SAGE Research Methods Cases

Medicine & Health

Submission for Consideration

Case Title

Designing a Robust Evaluation of Health Service Innovation: an Observational Study Comparing the Outcomes of Patients Referred to Integrated Care Teams with a Retrospectively Matched Control Group

Author Name(s)

Therese Lloyd, Arne T. Wolters

Author Affiliation & Country of Affiliation

The Health Foundation, United Kingdom

Lead Author Email Address

Email: therese.lloyd@health.org.uk

Discipline

Public Health [D26]

Sub-discipline

Health Services Research [SD-PH-3]

Academic Level of intended readership

Advanced Undergraduate

Contributor Biographies

Therese Lloyd is a Senior Statistician within the Improvement Analytics Unit, a partnership between the Health Foundation and NHS England and NHS

Improvement. She joined the Heath Foundation in October 2014 and has in recent years evaluated several interventions relating to older people, including care home initiatives, integrated care teams and the national introduction of a named accountable GP for patients aged 75 or over. She has a Bachelor's and a Master's degree in Statistics from Lund University, Sweden.

Arne Wolters is Head of the Improvement Analytics Unit. He joined the Health Foundation in November 2014, having previously worked in various data roles, enabling safe use of sensitive data to support research and evaluation. Arne has been involved in the evaluation of various health services interventions in urgent and emergency care, primary care, social care as well as the integration of care services. Arne completed both his undergraduate and graduate studies in Econometrics and Operational Research at the University of Groningen in the Netherlands.

Published Articles

Lloyd, T., Brine, R., Pearson, R., Caunt, M., & Steventon, A. (2018). *The impact of integrated care teams on hospital use in North East Hampshire and Farnham*. The Health Foundation. Retrieved from www.health.org.uk/publications/impact-integrated-care-teams-hospital-use-north-east-hampshire-and-farnham

Lloyd, T., Brine, R., Pearson, R., Caunt, M., & Steventon, A. (2018). *Technical appendix: The impact of integrated care teams on hospital use in North East Hampshire and Farnham*. The Health Foundation Retrieved from https://www.health.org.uk/sites/default/files/2018-10/iau_nehf_technical_appendix.pdf

Lloyd, T., & Steventon, A. (2017). *Statistical Analysis Protocol for an evaluation of integrated care teams as part of a primary and acute care system (PACS) in North East Hampshire and Farnham*. Retrieved from https://www.health.org.uk/sites/default/files/2018-10/iau_nehf_sap_addendum.pdf

Abstract

Using our evaluation of a health care intervention as a case study, we discuss some of the study design choices we made. We explain the statistical issue of regression to the mean and show, with a practical example, why a before-and-after analysis can give misleading results. We introduce the concept of a counterfactual analysis as a more robust methodology and describe how we selected a similar, matched control group and compared their outcomes with those of the intervention group using multivariable regression. We discuss the limitations of our study design, in particular the risk of unobserved confounding, and how we took these limitations into account when drawing conclusions from our results.

Learning Outcomes

By the end of this case, students should be able to:

- Describe why counterfactual analysis is important in observational studies
- Describe the key decisions and challenges in designing a matched control study
- Discuss the limitations of an observational study, and the importance of interpreting findings in light of these limitations

---

Project Overview and Context

In England, the National Health Service (NHS) provides health services that are free at the point of care. It is commissioned by NHS England, while NHS Improvement is responsible for the quality of care of hospital providers.

In 2015, NHS England funded 50 New Care Model (NCM) vanguards across England, to lead on developing and trialing new models of care to improve care and outcomes for the population. One of these was based in North East Hampshire and Farnham (NEHF), where they implemented a broad range of health care initiatives between 2015 and 2018, one of the core initiatives being integrated care teams (ICTs).

Integrated care teams

Integrated care teams are multidisciplinary teams who work together to proactively provide more coordinated care for their patients. There were five ICTs established in NEHF, each one consisting of a multidisciplinary team comprising professionals from across primary, community, mental health, social care and voluntary services. The ICTs met weekly to develop a single, coordinated care plan for each patient referred to its services. The main objectives of the ICTs were to reach patients in highest need and at risk of going into crisis and to improve patients' health, health confidence, experience and wellbeing and reduce A&E attendances and emergency admissions. Patients could be referred to an ICT by any health professional, although the majority of referrals were made by GPs. During the period of our study, the vast majority of referrals were made on the basis of clinical judgement.

Understanding the impact of ICTs

As the vanguards received funding to specifically test new models of care, there was a need to evaluate the effect of the initiatives and each vanguard received funding to

commission an evaluation. The Improvement Analytics Unit (IAU) committed to doing an evaluation of the ICT element of the NEHF vanguard programme in 2017.

About the Improvement Analytics Unit

The Improvement Analytics Unit is an innovative partnership between NHS England & NHS Improvement (NHSE-I) and the Health Foundation, an independent charity committed to bringing about better health and health care for people of the UK. The IAU evaluates complex initiatives in health care to improve quality of care, effectiveness of services, and patients' health. We do this by applying robust statistical methods; in particular, we ensure we always compare the intervention group with a comparison group that aims to reflect what would have happened in the intervention group, had they not received the intervention.

The IAU has access to pseudonymized routine administrative data, such as patient hospital records. As the IAU evaluates health care services, i.e. the direct care of patients is not affected through this work, no ethics committee approval is required for our analyses.

Section summary:

- ICTs are multidisciplinary teams who work together to proactively provide more coordinated care for their patients
- These were implemented in an area of England, as part of the New Care Models programme

Research Design

1. An observational study using a 'counterfactual'

Innovations in medicine are often tested out using randomized controlled trials (RCT). RCTs randomly assign patients to either the intervention group (those receiving a new treatment) or the control group (those receiving no or standard treatment). Random assignment of the treatment reduces the risk of differences, and therefore of bias, between the two groups and allows the impact of the treatment to be estimated by looking at the difference between the two groups.

The implementation of innovations in health service delivery is often difficult to randomize, for example because the changes affect services that are available to a whole population or because it may be unethical to provide differing services to patients. Therefore, other methods of evaluation are needed. Ideally, you would compare the health outcomes of patients receiving the new service (the 'intervention group') with a 'counterfactual', i.e. what would have happened if these patients would have received standard care. As it is not possible to simultaneously observe what happens to a person when they are both receiving and not receiving an intervention, the next best thing is to try to replicate a counterfactual. There are several ways of doing so, depending on the context and data available, for example by using a regression discontinuity design or creating a synthetic control. For more information on potential methods, see Clarke, Conti, Wolters & Steventon (2019) paper on how to evaluate the impact of health care interventions, under 'Further reading'. In this study, we used a matched control method, which selects a control group of patients

that are as similar as possible to the intervention group on a range of observable characteristics (the 'matched control group').

Why a counterfactual analysis is often preferable to a before-and-after analysis

Although all methods have limitations, a counterfactual analysis, if implemented properly, is more reliable than for example a before-and-after analysis. In a before-and-after analysis (also known as pre-post analysis), one compares patients' outcomes before and after receiving an intervention. However, this method can suffer from something called 'regression to the mean', which means that after an unusually high or low value, the next value will often be closer to the average, even in the absence of an intervention. This particularly applies to evaluations of health care initiatives if patients are chosen to receive the intervention because they had recently had deteriorating health or an emergency hospital admission, as a number of these patients will naturally recover from their illnesses over time.

2. Data

Observational studies often use administrative patient data, for instance hospital records, allowing access to large amounts of data that would otherwise be difficult to collect.

The IAU had access to these patient-level hospital records for all of England, as well as patient registration data, which has information on all patients registered with a general practice, including date of birth and death (although we only access month and year of these events). For this study we also had access to the list of patients who had been referred to an ICT and the date of the referral. All data processed by the IAU are pseudonymised, which means that all identifiable fields, such as names, full

date of birth and address have been removed and the NHS number has been replaced by a pseudo-identifier.

3. Outcomes

In order to determine whether the intervention is having the intended effect, we want to evaluate the outcomes (or 'endpoints') that reflect the aims of the intervention. The choice of outcomes may, however, also be dependent on what outcomes you can measure from your available data. Interventions in health services often aim to improve the quality of care, which can be difficult to measure. Sometimes it might be necessary to evaluate proxies instead. For example, if assessing the quality of care in primary care, we might instead assess the number of emergency hospital admissions. It can also be appropriate to assess other outcomes, e.g. to check for unintended consequences (positive or negative).

Our main outcomes were A&E attendances and emergency admissions, as one of the main aims of the ICTs was to reduce these outcomes. We also looked at some other measures that either the local team or NHS England were interested in, for example total number of bed days following emergency admissions and proportion of deaths in hospital, as a proxy for not dying in your preferred place of death. Due to lack of data, we could not assess other aims of the ICTs, such as patients' health confidence or experience.

For the purpose of this case study we have only reported outcomes for A&E attendances and emergency admissions. Please refer to the full report and technical appendix, both by Lloyd, Brine, Pearson, Caunt & Steventon (2018), for results on all study outcomes and other additional information.

4. Defining the intervention group and study period

The intervention group in our study included those patients who were registered in the area and who were referred to an ICT between July 2015 and May 2017, although there were some exclusion criteria (see full briefing for details). There were 774 intervention patients included in our study. The period of the evaluation (study period) was July 2015, which was when the ICTs were launched, to June 2017, which was the latest period for which we had reliable data at the start of our evaluation. We allowed each patient to have an individual start and end date within the study period, depending on when they were referred (or for the control group, selected) to the study and whether they died or moved away before the end of the study. This meant that the patients enrolled early could be followed for up to 23 months, but patients who were referred later could only be followed up for a shorter amount of time. As the ICTs met weekly and newly referred patients were discussed at the first meeting after referral to agree an action plan, we used the date of referral as each referred patient's start date.

5. Selecting a pool of potential controls - locally or a nationally?

When selecting a matched control group, the first decision is whether to use a local or national control group. Both have advantages and disadvantages: for example, it is often easier to get good balance (i.e. similarity) on observed individual-level characteristics with a national comparison group; however, choosing a local control group means that the groups are often more similar on unobserved higher-level characteristics, such as hospital admission thresholds. Sometimes, of course, it won't be possible to choose a local control group if the entire target population has received the intervention. A paper by Steventon, Grieve & Sekhon (2015) provides a comparison of these alternatives.

We wanted to evaluate just the ICTs, but there were other interventions implemented in NEHF that the ICT patients may also have benefitted from. This included other initiatives that were part of the NCMs, such as a rapid home response team by specially trained community paramedics, and other concurrent changes, such as the opening of an ambulatory emergency care unit within the local hospital. There were no available data on which patients received these other services. Therefore, if we chose a comparison group from outside NEHF, we would not be able to know if any difference in outcomes was due to the ICTs or any of the other initiatives. By choosing the control group from the local area, we could assess the effect of ICTs over and above the other services. However, if the referrers were successful in identifying the patients most in need, then all these patients would already be in the intervention group, so the control group would by definition consist of patients with lower need. If we're comparing against a less sick population, then this might bias the results.

This evaluation was at the early stages of the implementation: only 1,039 patients out of a population of 225,000 had been referred to an ICT during our study period, when the initial aim was to reach the 2% of the adult population most at risk. We therefore thought it likely that there would still be patients eligible for referral that hadn't been referred, and chose a local control group.

The same eligibility criteria were applied to the control group as to the intervention group. There were ~78,000 potential control patients in the local area that met our eligibility criteria.

6. Selecting the matching method to identify a control group

There are several ways of selecting a matched control group, for example propensity score matching or entropy balancing. For a useful guide, see Stuart's (2010) paper on matching methods, under 'Further reading'. We used a method called genetic matching, which is a computer intensive search algorithm that iteratively assesses the balance between the intervention and different versions of the matched control group until balance has been optimized. This method produces more similar groups than some of the commonly used methods such as nearest neighbour matching or the propensity score. For more information on genetic matching, refer to Diamond and Sekhon (2013), under 'Further reading'.

We allowed matching 'with replacement', which means that the same control patient can be matched several times to different intervention patients. This helps optimize the similarity of the two groups, especially if your pool of potential controls is limited. We match each intervention patient to a control. Sekhon's (2019) documentation, under 'Further reading', provides more information about the R package 'GenMatch' that we used.

7. Matching variables

When matching, it is important to make sure that the groups are as similar as possible on those characteristics that could affect the outcomes – as we want to compare 'like for like'. However, we want to ensure that we don't match on anything that could have been affected by the intervention, to not underestimate the effect. Therefore, we only match on characteristics recorded before or at the time of the referral ('baseline characteristics'). Thinking through what variables these should be and whether we can identify them in our data is an important part of the design.

The characteristics that we identified as important – and that were observable in our data – were:

- Demographics & socio-demographics, e.g. age, gender, ethnicity and level of socio-economic deprivation
- Prior hospital use, e.g. number of A&E and emergency admissions in the previous 60 days and previous year; number of missed outpatient visits
- Health characteristics, including conditions linked to frailty (e.g. cognitive impairment, pressure ulcers) and conditions commonly used in risk prediction models (Elixhauser and Charlson list of conditions – see papers by Elixhauser, Steiner, Harris, & Coffey (1998) and Charlson, Pompei, Ales, & MacKenzie (1987))
- Proxy for level of daily support (living in a care home or not)
- Time period of referral – this is because:
    - the period when you are referred affects how long you can receive the intervention within the study period
    - the range of other interventions available in the area changed over time
    - the time of year may also affect the likelihood of being admitted to hospital

8. Statistical inference

Following the matching, we used multivariate regression analysis to estimate the causal difference in outcomes between the two groups, after adjusting for any remaining observed differences between the ICT patients and the matched controls.

Combining matching and regression usually produces more accurate estimates of the effect of an intervention than using only one method, as we have in effect two opportunities to remove any observed differences between the intervention and comparison group. We also decrease the dependency on the correct model specification, i.e. the results will not be as dependent on having chosen the correct regression model. See Ho, Imai, King & Stuart (2007) paper for more details.
For each outcome, we fitted a regression model that was appropriate to the type of outcome and the distributional properties of the data. For example, to model count data (e.g. the number of emergency admissions per patient), we used a Poisson or Negative Binomial model. We accounted for patients' differing amounts of time in the study by introducing an offset into the regression model. Where possible, all baseline characteristics were adjusted for; however, this was not always possible due to multicollinearity (i.e. where two or more variables are interrelated) or sparse data (i.e. where fitting too many variables would lead to overparametrisation).

Section summary

- A counterfactual analysis is often more reliable than a before-and-after analysis
- We selected a local matched control group of patients who were similar in observable characteristics to those who were referred to ICTs
- We compared the hospital activity of the intervention and matched control groups using multivariate regression to evaluate the effect of the ICTs on hospital use

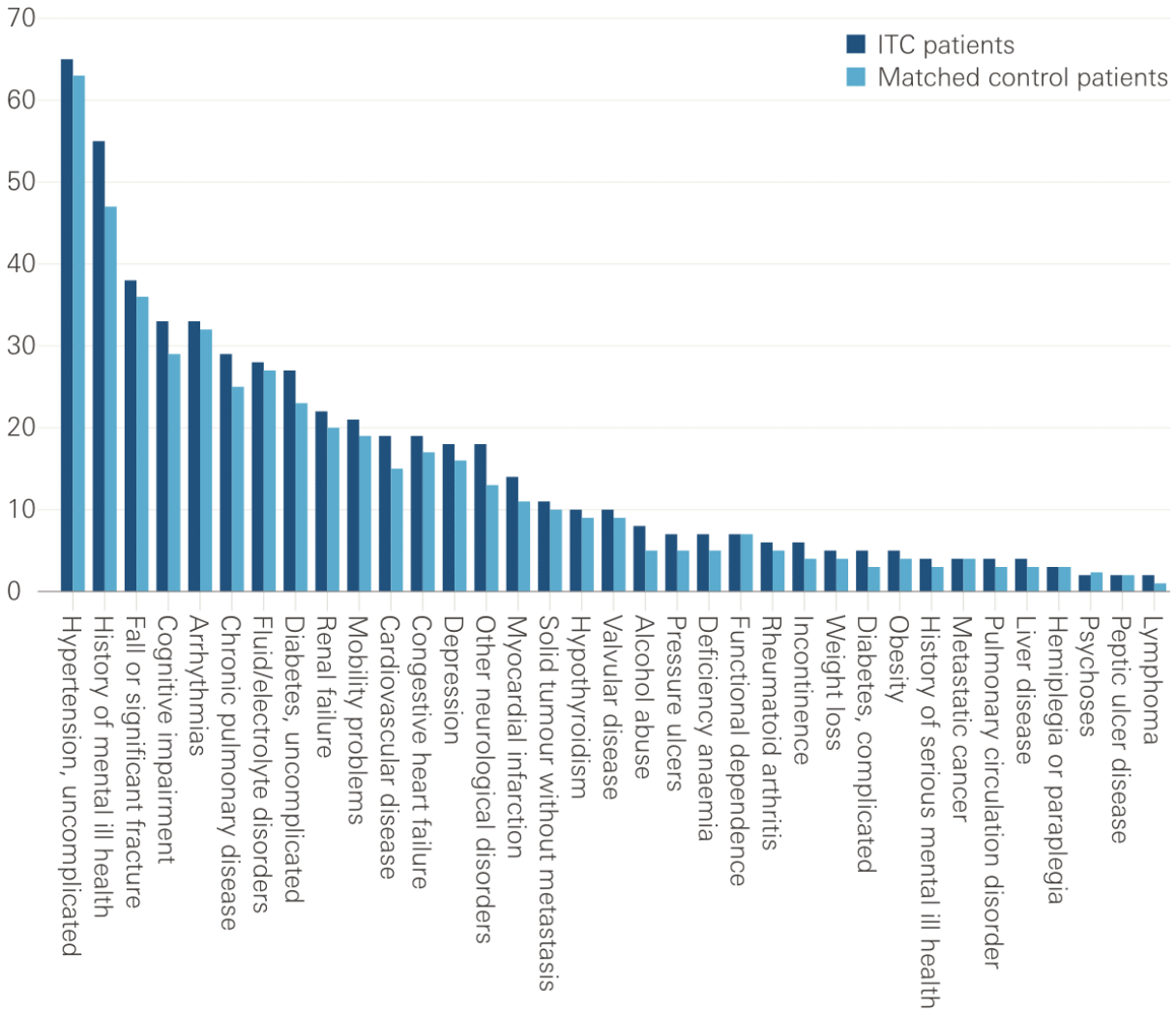
Methods in Action

Checking if the intervention and matched control groups were similar

Once we had completed the matching algorithm and selected 774 matched control patients, we checked whether the intervention and matched control groups were similar, by calculating the standardized mean difference (SMD) for each baseline characteristic. The SMD is the difference in means as a proportion of the standard deviation in the intervention group – see Austin's paper for more details (2009). If the SMD was within +/- 10%, the groups were considered balanced. We calculated SMDs rather than calculating p-values to assess balance, as the SMD doesn't depend on the size of the groups – with p-values, you are less likely to detect a significant difference if your sample size is small.

We found that in this study, patients in the matched control group had broadly similar characteristics to the ICT patients (i.e. with an SMD smaller than +/- 10%). There was, however, a pattern of the ICT patients tending to have underlying health conditions more often (see Figure 1) and historically experiencing more emergency hospital use than the matched control patients. However, we were able to adjust for some remaining differences between the groups at the regression stage.

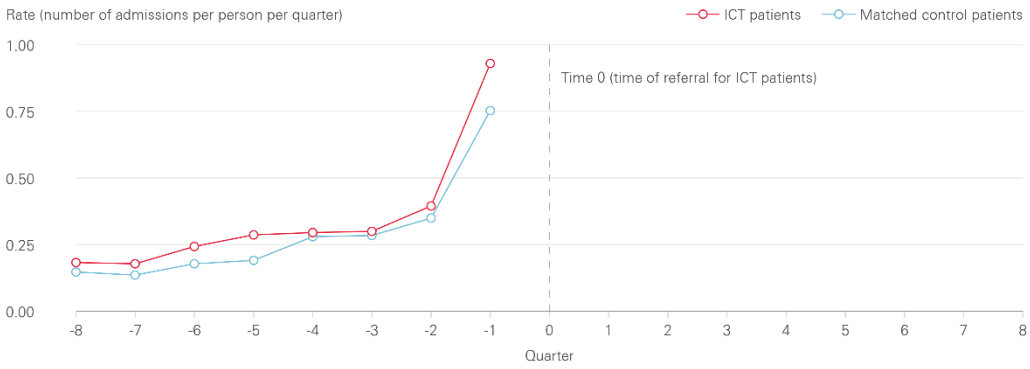
Figure 1. Percentage of patients with various health conditions at the time of referral



Source: Lloyd, Brine, Pearson, Caunt & Steventon (2018). The impact of integrated care teams on hospital use in North East Hampshire and Farnham

We also looked at the trends in hospital use over time, as another check of whether the intervention and matched control groups were similar before referral. Figure 2 presents the rates of emergency admissions in the two years prior to referral/selection and shows that ICT patients have a sharp increase in emergency admissions in the three months before referral - which is not surprising, as health professional may have chosen to refer based on this. We can also see that the matching selected control patients with similar trends in emergency hospital use in the pre-period.

Figure 2. Rates of emergency admissions: intervention and matched control patients before referral



Source: Lloyd, Brine, Pearson, Caunt & Steventon (2018). The impact of integrated care teams on hospital use in North East Hampshire and Farnham

Results and drawing conclusions

After adjustment, we estimated that ICT patients had 33% more A&E attendances compared with the matched controls. The 95% confidence interval suggests this difference is likely to be between 16% and 54% more A&E attendances for ICT patients. We also found that ICT patients had 43% higher rates of emergency admission (95% confidence interval of 23% to 57% higher). Both these results were statistically significant. For both A&E attendances and emergency admissions this difference translates into approximately 0.5 more visits/admissions per patient per year.

But before drawing any conclusions, it is important to consider that there may be unobserved differences between the groups, i.e. differences between the groups that we can't see in our data. If these differences affect the outcomes, this is called

unobserved confounding. There is always a risk of unobserved confounding in observational studies, but this might be a particular concern in this study, as patients were referred based on clinical judgement, rather than on an objective measure such as a risk score. There may be a reason why a health professional referred patient A but not patient B, even though according to our data they have similar health profiles. It may be that patient A lacks family support or isn't managing their long-term conditions as well – or that the GP thinks that patient A will be more responsive to the additional support. None of these characteristics are typically recorded in administrative data.

We can't statistically adjust for characteristics that are not recorded in the data and the additional uncertainty from unobserved differences between the groups isn't reflected in the 95% confidence intervals. Therefore, we can't determine with certainty whether such differences in characteristics could account for the differences in outcomes and we need to consider this when drawing conclusions from our results.

In this study, we looked at the size of the difference in outcomes and considered that, while unobserved confounding may explain some of the higher rates, they are unlikely to account for all the difference. Furthermore, it is very unlikely that any unobserved differences could mask an actual decrease in hospital use. We therefore concluded in our report that the findings imply that the ICTs did not reduce A&E attendances and emergency admissions and may even have led to increases.

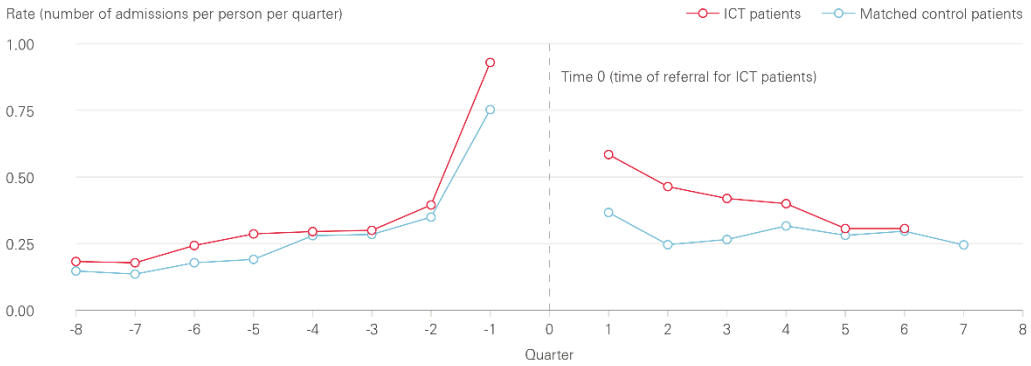
An example of regression to the mean and the risk of doing a before-and-after analysis

Figure 3 shows emergency admission rates, both before referral/selection and during the study period (right-hand side of 'time 0'). Looking at just the intervention group, we can see there is a drop in emergency admission rates after referral. If we were

comparing patients' hospital use before and after referral (i.e. a before-and-after analysis), we might conclude that the ICTs led to an immediate decrease in emergency admissions. But if we look at the matched control group, we see a drop in emergency admissions too – even though the matched control group didn't receive any intervention at 'Time 0': 'Time 0' was an arbitrary date that was selected so that the control group would have similar characteristics to the intervention group in the pre-period.

This drop in rates in the matched control group is due to regression to the mean – as we selected patients with recent high emergency hospital use, on average people got better, even in the absence of an intervention. In a before-and-after analysis, it isn't possible to assess to what extent a change is due to the intervention or to regression to the mean, which is why a before-and-after analysis can give misleading results. In contrast, a counterfactual analysis allows us to estimate the effect of the intervention without risk of regression to the mean.

Figure 3. Rates of emergency admissions: intervention and matched control patients before and after referral



Source: Lloyd, Brine, Pearson, Caunt & Steventon (2018). The impact of integrated care teams on hospital use in North East Hampshire and Farnham

Section summary

- The intervention and matched control group were broadly similar in observable characteristics
- The intervention patients had higher emergency hospital use than the matched control group
- There may be unobserved differences that could be biasing the results

---

Practical Lessons Learned

Findings from the early stages of implementation

An important decision in the study design process is to determine how long after an intervention has been introduced you should carry out an evaluation. The most appropriate study length will depend on the type of intervention, how long time it will take to embed the intervention, what outcome measures the intervention is likely to have an impact on and in what timeframe. However, this sometimes needs to be balanced with the need to provide evidence to inform decision making, e.g. of whether to make changes to the service; continue funding an initiative; or whether to roll out a model nationally.

This study measured the early impact of the ICTs on patients in North East Hampshire and Farnham in order to provide insights that, combined with other local evidence, could inform the development of the ICTs in NEHF. It is important that the

findings from the study are considered within this context, rather than drawing conclusions about the overall effectiveness of ICTs.

However, looking at other published evidence, the results were maybe not that surprising. Integrated care initiatives often aim to reduce hospital use, by providing more proactive than reactive care but, in practice, these initiatives often result in increased hospital activity by identifying unmet need and providing more timely access to care. This could result in an increase in hospital use in the short term, but with potentially longer-term benefits materializing over the course of several years. Although our study period ran for nearly 2 years, the average length of time that a patient was in the study was only approximately 7 months, which may be too short a period for any gains to materialize.

Ideally, we would have followed up with a further evaluation, looking at a longer study period.

Limited outcome metrics

Our evaluation focused only on hospital use outcomes, as those were the outcomes from the list of aims of the ICTs that we could measure with our data and evaluate using a counterfactual analysis.

The question is whether it is realistic to expect ICTs to reduce emergency hospital use, if referred patients are at highest risk and need. It might be instead that the value of ICTs lies in improving patients' quality of life, by providing better coordinated care and managing health problems in a timely and effective manner. In NEHF, the ICTs also aimed to improve patients' health confidence, experience and wellbeing, for which there was some promising local evidence.

Unobserved confounding

The challenge in any observational study is unobserved confounding. This was a particular concern in this study as patients were referred based on clinical judgement and we selected the control patients from the same area. As one check for unobserved differences, we compared mortality rates between the groups. We did not expect that the intervention would affect mortality rates, and so a difference would have suggested that the group were different in overall health levels. We did not find a statistically significant difference in mortality.

There are some – more refined – methods, which assess the sensitivity to unobserved confounding, for example by allowing you to understand how strong a relationship needs to exist between a confounder and the outcome measure to change the outcome of your model – see Rosenbaum & Rubin (1983). This would allow you to judge whether it is likely than any unobserved confounding could explain the results. However, current applications of this work are limited to standard linear models, and so could not be used in this study.

In other ways, having a local control group can lessen the risk of unobserved confounding. In our study, both groups had access to the same area- and hospital-level services and were subjected to e.g. the same hospital coding practices – things that could otherwise bias any comparison.

Section summary

- The study presents early findings of the intervention. A follow-on analysis would provide an opportunity to see the long-term effect of the ICTs, which may differ from the early findings
- The study was limited in the outcome measures it could evaluate. Patients might have benefited from the ICT in other ways
- It is difficult to assess unobserved confounding

Conclusion

The local team wanted to know the effect of ICTs, over and above the other services available in the area. To evaluate this, we designed a study that aimed to be as robust as possible. However, the approach did have some limitations, in particular the risk that the control group was not similar enough to the intervention group to draw any reliable conclusions. The risk of this being the case was not quantifiable. However, we were upfront about this limitation and took this into consideration when interpreting the findings. All things considered, we argue that this study used the most robust and appropriate methods to evaluate the ICTs.

The other main question about this evaluation was whether the evaluation was carried out too early. The evaluation was carried out at that timepoint as the local team wanted to understand the effect of the initiative; however, there would have been real value in doing another follow-on study with a longer study period, to see whether the effect changed over time. This would have been relatively easy to do, as the code for doing the data manipulation and analysis already existed. Within the IAU, we have now started incorporating follow-on analyses into our study designs, in order to do just that.

Lastly, the evidence suggests that it may be difficult for ICTs to reduce emergency use. The question is whether it is realistic to expect ICTs to reduce emergency hospital use, given that referred patients are at highest risk and need. It might be instead that the value of ICTs lies in improving patients' quality of life, by providing better coordinated care and managing health problems in a timely and effective manner and that therefore these are the outcomes to measure and evaluate. For this to be possible, we need routinely and consistently collected data, made available to both care providers and analysts.

Classroom Discussion Questions

- Are there limitation of the methods used other than the ones highlighted in this case study, and if so, what are they?
- Would you have chosen a different methodology, and if so why?
- Do you think that the results are biased? Why? If yes, is it still possible to draw conclusions from this study?

Multiple Choice Quiz Questions

Which is an example of a counterfactual analysis?

A. Matched control study - CORRECT

B. Before-and-after analysis

C. Time trend analysis

What does ‘regression to the mean’ signify?

A. A statistical method for calculating an average

B. After a high or low value, the next value will often be closer to the average – CORRECT

C. The use of a regression model to provide an estimate of the mean strength of a relationship between two variables

The risk of unobserved differences between an intervention and matched control group is…

A. Negligible after matching

B. Reflected in the 95% confidence interval

C. Difficult to quantify but there are some ways of exploring the risk – CORRECT

Declaration of Conflicting Interests

The Authors declare that there is no conflict of interest.

Further Reading

Clarke, G. M., Conti, S., Wolters, A. T., & Steventon, A. (2019). Evaluating the impact of healthcare interventions using routine data. *BMJ (Online)*, *365*(June), 1–7. https://doi.org/10.1136/bmj.l2239

Diamond, A., & Sekhon, J. S. (2013). Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. *Review of Economics and Statistics*. https://doi.org/10.1162/REST_a_00318

Sekhon, J. S. (2019). *Package ' Matching .'* Retrieved from https://cran.r-project.org/web/packages/Matching/Matching.pdf

Stuart, E. a. (2010). Matching methods for causal inference: A review and a look forward. *Stat Sci*, *25*(1), 1–21. https://doi.org/10.1214/09-STS313.Matching

References

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*. https://doi.org/10.1002/sim.3697

Charlson, M. E., Pompei, P., Ales, K. L., & MacKenzie, R. (1987). A new method of classifying prognostic in longitudinal studies: development and validation. *Journal of Chronic Diseases*, Vol. 40, pp. 373–383. https://doi.org/0021-9681/87

Elixhauser, a, Steiner, C., Harris, D. R., & Coffey, R. M. (1998). Comorbidity measures for use with administrative data. *Medical Care*, *36*(1), 8–27. https://doi.org/10.1097/00005650-199801000-00004

Ho, D. E., Imai, K., King, G., & Stuart, E. a. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, *15*(3), 199–236. https://doi.org/10.1093/pan/mpl013

Lloyd, T., Brine, R., Pearson, R., Caunt, M., & Steventon, A. (2018a). *Technical appendix: The impact of integrated care teams on hospital use in North East Hampshire and Farnham Consideration of findings from the Improvement Analytics Unit*. Retrieved from https://www.health.org.uk/sites/default/files/2018-10/iau_nehf_technical_appendix.pdf

Lloyd, T., Brine, R., Pearson, R., Caunt, M., & Steventon, A. (2018b). *The impact of integrated care teams on hospital use in North East Hampshire and Farnham*.

Retrieved from www.health.org.uk/publications/impact-integrated-care-teams-hospital-use-north-east-hampshire-and-farnham

Quan, H., Sundararajan, V., Halfon, P., Fong, A., Burnand, B., Luthi, J.-C., … Ghali, W. a. (2005). Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Medical Care*, *43*(11), 1130–1139. https://doi.org/10.1097/01.mlr.0000182534.19832.83

Rosenbaum, P. R., & Rubin, D. B. (1983). Assessing Sensitivtiy to an Unobserved Binary Covariate in an Observational Study with Binary Outcome. *Journal of the Royal Statistical Society, Series B*, *45*(2), 212–218.

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