PUBLIC HEALTH XXX (2015) 1–10

Available online at www.sciencedirect.com

Public Health

journal homepage: www.elsevier.com/puhe

Original Research

Propensity score matching for selection of local areas as controls for evaluation of effects of alcohol policies in case series and quasi case–control designs

F. de Vocht [a,b,*], R. Campbell [a,b], A. Brennan [a,c], J. Mooney [a,c], C. Angus [a,c], M. Hickman [a,b]

[a] NIHR School for Public Health Research (SPHR), UK
[b] School of Social and Community Medicine, University of Bristol, Bristol, UK
[c] ScHARR, School of Health and Related Research, University of Sheffield, Sheffield, UK

ARTICLE INFO

Article history:
Received 23 March 2015
Received in revised form
23 October 2015
Accepted 29 October 2015
Available online xxx

Keywords:
Propensity score matching
Natural experiments
Methodology

ABSTRACT

Objectives: Area-level public health interventions can be difficult to evaluate using natural experiments. We describe the use of propensity score matching (PSM) to select control local authority areas (LAU) to evaluate the public health impact of alcohol policies for (1) prospective evaluation of alcohol policies using area-level data, and (2) a novel two-stage quasi case–control design.

Study design: Ecological.

Methods: Alcohol-related indicator data (Local Alcohol Profiles for England, PHE Health Profiles and ONS data) were linked at LAU level. Six LAUs (Blackpool, Bradford, Bristol, Ipswich, Islington, and Newcastle-upon-Tyne) as sample intervention or case areas were matched to two control LAUs each using PSM. For the quasi case–control study a second stage was added aimed at obtaining maximum contrast in outcomes based on propensity scores. Matching was evaluated based on average standardized absolute mean differences (ASAM) and variable-specific *P*-values after matching.

Results: The six LAUs were matched to suitable control areas (with ASAM < 0.20, *P*-values >0.05 indicating good matching) for a prospective evaluation study that sought areas that were similar at baseline in order to assess whether a change in intervention exposure led to a change in the outcome (alcohol related harm). PSM also generated appropriate matches for a quasi case–control study – whereby the contrast in health outcomes between cases and control areas needed to be optimized in order to assess retrospectively whether differences in intervention exposure were associated with the outcome.

* *Corresponding author*. School of Social and Community Medicine, University of Bristol, Canynge Hall, 39 Whatley Road, Bristol, BS8 2PS, UK. Tel.: +44 (0)117 928 7239.
E-mail address: frank.devocht@bristol.ac.uk (F. de Vocht).

http://dx.doi.org/10.1016/j.puhe.2015.10.033

Conclusions: The use of PSM for area-level alcohol policy evaluation, but also for other public health interventions, will improve the value of these evaluations by objective and quantitative selection of the most appropriate control areas.

Introduction

The need for a better evidence base for public health policies is widely acknowledged.[1] Interventions should at the very least be evaluated on: (a) whether the research is of sufficient quality to support a decision on implementation of the intervention; (b) what the research outcomes are; and (c) whether the research findings are generalizable to potential recipients of the intervention.[2] However, evaluation of public health interventions is often more difficult than for clinical interventions because they generally aim at achieving population rather than individual level impact, and the intervention may be complex,[3] programmatic, and context dependent.[2]

Rigorous evaluation requires an exposed and a non-exposed group to be compared (or compared across different levels of exposure). Additional design elements that further strengthen causal inferences include multiple pre/post measures, multiple exposed and unexposed groups, and accurate measurement of exposures.[3] Ideally, (public health) interventions should be evaluated using a randomized controlled design, but these may not always be feasible or fail to be considered when interventions are implemented.[4] Alternatively, 'natural experiments' can be created and used to study the relation between external changes and effects on population disease patterns. A famous example is John Snow's analysis and intervention to prevent the spread of cholera. More recently natural experiments have been used to evaluate the introduction of smoke-free legislation.[5–7]

However, in contrast to external shocks, beneficial events generally have a much less pronounced impact, and it may also take longer for an effect to emerge, making evaluation harder to study and more susceptible to bias.[3] It is thus important from an evidence-based policy perspective to qualitatively and quantitatively study variations in the delivery of interventions, either temporally or spatially, to evaluate their impact on population health.

When evaluating the impact of public health policies an important methodological consideration is how to select appropriate control areas in such a way as to strengthen causal inference. In contrast to studies with individual-level data (e.g. with participants), the control areas for studies on policies are often opportunistically chosen and may include neighbouring local authority areas or other, broadly comparable areas to which the research team has access.

In this paper we describe the use of propensity scores to match 'case areas' to 'control areas' so that a subsequent qualitative and/or quantitative evaluation of a policy is demonstrable between local areas comparable for the domain under study. It has been shown that propensity score matching (PSM) can be an effective methodology to minimize bias by matching cases to controls based on a set of baseline covariates,[9] and has been used in health services research, pharmaco-epidemiology[10–13] and health economics.[14] However, there has been limited use in public health to demonstrate the effect of certain interventions at the individual level (for example:[3,15,16]). We further demonstrate a novel two-stage propensity score matching (PSM) design aimed at mimicking a traditional case control study that has not been used previously.

Expanding on standard PSM methodology, this manuscript deals specifically with the use of PSM for local area-level data of both intervention and outcome data for which PSM is not often, if ever, used. We will describe this in the context of the evaluation of local authority public health interventions aimed at reducing alcohol-related harm where detailed data collection and in-depth analysis are required, which prohibits inclusion of all 353 local authority units (LAU) in England.

Methods

Motivating example

Alcohol related harm varies by geographical area. In order to test whether the intensity of local alcohol policies is associated with changes in alcohol related harm requires additional data collection because the level of intervention delivered locally is not available from routine administrative datasets. We consider a case study to determine the most appropriate control Local Authority Unit (LAU) for six areas where data are being collected: Blackpool, Bradford, Bristol, Ipswich, Islington, and Newcastle-upon-Tyne. We consider two potential evaluation designs: a) a cohort (prospective or retrospective) where we aim to test whether the introduction of new local alcohol polices are associated with change in alcohol related harm; and b) a case control where we aim to test whether sites with contrasting levels of alcohol related harm differ in relation to intensity of local alcohol interventions.

Ideally, we would prefer to obtain the same data for an area where a specific intervention will be introduced (i.e. the 'case area') but also for that same area as if the intervention had not been introduced (i.e. the 'counterfactual case').[17] When evaluating the impact of a new policy we cannot simultaneously measure its effectiveness in a specific area where the policy is introduced and in the same area where it is not introduced, and we are forced to compare it to another area. Thus the choice of an appropriate control is essential to achieving an unbiased assessment of effectiveness. In situations where randomization is not possible, we need a

Please cite this article in press as: de Vocht F, et al., Propensity score matching for selection of local areas as controls for evaluation of effects of alcohol policies in case series and quasi case–control designs, Public Health (2015), http://dx.doi.org/10.1016/j.puhe.2015.10.033

method to ensure that selected control areas are as similar as possible to intervention areas at baseline such that, in theory, it would not have made a difference to evaluation of the efficacy which of the areas would receive the intervention or which would be the control (e.g. to mimic as much as possible the counterfactual case). If we have baseline information for all local areas, we need a method to choose the most closely related area across a (large) number of key variables.

We propose the use of Propensity Scores for matching (PSM), which is in essence a model to estimate the probability/propensity that a study unit which has not received the intervention (usually a study participant) is similar at baseline to another unit from the 'intervention group', based on a set of key characteristics. As such, it reduces the problem of comparison across large numbers of key variables to a 1-dimensional problem; i.e. the minimization of the difference, or distance, between case and control propensity scores. In the context of our study this matching is done to select controls for six local areas in England instead of individual study participants for which PSM is generally used. Exact matching in this context would not be feasible since many of the indicators are population rates or other continuous variables, and exact matching on these for some indicators will only occur in a few instances, while exact matching on all covariates in this case is impossible. Here we extend the PSM methodology to enable the study of alcohol-related harm in England at the level of local authority units (LAU).

Data

LAU indicator data for comparison between areas were obtained from the Local Alcohol Profiles for England (LAPEs) 2014 update (http://www.lape.org.uk/data.html),[18] and were linked to 2013 Health Indicator data (http://www.apho.org.uk/resource/view.aspx?RID=142075) produced by the Northwest Knowledge and Intelligence Team and LAU-level data from the UK Office of National Statistics. This resulted in a set of 93 indicator variables which were used by the experts (online supplementary material).

Selection of indicator variables

A variety of approaches for covariate selection ranging from sole reliance on subject-matter specification of variables to completely data-driven approaches has been described, but it has been shown empirically that in cases with a limited number of 'units' sole reliance on a selection algorithm may increase the likelihood of bias and that in these situations expert selection of variables combined with empirical specification may be beneficial.[19] Therefore, we used expert variable selection (in this example by the authors), using a modified Delphi approach, with and without subsequent data-driven optimization strategies. A list of 93 indicator variables (online supplementary material) was used by all authors (e.g. the experts) independently to select a set of five (set 1) and 12 (set 2) covariates which they considered to be key factors influencing baseline levels of alcohol-related burden, and potentially predicting uptake of, and response to, the introduction of new alcohol-related policies. All responses were subsequently collated and a *narrow* and a *wider* dataset were derived by including the 5 and 12 (initially 5 and 15 were selected, but 3 of selected 15 overlapped) indicators most often selected (see Tables 1 and 2 for a list of the variables).

Matching strategies

Subsequently, we calculated propensity scores to match each case LAU to control LAUs for two different purposes:

(a) a prospective study in which cases are matched to two controls each based on a propensity score model of selected key baseline factors, and
(b) a two-stage, quasi case–control design for retrospective evaluation of natural experiments.

The first purpose (a) is to ensure that effects from policies can be compared prospectively by comparing control and case areas that are comparable at baseline. This is the standard way PSM is used in epidemiology, but at area-level instead. We explored three different empirical strategies to *a priori* decide on the explanatory variables entered into the PS model: model 1 based on the five key baseline characteristics of set 1 to evaluate matching on a small set of indicators; model 2 based on the 12 key variables identified in set 2 to evaluate matching on a large set of indicators; and a hybrid method (model 3) in which set 2 was reduced to improve statistical estimation by reducing multicollinearity. This was achieved by working backwards from model 2 by removal of variables for which the generalized variance inflation factor (calculated as: $GVIF^{1/(2*Df)2}$) was 10 or higher.[20] This approach was included because it allows for a larger set of key variables to initially base the PSM on, but uses a data driven approach to improve statistical estimation of the propensity scores by removal indicators that add little new information.

For our second intended purpose (b), we describe a novel, quasi case–control design for retrospective evaluation of natural experiments in which ideally we'd want control areas that were similar to the case areas prior to the intervention, but that are now as different as possible for outcome measures. This approach is applicable to a situation in which we are interested in maximized differences in outcome between case and control areas, but we were unable to prospectively obtain data when an intervention was introduced. Similar to a case control study, we can then compare the policies (i.e. the exposure) that were in place across that period (either quantitatively or qualitatively) and assess whether there is evidence of an association between intervention exposure and case/control status. However, for valid inferences to be made using this design the case and control areas should have been comparable at baseline, and to allow for this a 2-stage design is required. LAUs are PS-matched such that they are comparable at baseline for the set of confounders (stage 1), and subsequently in a 2nd stage differences in the outcome (e.g. measures of alcohol-related harm in our example) are maximized. Stage 2 matching therefore can be considered a form of 'maximum variation sampling' in that from the controls, those with maximum difference in outcomes of interest to the

Please cite this article in press as: de Vocht F, et al., Propensity score matching for selection of local areas as controls for evaluation of effects of alcohol policies in case series and quasi case–control designs, Public Health (2015), http://dx.doi.org/10.1016/j.puhe.2015.10.033

Table 1 – Details of six matched LAU sets using the 'wider' set of key variables and multicollinearity reduction for propensity score model 3.

Local authority unit	Annual admission episodes for alcohol attributable conditions	Alcohol-related recorded crimes (crude rate per 1000 population)	Annual admission episodes for alcohol attributable conditions (under 18)	IR&HR drinkers (% in the drinking population)[a]	ONS supergroup[b]	Alcohol-related mortality[c] (males)	Alcohol-related mortality[c] (females)	Bar employees (% of all employees)	Binge drinking (% in the drinking population)[a]	Distance	Matched set
Bristol, City of	**2435**	**8.1**	**57.4**	**23.3**	**B**	**22.5**	**6.0**	**1.6**	**26.3**	**−1.59**	**1**
Nottingham	2398	9.7	43.0	21.0	B	26.3	8.2	1.2	23.9	−1.78	1
Bournemouth	2373	7.3	63.2	23.5	B	32.5	7.6	2.2	25.5	−1.48	1
Islington	**2658**	**10.9**	**71.5**	**22.0**	**C**	**19.6**	**6.5**	**1.0**	**21.1**	**−0.84**	**2**
Harlow	2380	8.1	25.5	20.7	E	16.6	3.3	1.5	19.6	−1.23	2
Burnley	3245	8.4	121.4	20.7	B	17	6.8	2.3	23.9	−1.06	2
Ipswich	**2009**	**7.9**	**49.6**	**21.7**	**E**	**13.8**	**4.8**	**2.1**	**17.0**	**−4.11**	**3**
Plymouth	2265	8.1	92	23.4	B	15.2	5.4	2.0	23.4	−4.14	3
North Devon	1920	4.9	74.3	23.1	D	12.0	9.5	3.5	19.1	−4.14	3
Newcastle upon Tyne	**2575**	**5.2**	**76.9**	**22.9**	**B**	**20.1**	**9.1**	**2.6**	**33.7**	**−0.55**	**4**
City of London	1912	31.3	39.1	20.6	C	18.2	5.9	0.6	25.3	−1.42	4
Blackburn with Darwen	3163	6.5	74.6	20	B	18.2	3.9	1.2	18.9	−0.62	4
Bradford	**2565**	**6.2**	**49.5**	**19.7**	**B**	**16.4**	**8.9**	**1.4**	**18.8**	**−2.75**	**5**
Cambridge	2190	5.0	57.7	24.5	B	17.8	4.1	1.6	26.3	−2.70	5
Craven	1719	2.8	45.2	23.5	D	8.79	5.5	3.1	25.6	−2.75	5
Blackpool	**2950**	**11.9**	**113.8**	**22**	**D**	**40.5**	**12.6**	**3.2**	**23.7**	**0.16**	**6**
Hammersmith and Fulham	2554	10.2	59.5	22.9	C	20.6	7.2	1.3	22.6	−0.40	6
Manchester	3276	9.0	76.7	21	B	33.6	12.9	1.6	29.0	1.18	6
P-value t-test difference after matching	0.12	0.45	0.31	0.83	–	0.43	0.13	0.24	0.98	0.07	

[a] Binge drinking, Increasing Risk (IR) and Higher Risk (HR); synthetic estimate.[1]
[b] (A) Mining and Manufacturing, (B) Cities and Services, (C) London Centre, (D) Coastal and Countryside, (E) Prospering UK.
[c] per 100,000 population.

Please cite this article in press as: de Vocht F, et al., Propensity score matching for selection of local areas as controls for evaluation of effects of alcohol policies in case series and quasi case–control designs, Public Health (2015), http://dx.doi.org/10.1016/j.puhe.2015.10.033

Table 2 – Results of six 'matched' (based on maximum variation matching (step 2)) LAU sets for retrospective quasi case–control evaluation.

Local authority	Annual admission episodes for alcohol attributable conditions	Alcohol-related recorded crimes (crude rate per 1000)	Annual admission episodes for alcohol attributable conditions (under 18 years of age)	Alcohol-related mortality (males)	Alcohol-related mortality (females)	Propensity score
Bristol, City of	**2435**	**8.1**	**57.4**	**22.5**	**6.0**	**0.287**
Gloucester	2043	6.6	54.3	17.6	9.0	0.064
Richmondshire	1644	2.6	70.7	5.0	6.2	0.012
Islington	**2658**	**10.9**	**71.5**	**19.6**	**6.5**	**0.394**
Kensington and Chelsea	1353	8.5	46.9	9.5	5.6	0.047
Cheltenham	1903	5.3	85.9	14.3	5.7	0.046
Brighton and Hove	1987	6.6	88.5	21.1	11.3	0.030
Ipswich	**2009**	**7.9**	**49.6**	**13.8**	**4.8**	**0.152**
Worcester	1848	6.6	96.3	10.4	11.2	0.013
Newcastle upon Tyne	**2575**	**5.2**	**76.9**	**20.1**	**9.1**	**0.091**
York	1413	4.9	65.1	13.3	6.8	0.019
Weymouth and Portland	1703	6.0	79.6	25.5	10.2	0.030
Bradford	**2565**	**6.2**	**49.5**	**16.4**	**8.9**	**0.129**
Blackburn with Darwen	3163	6.5	74.6	18.2	3.9	0.474
Selby	1382	3.7	57.4	11.6	6.0	0.018
Blackpool	**2950**	**11.9**	**113.8**	**40.5**	**12.6**	**0.387**
Bury	2272	5.6	78.3	17.7	11.1	0.037
Salford	3192	6.2	125.5	21.1	12.0	0.091

case areas are matched to each case.[23] When implementing the two stage approach in our example, Stage 1 was comparable to the methodology outlined above for the normal, prospective evaluation: based on the wider set of key variables, but without the alcohol-related outcomes because these are required for maximum variation sampling in stage 2 (N = 7).

In all PS modelling 'control LAUs' were matched to 'case LAUs' using *nearest neighbour optimal matching*, which has been shown to be comparable to *'greedy' matching*, but can result in better overall minimization of the distance between pairs.[9,21] *Nearest neighbour optimal matching* aims to generate matched pairs by minimizing overall average within-pair difference in propensity scores across all pairings, while *greedy matching* selects the control with the closest propensity score for each case.[22] Two control LAUs were matched to each case area and were not re-used, which resulted in a total of 18 different LAUs in each analysis. Alternatively, it is also possible (and may be beneficial if time and budget constraints are of concern) to allow for areas to be controls for multiple cases.

For the second purpose that requires two stage sampling, each case area was initially matched to six control LAUs (resulting in 42 LAUs for stage 2) in the first stage (to generate enough matched potential controls for selection in stage 2), and subsequently, in stage 2, PS were calculated based on alcohol-related outcome measures. Within each set of one case and six control LAUs, cases were then matched to two controls each by selecting the two LAUs with the largest PS distance from the case. All cases were kept and unmatched controls were discarded.

Analyses were done using the *MatchIt* package[24] in *R* version 3.0.1.[25] Average standardized absolute mean differences (ASAM) were calculated as a global measure of matching, using values much larger than 0.20 to indicate possible matching problems.[26]

Results

(a) Prospective design

Because of missing information on one or more indicators, of the total of 353 LAUs in England, 338 (e.g. six case areas and 332 potential control areas) could be included for model 1, 334 for model 2 and 332 for model 3.

The result of the matching can be seen for the best matching strategy (model 3) in Table 1 (results for models 1 and 2 are presented in Tables S1 and S2 in Online Supplementary Material, respectively). Model 1 would require only few key variables for matching, but evaluation of global matching indicated that it was only borderline (un)acceptable (ASAM = 0.28); for example the distances between Ipswich and the selected controls; Great Yarmouth and Nuneaton and Bedworth, is much smaller than for Blackpool, Salford and Middlesbrough (Table S1).

Although it has been suggested to add as much information as possible to propensity score models, model 2 based on the 12 key variables (Table S2) similarly resulted in unacceptable differences between the quality of the matched sets (ASAM = 0.39 with post matching *P*-values <0.05).

As shown in Table 1, minimizing multicollinearity in model 3 results in acceptable matching (ASAM = 0.20). Balance was also good within matched sets (*P*-values > 0.05). In addition to quantitative evaluation of balance, qualitatively (based

Please cite this article in press as: de Vocht F, et al., Propensity score matching for selection of local areas as controls for evaluation of effects of alcohol policies in case series and quasi case–control designs, Public Health (2015), http://dx.doi.org/10.1016/j.puhe.2015.10.033

on knowledge of the local areas) the matched sets seem acceptable too. Overlap across all (scaled) key variables within each set has been shown graphically in Fig. 1.

(b) Quasi case–control

Of 353 LAUs, 334 could be used in this example. The first stage of the quasi case–control approach resulted in borderline acceptable global matching (ASAM = 0.23, and all post matching variable-specific P-values >0.05; details provided in Table S3 in Online Supplementary Materials); mainly as a result of the need to oversample controls to achieve maximum variation on alcohol-related key outcome measures in stage 2. In stage 2, two controls were matched to each case LAU (from the six selected in stage 1) based on maximum PS distance between case and control LAUs (Table 2). All key outcome variables have again been scaled and shown graphically for each set in Fig. 2, and indicate that the data space of the indicators for the case area (in grey) hardly overlaps with that of the two control areas (the black lines). In direct comparison with the overlap in Fig. 1, it is clear that the differences between the case LAUs and the controls are larger and are prominent for more of the indicators.

The two stage quasi case control methodology outlined here has resulted in matched sets with comparable key statistics at baseline but with maximum contrast in outcome indicators (of alcohol-related harms). This is similar to what would be expected for a matched case–control study. The impact of policies that were introduced or were in place between the case and control areas can now be quantitatively or qualitatively evaluated.

Discussion

We demonstrate a quantitative and transparent framework for selection of controls areas in order to develop a natural experiment to evaluate public health policies. In addition, we described a novel 2-stage approach in which PS matching can be used to mimic a case control design; thus enabling evaluation of (public health) policies retrospectively where information on exposure needs to be collated. In this example, we specifically focussed on selection of control LAUs in England for the evaluation of the impact of alcohol policies in a set of case areas. Both statistically and theoretically our analyses demonstrate how a framework combining *a priori* key indicators and quantitative propensity score matching can be used to select appropriate control areas for prospective or retrospective evaluation of the impact of public health interventions at a population level.

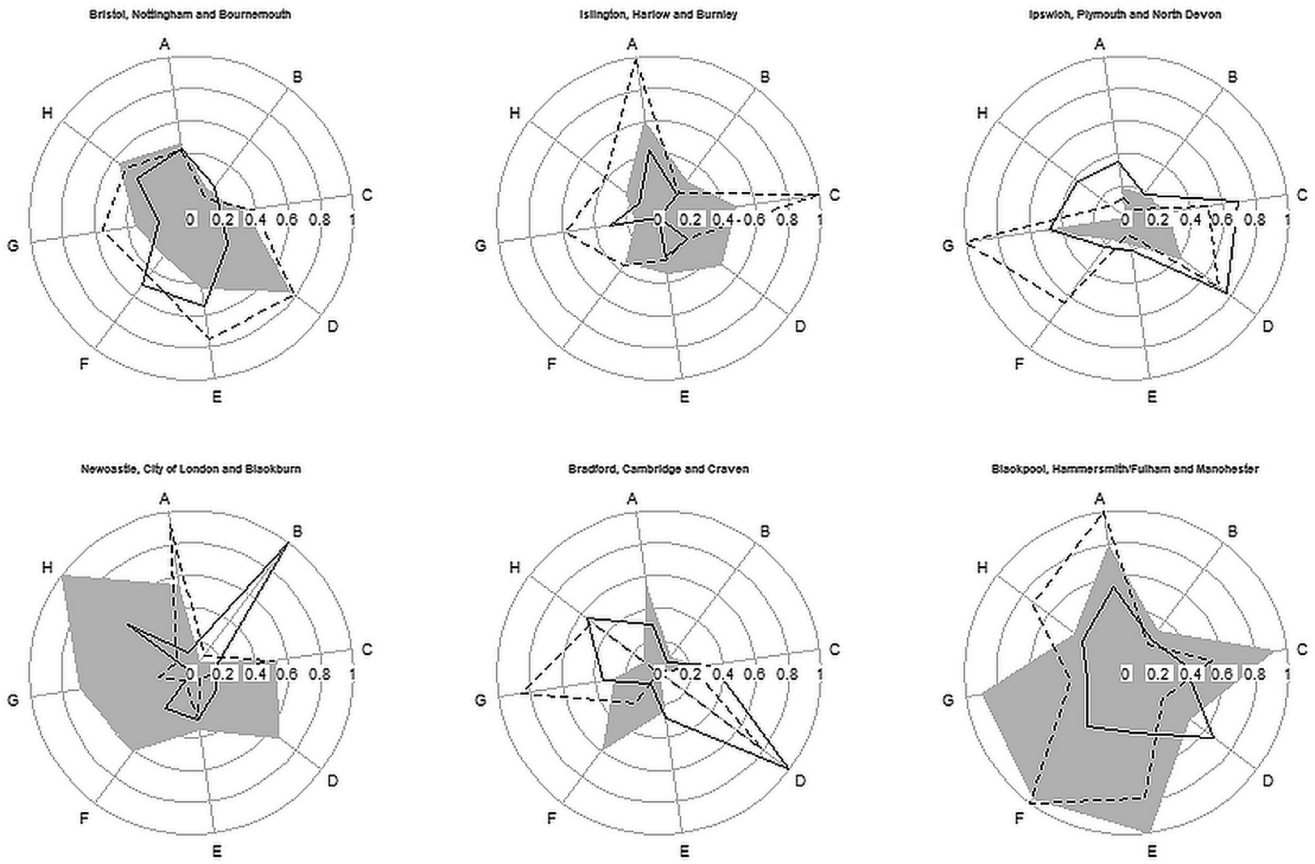

Fig. 1 – Radial plots of each case study (grey) and its two matched control areas (line and dotted line) for each of the variables included in the propensity score model: HES alcohol-attributable conditions (A), alcohol-related crime rate (B), HES alcohol-attributable conditions (<18 years) (C), percentage increasing risk and high risk drinkers (D), alcohol-related mortality (males) (E), alcohol-related mortality (females) (F), percentage of bar employees (G), percentage regular binge drinkers (H). Note that all variables have been scaled between 0 and 1 and that variable ONS Supergroup is not included in plots.

Please cite this article in press as: de Vocht F, et al., Propensity score matching for selection of local areas as controls for evaluation of effects of alcohol policies in case series and quasi case–control designs, Public Health (2015), http://dx.doi.org/10.1016/j.puhe.2015.10.033

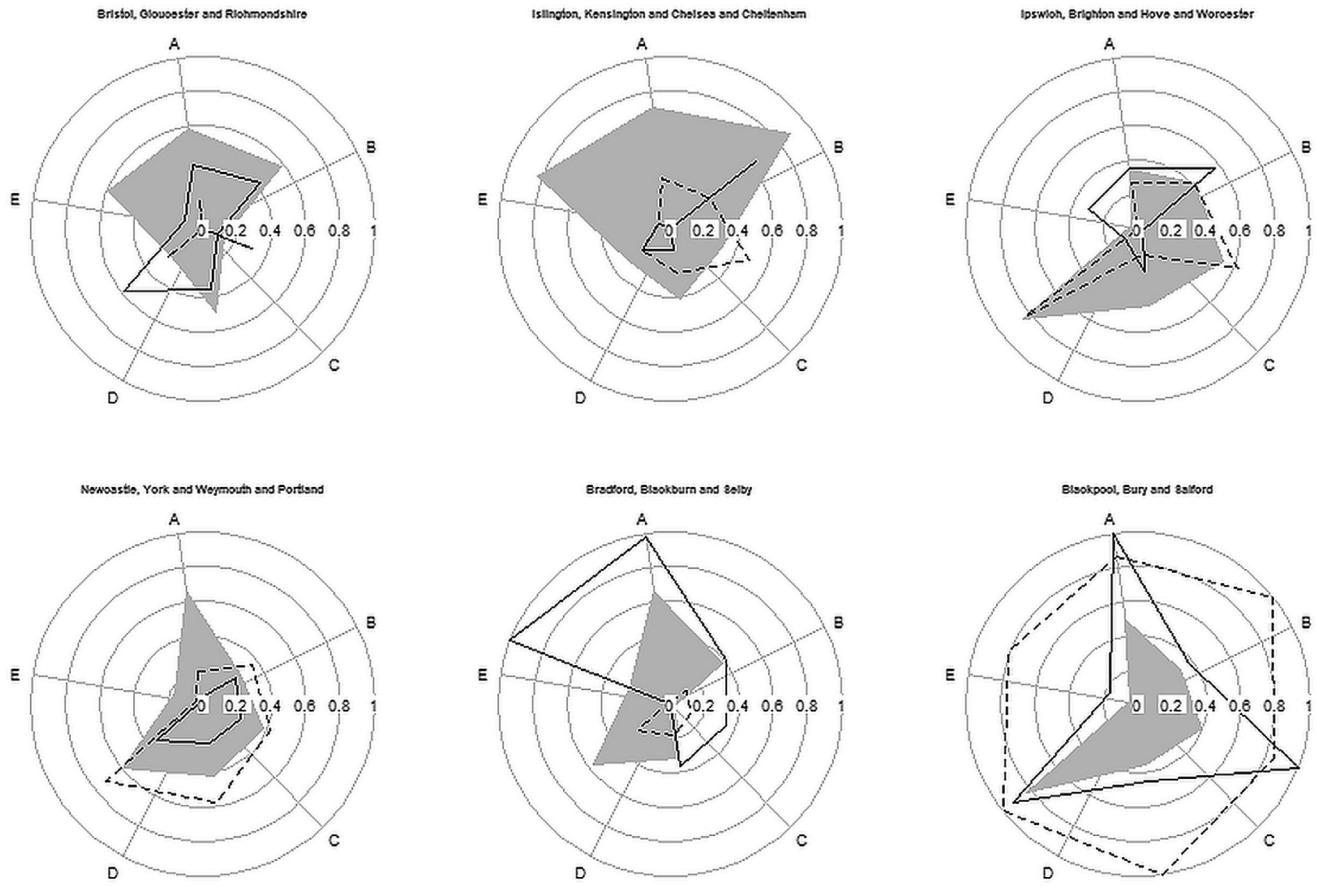

Fig. 2 – Radial plots of each case study (grey) and its two matched control areas (line and dotted line) for each of the outcome variables included in the 2nd stage propensity score model: number of adult alcohol-related hospital admissions (A), alcohol-related crime rate (B), number of alcohol-related hospital admissions for under 18 years of age (C), Alcohol-specific mortality rate (males) (D), Alcohol-specific mortality rate (females) (E). Note that all variables have been scaled between 0 and 1.

Often no formal matching is attempted when the impact of new policies is evaluated either at an individual[27–34] or aggregated (area) level, even though appropriate use of controls is essential for valid evaluation of the impact of new policies. The framework set out in this paper formalizes the concepts set out in generic terms in recent MRC Guidance for the evaluation of natural experiments,[3] and the argument for a data-driven procedure for control selection is analogous to that of Abdie and Gardeazabal.[35] Propensity Scores have not been used to select control areas in public health, although they have been used previously to match at an individual level; for example to study the effect of a new public transit system to connect low-income areas with the urban centre on violence.[36]

This study's results showed that by adopting a formal and quantitative framework case and control areas can be matched effectively using, importantly, a transparent approach to select control LAUs at an aggregated, instead of individual, level. We further described how, using a combination of matching and maximum variation matching, a quasi case–control design can be created for retrospective evaluation of a natural experiment. An important benefit of these approaches over, for example, qualitative selection of control areas is that, in the absence of the possibility to randomize case and control areas, this approach nonetheless approximates a randomized block experiment (with respect to the covariates used).[37] It further avoids the *de facto* use of neighbouring areas as controls for convenience, which may be subject to 'spill-over effects' because of the introduction of a new policy in its neighbouring area. Of course it is still possible that the PS model will result in matching of a case area to a neighbouring area, but matching is now based on a set of *a priori* defined criteria rather than just convenience (while it is of course always possible to explicitly exclude such matching). A further benefit is the transparency of the methodology that allows for independent evaluation of matching indicators.

We do not argue that this is the only methodology, and alternatives have been developed. For example, Abadie et al.[38] describe an alternative methodology of using 'synthetic' controls, which are weighted averages of potential controls with weights chosen to mimic indicators in the case area. This methodology has distinct benefits in circumstances where the intervention may be be a unique event or in situations where suitable control sites exposed to differing levels of the intervention are unavailable, but inference does depend on the reliability of a regression model to estimate the outcome and these

Please cite this article in press as: de Vocht F, et al., Propensity score matching for selection of local areas as controls for evaluation of effects of alcohol policies in case series and quasi case–control designs, Public Health (2015), http://dx.doi.org/10.1016/j.puhe.2015.10.033

can be very uncertain. Regardless, by adopting one of these, or another similar, quantitative methodology for selection of control areas, evaluation of public health policies can now be evaluated based on a better and more transparent methodology, instead of relying on unvalidated, convenience control areas.[38] An additional benefit is that the selection of two controls per case as done in this example, resulted in only 18 LAUs included in the study, which now enables collection of more detail information from these areas (if required). This would not have been feasible if all LAUs in England had been included.

The bias-reducing potential of PSM depends critically on the choice of the key variables used in the matching model.[39] Indeed, the *a priori* selection of the most important indicators from an available set of 93 indicators in our example will have involved subjective assessment. Some empirical studies have shown that after PSM substantial bias can still be present when compared to the results from RCTs,[40] although others showed good agreement between the two.[41] By utilizing a strategy analogous to that described at individual level by Patorno et al.[19] we aimed to mitigate this possibility by (a) selection of the most important key variables by experts independently and prior to development of the PS model, and specifically describing the approach for selection and evaluation[42] so that the PS model would have a good theoretical foundation, and (b) by specifying the PS model specifically for public health policies on alcohol. The latter is important since minimizing bias for one problem does not preclude the absence of bias in another problem.[39] Nonetheless, only key covariates selected by the team of experts were included, while data also had to be available for inclusion in the PS models. For example some of the experts would have liked to include an indicator of the 'nighttime economy',[43] but these were not available (although the percentage of bar workers in an area was included, which could be a proxy for the nighttime economy in an area).

Similarly, most alcohol policies are aimed to directly affect drinking habits, and thereby aim to indirectly affect harm. As such, inclusion of measured consumption at LAU aggregation would have been preferable, but these data are not available. In the Local Alcohol Profiles for England (LAPE) data consumption is only available as 'synthetic estimates' inferred from modelling of various indicators.[44]

Another limitation of the proposed methodology is the relatively small number of LAUs, which prohibits more detailed matching models. It has been shown that PSM works better with a large number of potential controls to select from.[40] However, there is only a limited number of LAUs (353 at most) and the main interest is comparing policies at this level (note that describing a methodology to deal with this formally was the aim of this work). As such we suggest a qualitative, in addition to quantitative, additional evaluation of the matched sets to evaluate that matched pairs seem appropriate, as well as some flexibility in the use of ASAM cut-off values and other quantitative measures. The number of units could be increased by conducting the analysis using smaller spatial units, which would have the benefit of evaluation of policies or natural experiments in parallel between and within the different areas in a LAU. Although preferable, the main limitation of this approach is that data required for matching will likely not be collected at such small level. Since no additional LAUs exist, an alternative approach could be to reduce the number of variables in the PS model by incorporating an additional dimensionality-reduction step such as to use a factor analysis methodology on the raw data and inclusion of the main factors in the PS model. However, this would include another level of assumptions for which the theoretical basis is unclear.

In conclusion, to evaluate the impact of new (alcohol) policies on public health ideally randomized experiments should be conducted. If this is not possible, for example because new policies have been introduced in certain LAUs prior to researcher involvement, because evaluation can only be done retrospectively, and/or outcome data are not available yet (such as when the researchers rely on routinely collected data as the outcome of interest), we described two methods that makes use of propensity score matching to nonetheless select the most appropriate control areas thereby at least approximating the *a priori* randomization procedure.

Author statements

Acknowledgements

This article presents independent research funded by the National Institute for Health Research (NIHR). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. The authors would like to thank Dr Matt Egan (LSHTM, NIHR SPHR) for valuable input.

Ethical approval

No ethical approval was required for this study because the analyses were based on open access databases of aggregated data.

Funding

This work was funded by the National Institute for Health Research School for Public Health Research (NIHR SPHR). NIHR SPHR is a partnership between the Universities of Sheffield, Bristol, Cambridge, Exeter, UCL; The London School for Hygiene and Tropical Medicine; the LiLaC 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.

Competing interests

None declared.

REFERENCES

1. Campbell R, Bonell C. *Development and evaluation of complex multicomponent interventions in public health. Oxford textbook of public health.* 6 ed. OUP Oxford; 2014751–60.

2. Rychetnik L, Frommer M, Hawe P, Shiell A. Criteria for evaluating evidence on public health interventions. *J Epidemiol Community Health* 2002;**56**:119–27.
3. 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. *J Epidemiol Community Health* 2012;**66**:1182–6.
4. Ziglio E. How to move towards evidence-based health promotion interventions. *Promot Educ* 1997;**4**:29–33.
5. Brown A, Moodie C, Hastings G. A longitudinal study of policy effect (smoke-free legislation) on smoking norms: ITC Scotland/United Kingdom. *Nicotine Tob Res* 2009;**11**:924–32. official journal of the Society for Research on Nicotine and Tobacco.
6. Lee PN, Fry JS, Forey BA. A review of the evidence on smoking bans and incidence of heart disease. *Regul Toxicol Pharmacol RTP* 2014;**70**:7–23.
7. Semple S, van Tongeren M, Galea KS, MacCalman L, Gee I, Parry O, et al. UK smoke-free legislation: changes in PM2.5 concentrations in bars in Scotland, England, and Wales. *Ann Occup Hyg* 2010;**54**:272–80.
9. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivar Behav Res* 2011;**46**:399–424.
10. Sturmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J Clin Epidemiol* 2006;**59**:437–47.
11. D'Agostino Jr RB. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics Med* 1998;**17**:2265–81.
12. Rosenbaum P, Rubin D. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;**70**:15.
13. Tannen RL, Weiner MG, Xie D. Use of primary care electronic medical record database in drug efficacy research on cardiovascular outcomes: comparison of database and randomised controlled trial findings. *BMJ* 2009;**338**:b81.
14. Böckerman P, Ilmakunnas P. Unemployment and self-assessed health: evidence from panel data. *Health Econ* 2009;**18**:161–79.
15. Ma X, Fleischer NL, Liu J, Hardin JW, Zhao G, Liese AD. Neighborhood deprivation and preterm birth: an application of propensity score matching. *Ann Epidemiol* 2015;**25**:120–5.
16. Yanovitzky I, Zanutto E, Hornik R. Estimating causal effects of public health education campaigns using propensity score methodology. *Evaluation Program Plan* 2005;**28**:11.
17. Pearl J. *Causality. Models, reasoning, and inference*. 2 ed. New York: Cambridge University Press; 2009.
18. Beynon C, Jarman I, Perkins C, Lisboa P, Bellis M. *Topography of drinking behaviors in England*. NorthWest Public Health Observatory. Available from: http://www.lape.org.uk/downloads/alcoholestimates2011.pdf; 2011 (last accessed 24 August 2015).
19. Patorno E, Glynn RJ, Hernandez-Diaz S, Liu J, Schneeweiss S. Studies with many covariates and few outcomes: selecting covariates and implementing propensity-score-based confounding adjustments. *Epidemiology* 2014;**25**:268–78.
20. Kleinbaum D, Kupper L, Muller K, Nizam A. *12-5-2. Collinearity concepts. Applied regression analysis and other multivariate methods*. 3rd ed. Pacific Grove: Wadsworth Publishing Co Inc; 1997240–5.
21. Gu X, Rosenbaum P. Comparison of multivariate matching methods: structures, distances, and algorithms. *J Comput Graph Statistics* 1993;**2**:15.
22. Austin PC. A comparison of 12 algorithms for matching on the propensity score. *Statistics Med* 2014;**33**:1057–69.
23. Vitcu A, Lungu E, BVitcu L, Marcu A. Multi-stage amximum vaiation sampling in health promotion prorgams' evaluation. *J Prev Med* 2007;**15**:13.
24. Ho D, Imai K, King G, Stuart E. *MatchIt: nonparametric preprocessing for parametric causal inference*; 2011.
25. Team RDC. *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing; 2008.
26. McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol Methods* 2004;**9**:403–25.
27. Humphreys DK, Eisner MP. Do flexible alcohol trading hours reduce violence? A theory-based natural experiment in alcohol policy. *Soc Sci Med* 2014;**102**:1–9.
28. Egan M, Katikireddi SV, Kearns A, Tannahill C, Kalacs M, Bond L. Health effects of neighborhood demolition and housing improvement: a prospective controlled study of 2 natural experiments in urban renewal. *Am J Public Health* 2013;**103**:e47–53.
29. Stronks K, Mackenbach JP. Evaluating the effect of policies and interventions to address inequalities in health: lessons from a Dutch programme. *Eur J Public Health* 2006;**16**:346–53.
30. Fone D, Dunstan F, White J, Webster C, Rodgers S, Lee S, et al. Change in alcohol outlet density and alcohol-related harm to population health (CHALICE). *BMC Public Health* 2012;**12**:428.
31. Herttua K, Makela P, Martikainen P. The effects of a large reduction in alcohol prices on hospitalizations related to alcohol: a population-based natural experiment. *Addiction* 2011;**106**:759–67.
32. Livingston M. Alcohol outlet density and harm: comparing the impacts on violence and chronic harms. *Drug Alcohol Rev* 2011;**30**:515–23.
33. Pridemore WA, Chamlin MB, Andreev E. Reduction in male suicide mortality following the 2006 Russian alcohol policy: an interrupted time series analysis. *Am J Public Health* 2013;**103**:2021–6.
34. Wicki M, Gmel G. Hospital admission rates for alcoholic intoxication after policy changes in the canton of Geneva, Switzerland. *Drug Alcohol Dependence* 2011;**118**:209–15.
35. Abadie A, Gardeazabal J. The economic costs of conflict: a case study of the Basque Country. *Am Econ Rev* 2003;**93**:112–32.
36. Cerda M, Morenoff JD, Hansen BB, Tessari Hicks KJ, Duque LF, Restrepo A, et al. Reducing violence by transforming neighborhoods: a natural experiment in Medellin, Colombia. *Am J Epidemiol* 2012;**175**:1045–53.
37. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics Med* 2007;**26**:20–36.
38. Abadie A, Diamond A, Hainmueller J. Synthetic control methods for comparative case studies: estimating the effect of California's Tobacco Control Program. *J Am Stat Assoc* 2010;**105**:493–505.
39. Pearl J. Understanding propensity scores. In: Pearl J, editor. *Causality models, reasoning, and inference*. 2 ed. New York: Cambridge University Press; 2009.
40. Peikes D, Moreno L, Orzol S. Propensity score matching: a note of caution for evaluators of social programs. *Am Statistician* 2008;**62**:9.
41. Dehejia R, Wahba S. Causal effects in nonexperimental studies: re-evaluating the evaluation of training programs. *J Am Stat Assoc* 1999;**94**:10.
42. Ali MS, Groenwold RH, Belitser SV, Pestman WR, Hoes AW, Roes KC, et al. Reporting of covariate selection and balance

Please cite this article in press as: de Vocht F, et al., Propensity score matching for selection of local areas as controls for evaluation of effects of alcohol policies in case series and quasi case–control designs, Public Health (2015), http://dx.doi.org/10.1016/j.puhe.2015.10.033

assessment in propensity score analysis is suboptimal: a systematic review. *J Clin Epidemiol* 2015;**68**:112–21.
43. Hobbs D, Lister S, Hadfield P, Winlow S, Hall S. Receiving shadows: governance and liminality in the night-time economy. *Br J Sociol* 2000;**51**:701–17.
44. Public Health England (PHE). *User guide: local alcohol profiles for England*. Available from: http://www.lape.org.uk/downloads/lape_guidance_and_methods.pdf; 2014 (last accessed 24 August 2015).

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.puhe.2015.10.033.

Please cite this article in press as: de Vocht F, et al., Propensity score matching for selection of local areas as controls for evaluation of effects of alcohol policies in case series and quasi case–control designs, Public Health (2015), http://dx.doi.org/10.1016/j.puhe.2015.10.033