. Methods for the development of NICE public health guidance. 3rd edition, London: National Institute for Health and Care Excellence, 2012.
13. Balshem H, Helfand M, Schunemann HJ, Oxmand AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol. 2011;64(4):401-6. doi: 10.1016/j.jclinepi.2010.07.015.
14. List of GRADE working group publications and grants. 2012 [cited 10 February 2012]. Available from: http://www.gradeworkinggroup.org/publications/index.htm.

15. Hill AB. The environment and disease: association or causation? Proceedings of the Royal Society of Medicine. 1965;58:295-300.
16. Howick J, Glasziou P, Aronson JK. The evolution of evidence hierarchies: what can Bradford Hill's 'guidelines for causation' contribute? J R Soc Med. 2009;102:186-94. doi: 10.1258/jrsm.2009.090020.
17. Pope DP, Mishra VK, Thompson L, Siddiqui AR, Rehfuess E, Weber M, et al. Risk of low birth weight and stillbirth associated with indoor air pollution from solid fuel use in developing countries. Epidemiol Rev (2010) 32 (1): 70-81. doi: 10.1093/epirev/mxq005.
18. Bonjour S, Adair-Rohani H, Wolf J, Bruce N, Metha S, Prüss-Ustün A, et al. Solid Fuel Use for Household Cooking: Country and Regional Estimates for 1980-2010. Environmental Health Perspectives. 2013;121(7):784-90.
19. Harden A, Brunton G, Fletcher A,Oakley A. Teenage pregnancy and social disadvantage: systematic review integrating controlled trials and qualitative studies. Br Med J, 2009;339(b4254). doi: 10.1136/bmj.b4254.
20. Atkins C,Sampson J. Critical appraisal guidelines for single case study research. In: Wrycza S, editor. 10th European Conference on Information Systems (ECIS); Gdansk, Poland.2002.