Natural experiments for the evaluation of place-based public health interventions: a methodology scoping review

Patricia N Albers[1], Chiara Rinaldi[2], Heather Brown[3], Kate E Mason[4,5], Katrina d'Apice[1], Elizabeth McGill[2], Cheryl McQuire[1,6], Peter Craig[7], Anthony A Laverty[8], Morgan Beeson[9], Mhairi Campbell[7], Matt Egan[2], Marcia Gibson[7], Maxwell Fuller[1], Amy Dillon[1], David Taylor-Robinson[10], Russell Jago[1,11,12,], Kate Tilling [1,13], Benjamin Barr[14], Falko F Sniehotta[15,16], Matthew Hickman[1], Christopher J Millett[8], Frank de Vocht[1,11]

[1]Population Health Sciences, Bristol Medical School, the University of Bristol. Bristol, United Kingdom
[2]Department of Public Health, Environments and Society, London School of Hygiene and Tropical Medicine, United Kingdom
[3]Health Research, Lancaster University, United Kingdom
[4]Department of Public Health Policy & Systems, University of Liverpool
[5]Centre for Health Policy, University of Melbourne
[6]National Institute for Health and Care Research (NIHR) School for Public Health Research (SPHR)
[7]MRC/CSO Social and Public Health Sciences Unit, School of Health and Wellbeing, University of Glasgow.
[8]School of Public Health, Imperial College London
[9]Newcastle University Business School, Newcastle University
[10]Department of Public Health, Policy and Systems. University of Liverpool
[11]NIHR Applied Research Collaboration West (NIHR ARC West)
[12]Centre for Exercise, Nutrition & Health Sciences, School for Policy Studies, University of Bristol
[13]MRC Integrative Epidemiology Unit, University of Bristol
[14]Institute of Population Health, University of Liverpool
[15]NIHR Policy Research Unit Behavioural Science, Newcastle University
[16]Department of Public Health, Social and Preventive Medicine, Medical Faculty Mannheim, Heidelberg University.

Corresponding author: Dr Frank de Vocht. Population Health Sciences, Bristol Medical School, University of Bristol. Canynge Hall, 39 Whatley Road, Bristol, United Kingdom. BS8 2PS. Email: frank.devocht@bristol.ac.uk. Tel: +44.(0)117.928.7239. Fax: N/A

ABSTRACT

Place-based public health evaluations are increasingly making use of natural experiments. This scoping review aimed to provide an overview of the design and use of natural experiment evaluations (NEEs), and an assessment of the plausibility of the *as-if* randomisation assumption. A systematic search of three bibliographic databases (Pubmed, Web of Science and Ovid-Medline) was conducted in January 2020 to capture publications that reported a natural experiment of a place-based public health intervention or outcome. For each, study design elements were extracted. An additional evaluation of *as-if* randomisation was conducted by 12 of this paper's authors who evaluated the same set of 20 randomly selected studies and assessed '*as-if*' randomisation for each. 366 NEE studies of place-based public health interventions were identified. The most commonly used NEE approach was a Difference-in-Differences study design (25%), followed by before-after studies (23%) and regression analysis studies. 42% of NEEs had likely or probably *as-if* randomisation of exposure (the intervention), while for 25% this was implausible. An inter-rater agreement exercise indicated poor reliability of *as-if* randomisation assignment. Only about half of NEEs reported some form of sensitivity or falsification analysis to support inferences. NEEs are conducted using many different designs and statistical methods and encompass various definitions of a natural experiment, while it is questionable whether all evaluations reported as natural experiments should be considered as such. The likelihood of *as-if* randomisation should be specifically reported, and primary analyses should be supported by sensitivity analyses and/or falsification tests. Transparent reporting of NEE designs and evaluation methods will contribute to the optimum use of place-based NEEs.

Key Words: Public Health; Public Health Policy; Natural Experiments; Quasi Experiments; Evaluations; Place-based

1. INTRODUCTION

'Place-based public health interventions' can be described as activities, programmes, or policies delivered at a local, regional, or national level, that aim to improve health or reduce health inequalities of the community. 'Place' is study of space or places, and in health often specifically refers to the places where people are born, grow, live, work, and age [1]. It is well established that contextual factors such as 'place' play a role in creating and perpetuating health inequities [2]. Moreover, place-based interventions are an important policy lever for influencing the wider determinants of health [3]. Community and place-based public health interventions are therefore common: a recent umbrella review found 13 systematic reviews, covering 51 unique studies (published since 2008) looking at the effectiveness of place-based interventions to improve public health and reduce health inequalities [4]. One of the most famous examples of a place-based natural experiment in public health is the that of John Snow and the Broad Street pump, where he showed that people obtaining water from that pump were being infected by cholera [5]. Other well-known examples of place-based natural experiments at the national level are the evaluation of the effects of the 'Dutch Hunger Winter' at the end of World War 2, in which food was scarce in the occupied West of the Netherlands, but not in the liberated South [6]; the effects of the US blockade of Cuba following the collapse of the Soviet Union [7]; and the impact of a drastic improvement in air pollution during the Beijing Olympics as a result of various place-based interventions such as reducing traffic [8]. These examples all demonstrate how the spatial location, or place, as the target of an intervention can positively impact on population health. A recent local level example is the evaluation of the impact of Transport for London's ban of fast food advertising on household purchases [9]. , or relies on observational studies [10, 11]. The missed opportunities (where place-based interventions were implemented but could not be evaluated by a randomised controlled trial (RCT)) have resulted in a lack of evidence-based policy, with many public health interventions only being supported by descriptive evidence from observational studies [12]. Indeed, for the above examples, as well as for many other evaluations of place-based public health interventions, it is not possible to conduct the traditionally preferred method of an RCT. RCTs have the distinct advantage for causal inference that, when properly carried out, randomisation of exposure allocation

minimises or eliminates residual confounding of the exposure-outcome association. Causal inference aims to deduce the influences of observed patterns under changing conditions for instance exploring the changes brought about by an external intervention. this goes beyond simply exploring associations [13]. However, whereas RCTs are traditionally considered the 'gold standard' in medicine due to their (at least theoretically) high internal validity, in other fields RCTs are less favoured due to low external validity.

For many (place-based) policy interventions therefore, there remains the opportunity for evaluation using quasi-experimental or natural experimental designs, which we refer to collectively as natural experiment evaluations (NEEs). Although the terms quasi and natural experiments are often used interchangeable, a stricter view of the two can be used, whereby in quasi-experiments there is no assumption of randomisation or *as-if* randomisation, while natural experiments take advantage of random or *as-if* random assignment [15]. The interest of public health in NEEs is relatively recent, but in cognate fields (such as economics) the use and development of NEE methodology has been much more pronounced [12]. NEEs can have additional advantages over RCTs for the evaluation of place-based public health interventions in that interventions may have already been implemented or are to be implemented by, for example, local councils, making it unethical or impractical to randomise. Additionally, the cost and scale of an RCT may be prohibitive. The value that NEEs have in the evaluation of public health policies, especially where they can be aligned to the policy cycle, has resulted in their increased use to inform public health policy [16].

NEEs combine features of i) RCTs, in which an event or process can be used to divide the population in exposed and unexposed groups (or different levels of exposure), and ii) observational studies, as the allocation of participants to a study group is not under the control of the researcher. The former, at least in theory, makes NEEs less susceptible to confounding than many other observational designs [17, 18]. Although there are a few variations in the definition of a natural experiment, an important commonality is that the exposure or intervention (and as such group allocation) are not controlled by the researcher [17, 19]. This lack of control over exposure assignment may complicate attempts to draw causal inferences, as it may be more difficult to determine whether the allocation of individuals or groups to

intervention or control group(s) is truly independent of the outcome [17, 19]. To improve the strength of causal statements from NEEs, Dunning [15] proposed a stricter definition of NEEs in which a study is only considered an NEE if the allocation of exposure is *as-if* random. *The as-if random assumption* posits that in some circumstances where exposure allocation has not been controlled by the researchers, allocation can be considered to originate from a random process, thereby replicating randomisation in an experiment (with studies having implausible *as-if* randomisation categorised as quasi experiments) [15]. A classic example of this is the US draft lottery which took place during the Vietnam War where men born between certain periods were assigned, at random, a lottery number based on their date of birth. All men with birth dates corresponding to the first 195 numbers were called to service, a similar process ran three times during the war period [15, 20]. The *as-if* random allocation of exposure however, is not easily, if ever, verifiable [14, 19], which has important implications for the extent to which causal claims can be made [17].

In its guidance on the use of natural experiments [16], the UK Medical Research Council (MRC) does not recommend particular designs or analytic methods for NEEs. In practice, the evaluation of NEEs is typically conducted using several different evaluation designs, including post-intervention comparison studies, pre-post studies, difference-in-differences (DiD) studies, regression discontinuity studies, or (controlled) interrupted time series studies [21], and various different statistical methods to analyse these, including regression adjustments, multilevel methods, propensity score matching or weighting, synthetic controls or the use of instrumental variables [21]. For illustration, in a recent review exploring the use of NEEs to improve public health evidence for obesity, 32% of included studies were pre-post longitudinal studies with a comparison group and four (8.5%) were interrupted time series studies, with the remainder using mostly descriptive methods such as repeated cross-sectional studies, post-test observational studies only, and longitudinal pre/post single group cohort studies [22].

To support causal inferences NEEs can benefit from additional sensitivity and falsification analyses, such as temporal falsification methods where intervention implementation times are adjusted, or spatial falsification methods using different groups (e.g. different geographical areas, for example) regardless

of the exposure, and/or by conducting methodological triangulation of using different statistical analysis [17]. Negative controls, where no impact from the intervention is expected, are also a valuable tool for identifying and correcting bias [23].

Because of the increasing importance of NEEs in the context of the evaluation of (placed-based) public health interventions, it is important to evaluate the methodological strength on which this evidence of effectiveness is based. We aimed to explore the practices and methods used, trends and developments in the use of methods and concepts, their strength for causal inference, and provide recommendations for addressing the main weaknesses in published place-based NEEs in public health.

2. MATERIALS & METHODS

Scoping reviews are typically chosen when the purpose of the review is to explore gaps in the subject matter, understand the extent of a body of literature, explain theories or constructs, or examine specific research practices [24, 25]. Scoping reviews are a process of summarising a body of evidence [26], but with key differences to systematic reviews that the quality of the included papers is not normally assessed formally and that no quantitative synthesis is conducted [26]. The current scoping review aims to understand the extent of the body of literature on place-based NEEs in public health, examine specific research practices, and provide recommendations for future improvements of such evaluations.

A systematic search of three bibliographic databases (Pubmed, Web of Science and Ovid-Medline) was conducted in January 2020 to capture publications that reported a natural experiment of a place-based public health intervention or outcome. The search strategy used a modified version of the one used in a similar study on the use of complex systems in place-based public health [27], but systems-level terms were replaced with natural experiment terms. Three separate searches using similar terms were conducted (details provided in Online Supplement Materials (OSM) Table S1). Search 1 included natural experiment and public health terms. Search 2 included natural experiment and evaluation and

public health terms, in order to identify additional studies classified by evaluation specific items not covered in Search 1. Search 3 similarly aimed to identify additional papers missed by Searches 1 and 2 by including public health terms with additional terms related to common study design and analytic methods used in NEEs. All three searches were run on all three databases, where titles and abstracts where searched. Due to resource limitations we limited the scoping review to studies published in the English language. For this review we defined public health as efforts to prevent illness, increase life, improve health and address health inequalities. Results were merged and duplicates removed before being added to Rayyan (https://www.rayyan.ai/).

Titles and abstracts were screened and marked for in/exclusion by a minimum of two reviewers. In the event of a difference of opinion a third reviewer made the final decision. Papers were excluded if the title or abstract included clinical and not public health outcomes, were based on Mendelian randomisation, or if the paper was a review, protocol, policy debate, or discussion piece. A study was included if it met the following eligibility criteria:

- Study self-identified as being an NEE, regardless of the actual methods use, allowing us to explore the scope of self-defined NEEs, and the predominant practices among them.
- Study evaluated a place-based public health intervention. We defined this as activities, programmes, or policies delivered at a local, regional, or national level, that aimed to improve health or reduce health inequalities of the community. Examples of such interventions include the opening of a fast food restaurant [28], or other changes to a community, or policy changes, for instance a change in policy on access to emergency contraception [29].
- Study reported an intervention which addressed a public health elevant topic. Some examples of health relevant topics include; substance use, physical activity, weight, nutrition, and mental health. We excluded clinical outcomes and papers using Mendelian randomisation as the latter are typically not conducted as part of place-based public health policy evaluation.
- Study needed to report empirical findings rather than hypothetical scenarios. Therefore reviews, protocols, policy debates, or discussion pieces we also not included.

Studies from any country were eligible, however, we limited the search to studies published in the English language.

A data extraction template was populated in Microsoft Excel and tested by PA and FdV using four papers, after agreement about the process, the remaining extraction was conducted using the final template. One member of the study (PNA, FdV, CR, HB, KEM, KD, EM, CM, PC, AAL, MB, MC, ME, MG, MF, AD, or DTR) team extracted an allocation of papers after which PA checked the completeness of the extractions and in places where the extracted data were incomplete or the author was unsure, PA completed these.

The extraction template with an explanation of what was collected in each field is provided in Table 1.

Table 1. Extraction template with details of what was collected in the different fields

This template collected the following data; *Year*, *Population*, *Intervention/Exposure evaluated*, *Individual/Aggregated data*, *As-if randomization* (Implausible/Possible/Probably/Likely), *Approach* (Instrumental variable/Synthetic control /Difference-in-differences (DiD)/Interrupted time series/Before-after comparison/After comparison/Other), *Comparator*, *Definition of comparator*, *Sensitivity/ Falsification* (No/Temporal (different timepoints)/Spatial (different cases/controls)/Both/Other), *Evaluation* (Local policy/National policy/Other), and *Aim* (Physical activity / Miscellaneous outcomes /Mental health/Nutrition/ Substance use/ Weight/Well-being/Other). 'Miscellaneous outcomes' was provided as a category for studies that could not easily be categorised or did not address a specific single outcome, and included for instance 'hospital admissions' or 'influenza prevention'. A brief description of the various approaches is provided below:

- Instrumental variable: using a variable that is only associated with the exposure and outcome and not other confounders or unmeasured error, these variables can often replicate randomisation.
- Synthetic control: allows a counterfactual to be constructed based on a weighted average of the outcome from a group similar to the exposed.

- Difference-in-differences: explores changes in the outcome from pre to post between the exposed and unexposed group but taking account of the trend of the unexposed group.
- Interrupted time series: chronological sequence of outcome data which is then interrupted by the exposure or intervention, the trend after the interruption is compared to the counterfactual.
- Before- after comparison: measuring the outcome in a group of participants before and after an intervention.
- After comparison: comparison of outcomes in an exposed and unexposed group.

We collected additional data on the statistical methodology using an open field for further details. Finally, we included a field for any additional notes of relevance not covered by the pre-specified categories. For some sections on the template the categories were not mutually exclusive, however, those carrying out the extraction selected the category the best aligned with how the author/s of that individual paper described their own study. For most sections we provided an '*Other*' category intended for use when a reviewer felt the study did not match one of the provided categories. We also provided an open-ended text field where the reviewer could write the details for the particular study. For example, if a reviewer believed the study was a time series study rather than an interrupted time series study.

The *Population, Intervention/Exposure evaluated, Comparator,* and *Definition of comparator* fields, as well as all the '*Othe*r' fields were all open-ended. These fields were manually coded into groups and themes that directly corresponded to the topics. Where a field did not completely match an existing code a new code was created. After coding, all the new codes were compared and where the underlying theme was the same, these were merged. When one of these open-ended fields were left blank they were coded as missing. Using the *Population* field, we coded the countries from where the studies reported data.

For the comparator and definition of comparator fields the open-ended responses were used to create a coded variable which included Control groups/ Control areas/ Pre-intervention (same population)/ Temporal controls (different population and different time points)/ Matched controls/ Synthetic controls or counterfactuals/ Sibling controls or control products/ No comparator. Where noted, we used the

terminology stated by the authors. Control groups refers to comparison groups that were not specifically matched, they were a selection of other (possibly related) groups, for example schools in the same region. Control areas refers to a selection of individuals from a particular area which was different to the primary study area, for example different countries. Pre-intervention controls refers to data from the same sample of individual before the intervention began. Temporal controls refers to a sample of people before the intervention, however, not necessarily the same sample of people. For example adult respondents of a survey before a particular event, compare to an adult sample of respondents to the same survey but after the event. Matched controls refers to controls where the authors specifically stated that they were matched on various criteria. Counterfactuals refers to controls whereby the study authors specifically explained the creation of a counterfactual based on pool of various control options. No comparator refers to studies that did not include any comparator or studies where authors compared groups of people with different levels of exposure.

The methodology used was categorised based on how the authors of each study described their approach (Instrumental variable/Synthetic control/DiD/Interrupted time series/Before-after comparison/After comparison/Other), and was as a result an amalgamation of designs and statistical methods. Studies that did not use one of these approaches were categorised as 'Other'. Two studies defined themselves as before-after studies and were subsequently coded as DiD studies, because despite the authors referring to these as before-after studies, they in fact describe a DiD study.

The plausibility of *as-if* randomisation in each study was judged by the reviewers who carried out the extraction. In the data extraction template plausibility was categorised in four alternatives (Implausible/Possible/Probably/Likely). *As-if* randomisation was assessed by each reviewer independently, where this was not provided by the study authors. In some situations where reviewers were uncertain an additional reviewer (PA) independently assessed this. There are currently no specific criteria or thresholds to establish '*as-if*' randomisation of exposure; therefore, we relied on the traditional methods used to assess the strength of observational studies to establish causality, such as possible exposure (self)selection and other risks of bias [19]. These often have to be inferred from

manuscripts rather than that they are explicitly stated by the authors, leaving room for subjective assessments. One of the key considerations we used to assess *as-if* randomisation was that participants did not have the capacity, information, or incentive to self-select into a study group. And similarly policy makers or other individuals assigning participants to groups also did not have the capacity, information, or incentive to assign individuals to a particular group.

Given that *as-if*-random is an essential criterion impacting on the strength of any causal conclusions [30], we conducted an additional exploratory exercise. Twelve of this paper's authors evaluated the same set of 20 natural experiments, randomly selected from the list of identified studies, and assessed '*as-if*' randomisation for each of them using the same categories used in the extractions template (Implausible/Possible/Probably/Likely). Intra-class correlations were calculated to determine inter-rater agreement using the R *irr* package based on two-way random effects where absolute agreement is important for single raters [31].

3. RESULTS

A total of 1,526 titles and abstracts were screened, resulting in 396 studies included for data extraction. An additional 29 studies were later excluded at the data extraction stage primarily because they were not related to public health or were a policy commentary or methodological paper. One further study was excluded because the full-text was not available, resulting in a total of 366 papers included for extraction. The flow diagram of the literature search is depicted in Figure 1 below, the total number of excluded studies per categories does not equal the total number of those excluded because studies may have been assigned more than one label. A table with all included references is provided in OSM Table S2.

Figure 1. PRISMA Flow Diagram. Note that the total number of excluded studies per categories does not equal the total number of those excluded because studies may have been assigned more than one label.

The annual numbers of published place-based NEEs of public health interventions from 1987 to 2019 are shown in Figure 2 and indicate a pattern of increase in their use from about 2009 onwards. Our search was run in early January 2020, 6 studies were picked up from 2020 but are not show in the plot below. For reference, the publication of the MRC guidance in 2012 [19] is shown by the dotted line in the figure.

Figure 2. Number of studies identified as 'natural experiment evaluations' by their authors up to December 2019 (6 studies from early 2020 included in this study, but not shown in this Figure to avoid erroneous inferences on trends). Dottedline indicates publication MRC guidance were published [19].

Included studies were from 54 mostly English-speaking countries (Table 1). One third of studies originated from the USA (33.3%, n=122), followed by the UK (13%), Australia (7%) and Canada (6%). In the USA, studies were predominantly DiD studies, while in the UK, Australia and Canada studies used predominantly before-after comparison studies.

3.1 Study characteristics

Nearly 67% (245) of included studies investigated the effect of a policy (national or local), while the remaining third were classified as evaluations of community initiatives. The two most frequently described type of evaluation explored the impact of neighbourhood infrastructure changes (33%), such as the addition of outlets or green space, and nutrition and diet interventions (13%). These were

followed by schooling/school policy or school environment changes (8%), health care changes (7%), and economic changes (6%) (see Table 2).

Table 2. Study characteristics of the included studies

The specific aims of the studies were categorised as seeking to evaluate the impact of a policy or other change onphysical activity, miscellaneous outcomes, mental health, nutrition, substance use, weight, well-being or other. Substance use (18.8%) and physical activity (13.3%) were the primary areas of research, with various miscellaneous outcomes interventions evaluated in 15.7% of included studies. A high proportion (30.1%, n=110) of studies did not align with the prelisted categories and from this four additional categories were identified. These included the evaluation of policies or other interventions aimed at impacting primarily pregnancy-related outcomes (8.7%), access to healthcare (3%), crime (2.7%), and mortality (2.7%).

3.2 Methodological characteristics

An overview of the methodological characteristics of the included studies is provided in Table 3. Most NEE studies (68%, n=248) were based on individual-level data for their primary analyses, with the remaining 32% based on aggregate data such as for example counts of incident cases.

The most commonly used NEE approach was the DiD study design (25%, n=93). This was followed by before-after comparison studies (23%, n=84) and interrupted time series studies (11%, n=40). Statistical approaches that have become more frequently used included instrumental variables (used in 6% of NEEs) and synthetic controls (used in 1.6%). 28% of studies recorded something other than these approaches, with 68.9% of these reporting regression-based analysis studies where no specific change in exposure is modelled and 12.6% reporting time series analysis, to model variation in exposure but without an element of the specific analysis of a change in exposure [17]. The remaining studies used ‘other econometric methods’ (2.9%) or descriptive analysis (1.9%). Only one study reported using propensity score matching, captured under ‘other’.

Table 3 shows most studies included a comparator either in the form of control groups (34%), control areas (25%), or used the same population before the intervention (16%). 13% of studies had no comparison group at all, and only 3 studies (<1%) used synthetic control or formal counterfactuals.

42.3% of all included studies were assessed as having likely (19.1%) or probable (23.2%) *as-if* randomisation, while as many as a quarter (24.9%) of NEEs were assessed as having 'implausible *as-if* randomisation'. 30% of NEEs were classified as 'possible *as-if* randomisation'. A small proportion of studies (2.8%) provided insufficient details for evaluation and were coded as "Unsure or cannot tell". The majority of studies with likely *as-if* randomisation were evaluating national policy (44%), with local policy evaluations being the least likely to be categorised as having likely *as-if* randomisation. (18%). There was no obvious pattern between country of publication and plausibility of *as-if* randomisation in study design.

Table 3. Methodological characteristics of the included studies

Table 4 shows the *as-if*-random classification for each of the study types. NEEs based on synthetic controls (n=6) were assessed as having the highest proportion of probable or likely *as-if* randomisation (83.3%) followed by instrumental variable studies(n=22; 50.0%) and DiD studies(n=93; 47.4%). Other methods and 'after comparisons' (n=18) included about 44% of studies with plausible or likely *as-if* randomisation, while for interrupted time series studies and before-after comparison studies only about 32% had plausible or likely *as-if* randomisation. On the other hand, 33% of 'after comparison' studies and about 30% of 'other' studies, interrupted time series studies, and before-after comparison studies were assessed as having implausible *as-if* randomisation.

Table 4. As-if randomisation by different statistical or design approach

Given that there are no specific criteria for assessing *as-if* randomisation we specifically set out to assess inter-rater agreement for *as-if* randomisation as this is a subjective classification that has nonetheless important implications for causal inference. *As-if* randomisation was not reported in the

studies included in the exercise. The validation exercise was completed for 18 of the 20 randomly selected studies by all 10 assessors (there were incomplete responses for two papers, which were excluded from the analysis). The corresponding intraclass correlation coefficient (ICC; 95% confidence interval) was 0.35 (0.19-0.47) indicating poor reliability. Complete agreement on the 5-point scale across all ten authors did not occur, with agreement within one unit occurring for only 6% of studies (details in of all individual assessments are provided in OSM). As such it worth noting the results presented in Table 4 are based on subjective assessments made by different authors of this paper.

Half (50%) of evaluations reported some form of sensitivity or falsification analysis. The most frequently used method was spatial falsification (17%), in which different control groups, generally geographical areas where an intervention was not implemented, are included as intervention areas (to assess whether any effect is unique to the intervention area). 10% of studies included a temporal falsification analysis in which alternate time points for the intervention were included (to assess whether any effect was unique to the intervention timepoint) [17]. Some studies reported using both spatial and temporal falsification analyses (8%). In addition, 16% of NEEs reported "other" methods, which was mainly the redefinition of model variables, adding additional variables to their models, or changes in the model specifications.

Sensitivity and falsification analyses stratified by statistical approach are shown in Table 5. Within the categories, the proportion of studies that did not include any falsification or sensitivity analyses varied from 17% to 55%. In contrast, although only six studies were included, 50% of NEEs using synthetic control methods used both temporal and spatial methods. Of the instrumental variable studies, 18% (n=4) reported using redefining variables in the models as a form of sensitivity analysis. For DiD studies about one in four studies (24 of 93) used spatial methods. For interrupted time series studies nearly equal proportions used temporal (12.5%, n=5), spatial (15%, n=6) or both (15%, n=6). Before-after comparison studies, after comparison studies, and 'other' studies had the lowest proportions of additional sensitivity or falsification analyses to strengthen causal inference; 30%, 33%, and 45%,

respectively. *Table 5. Percentage of sensitivity or falsification approaches used with different statistical approaches*

3.3 Weaknesses and limitations of included studies

The primary weakness that was evident among the included studies was that none considered whether the as-if randomisation assumption was satisfied and it was also challenging for readers to assess from the provided information whether this assumption was met. An additional weakness that we identified was the limited use of sensitivity analysis and falsification tests, which are important for strengthening causal inference.

4. DISCUSSION

In this review we aimed to explore the practices and methods used in place-based NEE evaluations. This review of specific place-based NEEs of public health interventions has highlighted that their use as an evaluation tool in public health has accelerated since 2009. This trend echoes a recent similar observation in relation to social policies [32]. We examined trends and developments in the use of methods and concepts for place-based NEEs, and their strength for causal inference. It was evident that there is lack of consistency in how a natural experiment was defined, and that studies sometimes self-identify as a natural experiment based on the characteristic that the intervention or the allocation of exposure was out of the researchers' control, regardless of the analysis approach used. However, a number of studies evaluated variation in exposure (similar to an observational study), rather than a specific change in exposure as a result of the occurrence/implementation of an intervention (akin to an RCT). This approach potentially weakens the potential for robust causal inference in many of the place-based NEE studies that we identified. We additionally observed that no studies considered whether the *as-if* randomisation assumption was satisfied, while this was also difficult to assess from the provided information. Combined with the relatively limited use of sensitivity analyses and falsification tests, this

has important implications for the extent to which such evaluations can be deemed to support causal inferences, and suggests an important way in which the reporting of NEEs could be improved.

Despite nearly 70% of the included studies evaluating a local or national policy, they employed a wide array of different study designs and analytic methods. 71% of studies used one of the well-defined natural experiment designs such as DiD studies, interrupted time series studies, before-after comparison studies, after comparison studies, instrumental variable studies, or synthetic control studies, but 29% used some other method; most often regression-based studies that explored variation in exposure and the change in exposure in the intervention group explicitly.

The exposure allocation process is critical to the design of natural experiment studies. This should mimic a random process as closely as possible to avoid biased inferences [17]. However, authors rarely reported the plausibility of the *as-if* randomisation assumption of exposure [15], leaving it to readers to assess this implicitly. Given that the *as-if* randomisation assumption cannot be formally tested [19], we relied on reporting by the authors and by assessment of the reviewer based on the information provided. The research team classified only a minority (~ 2 in 5) of studies as having had plausible or likely *as-if* randomisation, while one in four NEEs was rated as having implausible *as-if* randomisation. These would not have been classified as natural experiments based on the definition by Dunning [15] (but might have based on the MRC definition [19]). Assessing the plausibility of *as-if* randomisation of exposure based on the information provided in the studies proved to be difficult, and the results of our small validation study of 18 NEEs, each assessed by 12 raters, had a poor inter-rater agreement. This indicates that *as-if*-random assessment is extremely difficult and unreliable, even for reviewers with particular experience in this topic. For comparison, a similar study was previously conducted to assess inter-rater agreement for risk of bias assessment using the Cochrane Collaboration Risk of bias tool. This study reported a higher ICC of 0.58 (95% CI 0.20-0.81) [33], compared to 0.35 (0.19-0.47) for *as-if* randomisation in the current study. It remains unexplored whether this poor agreement results from the insufficient information provided in the publications, the raters themselves, or whether the process of judging the plausibility of *as-if* randomisation itself is highly susceptible to error and bias. Given this

difficulty in assessing *as-if* randomisation, we suggest that the reporting of NEEs should include an assessment of the plausibly of this assumption by the authors, or should provide sufficient detail on the process of exposure allocation to enable readers to assess the likelihood of *as-if* randomisation with a reasonable degree of confidence. Additionally, given that subjective judgments in these assessments will to some degree always be inevitable, causal inference from NEEs can be substantially increased by the inclusion of additional sensitivity analyses and falsification tests, ideally where these address different aspects of possible bias [17]. This requires researchers to conduct such additional analyses, as the results of this review indicate only about half of the included studies reported any form of sensitivity or falsification analysis [12, 21].

Finally, this scoping review has further highlighted that there is limited standardisation in the use and reporting of natural experiments. The evaluation of non-randomised interventions in public health would benefit from a more uniform approach for designing and appraising these evaluations, this approach could include direction on when certain approaches are more appropriate and when an evaluation could be classified as a natural experiment. . This includes the use of the 'Target Trial Framework' , which has been adapted for NEEs to support the design, reporting, and appraisal of NEEs [17]. Furthermore, journals and reviewers could ensure publications adhere to these recommendations to enhance the quality of evidence identified as NEEs.

This extensive scoping review is the first to review the use of natural experiment evaluations of place-based interventions in public health, and as such the first review of its kind to demonstrate the range and trajectory of NEE use to support place-based evaluations in public health research. With over 350 studies included, this scoping review provides a comprehensive overview and insight into how place-based NEEs have been conducted and reported over the last two decades. Our review was conducted by a multidisciplinary review team with expertise in NEEs; systematic reviews, and intervention evaluation. We based our search strategy on a previously used strategy that looked to assess the use of systems science in evaluations of place-based public health interventions. In addition, we recognised

the difficulty of assessing the *as-if* randomisation of the assignment of exposure, and included the first validation exercise for the assessment of the '*as-if*-random' assumption.

This scoping review also has a number of limitations. We applied a cut-off date for our review of January 2020, and thus our data does not include additional NEEs published since. We reran the same search on 1 July 2021, which identified a further 2,167 studies for potential inclusion. Unfortunately, because of constraints on resources we were not able to include these additional studies. However, we note that the majority of these studies concerned evaluations of the many aspects of the Covid-19 pandemic, which we consider a unique and distinct topic area which is not comparable to the routine place-based public health interventions and policy evaluations identified here and would require its own review.

In addition to the use of the January 2020 cut-off date for inclusion, the review was limited to studies published in the English language. We recognised that the vast majority of studies will have been published in the English language and this review included NEEs from 54 countries, but nonetheless this is likely to have resulted in evidence from non-English speaking countries being underrepresented, including that from low and middle income countries. Additionally, search 3 included the most common study methods associated with natural experiments, however, we did not include all such methods. Although, search 3 was intended to pick up studies not already found in search 1 and 2, and we did not anticipate many to be found we acknowledge that by including only the most common study designs in search 3 we may have missed a few studies. Finally, for the inter-rater reliability exploration for the *as-if* randomisation rating, it worth noting that although all authors that completed the rating have a common interest in the use of natural experiments for the evaluation of public health interventions, all come with a variety of backgrounds and expertise within public health. This may be a possible explanation for the low inter-rater reliability. Nonetheless, our scoping review provides a foundation to explore the use of natural experiment methodology in this field and more broadly in public health. Because of the large number of papers that were included we had to make the decision to have each

NEE extracted by a single author (although PA double-checked a proportion of the coding in response to queries from others). This may have resulted in variations in coding.

Finally, we conducted an additional inter-rater assessment of *as-if* randomisation of exposure for 18 randomly selected NEEs by ten of the authors of this review. This showed poor agreement and demonstrated the difficulty of assessing this assumption from the information provided in the publications. In the absence of the 'correct answers' or at least a 'gold standard', and despite all ten authors having experience with conducting and appraising NEEs, we cannot exclude the possibility that the poor inter-rater agreement might have resulted from variability in the experience of the raters, or errors made in the process.

In addition to exploring the practices and methods used, trends and developments in the use of methods and concepts, and their strength for causal inference, we also aimed to provide recommendations for addressing the main weaknesses in published place-based NEEs in public health. The review provides a strong motivation for a consistent design and reporting of NEEs to facilitate the appraisal of the strengths and weaknesses of NEEs and thereby their potential to make causal claims. There are already a number of reporting guidelines that are applicable to NEEs or relevant specific study designs, or that could be adapted specifically for NEEs, for example STROBE, TREND [34, 35], TIDieR-PHP [36] and others [37, 38], and for design and reporting following the Target Trial Framework has also been proposed [17]. Akin to RCTs, pre-registration of NEEs would further enhance methodological rigour. There are different schools of thoughts on the importance of *as-if* randomisation for the definition of NEEs, but it is unquestionable that this is an important criteria for assessing the strength of any causal claims based upon such studies.

Although not directly addressed in the scoping review, NEEs can make important contributions to addressing health inequalities. For example, they can be important in investigating the determinants of health inequalities and in the identification of effective interventions to reduce inequalities [12]; effectively because they provide opportunities to study differential, subgroup effects. Especially for the

evaluation of place-based public health interventions this offers particular benefits as geographical location data are often already available from routine data.

5. CONCLUSIONS

In conclusion, the scoping review demonstrates that NEEs are conducted using many different study designs and statistical methods and encompass various definitions of a natural experiment. It is questionable whether all evaluations reported as natural experiments should be considered as such, suggesting that the field would benefit from more consistent design and reporting of NEEs. Given the difficulty in assessing the likelihood of *as-if* randomisation of the exposure, we recommend that researchers address this specifically in reports of NEEs by clearly demonstrating how social or political drivers behind the group assignment are indeed random. Further we recommend that authors make more use of sensitivity/ falsification tests or other robustness checks in the design and conduct of their studies, in order to improve their causal inference. We also recommend that reviewers should look for these design elements in the evaluation of funding applications and that research funders and journals should provide specific guidelines for what is expected in the design and reporting of NEEs.

Ethics approval and consent to participate: Not Applicable

Conflicts of Interest: The authors declare that they have no conflicts of interests

Funding: This study is funded by the National Institute for Health Research (NIHR) School for Public Health Research (Grant Reference Number PD-SPH-2015). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. The funder had no input in the writing of the manuscript or decision to submit for publication. The NIHR School for Public Health Research is a partnership between the Universities of Sheffield; Bristol; Cambridge; Imperial; and University College London; The London School for Hygiene and Tropical Medicine (LSHTM); LiLaC – a collaboration between the Universities of Liverpool and Lancaster; and Fuse - The Centre for Translational Research in Public Health a collaboration between Newcastle, Durham, Northumbria, Sunderland and Teesside Universities. FdV & RJ are partly funded by National Institute for Health Research Applied Research Collaboration West (NIHR ARC West) at University Hospitals Bristol NHS Foundation Trust. PC, MG and MC acknowledge funding from the Medical Research Council (MC_UU_00022/2) and Scottish Government Chief Scientist Office (SPHSU17). KT works in the MRC Integrative Epidemiology Unit, which is supported by the Medical Research Council (MRC) and the University of Bristol [MC_UU_00011/3]. DT-R is funded by the Medical Research Council (MRC) on a Clinician Scientist Fellowship (MR/P008577/1). CR is funded by an NIHR Doctoral Fellowship (NIHR301784).

Authors' contributions: FDV conceived of the study. PA conducted the literature searches. All authors were involved in data extraction and analysis of the results. PA and FDV wrote the first version of the manuscript. All authors provided input in various iterations of the manuscript. All authors approved the final version.

Acknowledgments Not applicable

REFERENCES

1. Tunstall HVZ, Shaw M, Dorling D. Places and health. Journal of Epidemiology and Community Health. 2004;58(1):6-10.
2. Cummins S, Curtis S, Diez-Roux AV, Macintyre S. Understanding and representing 'place' in health research: A relational approach. Social Science & Medicine. 2007;65(9):1825-38.
3. Weiss D, Lillefjell M, Magnus E. Facilitators for the development and implementation of health promoting policy and programs – a scoping review at the local community level. BMC Public Health. 2016;16(1):140.
4. McGowan VJ, Buckner S, Mead R, McGill E, Ronzi S, Beyer F, et al. Examining the effectiveness of place-based interventions to improve public health and reduce health inequalities: an umbrella review. BMC Public Health. 2021;21(1):1888.
5. Snow J. On the mode of communication of cholera: John Churchill; 1855.
6. Lumey L, Stein AD, Kahn HS, van der Pal-de Bruin KM, Blauw G, Zybert PA, et al. Cohort Profile: The Dutch Hunger Winter Families Study. International Journal of Epidemiology. 2007;36(6):1196-204.
7. Franco M, Bilal U, Orduñez P, Benet M, Morejón A, Caballero B, et al. Population-wide weight loss and regain in relation to diabetes burden and cardiovascular mortality in Cuba 1980-2010: repeated cross sectional surveys and ecological comparison of secular trends. BMJ : British Medical Journal. 2013;346:f1515.
8. Rich DQ, Liu K, Zhang J, Thurston SW, Stevens TP, Pan Y, et al. Differences in Birth Weight Associated with the 2008 Beijing Olympics Air Pollution Reduction: Results from a Natural Experiment. Environmental Health Perspectives. 2015;123(9):880-7.
9. Yau A, Berger N, Law C, Cornelsen L, Greener R, Adams J, et al. Changes in household food and drink purchases following restrictions on the advertisement of high fat, salt, and sugar products across the Transport for London network: A controlled interrupted time series analysis. PLOS Medicine. 2022;19(2):e1003915.
10. Hunter RF, Cleland C, Cleary A, Droomers M, Wheeler BW, Sinnett D, et al. Environmental, health, wellbeing, social and equity effects of urban green space interventions: A meta-narrative evidence synthesis. Environment International. 2019;130:104923.
11. Mayne SL, Auchincloss AH, Michael YL. Impact of policy and built environment changes on obesity-related outcomes: a systematic review of naturally occurring experiments. Obesity Reviews. 2015;16(5):362-75.
12. Petticrew M, Cummins S, Ferrell C, Findlay A, Higgins C, Hoy C, et al. Natural experiments: an underused tool for public health? Public Health. 2005;119(9):751-7.
13. Pearl J. Causal inference in statistics: An overview. 2009.
14. Meyer BD. Natural and Quasi-Experiments in Economics. Journal of Business & Economic Statistics. 1995;13(2):151-61.
15. Dunning T. Natural Experiments in the Social Sciences: A Design-Based Approach: Cambridge University Press; 2012.
16. Ogilvie D, Adams J, Bauman A, Gregg EW, Panter J, Siegel KR, et al. Using natural experimental studies to guide public health action: turning the evidence-based medicine paradigm on its head. Journal of Epidemiology and Community Health. 2020;74(2):203-8.

17. de Vocht F, Katikireddi SV, McQuire C, Tilling K, Hickman M, Craig P. Conceptualising natural and quasi experiments in public health. BMC Medical Research Methodology. 2021;21(1):32.
18. Rosenbaum PR. How to see more in observational studies: Some new quasi-experimental devices. Annual Review of Statistics and Its Application. 2015;2:21-48.
19. Craig P, Cooper C, Gunnell D, Haw S, Lawson K, Macintyre S, et al. Using natural experiments to evaluate population health interventions: new Medical Research Council guidance. Journal of Epidemiology and Community Health. 2012;66(12):1182-6.
20. Angrist JD. Lifetime earnings and the Vietnam era draft lottery: evidence from social security administrative records. The american economic review. 1990:313-36.
21. Craig P, Katikireddi SV, Leyland A, Popham F. Natural Experiments: An Overview of Methods, Approaches, and Contributions to Public Health Intervention Research. Annual Review of Public Health. 2017;38(1):39-56.
22. Crane M, Bohn-Goldbaum E, Grunseit A, Bauman A. Using natural experiments to improve public health evidence: a review of context and utility for obesity prevention. Health Research Policy and Systems. 2020;18(1):48.
23. Shi X, Miao W, Tchetgen ET. A Selective Review of Negative Control Methods in Epidemiology. Current Epidemiology Reports. 2020;7(4):190-202.
24. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Medical Research Methodology. 2018;18(1):143.
25. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. International Journal of Social Research Methodology. 2005;8(1):19-32.
26. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implementation Science. 2010;5(1):69.
27. McGill E, Er V, Penney T, Egan M, White M, Meier P, et al. Evaluation of public health interventions from a complex systems perspective: A research methods review. Social Science & Medicine. 2021;272:113697.
28. Thornton LE, Ball K, Lamb KE, McCann J, Parker K, Crawford DA. The impact of a new McDonald's restaurant on eating behaviours and perceptions of local residents: a natural experiment using repeated cross-sectional data. Health & Place. 2016;39:86-91.
29. Atkins DN, Bradford WD. The effect of changes in state and federal policy for nonprescription access to emergency contraception on youth contraceptive use: a difference-in-difference analysis across New England states. Contemporary Economic Policy. 2015;33(3):405-17.
30. Dunning T. Improving Causal Inference:Strengths and Limitations of Natural Experiments. Political Research Quarterly. 2008;61(2):282-93.
31. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. 2016;15(2):155-63.
32. Matthay EC, Glymour MM. Causal Inference Challenges and New Directions for Epidemiologic Research on the Health Effects of Social Policies. Current Epidemiology Reports. 2022;9(1):22-37.
33. Armijo-Olivo S, Stiles CR, Hagen NA, Biondo PD, Cummings GG. Assessment of study quality for systematic reviews: a comparison of the Cochrane Collaboration Risk of Bias Tool and the Effective Public Health Practice Project Quality Assessment Tool: methodological research. Journal of Evaluation in Clinical Practice. 2012;18(1):12-8.

34. What Works Clearinghouse. Reporting Guide for Study Authors: Regression Discontinuity Design Studies. Institute of Education Sciences; 2018.
35. Wing C, Simon K, Bello-Gomez RA. Designing Difference in Difference Studies: Best Practices for Public Health Policy Research. Annual Review of Public Health. 2018;39(1):453-69.
36. Campbell M, Katikireddi SV, Hoffmann T, Armstrong R, Waters E, Craig P. TIDieR-PHP: a reporting guideline for population health and policy interventions. BMJ. 2018;361:k1079.
37. equator network. Enhancing the QUAlity and Transparency Of health Research [Available from: https://www.equator-network.org/?post_type=eq_guidelines&eq_guidelines_study_design=observational-studies&eq_guidelines_clinical_specialty=0&eq_guidelines_report_section=0&s=+&eq_guidelines_study_design_sub_cat=0.
38. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol. 2011;64(4):383-94.