How can additional secondary data analysis of observational data enhance the generalisability of meta-analytic evidence for local public health decision-making?

Authors:

Dylan Kneale[a], James Thomas[a], Alison O'Mara-Eves[a], Richard Wiggins[a]

[a]Department of Social Science, UCL Institute of Education, University College London, 20 Bedford Way, London, WC1H 0AL, United Kingdom

Corresponding author

Dylan Kneale

Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre), Department of Social Science, UCL Institute of Education, University College London, 20 Bedford Way, London, WC1H 0AL, United Kingdom

Phone: +44 20 7612 6020

Email: d.kneale@ucl.ac.uk

Abstract

This paper critically explores how survey and routinely collected data could aid in assessing the generalisability of public health evidence. We propose developing approaches that could be employed in understanding the relevance of public health evidence, and investigate ways of producing meta-analytic estimates tailored to reflect local circumstances, based on analyses of secondary data.

Currently, public health decision-makers face challenges in interpreting 'global' review evidence to assess its meaning in local contexts. A lack of clarity on the definition and scope of generalisability, and the absence of consensus on its measurement, has stunted methodological progress. The consequence means that systematic review evidence often failing to fulfil its potential contribution in public health decision-making.

Three approaches to address these problems are considered and emerging challenges discussed: (i) purposeful exploration after a review has been conducted, and we present a framework of potential avenues of enquiry and a worked example; (ii) recalibration of the results to weight studies differentially based on their similarity to conditions in an inference population, and we provide a worked example using UK Census data to understand potential differences in the effectiveness of community engagement interventions among sites in England and Wales; (iii) purposeful exploration before starting a review to ensure that the findings are relevant to an inference population. The paper aims to demonstrate how a more nuanced treatment of context in reviews of public health interventions can be achieved through greater engagement with existing large sources of secondary data.

Introduction

Evidence from well-conducted systematic reviews and meta-analyses is considered to represent one of the most methodologically robust sources available for public and clinical health decision-making (for example Berlin and Golub, 2014); however it is rarely utilised in local public health decision-making as it fails to meet decision-making needs in terms of local salience (Kneale et al., 2017). Some systematic reviewers claim that the act of combining evidence from studies taking place in different contexts is enough to warrant claims of ‘generalisability’, although these knowledge claims do not address do not address decision-makers’ needs around the local salience of evidence, and have substantial limitations in addressing different domains of generalisability. Large sources of secondary data could be instrumental in helping researchers and decision-makers to assess the feasibility of interventions contained within systematic reviews, as well as in understanding potential differences in the magnitude and direction of intervention effects, although their potential remains untapped.

Generalisability is a critical issue for systematic reviewers and meta-analysts working in public health to be engaging with now, for at least two reasons. Firstly, the perceived usefulness of—and decision to use—review evidence by local decision makers is undermined by the incongruity between a review’s scope and their informational needs (Kneale et al., 2017). The expansion in the geographic scope of systematic reviews runs in parallel to the scope of public health decision-making, which is becoming increasingly localised. Secondly, where review evidence is adopted in the decision-making process, ‘global’ estimates of effect could lead to the implementation of interventions that are not appropriate for the intended setting whereby ineffective or detrimental public health interventions can pose significant harm and potentially on a large scale (see Bonell et al., 2014, Lorenc and Oliver, 2013). Further utilisation of existing large secondary datasets could begin to address the gulf between ‘global’ review evidence and localised (generalised) review evidence. The challenges around the generalisability of evidence described above are thought to be more acute in the case of public health decision-making compared to clinical decision-making, given that the influence of context in determining the effect size is greater, which may in turn reduce the certainty of the evidence (Orton et al., 2011). Here, ‘context’ influences the design and delivery of interventions substantially in terms of structures, delivery bodies, epidemiological factors and populations at risk, as well as numerous other factors that influence outcomes. Fundamentally, naturally occurring conditions of ‘usual care’ also differ substantially between settings, perhaps more so than in the delivery of the intervention itself. However, the impact of differences in ‘usual care’ are rarely accounted for within review evidence. Public health interventions tend to be complex in nature, where the degree to which context is critical for the likely outcome achieved and hence ability to project generalisations from one case to the next is substantial; where intervention effects are often less certain; and where the intervention itself can be viewed as an adaptive or learning system, evolving in response to the context (Chandler et al., 2017, Lewin et al., 2017). Such contextual factors influence the underlying causal mechanisms of interventions, although experimental evidence such as that obtained from trials and included in meta-analyses often provide limited evidence of these factors, sufficient only for establishing causal descriptions, as opposed to richer causal explanations (Shadish et al., 2002).

Here we conceptualise generalisability as the assessment of whether evidence from one setting or population can be used to make valid inferences about what would happen in another specified setting or population. The concept remains central to scientific enquiry although its expression may vary across studies and systematic reviews to include external validity, transferability, extrapolation, portability or applicability; all of which are found across disciplines in the empirical sciences (Shadish et al., 2002). These terms are frequently conceived differently by different researchers at different times to articulate different problems or situations; some focus on the generalisability of results in laboratory conditions to real-life situations (for example Guala and Mittone, 2005); others focus on the ability to generalise on the basis of one population or group to another (for example Donaldson et al., 2001); others still take a more multi-faceted approach to conceptualising 'generalisability' through recognising the existence of different 'domains' of generalisability (for example Wang et al., 2006). Some have attempted to emphasise a clearer distinction between closely related terms that are (erroneously) used interchangeably in particular the distinction between the generalisability of the intervention model (applicability) and the intervention effectiveness (transferability) (Burchett et al., 2011, Wang et al., 2006, Cambon et al., 2012). Using existing secondary data sources, this paper seeks to conceptualise methods that may useful for (i) assessing the generalisability of evidence (both applicability and transferability) as well as (ii) tailoring estimates. If assessing the generalisability of evidence pertains to establishing generic truths or generic properties of evidence that may apply across areas but where the exact setting of application is undefined or broad, in contrast the tailoring of evidence involves shaping evidence and assessing its application to a specific and defined area.

Current approaches in enhancing and assessing the portability of meta-analysis

Traditional statistical approaches

In meta-analysis, the problem of generalisability revolves around the extent to which the single aggregate 'effect' from multiple studies, each of varying levels of external validity, can be used to consider what might happen if an aggregate of the interventions included in the analysis were implemented within a 'hidden', unspecified situation with respect to time, context, geography and population. In reality, the generalisability of findings is highly dependent on having a well-defined population to which a researcher intends to extrapolate the findings (O'Muircheartaigh and Hedges, 2014, Hedges, 2013, Glass, 2000). Implicitly in the case of meta-analysis this is a challenge as a target (inference) population is often undefined (Hedges, 2013), which thus limits the applicability of review evidence to any single setting. Indeed Glass, one of the original main proponents of meta-analysis, has suggested that meta-analysis is well equipped to provide a 'big fact' but may struggle to provide a more 'sophisticated answer' (2000).

Some researchers have claimed that evidence from meta-analysis holds superior properties of generalisability compared to evidence from single trials by virtue of the plurality of evidence sources used to generate the meta-analytic dataset. Authors such as Donaldson and colleagues (2001) emphasised that through the inclusion of data from 'different participants in different situations, and using different research procedures, one is able to get a better estimate of the robustness or the external validity of a given finding or effect' (p451). Such assertions of the superiority of meta-analyses with regards to generalisability only start to be convincing whenever we observe little heterogeneity in effect size estimates across studies; only then might we be able to better establish the consistency of impact across populations included within the meta-analysis. Where substantial statistical heterogeneity is encountered, meta-analytic methods (e.g. sub-group analyses, meta-regression) are intended to understand the ‘generalisable’ properties of the meta-analysis through providing a more nuanced configurative account of the overall effect. However, meta-analysts using sub-group analysis, and to a certain extent meta-regression, are repeatedly cautioned against the extensive use of these forms of configurative analysis, both on the basis of the types of causal attribution that is frequently, and erroneously, made in their interpretation (Kneale et al., under review, Petticrew et al., 2011, Thompson and Higgins, 2005), as well as their frequent deployment in the absence of a theoretical basis (Kneale et al., under review, Petticrew et al., 2011, Thompson and Higgins, 2005). With respect to generalisability, perhaps one of the most pivotal current limitations is that configurative meta-analyses are frequently undertaken as extensions of bivariate analyses, with the number of studies available for meta-analyses precluding more complex meta-regression models being undertaken.

Assessment tools

Several tools exist to aid systematic reviewers to assess the quality of individual studies included within reviews, the strength and quality of the overall body of evidence in the review (for example the GRADE tool), or allow others to assess the methodological robustness (Harden and Gough, 2012). These could be used alongside meta-analysis through sensitivity analyses to explore the impact of study quality on effectiveness (Thomas et al., 2017), although many only give limited consideration to the generalisability of evidence (Burchett et al., 2011, Burchett et al., under review). Tools explicitly designed to assess generalisability tend towards ‘frameworks’ to aid systematic reviewers to exercise a balance of judgement in the applicability of review evidence (Rychetnik et al., 2002, Wang et al., 2006, Burchett et al., 2011), although meaningful assessment of applicability can only take place with knowledge of the setting in which the evidence is to be applied. Recent methodological reviews highlight that many items used to assess generalisability lack methodological justification or empirical basis (Dyrvig et al., 2014, Burchett et al., 2011, Burchett et al., under review). Furthermore, Ahmad and colleagues (2010) uncovered substantial variation in the reporting of factors relating to context in 98 systematic reviews of public health interventions, with, for example, over a third of reviewers failing to report and consider the country in which the intervention took place on the generalisability of evidence. While new tools under development may help to address some of the issues outlined above (for example (Nøkleby and Munthe-Kaas, 2017)), the current absence of a preferred and adequate tool reviewers may use inadequate tools, unstructured approaches, or fail to assess the generalisability of their findings entirely.

Clearly there is variation in the way in which systematic reviewers of public health interventions approach generalisability. It is less clear that the current approaches address the key issues raised in this paper; we therefore propose new approaches in the following sections.

Proposed approaches: using secondary data analysis of large or routine data sources to enhance generalisability

We propose that further analysis of existing secondary data sources is an overlooked vehicle for assessing the transferability of evidence although to date, few strategies or methods have been proposed that explicitly seek to utilise the wealth of these data collected in the UK and other settings. These data can potentially provide valuable insights into epidemiological patterns that may inform assessment of applicability, particularly regarding ‘control’ conditions. The definition of ‘secondary data’ is vast and here we use the term to include real world data collected administratively as well as survey data collected longitudinally and cross-sectionally (Boslaugh, 2007, Kneale et al., 2016). Some of these data may have been collected with a specific analytic intent pre-defined, while others may be more descriptive and wide-ranging in nature, where the analytic intent is user-defined (Sim, 2015).

The epistemological focus and purpose of large scale secondary datasets and those of quantitative data extracted for meta-analysis differ widely. Meta-analytic data often lack detail around context needed to fully evaluate why interventions succeed where they do, and whether failure or unanticipated outcomes reflect problems with implementation and setting, or mechanistic problems with the intervention (Rychetnik et al., 2002). In contrast, secondary data give a fuller account of context and many accompanying life course transitions and processes surrounding health but are (in the clear majority of cases) absent of information on the intervention itself. In some ways, we could consider secondary data to reflect information on naturally occurring controls within a specific population, with histories reflecting lifestyles and circumstances that may amount to different levels of susceptibility to a condition or disease (or risk of drop-out or non-adherence), but for whom we do not usually observe any form of intervention. There may be greater similarity between meta-analysis data and other secondary data in the case of individual participant data (IPD) meta-analysis; the latter captures individual-level accounts of changes among both the intervention and control groups, and usually allows for greater investigation of individual-level potential confounding effects (Riley et al., 2010). However, these data are difficult to source, and there are few examples of IPD meta-analyses of public health interventions.

If observational secondary data can be used with meta-analysis, then what are some of the questions that could be asked and approaches that could be undertaken, and when can these take place? We might consider trialling one or a combination of three main approaches:

1. **Purposeful exploration before starting a review**: using the results of preliminary secondary data analysis to guide systematic reviews/meta-analysis.
2. **Purposeful exploration after conducting a review**: using systematic reviews/meta-analysis to structure secondary data analysis and explore the generalisability of review findings
3. **Recalibration**: using secondary data analysis to change the estimates given in meta-analysis to enhance the generalisability of (or tailor) the results to specific populations

Purposeful exploration before starting a review

Initial purposeful exploration could be undertaken using secondary data analysis (SDA) in order to understand ambiguities in the population that may shape the review itself. This approach would help to safeguard the applicability of the review to the inference population from the outset and strengthen its underlying conceptual framework. This form of preliminary SDA is not therefore confined to enhancing the generalisability of reviews that include meta-analyses, but could be applied to other types of systematic review. Application of this approach may be most useful when it is known that the trial literature is likely to underrepresent the experiences of a particular population segment of interest. Use of SDA before a review can help to refine the focus of a review to ensure that the research questions or synthesis address particular mechanisms or outcomes of interest. A recent example includes a review underway, led by Hayanga (2017), on interventions to address social isolation among older people where the evidence was intended to inform work among Black and Minority Ethnic (BME) older people.

Here preliminary SDA was undertaken in response to pilot searches indicating the lack of BME specific trials in the UK, and the underrepresentation of BME people within universal or geographically targeted interventions. Preliminary correlational SDA helped to identify that older BME people's friendship networks were particularly geographically disparate and led to a refined research question focussed on the effectiveness of community-based interventions (as opposed to individual or online interventions). This approach may not be necessarily novel, but linkage between SDA and protocol development is not widely reported. However, a potential disadvantage of this approach may be that the purposeful exploration becomes too prescriptive and may not be relevant to, or reflective of, the 'global' literature or application of the review evidence elsewhere.

Purposeful exploration after conducting a review

This approach involves using systematic reviews and meta-analyses as the starting point to undertaking secondary data analysis to better understand the results in context. Few examples exist where secondary data analysis is undertaken to understand the applicability of the results, although we describe an exemplar undertaken by Verma et al. (2012) below. Here, we suggest that secondary data analysis may aid local public health decision-makers to explore the results of meta-analysis (and systematic reviews) in five main ways: (i) refining targeting strategies; (ii) enhancing our understanding of control conditions in local contexts; (iii) understanding the applicability of interventions; (iv) evaluation of existing policy decisions that were based on review evidence; and (v) testing or assessing parts of a conceptual model developed on the basis of review evidence (Table 1).

TABLE 1 AROUND HERE

A related example of purposeful exploration after a review - Population Impact Analysis

Verma and colleagues' (2012) 'Population Impact Analysis' (PIA) approach incorporates data from the inference population as well as an effect size estimate from meta-analysis to forecast the potential impact of implementing influenza vaccination among older people in decreasing hospitalisations. In the source review, the study populations included the USA, Spain and Canada, each with different social insurance and hospital usage patterns (Vu et al., 2002). The pooled statistic was used directly to predict the decrease in hospitalisations that could occur in Trafford (Greater Manchester, UK) and the forecast was also based in part on existing vaccination rates and demographic information.

A worked example of purposeful exploration after a review – focused on applicability

Katz et al. (2008) conducted a review of strategies for the prevention and control of obesity in school settings which included education to improve nutrition and physical activity and to decrease sedentary behaviour. The interventions, which were conducted in eight different countries, led to a pooled standardised mean difference of -0.29 (CI: -0.45 to -0.14) in children's body weight. A sub-group of studies that incorporated parental involvement (either alongside children or separately) led to a pooled standardised mean difference of -0.20 (CI: -0.41 to 0.00) in children's body weight. The authors concluded, based on the meta-analysis, that among several commonly included program components, parental or family involvement and classroom (or after-school) instruction were effective; such a strategy may also be congruent with National Institute of Health and Care Excellence recommendations on family-based strategies to address child overweight/obesity (NICE, 2013). However, reviews of process evaluation studies of school-based public health interventions across a range of health conditions find that while intervention designs may include provision for parental involvement, in practice achieving desired and meaningful levels of parental engagement can be challenging for a number of reasons (Harris et al., in press, Pike et al., 2016). One reason may include poor or underdeveloped existing relationships between parents and schools before intervention implementation (Harris et al., in press). A hypothetical decision-maker may be interested in the results of Katz et al.'s (2008) review, but may decide to undertake further secondary data analysis to assess the feasibility of the intervention based on the results of the meta-analysis. The example we choose here is of a decision-maker working in London who may be interested in evidence of whether parents of overweight or obese children are differentially likely to attend after school activities, as a proxy indicator of baseline levels of parent-school relationships.

In our example, the Millennium Cohort Study (MCS), which has collected data from over 18,000 children born in 2000/1, is the chosen dataset for investigating this issue, using the age 7 sweep (see Connelly and Platt, 2014). We examine whether the parents of overweight or obese children are more or less likely to have reported attending a parents' evening, as a proxy indicator of engagement with the child's school. Among the parents in London, the likelihood of not attending a parents' evening is over twice as high among those with obese/overweight children as among children of normal weight (6% vs 2.5%), although the level of non-attendance is very low overall. To investigate further, the decision-maker is interested in whether parents' work during the evenings and whether these patterns may differ by the child's weight. The MCS showed that the proportion of families where one (in lone parent families) or both (in couple families) parents regularly worked evenings (once a week or more) was lower in London than the UK average (17.4% vs 19.7%) and this was not differentially patterned by the child's weight. The decision-maker could also examine whether other factors, such as deprivation or social class might pattern evening working (although no pattern by area level deprivation was detected here). Finally, the decision-maker could examine whether parents who worked evenings regularly were more or less likely to have attended parents' evening, with MCS data suggesting no significant association exists. Based on this form of purposeful analysis, the decision-maker could conclude that the majority of parents were attending parents' evenings, and that at least one parent was not working regular evenings among over four-fifths of families in London. Furthermore, even among those parents who did work regularly in the evening, analyses showed this did not differentially pattern their ability to attend parents' evenings. This may provide indicative evidence that parental involvement in after-school activities is a feasible strategy from the perspective of capacity for the majority of overweight/obese children, and help to identify the characteristics of those where this is not the case.

The use of further secondary data analysis to explore questions of transferability and applicability may be an important aid for decision-makers in developing further understanding of the generalisability of meta-analytic evidence. This step may already be frequently undertaken by decision-makers, although this practice does not appear to be widely reported in these terms. However, from the perspective of utilising SDA to understand the generalisability of meta-analytic evidence, this approach keeps both sets of data separate and does not represent any form of integration, which may be important in improving understanding of the generalisability of meta-analytic evidence. This separation means that such an approach is not specific to meta-analytic evidence, but could be applied more broadly to review evidence. This approach is also reliant on the availability of systematic reviews that include a substantial number of studies and, by extension, heterogeneity across effect sizes that allows for configurative forms of meta-analysis to take place, which in turn form the basis of purposeful enquiry. This is often not the case in public health review literature, where heterogeneity in intervention design precludes configurative analysis. Furthermore, it could be argued that the example above may represent an oversimplification of decision-maker concerns in using meta-analytic evidence, which are complex and multifactorial (Kneale et al., 2017). More fundamentally, such an approach continues to treat evidence from all studies included within a meta-analysis as being equally generalisable to any given setting. Addressing issues around the generalisability of meta-analytic evidence further may instead involve developing approaches that can better account for different characteristics simultaneously, and which recognise that the studies that contribute to a pooled effect in meta-analysis are not all equally generalisable. The next sections explore these considerations further through exploring the potential of recalibration (and related approaches).

Recalibration

Recalibration could involve a procedure to recalculate the overall pooled effect size so that the estimate more closely resembled the expected effect for the inference population. Recalibration could account for situations where the population of studies that are included within a review may have high external validity (in that they mirror conditions and groups within the specific source population) but where the results nevertheless have low transferability. This is to the extent that some included study interventions may be applicable but that the contexts in which they take place differ in observed (and likely unobserved) ways from a given inference population or setting. Observed factors may include population, setting, and conditions of usual care or intervention implementation.

Recalibration involves the development of methods to account for the similarity between a specific population of interest and a study's control conditions (including levels of naturally occurring change in the phenomenon of interest), and to account for this similarity in calculating the effect size of interest. This is analogous in principle to reweighting secondary survey data to match key marginal distributions in the population (Kish, 1990). Work has been undertaken elsewhere to develop methods that allow the results from unrepresentative experiments to be generalised to wider populations, albeit with differing levels of success (O'Muircheartaigh and Hedges, 2014, Hedges, 2013); recalibration would involve developing a similar package of methods for assessing and reflecting the generalisability of the results through re-estimating the effect size to apply to a population that more closely resembles the inference population.

A worked example of what recalibration methods could entail – focused on transferability

The following is an example of the *kind* of recalibration that might be developed, although it has not undergone formal evaluation and therefore the results that follow are purely for illustrative purposes. Anderson et al. (2015) conducted a review on the impact of community coalitions in reducing health disparities among ethnic minority populations. In addition to publishing evidence that allowed effect sizes to be calculated for behavioural outcomes, the primary studies generally reported the proportion of black and minority ethnic people, the proportion with low levels of education, and the proportion of women in the intervention or resident in the population (see table 2). In this example, we obtained the population estimates for these variables from analyses of the UK 2011 census (ONS, 2016), for three Local Authorities and two each of their constituent wards; these are also included in Table 2. We sought to explore the impact of recalibrating the effect size and upwardly weight the contribution of studies that were similar with respect to these variables, and downwardly weight those studies that differed.

TABLE 2 AROUND HERE

In the original review, pooled analyses were not conducted formally because of concerns about clinical heterogeneity (Anderson et al., 2015); however, statistical heterogeneity in the model for behavioural outcomes was moderate (I^2) and the between study variance was not significantly different from zero. For the purposes of this example (not to be used as evidence of the effectiveness of community coalitions in changing behavioural outcomes), fixed-effect meta-analysis was conducted leading to a pooled standardised mean difference of 0.037 (95% CI: 0.006-0.069; referred to as global effectiveness from hereon in; Figure 1).

Next we sought to investigate the generalisability of this effect size with respect to population characteristics in different localities. We first estimated the dissimilarity between the studies and the Local Authorities (LAs), based on the three observed characteristics, and created a dissimilarity matrix which measured the Euclidean distance between each of our studies and the LAs. Euclidean distances are the basis of cluster analysis and sequence analysis and in this case, measure the root of sum-of-squares of differences between observations on variables. Cluster analysis would then progress to examine natural groupings in score values, although for our purposes the scores themselves are used in these analyses. To reduce the dominance of any single characteristic in the distance calculations, all three variables were first standardised. After constructing the matrix, the inverse of squared dissimilarity scores[1] (measuring the distance between a single study and an individual area) were then used as scaling factors and multiplied by the inverse variance of the effect size for the study, the latter being the sole component of study weights in fixed effects models. In this sense, the results represent a series of further fixed effects models where both greater precision of effect size and greater similarity to the population of interest upwardly inflate the study's contribution to the overall pooled effect.

TABLE 3 AROUND HERE

[1] In calculating the inverse of the squared dissimilarity score, we aimed to achieve a greater balance between the variance and generalisability components of study weights.

The results (Table 3) suggest that when the study effect sizes are recalibrated to give higher priority to studies with demographic characteristics that are more similar to the UK LAs, the expected effect size (already very small), is weighted downwards slightly for Conwy, and has a negligible difference on the point estimates for Waltham Forest and Kensington, although in both cases the confidence interval is widened, suggesting that the scaling factor has given greater priority to studies with less precise results or studies where the effect size is more ambiguous or negative.

In the analyses of LAs' wards, upper and lower values with regards to ethnic composition (Waltham Forest and Kensington and Chelsea) and educational level (Conwy) were used to select wards of interest. The results show that while the overall pooled estimate changes little (a reflection of similarity in effect size), the composition of studies and their contribution to the pooled effect did change. In Chingford Green (Waltham Forest), for example (Figure 2), due to a higher similarity score, Brownson (1996) was weighted at 27% in the recalibrated model compared to 2% in the standard model. While the likely impact of the intervention might be small, a decision-maker in Waltham Forest may conclude that the likelihood of success of this intervention in Chingford Green is less certain when population characteristics (transferability) are considered and studies taking place in settings with similar characteristics are upwardly weighted. A similar procedure could be implemented when undertaking a random effects meta-analysis, with the within study variance, between study variance estimate (tau squared) and distance metric contributing to weighting the pooled effect size.

FIGURE 1 AND FIGURE 2 AROUND HERE

Related and future work on recalibration and effect sizes

A number of issues, both conceptual and statistical, need to be explored in refining the approach, and the above example is intended to serve as a basis for further enquiry. Furthermore, we do not suggest that recalibration would supersede conventional estimates and do not envisage recalibrated estimates would be presented in isolation of overall pooled effect sizes. In addition, an alternative route could see the a form of enhanced subgroup analysis take place, where instead of using the information from a distance matrix as a scaling factor, this could be used as the basis for creating subgroups of studies that are more and less similar to a given population to be used within conventional meta-analyses.

Issues that need to be explored further in the example provided include the impact of standardising variables across sites within the same dataset simultaneously; exploration of different methods of estimating ‘distance’ between the included studies and the proposed setting; exploring ways of accounting for precision in the study generalisability estimates; exploring ways of weighting variables in distance calculations; refining the way in which distance estimates are incorporated into study weights; and exploring conceptual issues around the choice of variables in estimating ‘generalisability’. It is perhaps this latter consideration that ultimately remains most fundamental, given that causal generalisation will always be more complicated than assessing the likelihood that a relationship is likely to be causal (Shadish et al., 2002).

Applying differential weighting to reflect factors other than heterogeneity is not entirely without precedent. Turner and colleagues (2009) outlined a number of strategies for incorporating forms of bias including into study weighting (see also Martins and Yang, 2009 for an example involving incorporating journal rankings into study weights); although conducting such an exercise through a lens of generalisability and utilising existing data is a novel contribution of this paper.

While there are many avenues to explore, and likely several methodological obstacles to overcome, the ideas presented offer a viewpoint on how generalisability, which implicitly involves considering several variables simultaneously, can be operationalised to explore the results of meta-analyses. We also envisage that such an approach can be extended to considering how data from control arms in trials can be harmonised with data from ‘naturally occurring controls’ that are collected in large secondary datasets. Of particular interest would be to examine levels of change over time within these control populations, and their similarity to changes occurring in inference populations, as a source of data to aid making valid inferences. Where individual level data are collected, these may lend themselves to the use of propensity score matching; where aggregate data alone are collected, we may view populations that are more similar to one another (and an inference population) as belonging to the same unobserved latent group and those that differ as belonging to a different latent group, enabling a form of enhanced sub-group analyses.

Related work on the use of secondary data – survey data and ‘real world’ data – to further enhance, understand and explore the generalisability of meta-analysis has been conducted. For example, Kriston and Meister (2014) explored incorporating subjective judgements on applicability by clinicians into meta-analysis estimates, initially focussed on the diagnosis of included patients or form of comparator treatment. Propensity score matching on population characteristics has been proposed as a means of enhanced assessment of generalisability of a single trial to a defined population (Stuart et al., 2015). Recent work by Riley and colleagues (2016) explored the use of real-world data in the further validation of clinical prediction models, for example in exploring the consistency in prediction across clinically relevant subgroups. Finally, Efthimiou and colleagues (2017) have proposed utilising information from non-randomised studies alongside information from randomised controlled studies in network meta-analyses, in order to better understand the likely impacts of treatments in real-world settings and to develop an understanding of the effectiveness as well as the efficacy of treatment. These non-randomised studies included observational studies that aimed to establish the comparative effectiveness of treatment, as well as comparative clinical trials that do not employ randomization; this is distinct from our own concern here where the observational data may not (and are very likely not to) capture treatment status, but capture important contextual information pertinent to transferability and applicability. The ideas proposed here are also distinct from, but potentially complementary to, Bayesian meta-analysis, although given that the intervention would usually be unobserved in the population in which the evidence is to be applied, priors derived from such data and applied to Bayesian models may uninformative in these models (Lewis and Nair, 2015, Sutton and Abrams, 2001).

While distinct from the approaches advocated in this paper, these emerging approaches (and established in the case of Bayesian) are complementary to our own arguments, and contribute to the debate that developments in the availability of data should also be matched with developments in the methods to utilise these data in order to support decision-making and enhance the portability of evidence.

Summary and Conclusions

Synthesising effect sizes from individual studies through meta-analysis increases statistical power compared to individual studies, although this alone is not equivalent to improving the generalisability of evidence. Many situations may occur where the population of studies included within a review may be replicated across different study settings, but where these settings differ in observed and unobserved ways from a given inference population or setting. Understanding usual care and patterns of naturally occurring change is an important and overlooked component in understanding the mechanics of the intervention itself.

In addition, although advanced methods have been developed to correct for under- or over-representation of cases in primary research, considering whether the sample of studies reflects the intended population of studies (beyond considering the size/effect through assessing publication bias) and correcting for this potential disparity is not widely practiced in meta-analysis. The challenge raised here is to ensure that the generalisability of evidence is maximised so that the methodological development of approaches to refining our use of evidence is seen as a natural adjunct to the process of aiding the decision-making process. Accounting for generalisability of evidence and understanding its application may also mean challenging the whole emphasis of systematic reviewing, away from accounts of whether interventions 'work' and towards reporting 'what happened' when interventions were trialled (Petticrew, 2015). Furthermore, concern about the generalisability of meta-analytic evidence exposes tension about the very purpose of systematic reviews. Some may regard systematic reviews as an academic tool for 'taking stock' of the available evidence for a given research question, while others may regard systematic reviews as an applied tool to support decision-making practice. The methods proposed here implicitly locate systematic reviews as tools to be applied across different contexts, although whether establishing and enhancing the generalisability of review and meta-analytic evidence is a role solely for systematic reviewers, decision-makers, or a joint role, is a debate that is beyond the remit of the current paper. Where there is less doubt is that successful implementation of any of the approaches advocated here is likely to involve much closer collaboration in the production of systematic reviews than is often the case currently.

Much of the public health systematic review evidence may be produced primarily with the interests of the ‘global scientist’ at heart. For decision-makers in English public health and elsewhere, this contrasts with the prevailing trend of localised decision-making structures. In this paper, we have advocated the development of methods and approaches based on the analysis of existing secondary data sources to understand the way in which meta-analytic evidence is reflective of locality and context – including epidemiological patterns, population characteristics, agencies and structures. Further development of these approaches could bridge the gulf between the needs of decision-makers for evidence that is locally salient and meta-analytic evidence which is often abstract and global. The development of approaches that facilitate meta-analytic evidence to better meet the concerns of public health decision-makers is also aligned with other recently developed approaches in qualitative evidence synthesis that place decision-makers’ needs for evidence on ‘what works for whom and in what circumstances’ as a central focus of enquiry. This is an objective that is central to Realist Systematic Reviews (Pawson et al., 2005), although these approaches do not focus on quantifying the extent to which the magnitude of the effect size is expected to vary across different settings, as is the case for the methods propsed here.

We also call for the development of a unified and cohesive conceptual framework for assessing the generalisability of review evidence: it is notable that new frameworks for assessing the quality of reviews still do not mandate the assessment of the generalisability of findings (Whiting et al., 2015). There may be good reason for this, in that the application of tools to assess many components of generalisability may only be appropriate in the presence of a clearly defined target group. There may be more, however, that systematic reviewers can do to aid decision-making audiences to assess generalisability – for instance through clearly describing, disaggregating and synthesising (e.g. conducting sub-group analyses, where appropriate) evidence on usual care, control conditions and other aspects of context that tend to be sparsely reported. For example, it is common to describe ‘usual care’ as the control condition in systematic reviews (and primary trials) without describing what this usual care entails, providing little opportunity for understanding the applicability and transferability of results.

The recommendations above are dependent on the publication of primary trial results that include sufficient detail to allow readers and reviewers to explore differences in implementation, the relationship with context, and the impact of the design of the intervention in examining effectiveness. Previous calls for process evaluations to become integral elements in the reporting of trials (Bonell et al., 2006) appear to have gone unanswered among the majority of trialists, resulting in a paucity of reviews that synthesise process data. Ultimately, changing this status quo may require action from funders to limit funding only to those trials which undertake a process evaluation. While only discussed briefly here, individual participant data (IPD) meta-analysis is likely to offer significantly greater opportunity to understand how participant factors may interact with both engagement with public health interventions and how these factors moderate outcomes, thereby enhancing the potential to understand the generalisability of findings. However, currently, there is little incentive for public health trialists to publish individual data for subsequent reanalysis, and until this changes, the potential utility of IPD meta-analyses in enhancing the generalisability of evidence is likely to remain unfulfilled.

In an age where sources of public health data are proliferating, further secondary data analysis may represent an important but overlooked means of enhancing the generalisable properties of evidence from meta-analysis. Generalising evidence of effectiveness is a delicate process, one that is fraught methodologically and philosophically (Guala, 2010), but in the case of public health interventions, one that remains open to interpretation. Some of the approaches we discuss – and particularly using preliminary secondary data analysis to help shape the review – could ostensibly challenge the potential of systematic reviews and meta-analyses to become 'global' summaries. However, none of the approaches that we are suggesting should interfere with the process of discovery of evidence or in terms of drafting inclusion criteria. Similarly, none of the approaches advocate excluding studies that meet the inclusion criteria, but that otherwise may have low generalisability as this would restrict comparative analyses between studies with higher and lower levels of generalisability. Instead, we are calling for – using secondary data analysis as a primary vehicle – a more nuanced treatment of the contexts in which interventions take place and the development of methods that can utilise this information to enhance the portability of public health review evidence for localised decision-making.

References

AHMAD, N., BOUTRON, I., DECHARTRES, A., DURIEUX, P. & RAVAUD, P. 2010. Applicability and generalisability of the results of systematic reviews to public health practice and policy: a systematic review. *Trials,* 11**,** 20.

ANDERSON, L. M., ADENEY, K. L., SHINN, C., SAFRANEK, S., BUCKNER-BROWN, J. & KRAUSE, L. K. 2015. Community coalition-driven interventions to reduce health disparities among racial and ethnic minority populations. *The Cochrane Library*.

BERLIN, J. A. & GOLUB, R. M. 2014. Meta-analysis as evidence: building a better pyramid. *JAMA,* 312**,** 603-606.

BONELL, C., JAMAL, F., MELENDEZ-TORRES, G. J. & CUMMINS, S. 2014. "Dark logic": theorising the harmful consequences of public health interventions. *Journal of epidemiology and community health,* 69**,** 95-98.

BONELL, C., OAKLEY, A., HARGREAVES, J., STRANGE, V. & REES, R. 2006. Assessment of generalisability in trials of health interventions: suggested framework and systematic review. *BMJ,* 333**,** 346-349.

BOSLAUGH, S. 2007. *Secondary data sources for public health: A practical guide*, Cambridge University Press.

BROWNSON, C. A., DEAN, C., DABNEY, S. & BROWNSON, R. C. 1998. Cardiovascular risk reduction in rural minority communities: The Bootheel Heart Health Project. *journal of health education,* 29**,** 158-165.

BROWNSON, R. C., BAKER, E. A., BOYD, R. L., CAITO, N. M., DUGGAN, K., HOUSEMANN, R. A., KREUTER, M. W., MITCHELL, T., MOTTON, F. & PULLEY, C. 2004. A community-based approach to promoting walking in rural areas. *American journal of preventive medicine,* 27**,** 28-34.

BROWNSON, R. C., SMITH, C. A., PRATT, M., MACK, N. E., JACKSON-THOMPSON, J., DEAN, C. G., DABNEY, S. & WILKERSON, J. C. 1996. Preventing cardiovascular disease through community-based risk reduction: the Bootheel Heart Health Project. *American Journal of Public Health,* 86**,** 206-213.

BURCHETT, H., BLANCHARD, L., KNEALE, D. & THOMAS, J. under review. Assessing the applicability of public health intervention evaluations from one setting to another: A methodological study of the usability and usefulness of assessment tools and frameworks *Health Research Policy and Systems*.

BURCHETT, H., UMOQUIT, M. & DOBROW, M. 2011. How do we know when research from one setting can be useful in another? A review of external validity, applicability and transferability frameworks. *Journal of health services research & policy,* 16**,** 238-244.

CAMBON, L., MINARY, L., RIDDE, V. & ALLA, F. 2012. Transferability of interventions in health education: a review. *BMC public health,* 12**,** 497.

CHANDLER, J., THOMAS, J., SUTCLIFFE, K., KAHWATI, L. & KNEALE, D. 2017. Applying current philosophical insights on causality using Qualitative Comparative Analysis as an additional synthesis in systematic reviews to address complex interventions. *Gobal Evidence Summit.* Cape Town, South Africa.

CONNELLY, R. & PLATT, L. 2014. Cohort profile: UK millennium Cohort study (MCS). *International journal of epidemiology,* 43**,** 1719-1725.

DONALDSON, S. I., STREET, G., SUSSMAN, S. & TOBLER, N. 2001. Using meta-analyses to improve the design of interventions. *In:* SUSSMAN, S. (ed.) *Handbook of program development for health behavior research and practice.* Thousand Oaks, California: Sage.

DYRVIG, A.-K., KIDHOLM, K., GERKE, O. & VONDELING, H. 2014. Checklists for external validity: a systematic review. *Journal of evaluation in clinical practice*.

EFTHIMIOU, O., MAVRIDIS, D., DEBRAY, T., SAMARA, M., BELGER, M., SIONTIS, G., LEUCHT, S. & SALANTI, G. 2017. Combining randomized and non-randomized evidence in network meta-analysis. *Statistics in medicine*.

GLASS, G. V. 2000. *Meta-analysis at 25* [Online]. Available: http://www.gvglass.info/papers/meta25.html [Accessed 26th November 2014].

GUALA, F. 2010. Extrapolation, analogy, and comparative process tracing. *Philosophy of Science,* 77**,** 1070-1082.

GUALA, F. & MITTONE, L. 2005. Experiments in economics: External validity and the robustness of phenomena. *Journal of Economic Methodology,* 12**,** 495-515.

HARDEN, A. & GOUGH, D. 2012. Quality and relevance appraisal. *In:* GOUGH, D., OLIVER, S. & THOMAS, J. (eds.) *An introduction to systematic reviews.* London: Sage.

HARRIS, K. M., KNEALE, D., LASSERSON, T. J., MCDONALD, V. M., GRIGG, J. & THOMAS, J. in press. School-based self management interventions for asthma in children and adolescents: a mixed methods systematic review. *The Cochrane Library*.

HAYANGA, B. 2017. Are mainstream interventions that target social isolation and loneliness effective for older Black and Minority Ethnic individuals? *UCL Institute of Education Doctoral Conference.* London.

HEDGES, L. V. 2013. Recommendations for practice: justifying claims of generalizability. *Educational Psychology Review,* 25**,** 331-337.

KATZ, D., O'CONNELL, M., NJIKE, V. Y., YEH, M. & NAWAZ, H. 2008. Strategies for the prevention and control of obesity in the school setting: systematic review and meta-analysis. *International journal of obesity,* 32**,** 1780-1789.

KISH, L. Weighting: Why, when, and how. Proceedings of the survey research methods section, American Statistical Association, 1990. 121-130.

KLOEK, G. C., VAN LENTHE, F. J., VAN NIEROP, P. W., KOELEN, M. A. & MACKENBACH, J. P. 2006. Impact evaluation of a Dutch community intervention to improve health-related behaviour in deprived neighbourhoods. *Health & place,* 12**,** 665-677.

KNEALE, D., KHATWA, M. & THOMAS, J. 2016. Identifying and appraising promising sources of UK clinical, health and social care data for use by NICE EPPI Centre, UCL Institute of Education, University College London.

KNEALE, D., O'MARA-EVES, A. & THOMAS, J. under review. It's a mean world for generalisability. *Americal Journal of Public Health*.

KNEALE, D., ROJAS-GARCÍA, A., RAINE, R. & THOMAS, J. 2017. The use of evidence in English local public health decision-making. *Implementation Science,* 12**,** 53.

KRISTON, L. & MEISTER, R. 2014. Incorporating uncertainty regarding applicability of evidence from meta-analyses into clinical decision making. *Journal of clinical epidemiology,* 67**,** 325-334.

LARSON, C. O., SCHLUNDT, D. G., PATEL, K., WANG, H., BEARD, K. & HARGREAVES, M. K. 2009. Trends in smoking among African–Americans: A description of Nashville's REACH 2010 Initiative. *Journal of community health,* 34**,** 311-320.

LEWIN, S., HENDRY, M., CHANDLER, J., OXMAN, A. D., MICHIE, S., SHEPPERD, S., REEVES, B. C., TUGWELL, P., HANNES, K. & REHFUESS, E. A. 2017. Assessing the complexity of interventions within systematic reviews: development, content and use of a new tool (iCAT_SR). *BMC medical research methodology,* 17**,** 76.

LEWIS, M. G. & NAIR, N. S. 2015. Review of applications of Bayesian meta--analysis in systematic reviews. *Global Journal of Medicine and Public Health,* 4**,** 1-9.

LORENC, T. & OLIVER, K. 2013. Adverse effects of public health interventions: a conceptual framework. *Journal of epidemiology and community health***,** jech-2013-203118.

MARTINS, P. S. & YANG, Y. 2009. The impact of exporting on firm productivity: a meta-analysis of the learning-by-exporting hypothesis. *Review of World Economics,* 145**,** 431-445.

NICE 2013. Weight management: lifestyle services for overweight or obese children and young people (PH47). London: National Institute for Health and Care Excellence.

NØKLEBY, H. & MUNTHE-KAAS, H. 2017. Using the TRANSFER framework for assessing transferability of review findings – a case study. *Global Evidence Summit.* Cape Town, South Africa.

O'MUIRCHEARTAIGH, C. & HEDGES, L. V. 2014. Generalizing from unrepresentative experiments: a stratified propensity score approach. *Journal of the Royal Statistical Society: Series C,* 63**,** 195-210.

ONS. 2016. *2011 Census aggregate data* [Online]. UK Data Service (Edition: June 2016). Available: http://dx.doi.org/10.5257/census/aggregate-2011-1 [Accessed 12/9 2016].

ORTON, L., LLOYD-WILLIAMS, F., TAYLOR-ROBINSON, D., O'FLAHERTY, M. & CAPEWELL, S. 2011. The use of research evidence in public health decision making processes: systematic review. *PloS one,* 6**,** e21704.

PAWSON, R., GREENHALGH, T., HARVEY, G. & WALSHE, K. 2005. Realist review-a new method of systematic review designed for complex policy interventions. *Journal of health services research & policy,* 10**,** 21-34.

PETTICREW, M. 2015. Time to rethink the systematic review catechism? Moving from 'what works' to 'what happens'. *Systematic reviews,* 4**,** 36.

PETTICREW, M., TUGWELL, P., KRISTJANSSON, E., OLIVER, S., UEFFING, E. & WELCH, V. 2011. Damned if you do, damned if you don't: subgroup analysis and equity. *Journal of epidemiology and community health*.

PIKE, K. C., HARRIS, K. M. & KNEALE, D. 2016. Interventions for autumn exacerbations of asthma in children. *The Cochrane Library*.

PLESCIA, M., HERRICK, H. & CHAVIS, L. 2008. Improving health behaviors in an African American community: the Charlotte Racial and Ethnic Approaches to Community Health project. *American Journal of Public Health,* 98**,** 1678-1684.

RILEY, R. D., ENSOR, J., SNELL, K. I., DEBRAY, T. P., ALTMAN, D. G., MOONS, K. G. & COLLINS, G. S. 2016. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *bmj,* 353**,** i3140.

RILEY, R. D., LAMBERT, P. C. & ABO-ZAID, G. 2010. Meta-analysis of individual participant data: rationale, conduct, and reporting. *Bmj,* 340.

RYCHETNIK, L., FROMMER, M., HAWE, P. & SHIELL, A. 2002. Criteria for evaluating evidence on public health interventions. *Journal of epidemiology and community health,* 56**,** 119-127.

SCHLUNDT, D. G., NIEBLER, S., BROWN, A., PICHERT, J. W., MCCLELLAN, L., CARPENTER, D., BLOCKMON, D. & HARGREAVES, M. 2007. Disparities in smoking: data from the Nashville REACH 2010 project. *The Journal of ambulatory care management,* 30**,** 150-158.

SHADISH, W. R., COOK, T. D. & CAMPBELL, D. T. 2002. *Experimental and quasi-experimental designs for generalized causal inference*, Houghton, Mifflin and Company.

SIM, I. 2015. The Uneven Future of Evidence Based Medicine. *Cochrane Colloquium.* Vienna, Austria.

STUART, E. A., BRADSHAW, C. P. & LEAF, P. J. 2015. Assessing the generalizability of randomized trial results to target populations. *Prevention Science,* 16**,** 475-485.

SUTTON, A. J. & ABRAMS, K. R. 2001. Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research,* 10**,** 277-303.

THOMAS, J., O'MARA-EVES, A., KNEALE, D. & SHEMILT, I. 2017. Synthesis Methods for Combining and Configuring Quantitative Data. *In:* GOUGH, D., OLIVER, S. & THOMAS, J. (eds.) *An Introduction to Systematic Reviews.* London: Sage.

THOMPSON, S. G. & HIGGINS, J. P. 2005. Can meta-analysis help target interventions at individuals most likely to benefit? *The Lancet,* 365**,** 341-346.

TURNER, R. M., SPIEGELHALTER, D. J., SMITH, G. & THOMPSON, S. G. 2009. Bias modelling in evidence synthesis. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 172**,** 21-47.

VERMA, A., TORUN, P., HARRIS, E., EDWARDS, R., GEMMELL, I., HARRISON, R. A., BUCHAN, I. E., DAVIES, L., PATTERSON, L. & HELLER, R. F. 2012. Population Impact Analysis: a framework for assessing the population impact of a risk or intervention. *Journal of public health,* 34**,** 83-89.

VU, T., FARISH, S., JENKINS, M. & KELLY, H. 2002. A meta-analysis of effectiveness of influenza vaccine in persons aged 65 years and over living in the community. *Vaccine,* 20**,** 1831-1836.

WAGNER, E. H., KOEPSELL, T. D., ANDERMAN, C., CHEADLE, A., CURRY, S. G., PSATY, B. M., VON KORFF, M., WICKIZER, T. M., BEERY, W. L. & DIEHR, P. K. 1991. The evaluation of the henry j. kaiser family foundation's community health promotion grant program: Design. *Journal of clinical epidemiology,* 44**,** 685-699.

WAGNER, E. H., WICKIZER, T. M., CHEADLE, A., PSATY, B. M., KOEPSELL, T. D., DIEHR, P., CURRY, S. J., VON KORFF, M., ANDERMAN, C. & BEERY, W. L. 2000. The Kaiser

Family Foundation Community Health Promotion Grants Program: findings from an outcome evaluation. *Health Services Research,* 35**,** 561.

WANG, S., MOSS, J. R. & HILLER, J. E. 2006. Applicability and transferability of interventions in evidence-based public health. *Health promotion international,* 21**,** 76-83.

WHITING, P., SAVOVIĆ, J., HIGGINS, J. P., CALDWELL, D. M., REEVES, B. C., SHEA, B., DAVIES, P., KLEIJNEN, J. & CHURCHILL, R. 2015. ROBIS: A new tool to assess risk of bias in systematic reviews was developed. *Journal of clinical epidemiology*.

WICKIZER, T. M., WAGNER, E., CHEADLE, A., PEARSON, D., BEERY, W., MAESER, J., PSATY, B., VONKORFF, M., KOEPSELL, T. & DIEHR, P. 1998. Implementation of the Henry J. Kaiser family foundation's community health promotion grant program: A process evaluation. *The Milbank Quarterly,* 76**,** 121-147.

Table 1: Potential avenues of enquiry in exploring the results of systematic reviews and meta-analysis

	Evidence from Systematic Reviews	**Use of Secondary Data Analysis (SDA) in local population**
Refining targeting strategies	Systematic Review and/or Pooled Effect Size shows overall intervention effectiveness	Indicate impact of 'pooled' change estimate across groups in population.
	Sub-group analyses (shows different impacts across groups)	(i) examine occurrence of sub-groups; (ii) simulate/indicate impact of disaggregated change estimate across groups in population.
	Meta-regression (decomposes heterogeneity)	(i) examine occurrence of factors associated with reduction of heterogeneity; (ii) simulate/indicate impact of disaggregated change estimate across groups in population.
Understand control conditions in local contexts	Systematic Review and/or Pooled Effect Size shows overall intervention effectiveness	(i) establish usual care delivered and comparability; (ii) estimate comparability of natural change in local population.
Understanding the applicability of interventions	Systematic Review and/or Pooled Effect Size shows overall intervention effectiveness; review narratively describes procedures for implementing intervention	(i) assess applicability of evidence in terms of necessary pre-conditions for intervention; (ii) assess equity of access to intervention across social groups; (iii) explore risk profiles of non-adherent participants from review evidence; (iv)
	Sub-group analyses and meta-regression (shows different impacts across groups)	(i) assess equity of intervention impacts across groups
Evaluate existing policy based on systematic reviews and meta-analysis	Systematic Review and/or Pooled Effect Size shows overall intervention effectiveness	Examine if intervention had expected impacts and explore why the intervention impacts differed from those described in meta-analysis
Test or assessing part or all of the conceptual model underlying an intervention	Systematic Review and/or Pooled Effect Size shows overall intervention effectiveness and underlying conceptual model	(i) test a theory of change within an inference population; (ii) test a particular pathway or mechanism within an inference population

Table 2: Effect sizes from Anderson et al. (2015) and UK census data

Study	Effect Size	Standard Error	Black and Minority Ethnic (%)	Low Education (%)	Female (%)
(Brownson et al., 1996)[1]	-0.044	0.081	11.4%	25.1%	54.0%
(Brownson et al., 2004)	-0.006	0.057	32.8%	25.1%	75.4%
(Kloek et al., 2006)	0.086	0.047	23.7%	23.6%	53.0%
(Larson et al., 2009)[2]	0.032	0.014	52.2%	33.7%	65.2%
(Plescia et al., 2008)	0.129	0.048	95.0%	22.9%	63.4%
(Wagner et al., 2000)[3]	0.013	0.036	39.8%	40.6%	57.0%
(Wagner et al., 2000)[3] (2000b)	-0.01	0.093	8.8%	8.8%	57.0%
Waltham Forest			47.8%	20.8%	50.0%
Conwy			2.3%	25.6%	51.6%
Kensington & Chelsea			29.4%	10.1%	50.7%
Waltham Forest – Lea Bridge			64.8%	23.0%	49.8%
Waltham Forest – Chingford Green			14.7%	24.2%	52.1%
Conwy – Towyn			1.1%	41.1%	52.9%
Conwy – Uwch Conwy			1.3%	17.8%	47.8%
Kensington – Earl's Court			52.1%	8.4%	47.6%
Kensington – Royal Hospital			15.3%	9.0%	50.0%

Notes: [1]Demographic data based weighted averages of demographic data found in companion paper (Brownson et al., 1998); [2]Demographic data based on surveys collected in (Schlundt et al., 2007) with data for low education based on data for smokers. [3]Demographic data for first estimate focussed on adult sites A and G and described in (Wickizer et al., 1998, Wagner et al., 1991); for second estimate focussed on B, C and I as set out in original papers.

Table 3: Examples of recalibrated effect sizes

Model	Pooled estimate (SMD)	95% Confidence interval
Global effectiveness (Fixed effect)	*0.037*	*0.006-0.069*
Waltham Forest recalibrated	0.038	0.014-0.062
Waltham Forest (Lea Bridge) recalibrated	0.040	0.016-0.064
Waltham Forest (Chingford Green) recalibrated	0.040	-0.026-0.106
Conwy recalibrated	0.033	0.005-0.060
Conwy (Towyn) recalibrated	0.030	0.004-0.056
Conwy (Uwch Conwy) recalibrated	0.037	0.002-0.073
Kensington recalibrated	0.036	0.012-0.060
Kensington (Earl’s Court) recalibrated	0.040	0.015-0.065
Kensington (Royal Hospital) recalibrated	0.034	-0.001-0.070

Note. SMD = standardised mean difference.