evidence for
better health care

nuffieldtrust

A guide for evaluators

Evaluation of complex health and care interventions using retrospective matched control methods

Alisha Davies, Cono Ariti, Theo Georghiou and Martin Bardsley

August 2015

The NHS is undertaking a range of initiatives that are introducing new ways of delivering care services to patients. It is becoming increasingly important for policy-makers and decision-makers to understand what works, why it works, and what impact these changes are having on cost and patient outcomes.

One of the recurrent problems when evaluating the impact of new care models on outcomes is how to know 'what would have happened under a different approach to delivering care'. One approach that can be used is retrospective matched control analysis, whereby the impact of an intervention can be measured in terms of differences in the outcome relative to a matched control group. This method addresses the challenges of a simple time-trend analysis, or before and after comparison, and delivers a more robust evaluation to assess the impact of changes on outcomes and costs over time. The Nuffield Trust has applied this approach in multiple evaluations of health and care initiatives over recent years. In this guide we draw on that experience to highlight some of the key challenges in evaluation and introduce the retrospective matched study design as an alternative. We then outline ten steps towards retrospective matching to evaluate new health and care service models, which we hope will be of interest to those involved in evaluation at a local, regional or national level.

Ten steps to retrospective matched control analysis:

1. **Clarify the aims of the service and the evaluation**
   Is an evaluation needed? What is the intervention and who is the target population? What are the desired outcomes? How will the new service lead to the desired outcomes? How long will it take for an impact to be seen?

2. **Decide on the number of people needed to demonstrate an effect**
   Conduct a 'power calculation' – an estimate of how many patients will be needed in order to detect a given level of impact associated with a high level of certainty.

3. **Ensure permission is granted to access person-level datasets**
   Think about the important issues of anonymisation of data, linked patient datasets and gaining patient consent.

4. **Ensure there are data on who received the new services, and some information about the service received**
   Clearly identify whether or not patients in the datasets received the intervention being evaluated, and find out the date on which the intervention actively started, and ideally details about the specific components received.

5. **Identify the potential control population**
   Some options include: local people who did not receive the intervention; people resident in other similar geographical areas; or the national population.

6. **Create longitudinal patient-level histories of service use**
   Typically, at least two years of hospital data from before the start of the very first intervention are needed, as well as data to follow up a year or so after the intervention period.

7. **Identify matched controls**
   Usually undertaken by a specialist analyst, the variables for matching need to be identified; a decision taken as to whether matches are selected with/without replacements; whether to use multiple controls for each patient receiving the intervention; and a method selected to use for constructing a control group. The most important thing is to find a balance of variables that strongly predicts the outcome.

8. **Monitor outcome variables for those receiving the new service and matched controls**
   Having identified the individuals receiving the new service and controls, the next step is to monitor the outcome variables over time post-intervention.

9. **Undertake summative analysis**
   The conclusions could be based on a simple comparison of the difference between the intervention and control groups at the defined time points. Or, compare the changes within the two groups relative to their baseline – a difference-in-difference approach.

10. **Continuously monitor**
    For the majority of evaluations the results and interpretation are often more complicated than a simple yes/no answer, so continuous monitoring is often needed.

Our work on evaluation

The Nuffield Trust has developed evaluation methodologies that exploit the large amount of administrative information on individual patients that is available in the NHS and social care. We use these rich information sources to help policy-makers and professionals decide where to direct investment in the interests of patients and taxpayers.

Find out more about our evaluation work, including our evaluation projects looking at integrated care and telehealth/telecare, at: **www.nuffieldtrust.org.uk/our-work/evaluation** .

Introduction

As national and local drivers to improve quality and reduce cost in the NHS encourage the development of new models of health and care, the key questions for any initiative remain the same:

- Have patient outcomes and the quality of care improved?
- Has the patient or carer experience of care improved?
- Do the changes deliver better value (better use of resources for the outcome)?

Though historically many changes in health and care systems have been implemented without considering these questions, the past decade has seen increasing use of evaluations alongside the implementation of new models of care. Currently, for a range of initiatives such as the Better Care Fund, Integration Pioneers, the Prime Minister's Challenge Fund and those emerging from NHS England's Five Year Forward View (referred to as the vanguards), there is increasing recognition at both national and local levels about the importance of understanding what works, why it works, and to demonstrate impact on cost and patient outcomes.

> One challenge in many evaluations of health and care initiatives is interpreting the findings within the context of what would have happened if nothing had changed

One challenge in many evaluations of health and care initiatives is interpreting the findings within the context of what would have happened if nothing had changed (i.e. the counterfactual). This necessitates an appropriate control population where nothing changed. In practice this can be difficult, but there are alternatives. An innovative approach is to use routinely collected health and care data to generate a control population which can be matched to the intervention population on factors such as age, sex, level of deprivation, presence/absence of particular health conditions, prior use of hospital or risk of admission to hospital. This is a method known as retrospective matched control study design.

The aim of this Nuffield Trust guide is to raise awareness of retrospective matched control methods as one approach to evaluating complex health and care service change.

In this guide we highlight some of the key challenges with current models of evaluation, and introduce the retrospective matched control study design. We then outline ten steps towards retrospective matching in order to evaluate new health and care service models, which we hope those involved in evaluation at a local, regional or national level can build on. We also set out an example of how one area is working towards using retrospective matched control study designs in local evaluations.

This guide is aimed at those involved in the evaluation of service redesign at a local, regional or national level, and is likely to include health and care analysts, statisticians, public health professionals, commissioners and others with an interest in evaluating new models of care.

Many of the steps we set out in this guide do not need technical skills, but do need an understanding of health data and data systems. Steps 1 to 5 are useful for any quantitative evaluation, not just for matched control evaluative designs. An understanding of data analyses and statistics is needed from step 6, and technical statistical expertise will be required when putting the retrospective matching methods into practice. We acknowledge that this short guide cannot cover the complex statistical methods involved in matching, but we have included references to other key technical documents that will help.

The Nuffield Trust has developed and applied evaluation methodologies that make use of the large amount of administrative information on individual patients available in the NHS and social care. The use of matching techniques has been integral to our evaluative work and we have used these methods in many evaluations of community-based service innovations in the NHS (Bardsley and others, 2013).

Find out more about our evaluation work at: www.nuffieldtrust.org.uk/our-work/evaluation .

Why use retrospective matched control methods?

Challenges with common methods of evaluation

One of the recurrent problems when evaluating the impact of new care models on outcomes is how to know 'what would have happened under a different approach to delivering care'. There are three commonly used methods to address this:

1. **Randomising patients to an intervention and control group.** For many aspects of health care, the prospective randomised controlled trial (RCT) is held up as the best way to obtain evidence about the value of an intervention, and has also been used to evaluate health service initiatives (Steventon and others, 2012). However, RCTs can bring logistical and ethical difficulties in terms of recruitment of patients and organising interventions. There are also the limitations common to all RCTs, namely the generalisability to everyday practice and therefore how they can feed into policy decisions (Cartwright, 2007). For example, the strict inclusion and exclusion criteria in RCTs results in a very specific population being selected, and so the results may not apply to the 'real-world' population. Even with a pragmatic design without restrictive inclusion criteria (Roland and Torgerson, 1998), the trial may exclude certain patient subgroups or models of provision that are of key interest in decision-making, again limiting the generalisability of the findings (Gheorghe and others, 2013; McCarney and others, 2007; Rothwell, 2005). However, it is clear that other forms of evidence are valid and, in some cases, the use of observational studies (for example cohort or case control) are the only ones that are feasible (Black, 1996).

2. **Comparing changes over time.** In some cases, such as where the service change is focused on a specific population and point in time (for example implementation of a new vaccination programme on a specific date), a simple time-trend analysis may be useful. However, in the majority of complex service changes where multiple components of care and support are delivered over time (for example improved case management of frailty patients through a new complex care hub), and commonly within the local context of wider changes in health and social care, it is often difficult to determine whether a change in outcome is specifically linked to the intervention implemented.

   Another limitation of comparing changes over time, is 'regression to the mean' – demonstrating an effect which may have happened irrespective of any changes in care (Box 1).

3. **Comparing between areas.** Some evaluative studies compare outcomes at an area level (for example across local authorities). These ecological study designs are often used out of necessity, because those are the only data available. Area-level analyses do have their uses. For example, some interventions (such as changes to reimbursement rules) are not targeted at particular lists of patients, but to whole populations in a particular area – enabling comparison across areas with and without changes to reimbursement rules. In some cases interventions may consist of several strands, some of which operate at an area level – so it is difficult to unpick specific impacts on patients independently of what is happening around them. A key limitation of this 'ecological' approach to evaluation is the assumption that changes observed within a larger population are the product of the intervention. This assumption is more realistic when the number of patients in the study is large compared with the overall population. However, problems emerge if the number of patients involved is small. A further limitation is that the impact of new initiatives such as the integration of

new care models may not be visible at an area level, or at least not for some time, so an ecological study design is not appropriate. In these instances, methodological designs which follow specific groups of patients over time are often needed.

Box 1: Regression to the mean

The phenomenon of 'regression to the mean' can occur whenever something which varies over time is measured once and is then measured again at a later point in time. Observations made at the extreme the first time round will tend to come back to the population average the second time round. For example, the warmest place in the UK today is more likely to be relatively cooler tomorrow than warmer.

Regression to the mean is a particular challenge when an intervention is focused on particular types of patients (for example patients with high emergency care use). Say we look at people with frequent hospital admissions at present. On average, these individuals will have lower rates of unplanned hospital admissions in the future, even without intervention (illustrated by the control line in Figure 5). So, if a community matron is working with patients who are currently having frequent hospital admissions, they may notice how the patients have fewer admissions over time. However, this reduction might well have occurred anyway due to regression to the mean, and it cannot necessarily be attributed to the input of the community matron.

Regression to the mean occurs simply because after one extreme period, the next period is statistically likely to be less extreme.

What is retrospective matching and how might it help?

Comparisons over time or between geographical areas have their uses and limitations, but carrying out any evaluation without a comparison (control) group means we do not know what would have happened in the absence of the service changes.

RCTs also have their limitations (as discussed above) and identifying a control population can be challenging in practice, especially if the new service is made available to the entire target population at the same time (for example the reconfiguration of care pathways for all frail elderly patients, or implementation of a new community diabetes service for the population of a clinical commissioning group).

An alternative approach is to retrospectively identify a valid comparison (control) group from routinely collected, computerised, patient-level health and care data.

Over the last few years, researchers at the Nuffield Trust have used a method of identifying 'retrospective matched controls' to assess the impact of many service evaluations (Georghiou and Steventon, 2014; Steventon and others, 2011; Chitnis and others, 2012).

Retrospective matching is a way of creating a form of control group which can be used to judge whether changes in outcomes for people using a service were any different from what would have been expected anyway from usual care.

By using existing data (usually hospital data, or in some cases social care data), outcomes can be followed (for example emergency admissions to hospital) for a group of patients receiving a new service or intervention. Using the same data, the characteristics of the patient and prior history before the start of the intervention can also be looked at, and individuals found who look very similar – but who did not receive the intervention. The matching process can use quite complex statistical methods and can take account of many different variables. The control and intervention populations can be matched on factors such as age, sex, level of deprivation, presence/absence of particular health conditions, prior use of hospital, risk of admission to hospital and so on.

An important strength of matched control methods is the availability of a diagnostic test since it is possible to assess how similar the intervention and matched control groups are at baseline. Although not every variable can be observed and accounted for in the matching, examining of tables of baseline characteristics is useful to prompt a conversation about the similarity of the groups.

The end result is that the impact of the intervention can be measured in terms of differences in the outcomes relative to the matched control group.

This method addresses the challenges of a simple time-trend analysis or before and after comparison, and delivers a much more robust evaluation to assess the impact on outcomes and costs over time. Linking across existing datasets to construct individual patient histories and identify matched controls makes it a timelier and cheaper approach to evaluation compared with an RCT. By using routine datasets there is the added advantage of being able to provide interim results and feedback during the evaluation period – and to potentially help fine-tune the intervention and the measurement process.

Box 2: Benefits and limitations of a retrospective matched control study design

Benefits

- Uses existing datasets so is relatively cheap to implement and allows large volumes of cases to be studied.
- Can be carried out on services/interventions that are already in place.
- Can provide intermediate results to track progress over time.
- Can be used to study 'real life' care delivery rather than the artificially controlled environment of clinical trials.
- Can be used to look at a range of secondary outcomes.

Limitations

- Can only be used where the key outcome measure is routinely collected at person level.
- In some cases, not all members of the intervention group can be matched.
- The matching process can create very similar groups in terms of the data used to match, but there might be other, hidden factors that explain differences between the intervention and control groups (unobserved confounding).
- Requires access to data and permission to link data over time – and potentially across sectors of care delivery.

Ten steps towards retrospective matched control analysis

1. Clarify the aims of the service and the evaluation

What needs to be evaluated, the expected outcomes and the timeframe can all have a major impact on how the evaluation is designed. So, before designing the evaluation, consider the following questions:

- **Is an evaluation needed?** It is worth clarifying whether an evaluation is needed in the first place. In particular, how will the findings from an evaluation process, positive or negative, inform future decisions? Is there a local or national commitment to the service/initiative that will overpower any negative findings? Is the service/initiative being evaluated sufficiently well established to investigate whether it has achieved its aims? If it is changing over time, will an evaluation help inform development, or is the model inflexible to change?
- **What is the intervention and who is the target population?** It can be challenging to define the specific intervention/service change and the target population. For example, a study evaluating the impact of multi-disciplinary teams (MDTs) in four different clinical localities within a single CCG needs to consider: Is the MDT being implemented using a consistent approach in each area? Are eligible patients identified in the same way? Is it the MDT or what happened as a consequence which will have the impact on outcomes (patient, staff and costs)? Asking what is being evaluated helps provide clarity from the outset on what is and is not included in the intervention/service change.
- **What are the desired outcomes?** It is also important to explore the desired impact of the changes across a range of outcomes. This might include:
  - patient outcomes (clinical outcomes, population outcome measures)
  - user and carer satisfaction/experience
  - intermediate markers of patient knowledge, attitudes or behaviour, for example patient-reported outcome measures or patient activation measures
  - appropriateness of care
  - efficiency and cost
  - organisational impacts and staff perceptions
  - stakeholder perceptions.

  A comprehensive evaluation could encompass all of these aspects, requiring a number of different study designs and a range of investigators, and ultimately increasing the cost of the evaluation. In practice, there have to be choices made about which sets of outcomes are most important in a given situation.
- **How will the new service lead to the desired outcomes?** At the start of the design process it is important to be explicit about the potential impact of the new service on desired outcomes, and to identify the specific components/processes that will achieve the outcome. A clear understanding of the potential outcomes, where these will be noticed (for example at the patient level or service level) and over what timeframe, will help inform the evaluation design.

  Useful approaches include 'logic models'[1] or 'theory of change' (Connell and others, 1995) to document the relationship between service changes and outcomes. These can

[1] See for example: www.healthscotland.com/scotlands-health/planning/logic-models.aspx

also help inform a realistic view on the time needed to implement the changes, and the timeframe for the expected outcomes.

In practice, understanding exactly how the new service model will achieve the desired outcomes is often vaguely defined. The value of articulating the links between changes in process and ultimate effects are two-fold; first it can act as a check that the chains of causality are reasonable; and second, it may help identify shorter-term process measures which can indicate progress towards the longer-term outcomes.

For instance, in a number of our studies, we have noted that while a reduction in emergency admission was not seen, changes in outpatient and elective care were observed – the latter could be an early marker of change for emergency inpatient care (Bardsley and others, 2013; Roland and others, 2012). Other interim markers might be clinical, such as improved disease control, or reflect patient measures such as patient activation scores.

- **How long will it take for impact to be seen?** Bringing about change in health systems is challenging and requires significant time and resources (Best and Lewis, 2012; McNulty and Ferlie, 2002). Development of both the intervention and evaluation takes time. For example, one year of operation is unlikely to show much result beyond the process of initial set-up and implementation. With more complex initiatives, such as the development of integrated care models, it may take many years to see an impact. For example, many existing integrated care models (such as Kaiser Permanente and Geisinger in the United States) which the NHS looks to learn from have been established for many years, if not decades, with stable leadership over that period.

  In many cases, tight financial controls and a desire to show a short-term return on investment in new service models puts pressure on evaluators to deliver results over much shorter time frames – and is sometimes accompanied with disappointment about the lack of effect.

  It is difficult to suggest what might be 'enough time' to demonstrate an effect, as a realistic timescale for evaluation will depend on the complexity of the intervention, the outcomes and the context within which it is being implemented (see Box 3).

Box 3: Factors to be considered when estimating the time needed for impact to be seen

1. The need to communicate changes to staff, and gain their support to make changes.
2. Providing resources and materials to help staff implement the changes in care.
3. Whether the new care pathways/processes are being implemented in different ways across the local health and care providers, thereby adding complexity.
4. Whether the new care models are being implemented across all providers in the area, and whether they impact on contractual agreements.
5. Allowing time to recruit/enrol a sufficient number of patients to actually receive care via the new pathways/processes.
6. Allowing time for a large enough sample of patients to have been actively managed in the new care model.
7. Allowing time for patients whose health has changed to demonstrate improved outcomes and show an overall statistically significant benefit.

As discussed, logic models may be useful in helping identify process indicators that are able to show change in the shorter term, and give an indication that the change is moving in the right direction towards the main outcome of interest.

2. Decide on the number of people needed to demonstrate an effect

It is standard practice at the beginning of an evaluation of a health model to conduct a 'power calculation': an estimate of how many patients will be needed in order to detect an assumed level of impact associated with a high level of certainty. If the evaluation does not include a sufficient number of patients, it may conclude that there was no evidence of an effect – not because no effect occurred, but because there were not enough people to be able to demonstrate an effect.

A statistician will be able to estimate the required sample sizes based on appropriate assumptions (Merrifield and Smith, 2012). The numbers of people required can be very high for studies where the outcome event is rare. For example, to detect a 20 per cent change in the number of emergency admissions per person over 12 months, 2,100 patients need to be recruited; to detect a smaller change (10 per cent), a large sample size (30,000 patients) needs to be recruited – and the equivalent number of controls identified in both cases. This assumes that:

- Power is 90 per cent: set at this level, we want there to be a 10 per cent or less chance that we will miss a real difference in the outcome (in this example a 20 per cent change in the number of emergency admissions).
- Type 1 errors are set at ≤0.05: there is a risk in any study of a false positive result (demonstrating an effect when actually no difference exists) and this is known as the type 1 error. At this level (0.05), we are willing to conclude that the intervention had an impact on emergency hospital admissions with a 5 per cent or less probability of that occurring by chance alone.
- A two-sided test is used: so the effect could be higher or lower than expected.
- A baseline rate calculated from appropriate historical patients (within hospital episode statistics (HES) or published estimates from the literature that represent the target group that the intervention is aimed at): for example, the mean admission rate for those with complex needs (defined as a record of diabetes, chronic obstructive pulmonary disease or heart failure with two or more previous emergency admissions) in HES was 0.97 per person and a standard deviation of 1.85. Note, however, that if the target group for the intervention has a relatively low baseline level of activity in terms of the outcome (for example, if the evaluation is looking at whether hospital admissions are reduced in a healthy group of younger people), relatively large sample sizes will be needed in order to show an effect.

In many studies the scale of the intervention is small (for example fewer than 1,000 patients), especially for a pilot project. In these cases, there are strategies that can be used to improve the probability of finding an effect, if one exists, with small sample sizes:

- **Tolerate a higher level of false positives.** The use of type 1 error of ≤0.05 as a measure of statistical significance is only due to custom. A smaller sample size will increase the type 1 error – increasing the chance of demonstrating an effect when no difference actually exists (false positive). However, for a pilot study it is important to retain a study's power (probability of finding a difference if one exists) at the expense of increasing the false positive rate. In a sense, the penalty for missing a true effect if it exists in a promising new intervention is much higher than a false positive result, which may result in simply lengthening the study for an ineffective intervention.

- **Choosing alternative outcomes.** The sample size will depend a lot on the natural variability in outcome. Emergency admission rates are quite variable because of many different factors, both systematic and random. It may be possible to use alternative outcome measures that are more statistically stable. For example, instead of a 'reduction in emergency admissions', it may be better to choose 'time to emergency event'.

3. Ensure permission is granted to access person-level datasets

As retrospective matched control methods involve linking individual patient records over time, and sometimes joining different datasets, there are important issues around information governance to address:

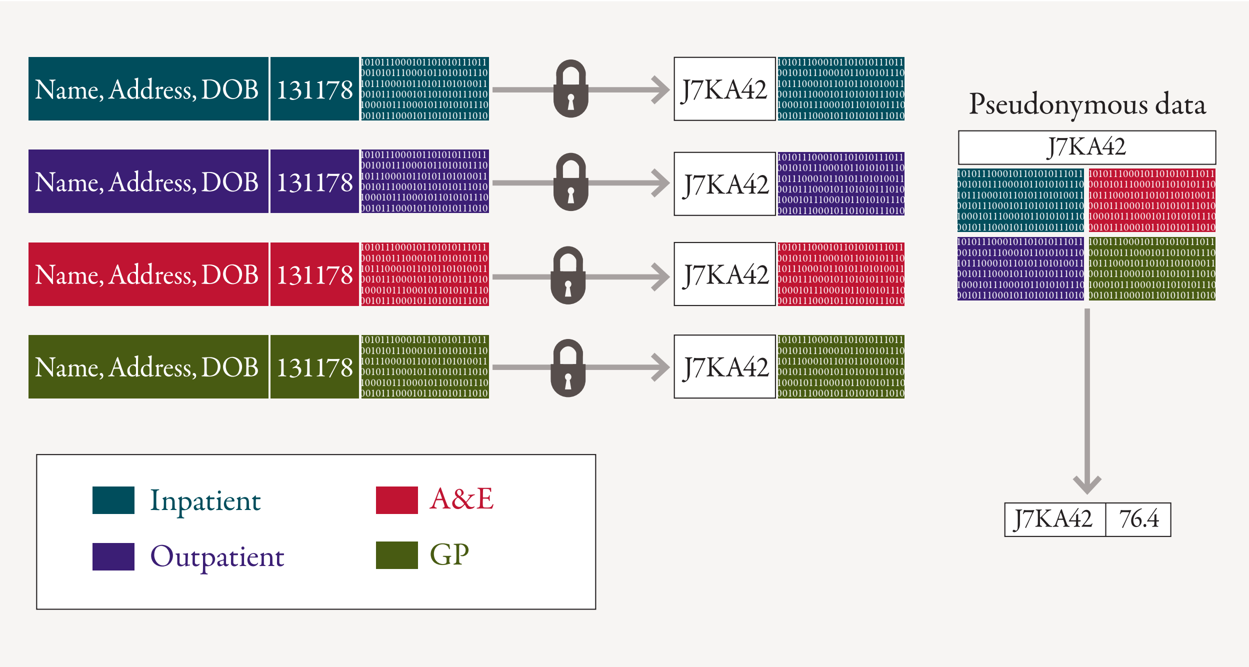
- **Anonymised data.** In some situations local datasets link patient records over time and between sectors using a series of anonymised keys (often called pseudonymised data; see below). With these linked data it is possible to track patients through health and care systems using a unique identifier. Although there are still important information governance issues around access and use of the datasets, once linked datasets have been created they make secondary analysis much easier.
- **Informed consent.** The general principle is that, where practicable, patient consent should be sought to share data. Although clearly desirable, for these retrospective studies patient consent can be problematic. If gaining consent is not possible, section 251 of the Health & Social Care Act can potentially be used, with approval from the Confidentiality Advisory Group (CAG). Section 251 is intended for use in exceptional circumstances where consent cannot be obtained and decisions on its application are made by the CAG (Health Research Authority, 2015).

One strategy that can be used with local data is to adopt an approach of pseudonymisation at source. Pseudonymous data have all personal identifiers (names, addresses, dates of birth, NHS numbers) removed, but each individual is allocated a unique code (or 'pseudonym').

Although anonymous for all intents and purposes, the pseudonym enables analysts to link together information relating to a particular individual from multiple databases (see Figure 1). It also allows the potential, under certain circumstances, for the manager of the database to re-identify each individual at a future time, usually via a 'key' that decodes the pseudonym back into the NHS number. In this sense, pseudonymous data are neither identifiable nor anonymous because all personal identifiers have been removed, but identification is still possible through the pseudonym and the key.

In England, obtaining linked data may be facilitated via the Health and Social Care Information Centre (HSCIC) or local Commissioning Support Units (CSUs), as long as the CSU is an accredited safe haven. For example, patient identifiers of those receiving the new service can be sent (with appropriate consent) to the HSCIC or CSU, who can carry out the data linkage and send the research team the pseudonymised identifiers of the HES records. This process may take up to three or four months if carried out via the HSCIC, but may be facilitated more quickly via the CSU, depending on local capabilities and capacity.

Figure 1: Linkage of pseudonymous data to calculate a person-level risk score



4. Ensure there are data on who received the new service, and some information about the service received

The ability to track events at person level is essential for retrospective matched control techniques. The main analysis will normally aim to use information from routine hospital administrative systems such as national HES or the Secondary Uses Service (SUS), and perhaps local GP practice and social services systems. These data will be used to identify matched controls and examine changes in outcome measures.

Before it is possible to do this, however, it is necessary to be able to identify whether or not patients in these datasets received the intervention (i.e. the service) being evaluated.

Those providing the service will need to collect some patient identifiers for people receiving the new services (i.e. those in the intervention group), for example name, date of birth, address/postal code (see Box 4, page 12). Depending on the information governance permissions in place and who will undertake the data linkage (see step 3), the evaluation team may not have direct access to this person-level information. However, this patient-identifiable information must be collected by the service so that the linkage to hospital data and other datasets can be done by either the HSCIC or CSU.

Some information about the service received is also needed. At a minimum this should include the date on which the intervention actively started for each individual, but might also include more detail about the specific components received, the number of contacts, etc. It is more helpful if this information can be collected electronically on existing patient-level data systems (for example by adding READ Codes specific to the service to document contacts within existing patient data management systems).

It can sometimes be important to be able to identify those who did not receive the intervention (for example if they were referred but did not meet the eligibility criteria, or if they were eligible but refused the service). When looking for controls from the local area it helps to ensure that any control patients have almost certainly not received the intervention, and that patients who have been identified as ineligible for the service are also excluded from the control group.

Box 4: Example of service level information about patients receiving a particular intervention

Identifiers that might be collected:

- NHS number (if available)
- Date of birth
- Sex
- Post code
- First name
- Last name

Service details:

- Start date
- End date (if available)
- Description of service including eligibility criteria for entry (if applied), referral routes etc.

5. Identify the potential control population

Before starting the matching process, the pool of potential controls must be identified. Ideally a control population should:

- reflect the population being compared
- be as much like the intervention group as possible (for example, if the intervention is targeted at patients aged 65+ years, then the control population should also be selected from those aged 65+ years)
- be from the same area (local), since selecting controls form other geographical areas risks bias from differences in the outcomes between areas (or measurement error)
- be as large as possible.

There are a number of choices for controls:

- **Local people who did not receive the intervention.** This approach has advantages in that a local control group will be affected by the same local factors as the intervention group. For example, a particular health economy might have a specific propensity to admit patients to hospital (based on local bed availability). The limitation with selecting controls from the local population is that the control group may include people who were 'rejected' from the intervention group for some specific reason. As a result the population to draw controls from may be contaminated by people who are 'different' in some important way from those who were eligible for the new service.
- **People resident in other similar geographical areas.** In this approach areas of the country are selected that are similar in terms of the demographics or health systems to the local population, but where they are not implementing the same intervention (this can be found out by reviewing published literature, or by phoning them and asking). This approach usually means being restricted to using only nationally available datasets such as HES to examine differences in outcomes. A further limitation is that the control area may change

its own practice and the evaluation team may not be aware of it. Selecting matched controls from non-local populations risks introducing biases into the study, due to unexplained variations in the outcomes between geographical areas (Steventon and others, 2015).

- **National population.** In this approach the potential controls are drawn from the national population, with no explicit attempt to find areas that are not implementing the same (or similar) intervention. Across the country many areas are implementing a mix of different initiatives to improve health and care, and outcomes in the control group are likely to be the product of many different services. Therefore, this method effectively tests whether local outcomes are deviating from average performance across the country. It is the most conservative in looking for success in that it would probably err toward not showing an effect.

Note that some of the differences between the intervention and control groups at baseline can have less of an impact if a 'difference-in-difference' method is used when analysing results. Further details are given in Box 6 (page 18).

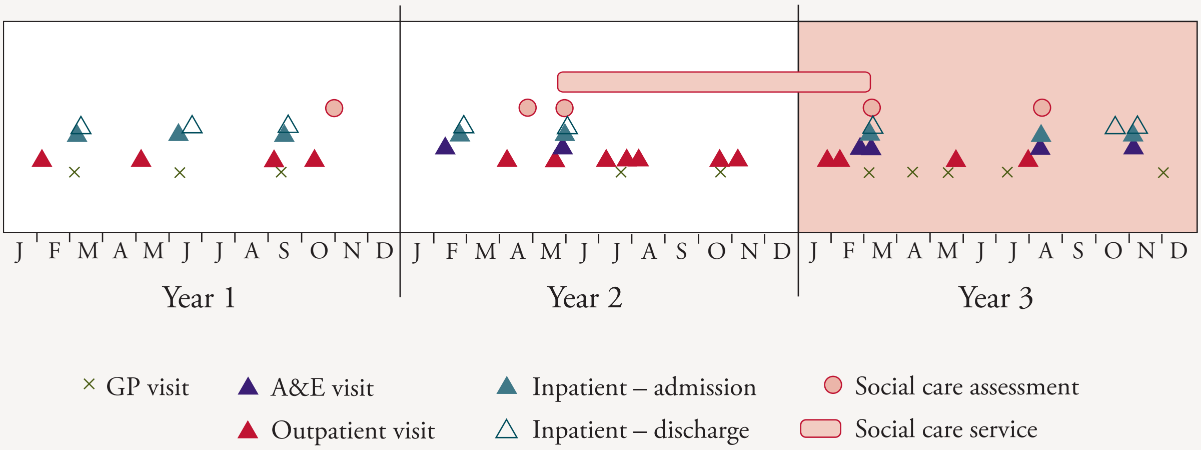
6. Create longitudinal patient-level histories of service use

When creating histories of patients' health and care activity over time, a typical approach is to use at least two years of data before the intervention start date, and to follow up a year or so after the intervention period – although shorter timescales can be feasible.

The hospital datasets that might be used include admissions and attendances for inpatient, outpatient and A&E activity. People will typically have one or more of these care events over a period of time so we need to combine the records of individual events into a person's care history. There are a number of ways to do this but the simplest is to build up a dataset with one row of records for each person in the potential control pool and the intervention group (this might be, for example, all over-65s in one area, if the intervention group are all over 65).

Illustrating an individual's health and social care history over a three-year period (Figure 2) can help bring the data to life, and is a useful way of gaining an understanding of the frequency and patterns of contacts with health and care providers for individuals in the study.

Figure 2: Diagrammatic illustration of an individual's health and social care history over a three-year period



Source: Bardsley and others, 2011.

In the example in Figure 2, in the first year the patient had four outpatient attendances and three hospital admissions, as well as some GP visits. A social care assessment was carried out towards the end of the year, but this appears not to result in any service being provided. In the following year, two social care assessments were carried out and a low-intensity package of home care was put into place. Using the information from years one and two, we can then predict likely care usage in year three. In this example, our model predicted an increase in the intensity of the home care packages or a care home admission. However, this was not observed: in the third year several unplanned hospital admissions occurred, as well as two social care assessments, but social care services did not continue past March in year three.

Comprehensive patient-level data such as this can then be used to generate variables on health and care usage for the retrospective matching, for example the average number of outpatient appointments in the past month, the number of social care assessments per year, and so on.

The raw data from the different datasets have to be structured and coded into a defined and consistent set of variables to put into the model. Each person in the dataset will need to be assigned a set of flags relating to particular events or attributes of that person – information that will be used in the control group matching process. This will include flags describing demographic factors (age, sex, deprivation measures), the presence or absence of diseases recorded in the person's inpatient history, and recent and more distant counts of hospital visits (relative to the intervention start date in the case of the intervention group, and to equivalent dates in the wider potential control pool). Identifying and generating the variables used for matching may be limited to what is available in the datasets. At the Nuffield Trust, we typically use variables such as those listed in Box 5, although the importance of individual variables may vary by study.

7. Identify matched controls

The actual process of matching is something that should probably be undertaken by a specialist analyst. However, the steps involved are outlined here.

Box 5. Examples of variables used in matching

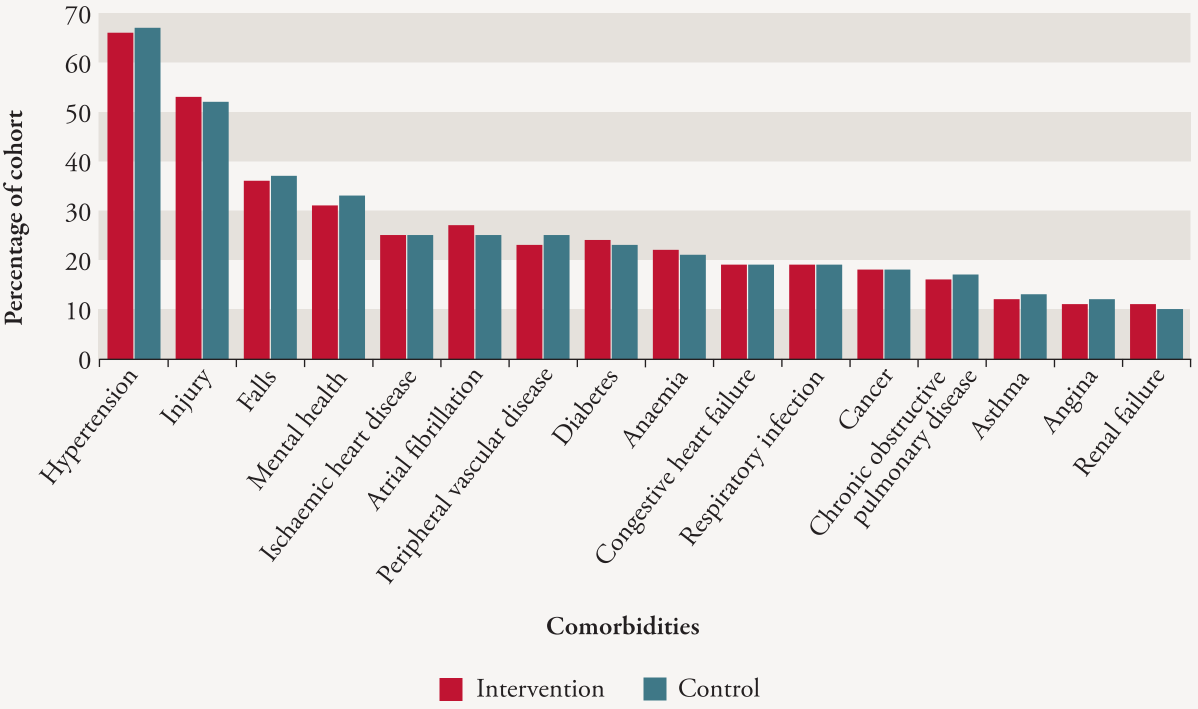
- Age
- Gender
- Index of Multiple Deprivation of local area
- Risk of admission score
- Presence of specific diseases
  - e.g. cancer, coronary heart disease, diabetes, chronic obstructive pulmonary disease, dementia
  - number of long-term diseases
- Prior hospital activity (1 month, 6 months, 3 years)
  - inpatient admissions
  - emergency admissions
  - A&E attendances
  - outpatient activity
  - bed days

Having selected the pool of potential controls (step 5) and variables for matching (step 6), there are then some basic decisions about whether matches are selected with/without replacements and whether to use multiple controls for each patient receiving the intervention (Stuart, 2010). These are largely technical questions that can impact on the statistical power. For example, 1:1 matching involves selecting one control patient for each treatment or intervention patient.

Then it is time to find control individuals that match most closely to the population receiving the intervention. The aim is for the control group to have the same distribution of relevant characteristics as the intervention group did just before the start of the intervention. The most important thing is to find balance on the variables that strongly predict the outcome. For example, if the outcome is hospital admissions, then matching on prior hospital admissions is crucial as this is a very important predictor of future admissions. There are several methods for constructing such a control group:

- Matching several of the underlying characteristics at once, without attempting to summarise them into a single figure, using Mahalanobis metric matching or genetic matching (Rosenbaum and Rubin, 1985; Diamond and Sekhon, 2013).
- Matching according to a *propensity* score. The propensity score summarises, as a single figure, characteristics that reflect the likelihood that a given person received the intervention. A control group is then determined by selecting people with similar propensity scores to those in the intervention group (Rosenbaum and Rubin, 1983).
- Matching according to a *prognostic* score. The prognostic score is a summary of the characteristics that reflect the likelihood that someone would experience the outcome

Figure 3: Example of how intervention and control groups are matched on a series of variables indicating the presence of prior disease



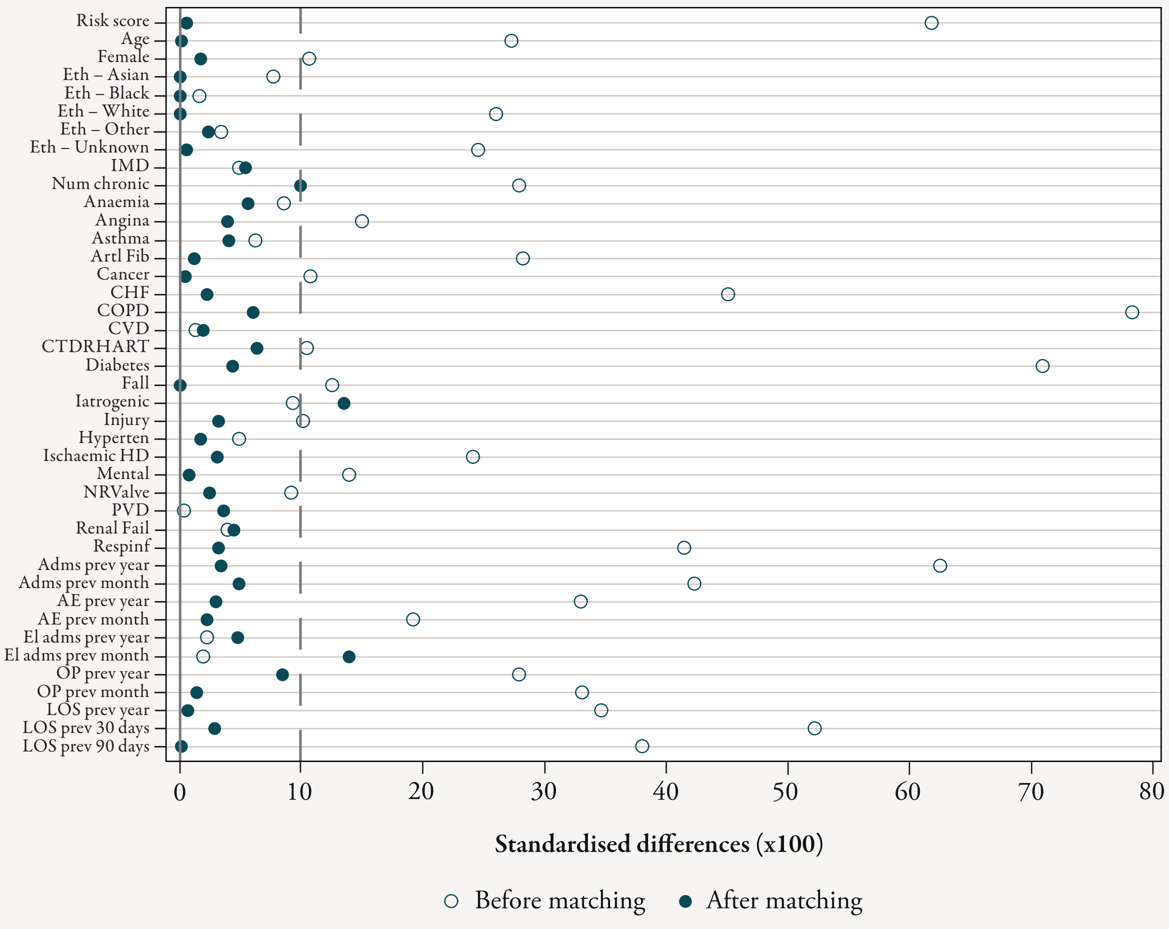
Source: Georghiou and Steventon, 2014.

of interest, in the absence of the intervention. So, for example, a control group might be selected to have exactly the same risk of future emergency admissions (Hansen, 2008).

At the Nuffield Trust, our preferred approach is based on a prognostic scoring technique – one that optimises the performance of the underlying predictive models. To derive our prognostic score, we develop predictive models focused on emergency hospital admissions (this is almost always the main outcome of interest). These models are similar to the Patients at Risk of Re-hospitalisation (PARR) model that has been widely used by the NHS in England. The models attribute a number between 0 and 100 for every person with a recent inpatient admission that reflects their probability of having an emergency hospital admission within 12 months. These models are calibrated on people who did not receive the intervention at any point. This helps us to derive an estimate of the probability of emergency hospital admission in the absence of receiving the intervention. Matching is an iterative process and the end point (or achieving successful matched controls) needs to be judged by the balance across a range of key variables in the control and intervention groups (Figure 3).

The success of matching is usually expressed in terms of the standardised difference (the difference in means as a proportion of the pooled standard deviation) – where smaller values indicate better matches. There is a general rule of thumb that standardised differences should be less than ten. However, if balance across the intervention group and controls cannot be reached, then the conclusion might be that the dataset is not adequate

Figure 4: Standardised difference across variables used for matching; before and after matching



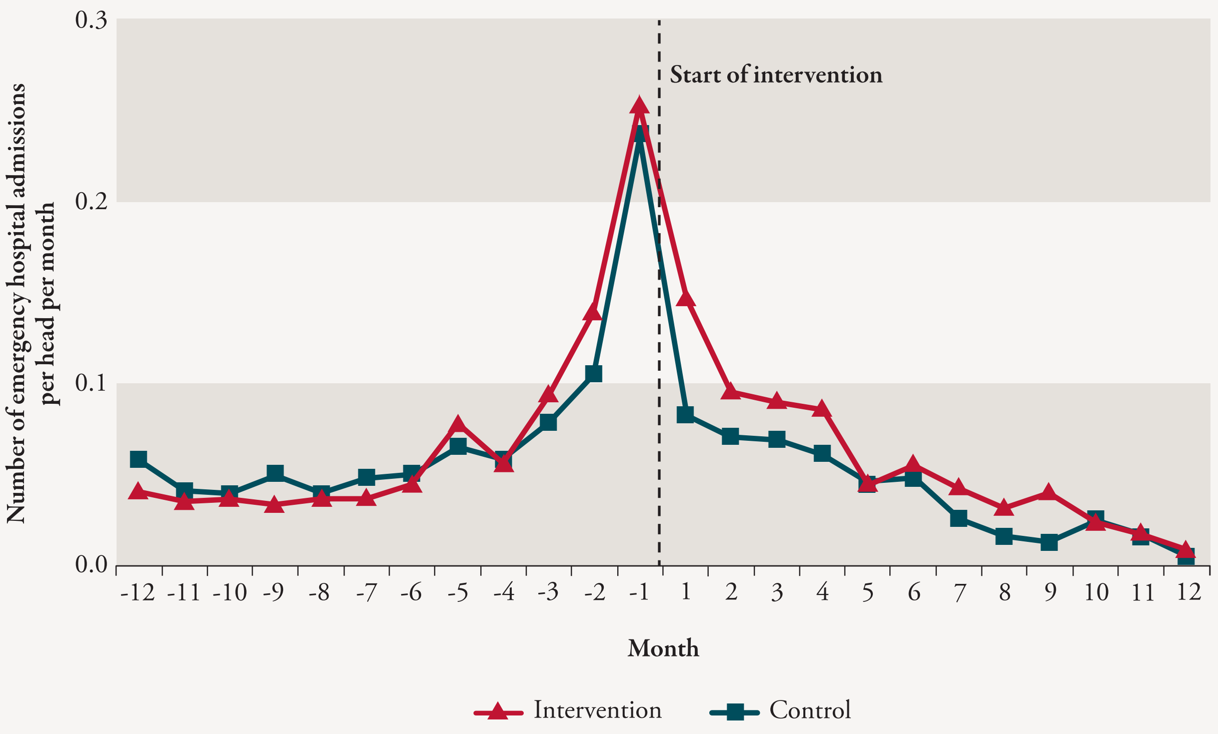
to answer the question being asked of it. Figure 4 compares standardised differences in the whole potential population before (open dots) and after matching (closed dots) across a range of variables used in one study. In most cases matching dramatically reduced the standardised differences – especially for more important variables (in this example these are 'overall risk score' and 'admissions in the previous year') – whereas in some cases the matching does not improve the standardised differences (in this example 'emergency admissions in the previous month', shown as 'AE prev month' here). It is important to remember that the idea is to achieve a balance across a range of variables.

8. Monitor outcome variables for those receiving the new service and matched controls

Having identified the individuals receiving the new service and controls, the next step is to monitor the outcome variables (as defined in step 1) over time post-intervention.

The example below (Figure 5) shows the results of a study that the Nuffield Trust evaluated which looked at a scheme to support older people who had been in hospital (Steventon and others, 2011). The intervention group showed a sharp increase in admissions before the start of the intervention, which is what would be expected when candidates for the service were people already in hospital. Following the intervention, the number of future emergency admissions fell away dramatically – but could this have been a result of the intervention, or alternatively regression to the mean (see Box 1)? To answer that question we made use of the retrospective matched control method to compare the level of activity pre- and post-intervention with a control group.

Figure 5: Comparing the number of emergency admissions pre- and post-intervention and for a matched control group – regression to the mean



Source: Steventon and others, 2011.

The controls were selected in part because they showed an almost identical rise in emergency admissions relative to specific months equivalent to the intervention start dates. In this way, they appeared to closely match the intervention group, but they were also matched on a range of other factors.

The 'future' emergency admissions of this control group appeared to decline even faster than the intervention group, suggesting that emergency admissions in the intervention group were indeed just regressing to the mean (Figure 5).

9. Undertake summative analysis

Difference-in-difference analysis
Having created a matched control group, the conclusions could be based on a simple comparison of the difference between the intervention and control groups. This would typically be the situation for comparisons of outcome measures such as mortality. However, if the same outcome is measured pre- and post-intervention, there are a number of more robust approaches. Despite matching there may remain slight underlying differences between the intervention and control groups, so it is generally better to compare the changes within the two groups relative to their baseline – and test whether the change in outcome found in the intervention group is greater than that found in the control group – a 'difference-in-difference' analysis (see Box 6).

This is a fairly standard approach, especially in econometric analysis, and it fits here for most cases. It also offers some reassurance about the success of the matching. For instance, when comparing differences in emergency admissions, if the baseline emergency admission rate in the matched control group is 20 per cent lower than the intervention group, then this may suggest that the matching was not successful and that there are other unobserved variables which need to be accounted for at the matching stage.

It is also important to consider how the relative effect of this unobserved variable on the outcome in the intervention group compared to the control group changes over time.

Considering hidden confounders
To take a more realistic example: suppose we compare emergency hospital admission rates for a group of patients receiving an intervention, but exclude people living in care homes. And say we know that living alone is associated with lower rates of admission, but when it comes to selecting controls we know very little about home arrangements so our control group may also include some people who live in care homes. Therefore, our baseline rate in the control will be slightly higher (by five per cent, say) than ideal due to this. So when we compare control and intervention after the intervention has taken effect, we look to see whether that difference has reduced – indicating a greater fall in admission in the intervention group. The comparison will only work if we are reasonably sure that the proportion of people living in care homes has stayed reasonably constant during the course of the analysis, i.e. a time-invariant confounder. This problem is more pronounced where we cannot use a difference-in-difference approach. For example, in a study of mortality using HES-based controls, there is no routinely collected data to generate a variable that measures smoking status. If there is an imbalance in the proportion of smokers between the intervention patients and controls, this may be the cause of mortality differences, not the intervention.

The problem of 'hidden confounders' creating bias is an important one to bear in mind when using retrospective matched control methods. One way to mitigate the risks is

Box 6. Example of difference-in-difference results table on secondary care utilisation

The table below gives an example of how a results table may look for a difference-in-difference analysis.

A difference-in-difference approach compares the change in outcome within the intervention group to the change in the outcome within the control group, over two time points.

	Intervention (N=556)			Control (N=556)			Intervention effect (g)
	Before (a)	After (b)	Change (c)	Before (d)	After (e)	Change (f)	
Emergency admissions per head	1.42 (1.40)	1.06 (1.54)	-0.35** (1.78)	1.38 (1.32)	0.80 (1.30)	-0.58** (1.47)	0.23** (1.95)
Emergency length of stay	9.45 (16.68)	19.63 (26.52)	10.18** (30.56)	10.20 (15.55)	12.27 (22.45)	2.06 (25.44)	8.11** (34.45)
Elective admissions per head	0.50 (1.05)	0.53 (0.96)	0.03 (1.27)	0.43 (0.96)	0.51 (1.07)	0.08 (1.14)	-0.05 (1.41)
Outpatient attendances per head	2.73 (4.14)	2.04 (2.87)	-0.69** (4.23)	2.49 (3.64)	2.42 (3.53)	-0.07 (3.79)	-0.62** (4.40)

Figures are based on the six months before/after intervention.

* Statistically significant at the 5% level; ** Statistically significant at the 1% level.

The change in emergency admissions for the intervention group, comparing rates before (a) and after (b) the intervention, was a 0.35 absolute reduction in emergency admissions (a-b=c).

In the control group, the change over the same periods before (d) and after (e) the intervention was a 0.58 absolute reduction in emergency admissions (d-e=f).

The difference-in-difference compares the difference between the change in the intervention group and the control group over the same two time points (c-f=g). So in this example the emergency admissions rate increased by 0.23 in the intervention group (after taking into account the change in the control group over the same time points).

to undertake sensitivity analyses to test the robustness of the findings to time-variant unobserved confounding. So, for example, test whether the intervention and control group differ for an outcome measure that was not expected to be influenced by the intervention. In a study of emergency admissions, test for differences in length of stay; or for a difference in ambulatory care sensitive admissions, compare inpatient admissions for hip fractures.

A more sophisticated approach is to assess how strong the unobserved confounding effect would have to be to alter the main conclusions from the analysis. So, for example, a statistician may simulate a hypothetical unobserved confounder and estimate the odds ratios required between this confounder and intervention status and outcome to alter the results and conclusion. The values obtained would be compared with estimates of odds ratios for unobserved confounders, based on another study. In the end it boils down to a judgement about how likely it is that such an effect will exist in the study group.

10. Continuously monitor

The majority of the evaluative studies we have completed at the Nuffield Trust have been summative in their design in order to ascertain whether the service had an impact (often on cost and patient outcomes) after a certain length of time. Innovators are hopeful the evaluation will evidence a positive outcome for patients and staff, to help justify the initial investment and potentially achieve future funding for continuation of the service.

Whereas the funders seek unambiguous evidence on the benefits – usually in terms of cost savings, yes or no, success or failure – in the majority of evaluations the results and interpretation are often more complicated than a simple yes/no answer. Scenarios include:

- Something partly worked (for example demonstrating improvements in patient-reported outcome measures, but not in emergency admissions).
- The original model changed, making it difficult to attribute changes in outcome to specific services, or affecting the sample size as changes to the model may result in people receiving different versions of the initial service.
- The implementation phase took longer than expected, so patients have not been in the service long enough to demonstrate any change in outcome.
- Even after implementation there may have been challenges in delivery or recruitment – reducing the number of individuals in the intervention group to demonstrate any change in outcome.
- Something in the external environment changed (for example money available) and as a result the proposal needed to be radically altered.

Also, there is often reason to question the reproducibility of the findings, since it is frequently not clear how much of the success was down to the energy and talent of a few motivated individuals, rather than the service design in isolation. It is also unclear whether rapid expansion of local schemes to the whole population/area would radically alter the characteristics of the individuals given the intervention – and therefore impact on the scale of benefits.

The challenges of trying to achieve complex service redesign and demonstrate measurable outcomes in practice does not mean that we have to give up on being more robust about how we judge success: it prompts us to think of innovative approaches to evaluation.

The alternatives to summative evaluation are formative methods which use real-time feedback from evaluative findings to inform and modify the pilot intervention (Nuffield Trust, 2013). These approaches are usually more intensive for the evaluators, but can bring about dividends in terms of both being a better fit between the emerging evidence and decisions about implementation.

Conclusion

The Nuffield Trust is involved in evaluating new service models and we continue to provide advice, support and guidance to practitioners and policy-makers on how best to evaluate new service developments.

As policy-makers, commissioners and providers look to innovate and develop new ways of delivering care, there is increasing recognition at both national and local levels about the importance of understanding what works, why it works, and to demonstrate impact on cost and patient outcomes. One of the recurrent problems when evaluating the impact of new care models on outcomes is how to know 'what would have happened under a different approach to delivering care'. There are some designs that can be used, including a randomised control trial, or comparisons over time or between geographical areas – but these are not without limitations.

An alternative approach that researchers at the Nuffield Trust use is retrospective matched control analysis, whereby routinely collected data is used to construct a matched control group, and the impact of the intervention (or service redesign) is measured in terms of differences in the outcomes relative to the matched control group.

This guide has set out why this evaluation method may be preferable to other techniques, and has set out ten steps towards the application of retrospective matching methods in evaluative studies of health and care services.

We hope this guide will be of interest to those involved in evaluation at a local, regional or national level, and help raise awareness and encourage wider use of retrospective matched control study designs as one approach in the evaluation of complex service change.

Appendix: A local example of use of the retrospective matched control method

NHS Islington Clinical Commissioning Group (CCG) is working towards using retrospective matched control designs to test different models of multidisciplinary working across sectors. Here Dan Windross, Integrated Care Commissioning Manager at the CCG, reflects on progress against the ten steps to retrospective matching in their local example (as of June 2015), demonstrating how this method can be applied in practice.

1. **Clarify the aims of the service and the evaluation**
   - *Is an evaluation needed?*
     In Islington, we are testing three different models of multidisciplinary working involving primary care, community health, acute providers, mental health providers, social care and the voluntary sector. This work started in December 2014 and is ongoing. The models share core similarities but have slightly different approaches, such as using videoconferencing or face-to-face meetings, so we wanted to try to understand differences in the models. The matched cohort analysis was part of a broader model of evaluation, including patient and professional feedback.
   - *What is the intervention and who is the target population?*
     The target population was defined by the Integrated Care Board as the top two per cent of people at risk of admission to hospital, plus any patient a health care professional wanted to discuss. The eligibility criteria was intentionally broad, but meant that very different individuals were eligible for the intervention. One of the models focused on a slightly different cohort (those currently or recently discharged), which added to the complexity.
   - *What are the desired outcomes?*
     The outcomes were set at the start by the Board and included reduction in A&E attendances, non-elective admissions and admissions to care homes, and reduction in risk scores. Other activity monitored, but not with an expected direction, was primary care appointments and referrals to mental health services.
   - *How will the new service lead to the desired outcomes?*
     We used a simple logic model to check this. We started with our outcomes as above, then described our inputs, i.e. the staff and resources we had. We then described the activities and outputs to try and link the two – so one chain would end up looking like this:

Inputs	Activities	Outputs	Outcomes
Patients' views	Professionals will work with patients to identify their goals and share these goals with others involved	Patients evaluate the process using surveys	Increased patient satisfaction and engagement

   - *How long will it take for impact to be seen?*
     We needed to roll out a system across Islington in 2015, so this set the time limits for the initial evaluation. More time will always help, but we were clear about the limits of a short period in terms of evaluating the project, and shared this with stakeholders. We will also continue to refine and develop the model.

2. **Decide on the number of people needed to demonstrate an effect**
   We knew we needed large numbers of patients in order to show a demonstrable effect. Ideally we would have had time to evaluate thousands of patients, but ended up with hundreds. We compensated for this by being clear about the limitations and by focusing on qualitative feedback from patients and professionals.

3. **Ensure permission is granted to access person-level datasets**
   We asked patients for verbal consent to do this. We drew up a short script for professionals to use and got this approved by our Caldicott Guardian and information governance lead at the CCG. This worked well, but later on we had some problems when we wanted other organisations to accept this verbal consent. We are setting up a cross-organisational Caldicott Group to develop shared solutions to these problems and developing a robust process for evidence of consent.

4. **Ensure there are data on who received the new service, and some information about the service received**
   Administrative staff from the community health care provider used a simple spreadsheet to keep track of patients. This worked well enough but had limitations around data accuracy. We will move to using existing primary care information systems to record those patients who have received the intervention.

5. **Identify the potential control population**
   As we were working with eight practices for the pilot phase, we decided to use patients from the 28 other practices in Islington as our local area control group. This works well for this phase, but we will need to consider what happens when we roll out the model across the borough. We will probably end up using people not in the intervention in Islington, but will then have selection bias problems. We are talking with neighbouring CCGs about this to develop practical solutions.

6. **Create longitudinal patient-level histories of service use**
   When we identified a patient who was receiving the intervention, we used a patient linked dataset, provided by our commissioning support unit (CSU) to collate their history of health care use (primary and secondary care activity etc) for the 12 months before their intervention start date. We will then track this for a year after the intervention to monitor the impact of the intervention and inform commissioning decisions.

7. **Identify matched controls**
   We matched people using gender (exact match), age (within two years) and risk score (within +/- 20 per cent of one standard deviation). Our CCG analytics team did this initially, then shared it with commissioners and public health colleagues. We created five matches for each patient who received the intervention. This approach created two problems and required a practical, iterative response: for the lower risk scores we had too many possible matches; and at the higher risk scores we had too few. For the lower risk scores we first sought to match on long-term conditions, and then took the closest matches in terms of age and risk scores. At the higher risk scores we expanded the age and risk-score range until we had five matches. In the longer term we want to try and get a simple 'push-button' solution, and are working with the CSU to develop this. There will be some limitations with a single solution, but we hope this will be compensated for as we increase the scale.

8. **Monitor outcome variables for those receiving the new service and matched controls**
   This work is under way, and we are expecting results in August 2015.

9. **Undertake summative analysis**
   We recognised that this work involved more statistical analytical skills and knowledge than we had locally. We approached our public health team early on for support with this, and they have agreed to support this work as part of the core offer to the CCG, testing the intervention cohort against the control for impact. Once we have the initial findings, we will start the process of assessing how robust this is: What are our hidden confounders? How does selection bias impact on this?

10. **Continuously monitor**
    We know we are going to keep changing our approach, as we want to adapt as we understand what works. We want to be iterative and flexible while creating a sensible matched cohort for our evaluation. This is a learning exercise for us and locally we will apply this methodology to other services as our understanding and capability to use it grows.

Contact: Dan Windross, Integrated Care Commissioning Manager, NHS Islington CCG (dan.windross@nhs.net).

First published 5 August 2015. Updated 6 August 2015.

References

Bardsley M, Billings J, Chassin L, Dixon J, Eastmure E, Georghiou T, Lewis G, Vaithianathan R and Steventon A (2011) *Predicting Social Care Costs: A feasibility study*. Nuffield Trust.

Bardsley M, Steventon A, Smith J and Dixon J (2013) *Evaluating Integrated and Community-Based Care*. Nuffield Trust.

Best A and Lewis S (2012) 'Large-system transformation in health care: A realist review'. *Milbank Quarterly* 90(3), 421–456.

Black N (1996) 'Why we need observational studies to evaluate the effectiveness of healthcare'. *BMJ* 312, 1215–1218.

Cartwright N (2007) 'Are RCTs the Gold Standard?'. *BioSocieties* 2, 11–20.

Chitnis X, Georghiou T, Steventon A and Bardsley M (2012) *The Impact of the Marie Curie Nursing Service on Place of Death and Hospital Use at the End of Life*. Nuffield Trust.

Connell J, Kubisch AC, Schorr LB and Weiss CH (1995) *New Approaches to Evaluating Community Initiatives. Concepts, Methods, and Contexts.* Aspen Institute for Humanistic Studies, New York, N.Y.

Diamond A and Sekhon J (2013) 'Genetic matching for estimating causal effects. A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics* 95(3), 932–945.

Georghiou T and Steventon A (2014) *Effect of the British Red Cross 'Support at Home' Service on Hospital Utilisation*. Nuffield Trust.

Gheorghe A, Roberts TE, Ives JC, Fletcher BR and Calvert M (2013) 'Centre Selection for Clinical Trials and the Generalisability of Results: A Mixed Methods Study'. *PLOS One* 8(2), 1–9.

Hansen BB (2008) 'The prognostic analogue of the propensity score'. *Biometrika* 95(2), 481–488.

Health Research Authority (2015) Section 251 and the Confidentiality Advisory Group (CAG). www.hra.nhs.uk/about-the-hra/our-committees/section-251/#sthash.tS3icThz.dpuf .

McCarney R, Warner J, Iliffe S, van Haselen R, Griffin M and Fisher P (2007) 'The Hawthorne Effect: a randomised, controlled trial'. BMC *Medical Research Methodology* 7(30), 1–8.

McNulty T and Ferlie E (2002) *Reengineering Health Care. The Complexities of Organizational Transformation*. Oxford University Press Inc.

Merrifield A and Smith W (2012) 'Sample size calculations for the design of health studies: a review of key concepts for non-statisticians'. *NSW Public Health Bulletin* 23(7-8), 142–147.

Nuffield Trust and Imperial College London (2013) *Evaluation of the First Year of the Inner North West London Integrated Care Pilot*. Nuffield Trust.

Roland M and Torgerson D (1998) 'Understanding controlled trials: What are pragmatic trials?'. *BMJ* 316, 285.

Roland M, Lewis R, Steventon A, Adams J, Bardsley M, Brereton L, Chitnis X, Staetsky L, Tunkel S and Ling T (2012) 'Case management for at-risk elderly patients in the English Integrated Care Pilots: observational study of staff and patient experience and secondary care utilisation'. *International Journal of Integrated Care* 12, e130.

Rosenbaum P and Rubin D (1985) 'Constructing a control group using multivariate matched sampling methods that incorporate the propensity score'. *The American Statistician*.

Rosenbaum P and Rubin D (1983) 'The central role of the propensity score in observational studies for causal effects'. *Biometrika* 70(1), 41–55.

Rothwell P (2005) 'External validity of randomised controlled trials: "To whom do the results of this trial apply?"'. *Lancet* 365(9453), 82–93.

Steventon A, Bardsley M, Billings J, Georghiou T and Lewis G (2011) *An Evaluation of the Impact of Community-Based Interventions on Hospital Use. A case study of eight Partnership for Older People Projects (POPP)*. Nuffield Trust.

Steventon A, Bardsley M, Billings J, Dixon J, Doll H, Hirani S, Cartwright M, Rixon L, Knapp M, Henderson C, Rogers A, Fitzpatrick R, Hendy J and Newman S (2012) 'Effect of telehealth on use of secondary care and mortality: findings from the Whole System Demonstrator cluster randomised trial'. *BMJ* 344, e3874.

Steventon A, Grieve R and Sekhon JS (2015) 'A comparison of alternative strategies for choosing control populations in observational studies'. *Health Services and Outcomes Research Methodology*. doi: 10.1007/s10742-014-0135-8

Stuart E (2010) 'Matching methods for causal inference: A review and a look forward'. *Statistical Science* 25(1), 1–21.

evidence for
better health care

nuffieldtrust

For more information about the Nuffield Trust, including details of our latest research and analysis, please visit www.nuffieldtrust.org.uk

Download further copies of this guide from
www.nuffieldtrust.org.uk/publications

Subscribe to our newsletter:
www.nuffieldtrust.org.uk/newsletter

Follow us on Twitter: Twitter.com/NuffieldTrust

Nuffield Trust is an independent health charity. We aim to improve the quality of health care in the UK by providing evidence-based research and policy analysis and informing and generating debate.

59 New Cavendish Street
London W1G 7LP
Telephone: 020 7631 8450
Email: info@nuffieldtrust.org.uk

www.nuffieldtrust.org.uk

Published by the Nuffield Trust.