BMJ Open

Understanding and Addressing the Challenges of Conducting Quantitative Evaluation at a Local Level

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2019-029830
Article Type:	Original research
Date Submitted by the Author:	13-Feb-2019
Complete List of Authors:	Hinde, Sebastian; York University, Centre for Health Economics Bojke, Laura; University of York, Centre for Health Economics Richardson, Gerry; University of York, Centre for Health Economics
Keywords:	HEALTH ECONOMICS, Health policy < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Health economics < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, STATISTICS & RESEARCH METHODS

Understanding and Addressing the Challenges of Conducting Quantitative Evaluation at a Local Level

Sebastian Hinde, Laura Bojke, Gerry Richardson

Centre for Health Economics, University of York

Corresponding author:
Sebastian Hinde,
Centre for Health Economics,
Alcuin 'A' Block,
University of York,
Heslington,
North Yorkshire,
YO10 5DD.
United Kingdom
Sebastian.Hinde@york.ac.uk
+44 (0)1904 321455

SH developed the research idea, lead the writing of the manuscript, and acts as the guarantor of the article. LB and GR gave input at all stages including the commenting on the manuscript. All authors are health economics at the Centre for Health Economics, University of York, with experience in working on evaluation for local decision makers at various levels. The simulated dataset and all analytical code is available upon request. Due to the simulated nature of the dataset no patients were involved in the study.

This article presents independent research by the National Institute for Health Research Collaboration for Leadership in Applied Health Research and Care Yorkshire and Humber (NIHR CLAHRC YH). www.clahrc-yh.nir.ac.uk. The views and opinions expressed are those of the authors, and not necessarily those of the NHS, the NIHR or the Department of Health.

Abstract

Objectives

In the context of tightening fiscal budgets and increased commissioning responsibility, local decision makers across the UK healthcare sector have found themselves in charge of the implementation and evaluation of a greater range of healthcare interventions and services. However, there is often little experience, guidance, or funding available at a local level to ensure robust evaluations are conducted. In this paper, we evaluate the possible scenarios that could occur when seeking to conduct a quantitative evaluation of a new intervention, specifically with regards to availability of evidence.

Design

We outline the full set of possible data scenarios that could occur if the decision maker seeks to explore the impact of the launch of a new intervention on some relevant quantifiable outcome. In each case we consider the implicit assumptions associated with conducting an evaluation, exploring possible situations where such scenarios may occur. We go on to apply the scenarios to a simulated dataset to explore how each scenario can result in different conclusions as to the effectiveness of the new intervention.

Results

We demonstrate that, across the full set of scenarios, differences in the scale of the estimated effectiveness of a new intervention and even the direction of effect, are possible given different data availability and analytical approaches.

Conclusions

When conducting quantitative evaluations of new interventions the availability of data on the outcome of interest and the analytical approach can have profound effects on the conclusions of the evaluation. While it will not always be possible to obtain a complete set of data and conduct extensive analysis, it is vital to understand the implications of the data used and consider the implicit assumptions made through its use.

Strengths and limitations of this study

- Highlights the risks of partial analysis of time series data used to evaluate the impact of a service
- Presents the assumptions implicitly made through the differential use of data to inform quantitative evaluation in a range of scenarios
- Demonstrates that even a well-designed analysis is only as good as the informative data
- Provides guidance aimed at local decision makers, who are typically overlooked in the published methodological guidance
- The use of simulated data allows for a clear demonstration of the scenarios but risks oversimplifying the nature of “real world” data

Introduction

Clinical Commissioning Groups (CCGs), Local Authorities, and other local decision makers are under increasing pressure to demonstrate the value of any newly commissioned activities given tightening fiscal budgets. While the Health and Social Care Act of 2012[1] was instrumental in allowing local decision makers to be responsive to the health needs of the population they serve, it provided little guidance on how to do so in an effective and cost-effective manner. As a result, local decision makers have found themselves caught between two worlds, neither being served by national evidence generation due to the decentralisation of funding, nor with the ability, finance, or structure to generate robust evidence, such as randomised trials.

Whilst collaborations between the Local Government Association, Department of Health, NHS England, and others has led to a number of guides for good evaluation and evidence generation,[2-4] these have had a broad focus on the theory of good research, rather than offering practical advice for analyses.

While in some cases, such as the Vanguard projects, [4] funding has been ring fenced for evaluation, it is more common that the decision to conduct a service evaluation by local decision makers comes at the detriment of the service provision itself. As a consequence, any evaluation may be limited in scope, and the ability to fund sufficiently robust data collection severely compromised. While there are inevitably risks of funding services based on inadequate evidence, as we will go on to demonstrate, there is little logic in funding sophisticated studies that threaten provision of the service itself.

It has been the experience of these authors (GR is the University of York representative on York Teaching Hospital's Council of Governors, GR and LB are members of the Vale of York CCG's Research Group, and GR, LB and SH have experience in evaluating a number of local interventions including the Harrogate and District CCG's Vanguard programme, a Core-24 hour mental health liaison service, and Tier 3 weight loss services) that these factors have resulted in either no quantitative evaluation of new service provision or evaluations that are based on limited interpretations of outcome measures and incomplete data collection. This is despite the move towards monitoring of services, both for quality and financial reasons, and falls in the cost of data generation, which have meant its collection and use, is no longer an insurmountable barrier to evaluation.

In this paper, we explore a range of different scenarios faced by a local decision maker depending on the availability of data and analytical approach taken. We go on to use a stylised case study to

explore the implications of each scenario on the estimated impact of the intervention and the likely conclusions. We focus on a quantitative evaluation but highlight the importance of a mixed-method approach in achieving a robust evaluation.

We take as a starting point a decision maker who is seeking to evaluate a new intervention, where *intervention* is used to describe any new or change in service, care pathway or treatment. They possess time series data on an outcome of interest over a series of time-points, which is hypothesised to be impacted by the new intervention. These data may be at an aggregated level (e.g. local population) or data for individuals (e.g. patients or households). Such a generalised situation is common, with the decision maker being anything from CCGs, Local Authorities, to mental health providers. While the possible set of outcomes of interest is wide, the need to generalise findings often results in focus being on broad process outcomes such as non-elective attendances, and length of stay, which are easily benchmarked. Such an analysis is expected to play a role in a decision making process informed by a number of other quantitative and qualitative considerations.

The Different Scenarios

In this section, we consider the full set of data scenarios and analytical approaches that may occur when seeking to evaluate the impact of the launch of a new intervention on a single outcome of interest. We explore the range of implicit assumptions that are made for each of the scenarios, and possible examples of how each may occur. The different cases are characterised as six overarching scenarios.

Scenario 1 – no pre-launch data

In its simplest form an evaluation may consist of only data collected after the launch of an intervention with no historical evidence, for example if the intervention was unplanned and data could not be collected retrospectively, such as a piece of hospital infrastructure being replaced. Such an analysis can therefore only comment on the trajectory of the data over the intervention period as there is no knowledge of the *counterfactual* (what would have happened had the intervention not occurred), and no data on which to base any estimation. If any estimation of the total impact of the intervention is required, assumptions or external evidence would be required to inform the counterfactual.

Scenario 2 – a single pre-launch data point is available

Secondly, we consider a situation where the decision maker has only historic data for the final period before the launch of an intervention. Such a situation may occur when the decision to conduct an evaluation occurs only a short time before the launch and data cannot be collected retrospectively. Depending on the aggregation and availability of data two sub-scenarios are available:

A. The evaluator only has data for the last period before launch and a single time point of the post-launch time series, a simple before and after statement is possible. In all cases, some implicit or explicit statement is required regarding the generalisability of the observed data and trends in the data over the intervening time-period. Such as case would occur if data were only available at set time points and only informative of a short time period, for example annually occurring surveys or audits.

B. If the data is available for the last period before launch and all post-intervention time points, an average change over the period from the first time point can be calculated with some additional knowledge of how the data changed over the period. This might occur if repeated data collection is possible prospectively, such as the collection of electronic patient data once relevant patients have been identified and consented.

Given the limited pre-launch data available in this scenario, we must assume that, had the intervention not been launched, the outcome would have stayed at the same level as in the last time point before launch. While this assumption is inevitable if no other data is available, it risks being misleading if there is some underlying trend in the outcome, or if it is subject to natural variation from one time point to the next.

Scenario 3 – Data is available covering the full pre and post-launch period

To overcome the limitations of scenario 2, historic data in the intervention area can be used to inform the baseline value and any underlying trends in the outcome over time by relaxing the assumption that outcome data would have remained static. As with scenario 2, alternative aggregation of the historic data can result in different implications:

A. In the first case the data, both pre- and post-launch, may only be available as average values aggregated over a long period, for example if the data access is limited to annual audit figures that cover the entire pre-launch period. This scenario implies that no consideration of the disaggregated trends are possible.

B. An alternative, and arguably the most common scenario used when disaggregated time series data is available, is when extensive disaggregated data is available both before and after the launch. This allows for the direct comparison of each post-launch time-period with some matched period in the pre-intervention data, for example comparing January in one year with January in the next. The matching is used to conduct the analysis at a more disaggregated level, as well as adjusting for other factors such as seasonality and budgetary cycles. While the average estimate of the impact of the intervention launch will be the same as part A, we now have the ability to investigate the change in trend over the time-period. Such a case would occur either when an evaluation and data collection was started some time before the intervention launch, or when data on the outcome is readily available retrospectively. For example, if the evaluation is concerned with emergency department attendances over time, historic data can typically be retrospectively collected.

Scenario 4 – data is available on a control area post-launch

The scenarios so far have described when data is only available for the area covered by the intervention. However, data is often available for comparator areas as the informative outcome is often routinely collected and available across multiple areas, through systems such as Hospital Episode Statistics (HES), or collection can be prospectively arranged. Such comparator areas can be local, regional, national, or a synthetic comparator created by combining a number of areas. To be an informative comparator the area must represent a good match to the intervention area in all relevant characteristics and not be impacted by the launch of the new service being evaluated.[5] The goodness of the match can be determined qualitatively or quantitatively by comparing the known features of the two areas.

The most common use of such control data is to directly compare the post intervention outcomes in the two areas, using the same approach as scenario 3 but with the contemporary control data is used instead of the historic intervention area data. As before, there are two considerations:

A. If only average data is available post intervention launch for the two areas. As in previous A scenarios, an example of this would be analyses based on audit data alone but now across multiple areas.
B. If disaggregated data points are available post intervention then as with scenario 3B, a disaggregated matched comparison can be made which again, results in the same total estimated impact as part A but gives us an understanding of the respective trends. This situation would occur where an intervention is only launched in one part of a larger

geographic area or patient group where the decision makers has access to the data of the full set prospectively, for example one GP practice in a CCG area.

Under this scenario, comparator area data is used either instead of or due to a lack of historic evidence as used in scenario 3. Using simple analytical techniques there is no way to incorporate both, which we will explore in scenario 6. There is no hard rule for whether historic or contemporary comparator evidence is more appropriate, as it is dependent on the situation. For example, if the intervention of interest was not the only change at the point of launch of the intervention, the control area data would likely be most appropriate if the second new service was launched in both areas, but not if it were only in the control area. A number of other factors must be considered, for example, what if comparator data is available but is not a good match, how does one define a suitable match, and what if there are multiple comparators potentially telling different stories?

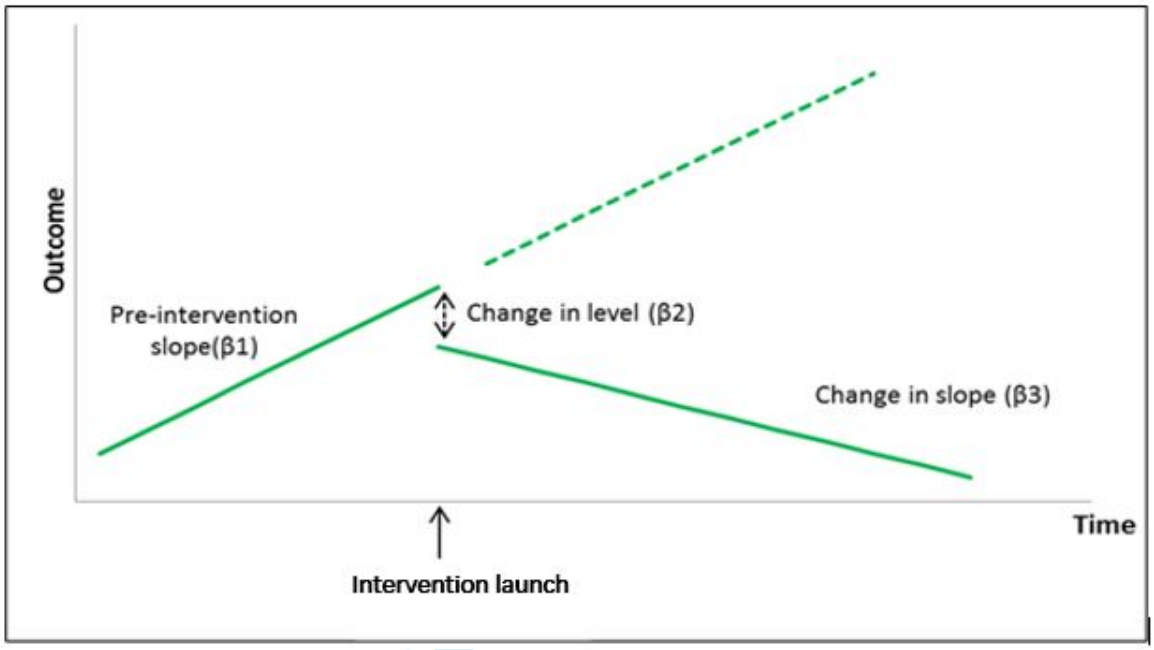
Scenario 5 – all pre and post-launch data is available

In this scenario and scenario 6 we explore the addition of more advanced analytical approaches to the analysis of the data, specifically the use of interrupted time series (ITS) or 'segmented regression' analysis. This approach has been well presented in the literature,[6-8] but in brief, the method considers the trend in an outcome of interest over time, segmenting it into the period before the intervention was launched, and after it. The example of using pre- and post-launch data for the intervention area is shown in the explanatory Figure 1, where the pre-launch data is used to infer a post-launch counterfactual case, with the nature of the change in the outcome define a-priori. Using the framework described by Bernal[8] it is possible to define the regression model using the following equation:

$$Y_t = \beta_0 + \beta_1 T_t + \beta_2 X_t + \beta_3 X_t T_t + \varepsilon_t$$

The application of such a regression model allows for the formal estimation of whether any change in the outcome of interest is statistically significant under a frequentist framework, and for any change to be quantified by estimating the area between the two regression lines, shown in Figure 1, over the analysis period.

Figure 1: ITS analytical method



The use of such method requires time series data both before and after the launch in the intervention area, as in scenario 3B.

Scenario 6 - data is available on both control and intervention areas pre- and post-launch

We demonstrated in scenario 4 that the addition of control area data typically implied the exclusion of historic intervention area data in informing the counterfactual. Using ITS methodology it is possible to formally incorporate comparator data, potentially from multiple areas or a synthetic area, alongside the full set of intervention area data. The method uses the pre-intervention data to formally test whether the comparator areas can be considered a good match. If so, the post-launch comparator data is then used to infer the post-launch counterfactual of the intervention area. Therefore, this approach assumes that the control area is indicative of what would have happened to the outcome in the intervention area had the launch not occured, much as we assumed in scenario 4 but with a formal assessment of the trend and reliability of the comparator. The equation detailed in scenario 5 can be extended to incorporate this analysis as detailed by Linden[7]:

$$Y_t = \beta_0 + \beta_1 T_t + \beta_2 X_t + \beta_3 X_t T_t + \beta_4 Z + \beta_5 Z_t T_t + \beta_6 Z X_t + \beta_7 Z X_t T_t + \varepsilon_t$$

Comparing the Scenarios

Each of the scenarios outlined above is characterised by a set of core assumptions, made implicitly or explicitly, if used to evaluate the impact of a new intervention on some outcome of interest. Similarly, the variability in the ease of implementation, and data and analytical requirements of each scenario implies a range of pros and cons associated with each. These are presented in Table 1, which highlights that the more analytically simple and data light the scenario the stronger the core assumption required about the nature of the interaction with the outcome and time trends in the data.

Table 1: Summary of the different analytical methods

Method	Core assumptions	Pros	Cons
Scenario 1, only data after launch in the intervention area	Only the change in the data after the launch is relevant to the evaluation.	Requires little data or technical knowledge.	Unable to comment on the change in the outcome of interest because of the intervention, only its trend after launch.
Scenario 2A, first and last time point of intervention period	The two data points are fully indicative of the change.	Requires little data or technical knowledge.	Highly dependent on small array of data. Risks loss of important details of data, intervention effect, or trends.
Scenario 2B, disaggregated change from starting period	Last pre-intervention period fully represents the counterfactual.	Only requires one pre-intervention data point. Analytically simple.	Highly dependent on small array of control data. No consideration of trend in counterfactual.
Scenario 3A, simple average of historic intervention area data	Simple averaging of before and after data incorporates all factors, there is no value in an assessment of the trends.	Only requires small amount of pre and post data. Analytically simple.	Fails to explore trends in data.
Scenario 3B, matched pre and post intervention	There is a repeating periodic fluctuation, e.g. seasonality, that impacts the outcome of interest and the trend over time is informative.	Simple means of adjusting for periodic fluctuations.	Result varies given matching approach. Blunt means of adjusting for periodic fluctuations that can result in incorrect estimates.
Scenario 4A, comparison of averages post intervention in control and intervention areas	The selected control area fully represents the counterfactual of the intervention area.	Allows for use of control area data. Only requires post-launch data.	Fails to explore trends in data. Makes no use of historic data. Difficult to determine if the control area represents a reasonable comparator.
Scenario 4B, matched post intervention control and intervention area	The selected control area fully represents the counterfactual of the intervention area and the trend over time is informative.	Allows for use of control area data. Explores trends in data without having to define a cycle length. Only requires post-launch data.	Makes no use of historic data. Difficult to determine if the control area represents a reasonable comparator.
Scenario 5, ITS analysis of intervention area	Regression of pre-intervention data fully represents post-intervention counterfactual and the trend over time is informative.	Allows for use of historic control data. Explores the trends.	Reliant on historic intervention area data being predictive of counterfactual.
Scenario 6, ITS analysis of control and intervention area	Control area fully represents counterfactual of intervention area but the match can be tested by exploring the pre-intervention	Allows for use of control area and exploration as to the closeness of the control and intervention areas.	Assumption that the control area continues to represent a good match after the intervention period.

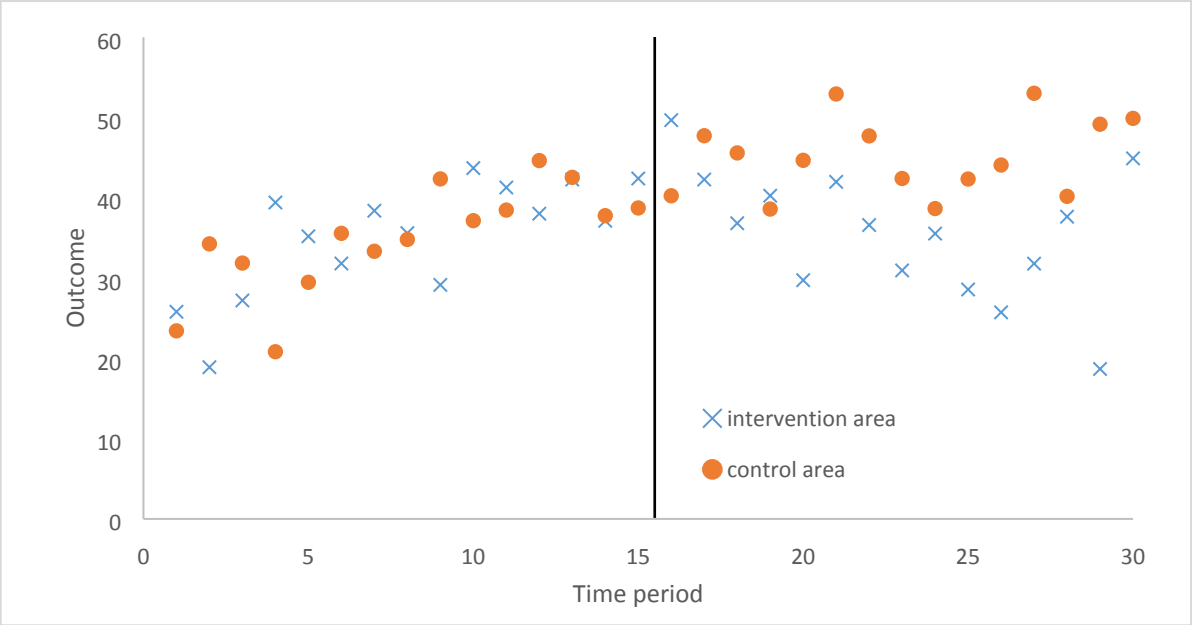
	data. The trend over time is informative.		

Case study

To explore the practical implications of the different scenarios, and demonstrate the potential for varied conclusions, we have created a case study to which each is applied. To inform the case study a time series dataset of an outcome unit of interest (e.g. bed days, hospital admissions, or indicators of quality and care outcome) has been simulated.

This data relates to two distinct groups (intervention and control) and a maximum of 30 observations are available over some defined time period at regular intervals (e.g. every week, month, or year). The data is structured such that in both areas the outcome was increasing for the first 15 observations at a rate of 4/3 per time period from a mean value of 20 units at time 1, after which point the intervention is implemented in the intervention area but not the control. From time point 15 onwards in the intervention area the outcome decreases at the same rate of 4/3 units per period, while in the control area the outcome levels off, assumed to be due to factors unrelated to the intervention. All time points are subject to some level of variation to mimic what is observed in real world data, simulated using a normal distribution (mean 45 and SD 5). We assume that after launch (t=15) the new service becomes fully operational, with no run in period. The last time point in the intervention area (t=30) was set as an extreme outlier (estimated as occurring with a probability of 0.99999 on the simulated distribution) to explore its impact on the results, for example if an exogenous factor affected the intervention such as failure of a key piece of machinery. Figure 2 shows the fabricated data in full, with each data point representing the time period before, such that data point 1 being the total outcome over times 0 to 1.

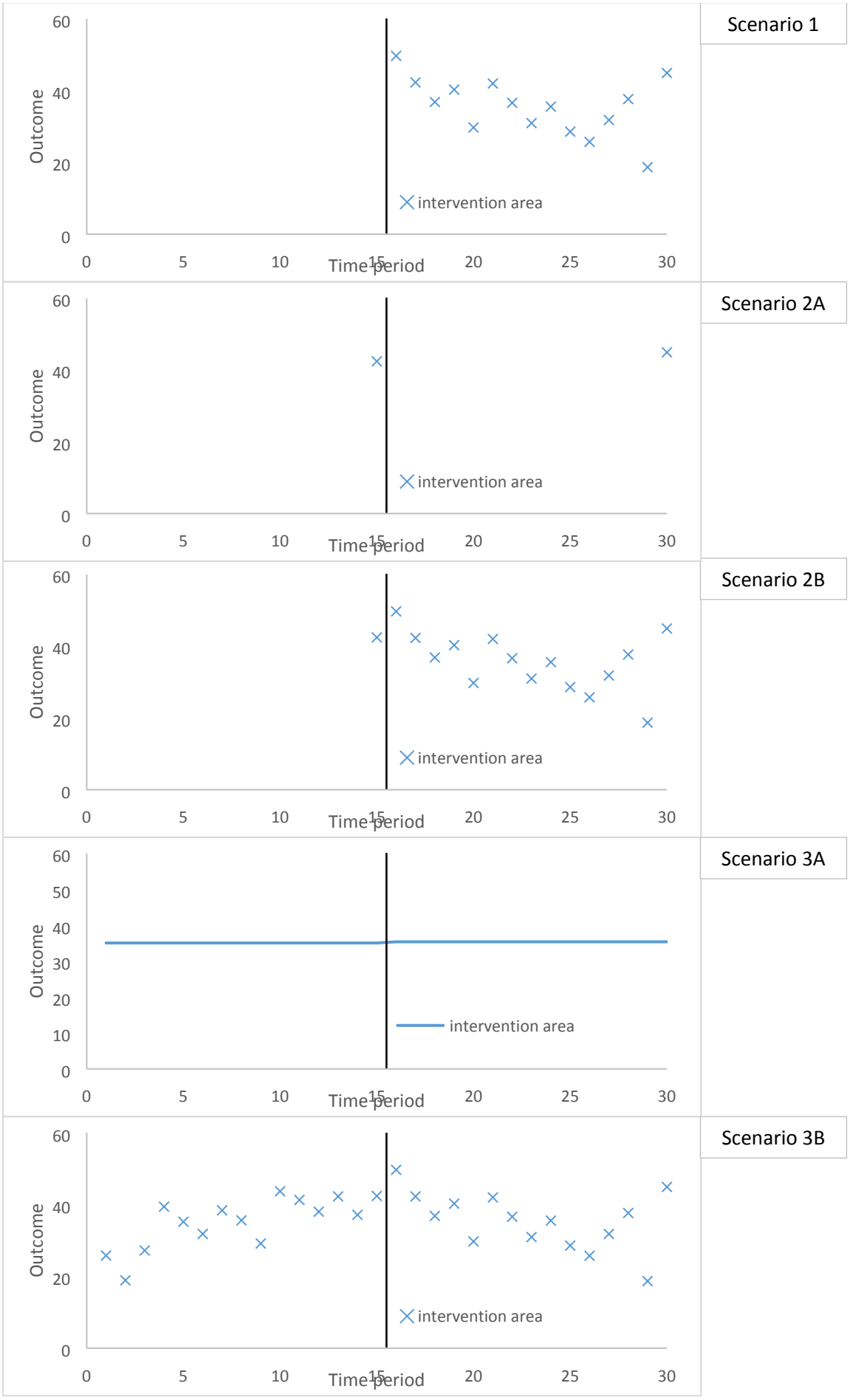
Figure 2: Fabricated time series data



Using the informative structure of the simulated case study it is possible to estimated two possible underlying effect values. If the control area is the best indicator of the counterfactual the intervention resulted in a reduction of 151 units over the period, if the historic intervention area is best, a reduction of 324 units. While these values can help us to understand the results of the different scenarios they must be interpreted with caution, as while they inform the underlying trend used to simulate the data the case study time points were sampled independently.

In the next part we explore what the data availability would look like under each of the scenarios outlined in the previous section, estimating what the impact and conclusions would be regarding the effectiveness of the intervention. As outlined earlier, in many of the cases only a limited set of the data is available, indeed it is only scenarios 4 and 6 where the full dataset is available to the decision maker. Figure 3 provides an overview of the data availability across all of the scenarios.

Figure 3: Data availability across the different scenarios of the case study



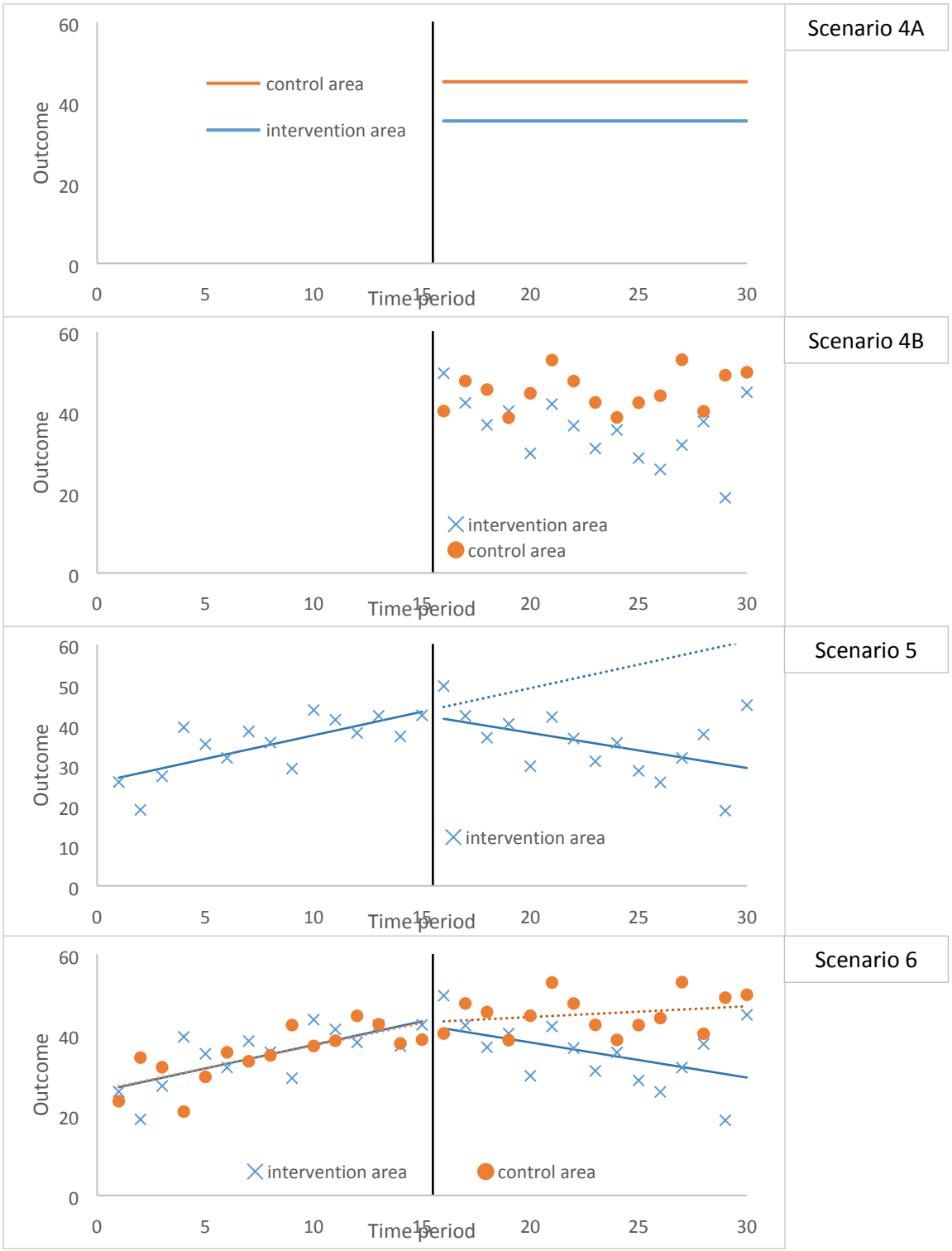

Scenario 4A
Outcome
Time period
control area
intervention area
Scenario 4B
Outcome
Time period
intervention area
control area
Scenario 5
Outcome
Time period
intervention area
Scenario 6
Outcome
Time period
intervention area
control area

Table 2 gives an overview of the results of the different possible scenarios and possible interpretations.

Table 2: Summary of the different scenarios results

Scenario	Possible interpretation of the result	Estimated change[1]
Scenario 1, only data after launch in the intervention area	The outcome of interest appears to have decreased over the post-launch time-period	not possible to estimate a change in the outcome
Scenario 2A, first and last time point of intervention period	There appears to have been an increase in the outcome from the pre-launch to post-launch period. Extrapolating the observed values over the entire 15 months of intervention suggests that the new intervention had increased the outcome by 37.6 units ((44.9-42.4)x15)	37.6
Scenario 2B, disaggregated change from starting period	The outcome of interest appears to have decreased over time from the pre-launch time-period, with an estimated change of -120.1 units over the period ((44.9-34.4)x15)	-120.1
Scenario 3A, simple average of historic intervention area data	There appears to have been little change from the pre- to post-launch periods in the outcome, with the average value going from 35.1 to 35.4	4.9
Scenario 3B, matched pre and post intervention	There appears to have been little change from the pre- to post-launch periods in the outcome, with the average value going from 35.1 to 35.4. However, it appears from the data that there was an increasing trend in the outcome before the intervention and a decreasing trend afterwards	4.9
Scenario 4A, comparison of averages post intervention in control and intervention areas	Compared to the control area the intervention area had a lower average level of the outcome after the launch of the intervention	-146.0
Scenario 4B, matched post intervention control and intervention area	Compared to the control area the intervention area had a lower average level of the outcome after the launch of the intervention. The control area appeared to have a flat trend in the outcome over the post-launch period compared to a decreasing trend in the intervention area	-146.0
Scenario 5, ITS analysis of intervention area	Compared to the pre-launch intervention area the post-launch saw a decrease in the trend over time in the outcome, from positive to negative, which was statistically significant	-258.8
Scenario 6, ITS analysis of control and intervention area	Both control and intervention areas saw a shallowing of the trend over time. The intervention area saw a greater decrease in the trend, being negative compared to the relatively flat trend in the control. This different was statistically significant. The control area was found to be a match to the intervention area in the pre-launch period	-146.0

[1]negative values indicate that the new service reduced the outcome

Figure 3 and Table 2 demonstrate the large potential for variation in the estimated impact of the intervention, and the overall conclusions that could be drawn given the different scenarios. Estimations of the change in the outcome vary from predicting the intervention increased the outcome by 37.6 units over the post-intervention period (scenario 2A), to decreasing it by 258.8 (scenario 5). Similarly the interpretations differ in their ability to identify the trends in the different areas and time periods, as well as the overall impact of the intervention.

In the case study presented here, with full access to the data and knowledge of the underlying trends in the simulated data, it is clear that several of these scenarios result is a very incorrect conclusion. However, the appropriateness of the scenarios and accuracy of their conclusions compared to any 'true' effects are clearly much harder to determine in the real world.

Discussion

In this paper we have explored a range of possible scenarios and analytical approaches available to a decision maker when evaluating the impact of a new intervention on an outcome of interest, highlighting the implicit assumptions made in each. Through our simulated case study we have demonstrated how these scenarios can yield very different estimates of effectiveness.

Comparison of the methods suggests that it is intuitively appealing to conclude that the approach outlined in scenario 6 , using the ITS methodology including the control area comparison, is the most accurate as it incorporates the most complete set of data whilst taking the most complete approach to statistical analysis. However, the actual optimal methodology may be driven by other factors, primarily the availability of informative data and the validity of the core assumptions detailed in Table 1.

Furthermore, the use of ITS analysis (scenarios 5 and 6) is not without assumptions, primarily relating to the suitability of the historic and control area data to inform the counterfactual, and the functional form of the trends modelled. It also requires a significant level of data and analytical ability to implement. However, the inability to observe exactly what would happen in the intervention area without the new service, necessitates such assumptions in order to estimate the impact of its launch. Fears about the robustness of such assumptions are likely to be best addressed by the identification of additional relevant evidence to either adjust the existing data or inform a new comparator. For example, methods are available to overcome concerns over additional service changes in the time period covered by the data,[7] to incorporate multiple control areas,[7] and to conduct a more rigorous selection of control area through matching.[9]

As with all such analyses, the ITS methodology can be extended to consider the significance of the findings beyond pure chance. This can be achieved through a frequentist framework, considering the statistical significance of the regression estimates, as discussed in Linden et al.,[7] or through a Bayesian framework.[10] Such considerations should play an important role in the decision making process, as a single estimate of the impact on an intervention can be misleading. Specifically, it fails to take account of the uncertainty of the informative data or the consequences of making an incorrect funding decision. Consideration of the impact on the quality of care should also be considered, where possible, to ensure improvements in activity levels does not come at the detriment of quality.

Analyses such as those presented here are most robust when combined with qualitative methodologies through a mixed-method approach, with the qualitative findings ideally facilitating a more detailed understanding of the trends seen in the data and informing the suitability of the different counterfactual scenarios. Furthermore, the use of robust methodologies, such as ITS analysis, does not replace the need for robust identification of relevant outcomes and data collection, alongside the prospective planning of evaluations, as any analysis can only be as robust as the data that informs it. Therefore, failure to prospectively design an intervention launch and evaluation which ensures the robust implementation of the new intervention, the required level of data collection, and sufficient consideration as to a contemporaneous control, will likely lead to an erroneous result whatever evaluative method is used.

Patient and public involvement

As the informative dataset was simulated there was no patient nor public involvement in this study, nor was consent required for access to patient data.

References

1. Health and Social Care Act 2012, avaiable at: http://www.legislation.gov.uk/ukpga/2012/7/contents/enacted (accessed 31/05/2018). **c.7.**
2. Bath Research & Development, *Research, evaluation and evidence: a guide for commissioners.* NHS Research and Development Forum, accessed: 31/-5/2018. **http://www.rdforum.nhs.uk/content/wp-content/uploads/2016/04/GUIDE-Evaluation-and-Research-for-CCGs-Final-version-April-2016.pdf**.
3. The Better Care Fund, *How to... understand and measure impact.* NHS England, 2015. **https://www.england.nhs.uk/wp-content/uploads/2015/06/bcf-user-guide-04.pdf.pdf accessed: 31/05/2018**(4).
4. NHS England, *Evaluation strategy for new care model vanguards.* NHS England, 2016. **https://www.england.nhs.uk/wp-content/uploads/2015/07/ncm-evaluation-strategy-may-2016.pdf accessed: 31/05/2018**.
5. Linden, A., *Conducting interrupted time-series analysis for single- and multiple-group comparisons.* The Stata Journal, 2015. **15, Number 2, pp. 480–500**.
6. Cruz, M., M. Bender, and H. Ombao, *A robust interrupted time series model for analyzing complex health care intervention data.* Statistics in Medicine, 2017. **36**(29): p. 4660-4676.
7. Linden, A., *Conducting interrupted time-series analysis for single- and multiple-group comparisons.* Stata Journal, 2015. **15**(2): p. 480-500.
8. Bernal, J.L., A. Gasparrini, and S. Cummins, *Interrupted time series regression for the evaluation of public health interventions: a tutorial.* International Journal of Epidemiology, 2016. **46**(1): p. 348-355.
9. Linden, A., *Improving causal inference with a doubly robust estimator that combines propensity score stratification and weighting.* J Eval Clin Pract, 2017. **23**(4): p. 697-702.
10. Spiers, G., et al., *Transforming community health services for children and young people who are ill: a quasi-experimental evaluation.* 2016. **4**(25): p. 1-222.

Supplementary Appendix: Regression output for ITS analysis (scenarios 5 and 6)

This appendix reports the regression outputs for the ITS analysis presented in scenarios 5 and 6 using the ITSA program in Stata.

Regression output for scenario 5

Outcome	Coef.	Newey-West Std. Err.	t	P>t	[95% Conf.	Interval]
β1	1.172109	0.2951369	3.97	0.001	0.565446	1.778772
β2	-2.93635	4.039829	-0.73	0.474	-11.2403	5.367642
β3	-2.04554	0.6483852	-3.15	0.004	-3.37832	-0.71276
β0	26.88192	2.872154	9.36	0	20.97813	32.78572
Treated (β1[_t]+β3[_x_t16])	-0.8734	0.5773	-1.5129	0.1424	-2.0601	0.3133

Regression output for scenario 6

outcome	Coef.	Newey-West Std. Err.	t	P>t	[95% Conf.	Interval]
β1	1.128589	0.2891484	3.9	0	0.548371	1.708808
β4	-0.23341	3.888218	-0.06	0.952	-8.03569	7.568873
β5	0.04352	0.4131738	0.11	0.917	-0.78557	0.872614
β2	-0.76556	3.07017	-0.25	0.804	-6.92631	5.395184
β3	-0.86161	0.3851036	-2.24	0.03	-1.63438	-0.08885
β6	-2.17078	5.074068	-0.43	0.671	-12.3527	8.011078
β7	-1.18393	0.7541274	-1.57	0.122	-2.6972	0.32934
β0	27.11533	2.620871	10.35	0	21.85617	32.37449
Treated (β1[_t]+β5[_z_t] +β3[_x_t16] +β7[_z_x_t16])	-0.8734	0.5773	-1.5129	0.1364	-2.0319	0.285
Controls (β1[_t]+β3[_x_t16])	0.267	0.2544	1.0496	0.2988	-0.2434	0.7774
Difference (β5[_z_t] +β7[_z_x_t16])	-1.1404	0.6309	-1.8077	0.0764	-2.4063	0.1255

BMJ Open

Understanding and Addressing the Challenges of Conducting Quantitative Evaluation at a Local Level, a worked example of the available approaches

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2019-029830.R1
Article Type:	Original research
Date Submitted by the Author:	23-Sep-2019
Complete List of Authors:	Hinde, Sebastian; York University, Centre for Health Economics Bojke, Laura; University of York, Centre for Health Economics Richardson, Gerry; University of York, Centre for Health Economics
Primary Subject Heading:	Research methods
Secondary Subject Heading:	Health services research
Keywords:	Health policy < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Health economics < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, STATISTICS & RESEARCH METHODS

Understanding and Addressing the Challenges of Conducting Quantitative Evaluation at a Local Level, a worked example of the available approaches

Sebastian Hinde, Laura Bojke, Gerry Richardson

Centre for Health Economics, University of York

Corresponding author:
Sebastian Hinde,
Centre for Health Economics,
Alcuin 'A' Block,
University of York,
Heslington,
North Yorkshire,
YO10 5DD.
United Kingdom
Sebastian.Hinde@york.ac.uk
+44 (0)1904 321455

SH developed the research idea, led the writing of the manuscript, and acts as the guarantor of the article. LB and GR gave input at all stages including the commenting on the manuscript. All authors are health economists at the Centre for Health Economics, University of York, with experience in working on evaluation for local decision makers at various levels. The simulated dataset and all analytical code is available as supplementary appendices, see "Fabricated data scenarios.xlsm", "Paper analysis.do" and "Fabricated data.dta". Please contact the corresponding author sebastian.hinde@york.ac.uk for full access. Due to the simulated nature of the dataset no patients were involved in the study.

This article presents independent research by the National Institute for Health Research Collaboration for Leadership in Applied Health Research and Care Yorkshire and Humber (NIHR CLAHRC YH) and the NIHR Applied Research Collaboration Yorkshire and Humber (ARC YH). The views and opinions expressed are those of the authors, and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Abstract

Objectives

In the context of tightening fiscal budgets and increased commissioning responsibility, local decision makers across the UK healthcare sector have found themselves in charge of the implementation and evaluation of a greater range of healthcare interventions and services. However, there is often little experience, guidance, or funding available at a local level to ensure robust evaluations are conducted. In this paper, we evaluate the possible scenarios that could occur when seeking to conduct a quantitative evaluation of a new intervention, specifically with regards to availability of evidence.

Design

We outline the full set of possible data scenarios that could occur if the decision maker seeks to explore the impact of the launch of a new intervention on some relevant quantifiable outcome. In each case we consider the implicit assumptions associated with conducting an evaluation, exploring possible situations where such scenarios may occur. We go on to apply the scenarios to a simulated dataset to explore how each scenario can result in different conclusions as to the effectiveness of the new intervention.

Results

We demonstrate that, across the full set of scenarios, differences in the scale of the estimated effectiveness of a new intervention and even the direction of effect, are possible given different data availability and analytical approaches.

Conclusions

When conducting quantitative evaluations of new interventions the availability of data on the outcome of interest and the analytical approach can have profound effects on the conclusions of the evaluation. While it will not always be possible to obtain a complete set of data and conduct extensive analysis, it is vital to understand the implications of the data used and consider the implicit assumptions made through its use.

Strengths and limitations of this study

- Highlights the risks of partial analysis of time series data used to evaluate the impact of a service
- Presents the assumptions implicitly made through the differential use of data to inform quantitative evaluation in a range of scenarios
- Demonstrates that even a well-designed analysis is only as good as the informative data
- Provides guidance aimed at local decision makers, who are typically overlooked in the published methodological guidance
- The use of simulated data allows for a clear demonstration of the scenarios but risks oversimplifying the nature of “real world” data

Introduction

Clinical Commissioning Groups (CCGs), Local Authorities, and other local decision makers are under increasing pressure to demonstrate the value of any newly commissioned activities given tightening fiscal budgets. While the Health and Social Care Act of 2012[1] was instrumental in allowing local decision makers to be responsive to the health needs of the population they serve, it provided little guidance on how to do so in an effective and cost-effective manner. As a result, local decision makers have found themselves caught between two worlds, neither being served by national evidence generation due to the decentralisation of funding, nor with the ability, finance, or structure to generate robust evidence, such as randomised trials.

Whilst collaborations between the Local Government Association, Department of Health, NHS England, and others have led to a number of guides for good evaluation and evidence generation, [2-4] these have had a broad focus on the theory of good research, rather than offering practical advice for analyses.

While in some cases, such as the Vanguard projects, [4] funding has been ring fenced for evaluation, it is more common that the decision to conduct a service evaluation by local decision makers comes at the detriment of the service provision itself. As a consequence, any evaluation may be limited in scope, and the ability to fund sufficiently robust data collection severely compromised. While there are inevitably risks of funding services based on inadequate evidence, as we will go on to demonstrate, there is little logic in funding sophisticated studies that threaten provision of the service itself.

It has been the experience of these authors (GR is the University of York representative on York Teaching Hospital's Council of Governors, GR and LB are members of the Vale of York CCG's Research Group, and GR, LB and SH have experience in evaluating a number of local interventions including the Harrogate and District CCG's Vanguard programme, a Core-24 hour mental health liaison service, and Tier 3 weight loss services) that these factors have resulted in either no quantitative evaluation of new service provision or evaluations that are based on limited interpretations of outcome measures and incomplete data collection. This is despite the move towards monitoring of services, both for quality and financial reasons, and falls in the cost of data generation, which have meant its collection and use is no longer an insurmountable barrier to evaluation.

In this paper, we explore a range of different scenarios faced by a local decision maker depending on the availability of data and analytical approach taken. We go on to use a stylised case study to explore the implications of each scenario on the estimated impact of the intervention and the likely

conclusions. We focus on a quantitative evaluation but highlight the importance of a mixed-method approach in achieving a robust evaluation.

We take as a starting point a decision maker who is seeking to evaluate a new intervention, where *intervention* is used to describe any new or change in service, care pathway or treatment. They possess time series data on an outcome of interest over a series of time-points, which is hypothesised to be impacted by the new intervention. These data may be at an aggregated level (e.g. local population) or data for individuals (e.g. patients or households). Such a generalised situation is common, with the decision maker being anything from CCGs, Local Authorities, to mental health providers. While the possible set of outcomes of interest is wide, the need to generalise findings often results in focus being on broad process outcomes such as non-elective attendances, and length of stay, which are easily benchmarked. Such an analysis is expected to play a role in a decision making process informed by a number of other quantitative and qualitative considerations.

The Different Scenarios

In this section, we consider the full set of data scenarios and analytical approaches that may occur when seeking to evaluate the impact of the launch of a new intervention on a single outcome of interest. We explore the range of implicit assumptions that are made for each of the scenarios, and possible examples of how each may occur. The different cases are characterised as six overarching scenarios. While there are few cases where data and analytical capability is not available to conduct all of the scenarios presented, as is explored alter in this manuscript, some of the more demanding cases require an element of forethought and buy-in from all stakeholders in order to ensure to facilitate the most appropriate scenario. It is the experience of these authors that it is most common for evaluation of an intervention to be done retrospectively or towards the end of a project, primarily due to a lack of evaluative experience and funding to embed evaluation from an early stage, however, there is a lack of reviews of the methodology applied by local decision makers in such setting.

Scenario 1 – follow-up data but no pre-launch data for the intervention area

In its simplest form an evaluation may consist of only data collected after the launch of an intervention with no historical evidence, for example if the intervention was unplanned and data could not be collected retrospectively, such as a piece of hospital infrastructure being replaced. Such an analysis can therefore only comment on the trajectory of the data over the intervention period as there is no knowledge of the *counterfactual* (what would have happened had the intervention not occurred), and no data on which to base any estimation. If any estimation of the

total impact of the intervention is required, assumptions or external evidence would be required to inform the counterfactual.

Scenario 2 – follow-up data and a single pre-launch data point for the intervention area

Secondly, we consider a situation where the decision maker has only historic data for the final period before the launch of an intervention. Such a situation may occur when the decision to conduct an evaluation occurs only a short time before the launch and data cannot be collected retrospectively. Depending on the aggregation and availability of data two sub-scenarios are available:

A. Data are only available for the last period before launch and a single time point of the post-launch time series, a simple before and after statement is possible. In all cases, some implicit or explicit statement is beneficial regarding the generalisability of the observed data and trends in the data over the intervening time-period. Such as case would occur if data were only available at set time points and only informative of a short time period, for example annually occurring surveys or audits.

B. Data are available for the last period before launch and all post-intervention time points, allowing an average change over the period from the first time point to be calculated with some additional knowledge of how the data changed over the period. This might occur if repeated data collection is possible prospectively, such as the collection of electronic patient data once relevant patients have been identified and consented.

Given the limited pre-launch data available in this scenario, we must assume that, had the intervention not been launched, the outcome would have stayed at the same level as in the last time point before launch. While this assumption is inevitable if no other data are available, it risks being misleading if there is some underlying trend in the outcome, or if it is subject to natural variation from one time point to the next.

Scenario 3 – data are available covering the full pre and post-launch period for the intervention area

To overcome the limitations of scenario 2, historic data in the intervention area can be used to inform the baseline value and any underlying trends in the outcome over time by relaxing the assumption that outcome data would have remained static. As with scenario 2, alternative aggregation of the historic data can result in different implications:

A. Both pre- and post-launch, may only be available as average values aggregated over a long period, for example if the data access is limited to annual audit figures that cover the entire pre-launch period. This scenario implies that no consideration of the disaggregated trends are possible.

B. Extensive disaggregated data are available both before and after the launch. This allows for the direct comparison of each post-launch time-period with some matched period in the pre-intervention data, for example comparing January in one year with January in the next. The matching is used to conduct the analysis at a more disaggregated level, as well as adjusting for other factors such as seasonality and budgetary cycles. While the average estimate of the impact of the intervention launch will be the same as part A, we now have the ability to investigate the change in trend over the time-period. Such a case would occur either when an evaluation and data collection was started some time before the intervention launch, or when data on the outcome is readily available retrospectively. For example, if the evaluation is concerned with emergency department attendances over time, historic data can typically be retrospectively collected.

Scenario 4 – data are available on a control area post-launch as well as the intervention area data

Scenarios 1-3 describe when data are only available for the area covered by the intervention. However, data are often available for comparator areas as the informative outcome is often routinely collected and available across multiple areas, through systems such as Hospital Episode Statistics (HES), or collection can be prospectively arranged. Such comparator areas can be local, regional, national, or a synthetic comparator created by combining a number of areas. To be an informative comparator the area must represent a good match to the intervention area in all relevant characteristics and not be impacted by the launch of the new service being evaluated.[5] The goodness of the match can be determined qualitatively or quantitatively by comparing the known features of the two areas.

The most common use of such control data are to directly compare the post intervention outcomes in the two areas, using the same approach as scenario 3 but with the contemporary control data are used instead of the historic intervention area data. As before, there are two categories:

A. Data are only available post intervention launch for the two areas. As in previous scenarios, an example of this would be analyses based on audit data alone but now across multiple areas.

B. Disaggregated data are available post intervention, allowing a disaggregated matched comparison can be made which again, results in the same total estimated impact as part A

but gives us an understanding of the respective trends. This situation would occur where an intervention is only launched in one part of a larger geographic area or patient group where the decision makers has access to the data of the full set prospectively, for example one GP practice in a CCG area.

Under this scenario, comparator area data are used either instead of or due to a lack of historic evidence as used in scenario 3. Using simple analytical techniques there is no way to incorporate both, which we will explore in scenario 6. There is no hard rule for whether historic or contemporary comparator evidence is more appropriate, as it is dependent on the situation. For example, if the intervention of interest was not the only change at the point of launch of the intervention, the control area data would likely be most appropriate if the second new service was launched in both areas, but not if it were only in the control area. A number of other factors must be considered, for example, what if comparator data are available but is not a good match, how does one define a suitable match, and what if there are multiple comparators potentially telling different stories?

Scenario 5 – all pre and post-launch data are available for the intervention area

In this scenario and scenario 6 we explore the addition of more advanced analytical approaches to the analysis of the data, specifically the use of interrupted time series (ITS) or 'segmented regression' analysis. This approach has been well presented in the literature,[6-8] but in brief, the method considers the trend in an outcome of interest over time, segmenting it into the period before the intervention was launched, and after it. The example of using pre- and post-launch data for the intervention area is shown in the explanatory **Error! Reference source not found.**, where the pre-launch data are used to infer a post-launch counterfactual case, with the nature of the change in the outcome define a-priori. Using the framework described by Bernal[8] it is possible to define the regression model using the equation detailed below, where Y is the aggregated outcome, β represents the relevant coefficients, T the time since the start of the study, t the specific time-point, X a dummy variable of the intervention, and ε the error term.

$$Y_t = \beta_0 + \beta_1 T_t + \beta_2 X_t + \beta_3 X_t T_t + \varepsilon_t$$

The application of such a regression model allows for the formal estimation of whether any change in the outcome of interest is statistically significant under a frequentist framework, and for any

change to be quantified by estimating the area between the two regression lines, shown in **Error! Reference source not found.**, over the analysis period.

The use of such method requires time series data both before and after the launch in the intervention area, as in scenario 3B.

[Figure 1 here]

Scenario 6 - data are available on both control and intervention areas pre- and post-launch

We demonstrated in scenario 4 that the addition of control area data typically implied the exclusion of historic intervention area data in informing the counterfactual. Using ITS methodology it is possible to formally incorporate comparator data, potentially from multiple areas or a synthetic area, alongside the full set of intervention area data. The method uses the pre-intervention data to formally test whether the comparator areas can be considered a good match. If so, the post-launch comparator data are then used to infer the post-launch counterfactual of the intervention area. Therefore, this approach assumes that the control area is indicative of what would have happened to the outcome in the intervention area had the launch not occurred, much as we assumed in scenario 4 but with a formal assessment of the trend and reliability of the comparator. The equation detailed in scenario 5 can be extended by incorporating a Z term as a dummy for assignment to the treatment or control population, as detailed by Linden[7]:

$$Y_t = \beta_0 + \beta_1 T_t + \beta_2 X_t + \beta_3 X_t T_t + \beta_4 Z + \beta_5 Z T_t + \beta_6 Z X_t + \beta_7 Z X_t T_t + \varepsilon_t$$

Comparing the Scenarios

Each of the scenarios outlined above is characterised by a set of core assumptions, made implicitly or explicitly, if used to evaluate the impact of a new intervention on some outcome of interest. Similarly, the variability in the ease of implementation, and data and analytical requirements of each scenario implies a range of pros and cons associated with each. These are presented in Table 1, which highlights that the more analytically simple and data light the scenario the stronger the core assumption required about the nature of the interaction with the outcome and time trends in the data.

Table 1: Summary of the different analytical methods

Method	Core assumptions	Pros	Cons
Scenario 1, only data after launch in the intervention area	Only the change in the data after the launch is relevant to the evaluation.	Requires little data or technical knowledge.	Unable to comment on the change in the outcome of interest because of the intervention, only its trend after launch.
Scenario 2A, first and last time point of intervention period	The two data points are fully indicative of the change.	Requires little data or technical knowledge.	Highly dependent on small array of data. Risks loss of important details of data, intervention effect, or trends.
Scenario 2B, disaggregated change from starting period	Last pre-intervention period fully represents the counterfactual.	Only requires one pre-intervention data point. Analytically simple.	Highly dependent on small array of control data. No consideration of trend in counterfactual.
Scenario 3A, simple average of historic intervention area data	Simple averaging of before and after data incorporates all factors, there is no value in an assessment of the trends.	Only requires small amount of pre and post data. Analytically simple.	Fails to explore trends in data.
Scenario 3B, matched pre and post intervention	There is a repeating periodic fluctuation, e.g. seasonality, that impacts the outcome of interest and the trend over time is informative.	Simple means of adjusting for periodic fluctuations.	Result varies given matching approach. Blunt means of adjusting for periodic fluctuations that can result in incorrect estimates.
Scenario 4A, comparison of averages post intervention in control and intervention areas	The selected control area fully represents the counterfactual of the intervention area.	Allows for use of control area data. Only requires post-launch data.	Fails to explore trends in data. Makes no use of historic data. Difficult to determine if the control area represents a reasonable comparator.
Scenario 4B, matched post intervention control and intervention area	The selected control area fully represents the counterfactual of the intervention area and the trend over time is informative.	Allows for use of control area data. Explores trends in data without having to define a cycle length. Only requires post-launch data.	Makes no use of historic data. Difficult to determine if the control area represents a reasonable comparator.
Scenario 5, ITS analysis of intervention area	Regression of pre-intervention data fully represents post-intervention counterfactual and the trend over time is informative.	Allows for use of historic control data. Explores the trends.	Reliant on historic intervention area data being predictive of counterfactual.
Scenario 6, ITS analysis of control and intervention area	Control area fully represents counterfactual of intervention area but the match can be tested by exploring the pre-intervention data. The trend over time is informative.	Allows for use of control area and exploration as to the closeness of the control and intervention areas.	Assumption that the control area continues to represent a good match after the intervention period.

Case study

To explore the practical implications of the different scenarios, and demonstrate the potential for varied conclusions, we have created a case study to which each is applied. To inform the case study a time series dataset of an outcome unit of interest (e.g. bed days, hospital admissions, or indicators of quality and care outcome) has been simulated. The data values and number of time points has been selected to best inform the characteristics of each of the scenarios described in Table 1 while representing the uncertain nature of real world data relevant to this setting. Please see the supplementary files ("Fabricated data scenarios.xlsm", "Paper analysis.do" and "Fabricated data.dta") for additional detail on the data and analyses conducted.

This data relates to two distinct groups (intervention and control) and a maximum of 30 observations are available over some defined time period at regular intervals (e.g. every week, month, or year). The data are structured such that in both areas the outcome was increasing for the first 15 observations at a rate of 4/3 per time period from a mean value of 20 units at time 1, after which point the intervention is implemented in the intervention area but not the control. From time point 15 onwards in the intervention area the outcome decreases at the same rate of 4/3 units per period, while in the control area the outcome levels off, assumed to be due to factors unrelated to the intervention. All time points are subject to some level of variation to mimic what is observed in real world data, simulated using a normal distribution (mean 45 and standard deviation 5). We assume that after launch (t=15) the new service becomes fully operational, with no run in period. The last time point in the intervention area (t=30) was set as an extreme outlier (estimated as occurring with a probability of 0.99999 on the simulated distribution) to explore its impact on the results, for example if an exogenous factor affected the intervention such as failure of a key piece of machinery. **Error! Reference source not found.** shows the fabricated data in full, with each data point representing the time period before, such that data point 1 being the total outcome over times 0 to 1.

[Figure 2 here]

Using the informative structure of the simulated case study it is possible to estimate two possible underlying effect values. If the control area is the best indicator of the counterfactual the intervention resulted in a reduction of 151 units over the period, if the historic intervention area is best, a reduction of 324 units. While these values can help us to understand the results of the different scenarios they must be interpreted with caution, as while they inform the underlying trend used to simulate the data the case study time points were sampled independently.

In the next part we explore what the data availability would look like under each of the scenarios outlined in the previous section, estimating what the impact and conclusions would be regarding the effectiveness of the intervention. As outlined earlier, in many of the cases only a limited set of the data are available, indeed it is only scenarios 4 and 6 where the full dataset is available to the decision maker. **Error! Reference source not found.** provides an overview of the data availability across all of the scenarios.

[Figure 3 here]

Table 2 gives an overview of the results of the different possible scenarios and possible interpretations.

Table 2: Summary of the different scenarios results

Scenario	Possible interpretation of the result	Estimated change[1]
Scenario 1, only data after launch in the intervention area	The outcome of interest appears to have decreased over the post-launch time-period	not possible to estimate a change in the outcome
Scenario 2A, first and last time point of intervention period	There appears to have been an increase in the outcome from the pre-launch to post-launch period. Extrapolating the observed values over the entire 15 months of intervention suggests that the new intervention had increased the outcome by 37.6 units ((44.9-42.4)x15)	37.6
Scenario 2B, disaggregated change from starting period	The outcome of interest appears to have decreased over time from the pre-launch time-period, with an estimated change of -120.1 units over the period ((34.4-42.4)x15)	-120.1
Scenario 3A, simple average of historic intervention area data	There appears to have been little change from the pre- to post-launch periods in the outcome, with the average value going from 35.1 to 35.4 ((35.4-35.1)x15)	4.9
Scenario 3B, matched pre and post intervention	There appears to have been little change from the pre- to post-launch periods in the outcome, with the average value going from 35.1 to 35.4. However, it appears from the data that there was an increasing trend in the outcome before the intervention and a decreasing trend afterwards ((35.4-35.1)x15)	4.9
Scenario 4A, comparison of averages post intervention in control and intervention areas	Compared to the control area the intervention area had a lower average level of the outcome after the launch of the intervention	-146.0
Scenario 4B, matched post intervention control and intervention area	Compared to the control area the intervention area had a lower average level of the outcome after the launch of the intervention. The control area appeared to have a flat trend in the outcome over the post-launch period compared to a decreasing trend in the intervention area ((35.4-45.1)x15)	-146.0
Scenario 5, ITS analysis of intervention area	Compared to the pre-launch intervention area the post-launch saw a decrease in the trend over time in the outcome, from positive to negative, which was statistically significant See the Supplementary Appendix for regression	-258.8
Scenario 6, ITS analysis of control and intervention area	Both control and intervention areas saw a shallowing of the trend over time. The intervention area saw a	-146.0

	greater decrease in the trend, being negative compared to the relatively flat trend in the control. This different was statistically significant. The control area was found to be a match to the intervention area in the pre-launch period See the Supplementary Appendix for regression	

[1]negative values indicate that the new service reduced the outcome

Error! Reference source not found. and Table 2 demonstrate the large potential for variation in the estimated impact of the intervention, and the overall conclusions that could be drawn given the different scenarios. Estimations of the change in the outcome vary from predicting the intervention increased the outcome by 37.6 units over the post-intervention period (scenario 2A), to decreasing it by 258.8 (scenario 5). Similarly the interpretations differ in their ability to identify the trends in the different areas and time periods, as well as the overall impact of the intervention.

In the case study presented here, with full access to the data and knowledge of the underlying trends in the simulated data, it is clear that several of these scenarios result is a very incorrect conclusion. However, the appropriateness of the scenarios and accuracy of their conclusions compared to any 'true' effects are clearly much harder to determine in the real world.

Discussion

In this paper we have explored a range of possible scenarios and analytical approaches available to a decision maker when evaluating the impact of a new intervention on an outcome of interest, highlighting the implicit assumptions made in each. Through our simulated case study we have demonstrated how these scenarios can yield very different estimates of effectiveness.

Comparison of the methods explored here suggests that it is intuitively appealing to conclude that the approach outlined in scenario 6, using the ITS methodology including the control area comparison, is the most accurate as it incorporates the most complete set of data whilst taking the most complete approach to statistical analysis. However, the most appropriate methodology may be driven by other factors, primarily the availability of informative data and the validity of the core assumptions detailed in Table 1.

Furthermore, the use of ITS analysis (scenarios 5 and 6) is not without assumptions, primarily relating to the suitability of the historic and control area data to inform the counterfactual, and the functional form of the trends modelled. It also requires a significant level of data and analytical ability to implement. However, the inability to observe exactly what would happen in the intervention area without the new service, necessitates such assumptions in order to estimate the impact of its launch. Fears about the robustness of such assumptions are likely to be best addressed

by the identification of additional relevant evidence to either adjust the existing data or inform a new comparator. For example, methods are available to overcome concerns over additional service changes in the time period covered by the data,[7] to incorporate multiple control areas,[7] and to conduct a more rigorous selection of control area through matching.[9]

As with all such analyses, the ITS methodology can be extended to consider the significance of the findings beyond pure chance. This can be achieved through a frequentist framework, considering the statistical significance of the regression estimates, as discussed in Linden et al.,[7] or through a Bayesian framework.[10] Such considerations should play an important role in the decision making process, as a single estimate of the impact on an intervention can be misleading. Specifically, it fails to take account of the uncertainty of the informative data or the consequences of making an incorrect funding decision. However, it is important to reflect that even if there is substantial uncertainty it is the mean estimate of the impact of the intervention that should be most informative to the commissioning decision, as argued by Claxton [11].

An intrinsic element to any analyses explored in this paper is an understanding of the data under interrogation, the application of robust methods is only helpful if the data being used is consistent and relevant to the question it is being used to answer. Prior to any analysis it is important to understand the data, answering questions such as how was it generated, is an estimate of the rate of an event more relevant than it's frequency, is it consistent over the time period of interest, what is the route of causality between the intervention of interest and the data, and when plotted do there appear to be any unexplainable outliers?

Analyses such as those presented here are most robust when combined with qualitative methodologies through a mixed-method approach, with the qualitative findings ideally facilitating a more detailed understanding of the trends seen in the data and informing the suitability of the different counterfactual scenarios. Such a mixed-methods approach may extend the quantitative incorporate health economic considerations, such that the generalisable cost-effectiveness of the intervention is considered.

Furthermore, the use of robust methodologies, such as ITS analysis, does not replace the need for robust identification of relevant outcomes and data collection, alongside the prospective planning of evaluations, as any analysis can only be as robust as the data that informs it. Therefore, failure to prospectively design an intervention launch and evaluation which ensures the robust implementation of the new intervention, the required level of data collection, and sufficient consideration as to a contemporaneous control, will likely lead to an erroneous result whatever evaluative method is used.

Patient and public involvement

As the informative dataset was simulated there was no patient nor public involvement in this study, nor was consent required for access to patient data.

Author contribution

SH devised the idea for the paper, generated the informative data, conducted the analysis, and led the drafting of the paper. LB and GR provided recommendations on the generation of the data and the analysis in addition to contributing to the drafting of the paper.

Competing interests

The authors have no competing interest to declare

Funding

This article presents independent research by the National Institute for Health Research Collaboration for Leadership in Applied Health Research and Care Yorkshire and Humber (NIHR CLAHRC YH) and the NIHR Applied Research Collaboration Yorkshire and Humber (ARC YH). The views and opinions expressed are those of the authors, and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Data availability statement

The simulated case study data is available upon request to the corresponding author sebastian.hinde@york.ac.uk.

References

1. Health and Social Care Act 2012, avaiable at: http://www.legislation.gov.uk/ukpga/2012/7/contents/enacted (accessed 31/05/2018). **c.7.**
2. Bath Research & Development, *Research, evaluation and evidence: a guide for commissioners.* NHS Research and Development Forum, accessed: 31/-5/2018. **http://www.rdforum.nhs.uk/content/wp-content/uploads/2016/04/GUIDE-Evaluation-and-Research-for-CCGs-Final-version-April-2016.pdf**.
3. The Better Care Fund, *How to... understand and measure impact.* NHS England, 2015. **https://www.england.nhs.uk/wp-content/uploads/2015/06/bcf-user-guide-04.pdf.pdf accessed: 31/05/2018**(4).
4. NHS England, *Evaluation strategy for new care model vanguards.* NHS England, 2016. **https://www.england.nhs.uk/wp-content/uploads/2015/07/ncm-evaluation-strategy-may-2016.pdf accessed: 31/05/2018**.
5. Linden, A., *Conducting interrupted time-series analysis for single- and multiple-group comparisons.* The Stata Journal, 2015. **15, Number 2, pp. 480–500**.
6. Cruz, M., M. Bender, and H. Ombao, *A robust interrupted time series model for analyzing complex health care intervention data.* Statistics in Medicine, 2017. **36**(29): p. 4660-4676.
7. Linden, A., *Conducting interrupted time-series analysis for single- and multiple-group comparisons.* Stata Journal, 2015. **15**(2): p. 480-500.
8. Bernal, J.L., A. Gasparrini, and S. Cummins, *Interrupted time series regression for the evaluation of public health interventions: a tutorial.* International Journal of Epidemiology, 2016. **46**(1): p. 348-355.
9. Linden, A., *Improving causal inference with a doubly robust estimator that combines propensity score stratification and weighting.* J Eval Clin Pract, 2017. **23**(4): p. 697-702.
10. Spiers, G., et al., *Transforming community health services for children and young people who are ill: a quasi-experimental evaluation.* 2016. **4**(25): p. 1-222.
11. Claxton, K., *The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies.* J Health Econ, 1999. **18**(3): p. 341-64.

Figure legends and captions

Figure 1: ITS analytical method

Figure 2: Fabricated time series data

Figure 3: Data availability across the different scenarios of the case study

Figure 1: ITS analytical method

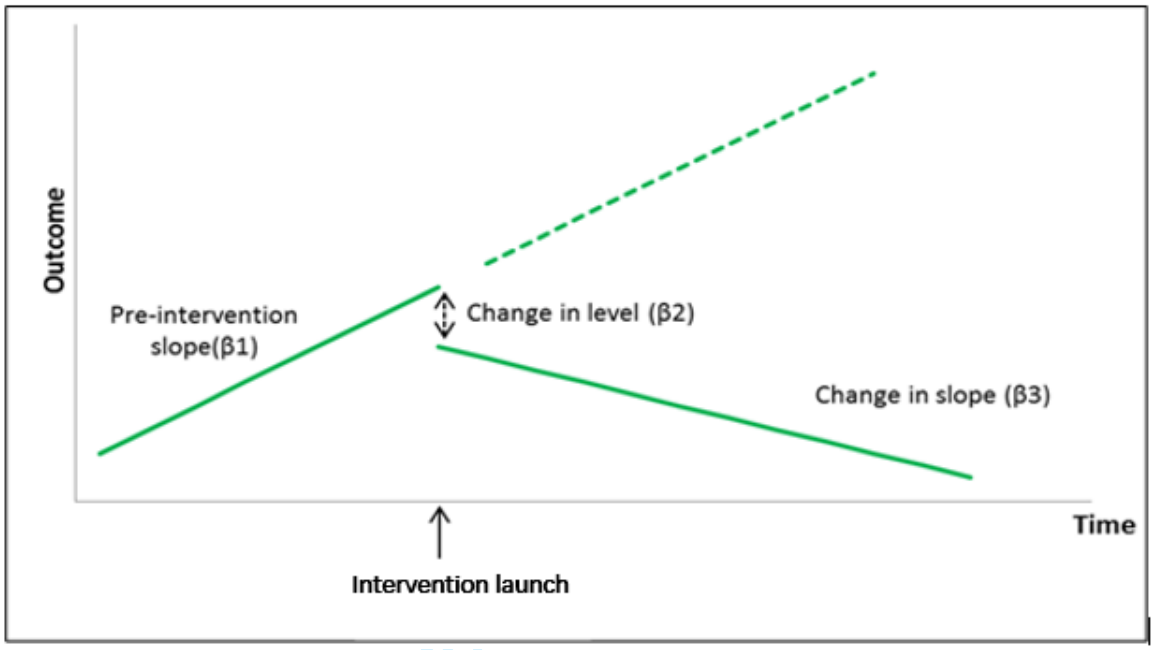


Figure 2: Fabricated time series data

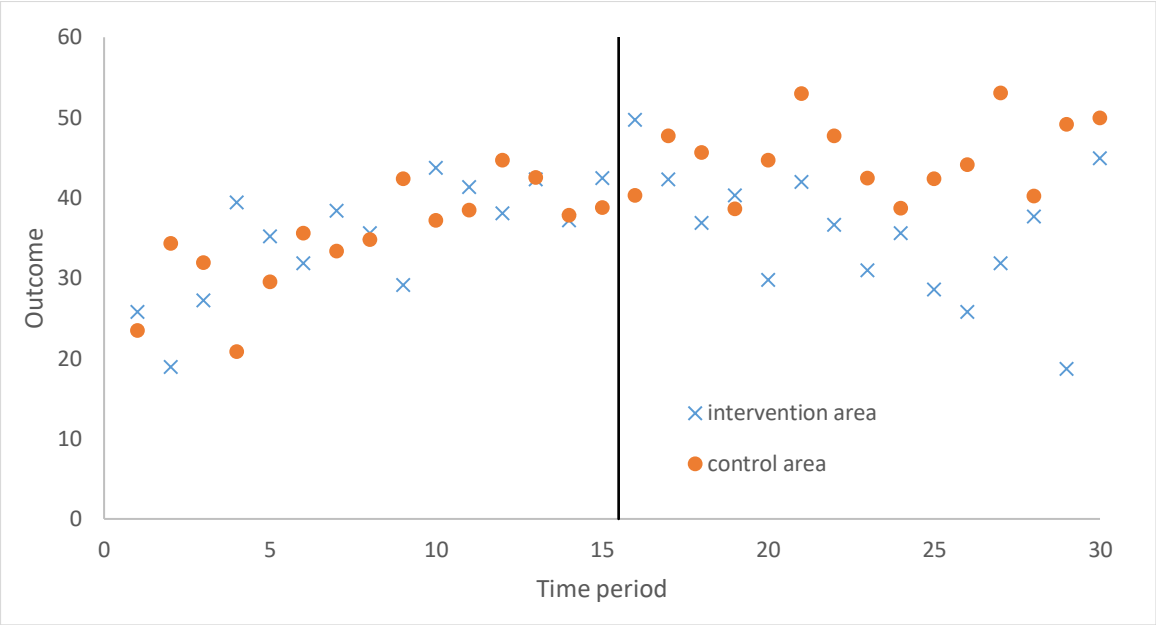


Figure 3: Data availability across the different scenarios of the case study

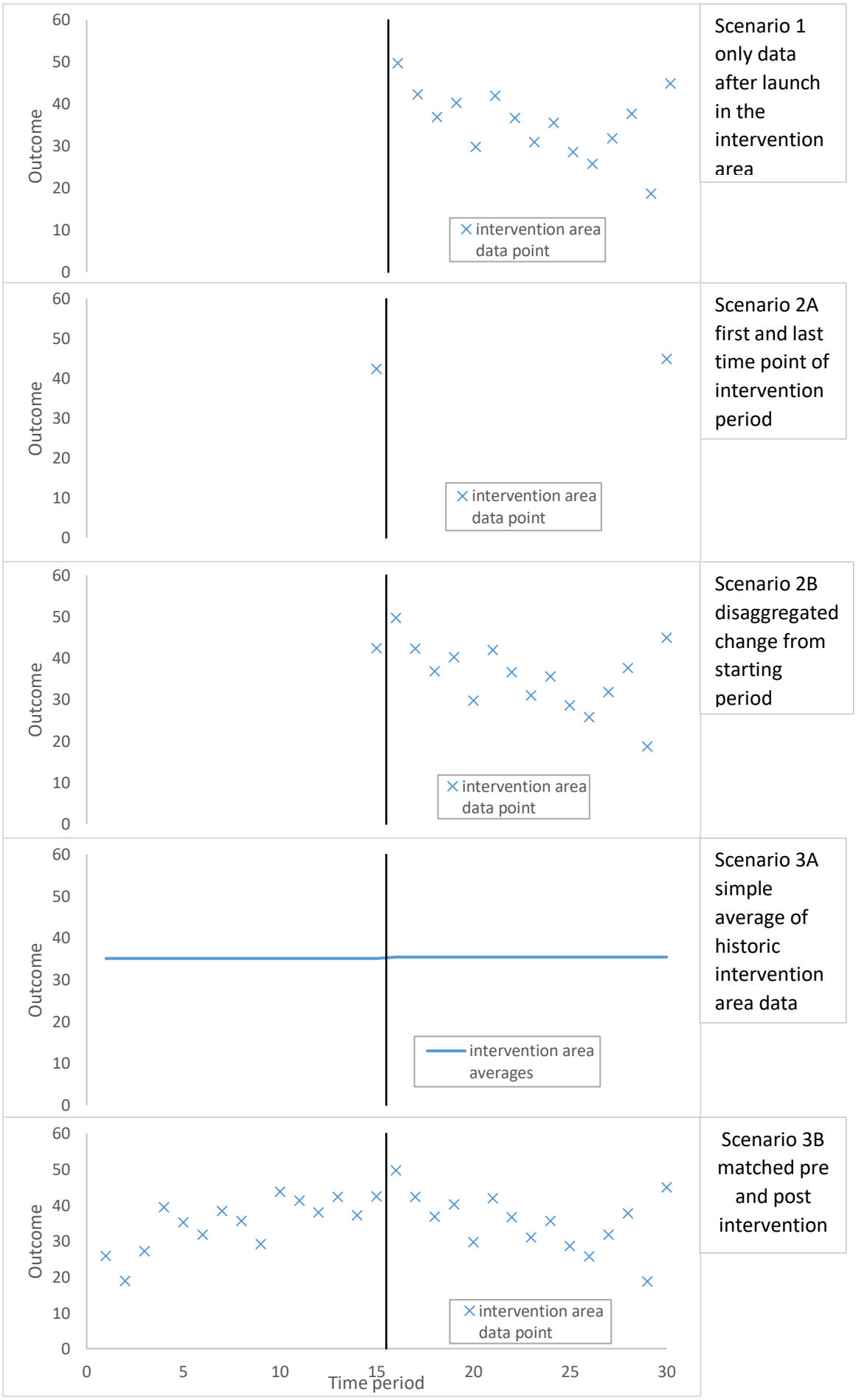
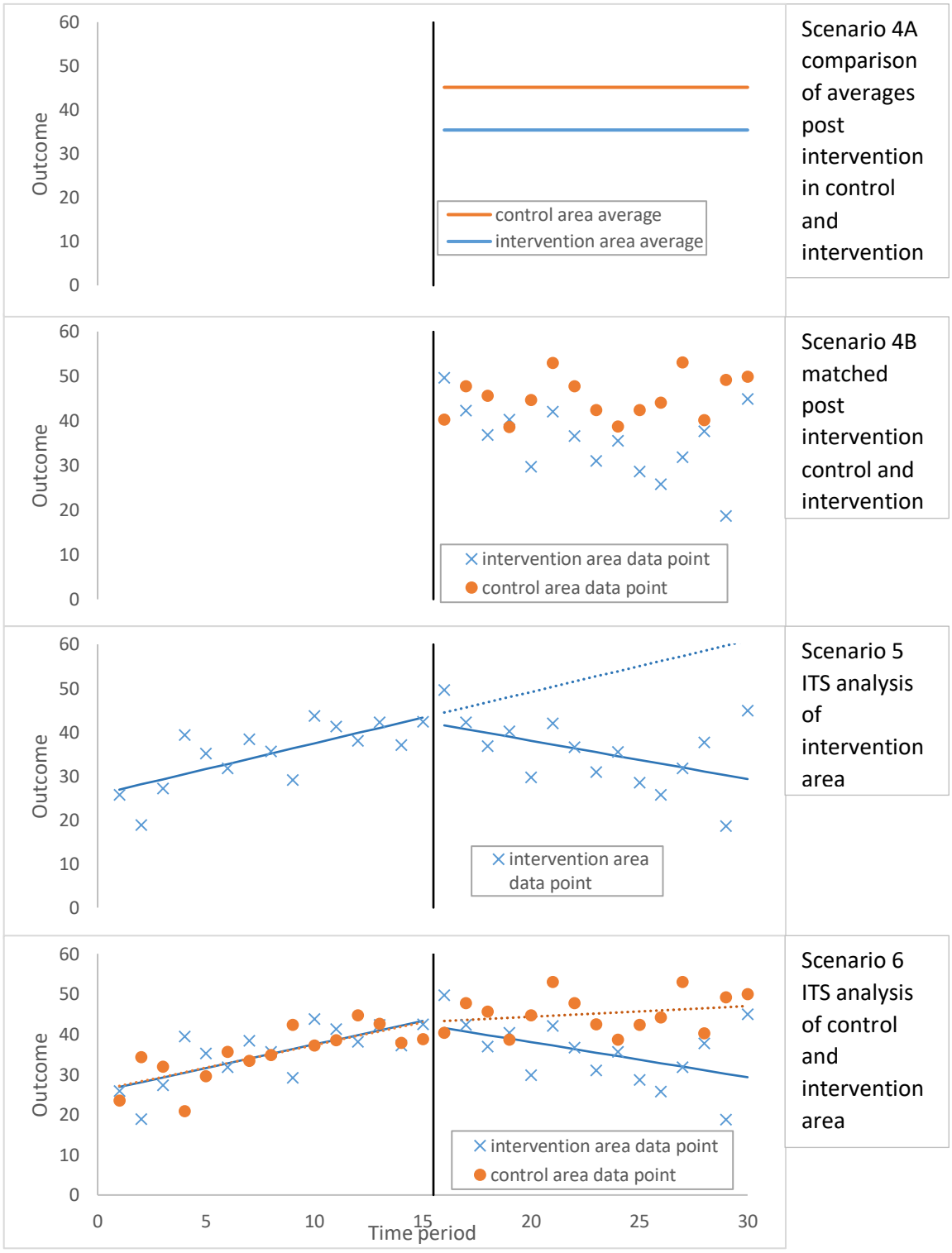


Figure 3: Data availability across the different scenarios of the case study



Supplementary Appendix: Regression output for ITS analysis (scenarios 5 and 6)

This appendix reports the regression outputs for the ITS analysis presented in scenarios 5 and 6 using the ITSA program in Stata.

Regression output for scenario 5

Outcome	Coef.	Newey-West Std. Err.	t	P>t	[95% Conf.	Interval]
β1	1.172109	0.2951369	3.97	0.001	0.565446	1.778772
β2	-2.93635	4.039829	-0.73	0.474	-11.2403	5.367642
β3	-2.04554	0.6483852	-3.15	0.004	-3.37832	-0.71276
β0	26.88192	2.872154	9.36	0	20.97813	32.78572
Treated (β1[_t]+β3[_x_t16])	-0.8734	0.5773	-1.5129	0.1424	-2.0601	0.3133

Regression output for scenario 6

outcome	Coef.	Newey-West Std. Err.	t	P>t	[95% Conf.	Interval]
β1	1.128589	0.2891484	3.9	0	0.548371	1.708808
β4	-0.23341	3.888218	-0.06	0.952	-8.03569	7.568873
β5	0.04352	0.4131738	0.11	0.917	-0.78557	0.872614
β2	-0.76556	3.07017	-0.25	0.804	-6.92631	5.395184
β3	-0.86161	0.3851036	-2.24	0.03	-1.63438	-0.08885
β6	-2.17078	5.074068	-0.43	0.671	-12.3527	8.011078
β7	-1.18393	0.7541274	-1.57	0.122	-2.6972	0.32934
β0	27.11533	2.620871	10.35	0	21.85617	32.37449
Treated (β1[_t]+β5[_z_t] +β3[_x_t16] +β7[_z_x_t16])	-0.8734	0.5773	-1.5129	0.1364	-2.0319	0.285
Controls (β1[_t]+β3[_x_t16])	0.267	0.2544	1.0496	0.2988	-0.2434	0.7774
Difference (β5[_z_t] +β7[_z_x_t16])	-1.1404	0.6309	-1.8077	0.0764	-2.4063	0.1255

BMJ Open

Understanding and Addressing the Challenges of Conducting Quantitative Evaluation at a Local Level: a worked example of the available approaches

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2019-029830.R2
Article Type:	Original research
Date Submitted by the Author:	30-Oct-2019
Complete List of Authors:	Hinde, Sebastian; York University, Centre for Health Economics Bojke, Laura; University of York, Centre for Health Economics Richardson, Gerry; University of York, Centre for Health Economics
Primary Subject Heading:	Research methods
Secondary Subject Heading:	Health services research
Keywords:	Health policy < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Health economics < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, STATISTICS & RESEARCH METHODS

Understanding and Addressing the Challenges of Conducting Quantitative Evaluation at a Local Level: a worked example of the available approaches

Sebastian Hinde, Laura Bojke, Gerry Richardson

Centre for Health Economics, University of York

Corresponding author:
Sebastian Hinde,
Centre for Health Economics,
Alcuin 'A' Block,
University of York,
Heslington,
North Yorkshire,
YO10 5DD.
United Kingdom
Sebastian.Hinde@york.ac.uk
+44 (0)1904 321455

SH developed the research idea, led the writing of the manuscript, and acts as the guarantor of the article. LB and GR gave input at all stages including commenting on the manuscript. All authors are health economists at the Centre for Health Economics, University of York, with experience in working on evaluation for local decision makers at various levels. The simulated dataset and all analytical code is available as supplementary appendices, see "Fabricated data scenarios.xlsm", "Paper analysis.do", and "Fabricated data.dta". Please contact the corresponding author sebastian.hinde@york.ac.uk for full access. Due to the simulated nature of the dataset no patients were involved in the study.

This article presents independent research by the National Institute for Health Research Collaboration for Leadership in Applied Health Research and Care Yorkshire and Humber (NIHR CLAHRC YH) and the NIHR Applied Research Collaboration Yorkshire and Humber (ARC YH). The views and opinions expressed are those of the authors, and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Abstract

Objectives

In the context of tightening fiscal budgets and increased commissioning responsibility, local decision makers across the UK healthcare sector have found themselves in charge of the implementation and evaluation of a greater range of healthcare interventions and services. However, there is often little experience, guidance, or funding available at a local level to ensure robust evaluations are conducted. In this paper, we evaluate the possible scenarios that could occur when seeking to conduct a quantitative evaluation of a new intervention, specifically with regards to availability of evidence.

Design

We outline the full set of possible data scenarios that could occur if the decision maker seeks to explore the impact of the launch of a new intervention on some relevant quantifiable outcome. In each case we consider the implicit assumptions associated with conducting an evaluation, exploring possible situations where such scenarios may occur. We go on to apply the scenarios to a simulated dataset to explore how each scenario can result in different conclusions as to the effectiveness of the new intervention.

Results

We demonstrate that, across the full set of scenarios, differences in the scale of the estimated effectiveness of a new intervention and even the direction of effect, are possible given different data availability and analytical approaches.

Conclusions

When conducting quantitative evaluations of new interventions, the availability of data on the outcome of interest and the analytical approach can have profound effects on the conclusions of the evaluation. While it will not always be possible to obtain a complete set of data and conduct extensive analysis, it is vital to understand the implications of the data used and consider the implicit assumptions made through its use.

Strengths and limitations of this study

- Highlights the risks of partial analysis of time series data used to evaluate the impact of a service
- Presents the assumptions implicitly made through the differential use of data to inform quantitative evaluation in a range of scenarios
- Demonstrates that even a well-designed analysis is constrained by the available data
- Provides guidance aimed at local decision makers, who are typically overlooked in the published methodological guidance
- The use of simulated data allows for a clear demonstration of the scenarios but risks oversimplifying the nature of “real world” data

Introduction

Clinical Commissioning Groups (CCGs), Local Authorities, and other local decision makers are under increasing pressure to demonstrate the value of any newly commissioned activities given tightening fiscal budgets. While the Health and Social Care Act of 2012[1] was instrumental in allowing local decision makers to be responsive to the health needs of the population they serve, it provided little guidance on how to do so in an effective and cost-effective manner. As a result, local decision makers have found themselves caught between two worlds, neither being served by national evidence generation due to the decentralisation of funding, nor with the ability, finance, or structure to generate robust evidence, such as randomised trials.

Whilst collaborations between the Local Government Association, Department of Health, NHS England, and others have led to a number of guides for good evaluation and evidence generation, [2-4] these have had a broad focus on the theory of good research, rather than offering practical advice for analyses.

While in some cases, such as the Vanguard projects, [4] funding has been ring fenced for evaluation, it is more common that the decision to conduct a service evaluation by local decision makers comes at the detriment of the service provision itself. As a consequence, any evaluation may be limited in scope, and the ability to fund sufficiently robust data collection severely compromised. While there are inevitably risks of funding services based on inadequate evidence, as we will go on to demonstrate, there is little logic in funding sophisticated studies that threaten provision of the service itself.

It has been the experience of these authors (GR is the University of York representative on York Teaching Hospital's Council of Governors; GR and LB are members of the Vale of York CCG's Research Group; and GR, LB and SH have experience in evaluating a number of local interventions including the Harrogate and District CCG's Vanguard programme, a Core-24 hour mental health liaison service, and Tier 3 weight loss services) that these factors have resulted in either no quantitative evaluation of new service provision or evaluations that are based on limited interpretations of outcome measures and incomplete data collection. This is despite the move towards monitoring of services, both for quality and financial reasons, and falls in the cost of data generation, which have meant its collection and use is no longer an insurmountable barrier to evaluation.

In this paper, we explore a range of different scenarios faced by a local decision maker depending on the availability of data and analytical approach taken. We go on to use a stylised case study to explore the implications of each scenario on the estimated impact of the intervention and the likely

conclusions. We focus on a quantitative evaluation but highlight the importance of a mixed-method approach in achieving a robust evaluation.

We take as a starting point a decision maker who is seeking to evaluate a new intervention, where *intervention* is used to describe any new or change in service, care pathway or treatment. They possess time series data on an outcome of interest over a series of time-points, which is hypothesised to be impacted by the new intervention. These data may be at an aggregated level (e.g. local population) or data for individuals (e.g. patients or households). Such a generalised situation is common, with the decision maker being anything from CCGs, Local Authorities, to mental health providers. While the possible set of outcomes of interest is wide, the need to generalise findings often results in focus being on broad process outcomes such as non-elective attendances, and length of stay, which are easily benchmarked. Such an analysis is expected to play a role in a decision making process informed by a number of other quantitative and qualitative considerations.

The Different Scenarios

In this section, we consider the full set of data scenarios and analytical approaches that may occur when seeking to evaluate the impact of the launch of a new intervention on a single outcome of interest. We explore the range of implicit assumptions that are made for each of the scenarios, and possible examples of how each may occur. The different cases are characterised as six overarching scenarios. It is the experience of these authors that it is most common for evaluation of an intervention to be done retrospectively or towards the end of a project, primarily due to a lack of evaluative experience and funding to embed evaluation from an early stage; however, there is a lack of reviews of the methodology applied by local decision makers in such setting.

Scenario 1 – follow-up data but no pre-launch data for the intervention area

In its simplest form an evaluation may consist of only data collected after the launch of an intervention with no historical evidence, for example if the intervention was unplanned and data could not be collected retrospectively, such as a piece of hospital infrastructure being replaced. Such an analysis can therefore only comment on the trajectory of the data over the intervention period as there is no knowledge of the *counterfactual* (what would have happened had the intervention not occurred), and no data on which to base any estimation. If any estimation of the total impact of the intervention is required, assumptions or external evidence would be required to inform the counterfactual.

Scenario 2 – follow-up data and a single pre-launch data point for the intervention area

Secondly, we consider a situation where the decision maker has only historic data for the final period before the launch of an intervention. Such a situation may occur when the decision to conduct an evaluation occurs only a short time before the launch and data cannot be collected retrospectively. Depending on the aggregation and availability of data two sub-scenarios are available:

A. Data are only available for the last period before launch and a single time point of the post-launch time series, a simple before and after statement is possible. In all cases, some implicit or explicit statement is beneficial regarding the generalisability of the observed data and trends in the data over the intervening time-period. Such as case would occur if data were only available at set time points and only informative of a short time period, for example annually occurring surveys or audits.

B. Data are available for the last period before launch and all post-intervention time points, allowing an average change over the period from the first time point to be calculated with some additional knowledge of how the data changed over the period. This might occur if repeated data collection is possible prospectively, such as the collection of electronic patient data once relevant patients have been identified and consented.

Given the limited pre-launch data available in this scenario, we must assume that, had the intervention not been launched, the outcome would have stayed at the same level as in the last time point before launch. While this assumption is inevitable if no other data are available, it risks being misleading if there is some underlying trend in the outcome, or if it is subject to natural variation from one time point to the next.

Scenario 3 – data are available covering the full pre and post-launch period for the intervention area

To overcome the limitations of scenario 2, historic data in the intervention area can be used to inform the baseline value and any underlying trends in the outcome over time by relaxing the assumption that outcome data would have remained static. As with scenario 2, alternative aggregation of the historic data can result in different implications:

A. Both pre- and post-launch, may only be available as average values aggregated over a long period, for example if the data access is limited to annual audit figures that cover the entire pre-launch period. This scenario implies that no consideration of the disaggregated trends are possible.

B. Extensive disaggregated data are available both before and after the launch. This allows for the direct comparison of each post-launch time-period with some matched period in the pre-intervention data, for example comparing January in one year with January in the next. The matching is used to conduct the analysis at a more disaggregated level, as well as adjusting for other factors such as seasonality and budgetary cycles. While the average estimate of the impact of the intervention launch will be the same as part A, we now have the ability to investigate the change in trend over the time-period. Such a case would occur either when an evaluation and data collection was started some time before the intervention launch, or when data on the outcome is readily available retrospectively. For example, if the evaluation is concerned with emergency department attendances over time, historic data can typically be retrospectively collected.

Scenario 4 – data are available on a control area post-launch as well as the intervention area data

Scenarios 1-3 describe when data are only available for the area covered by the intervention. However, data are often available for comparator areas as the informative outcome is often routinely collected and available across multiple areas, through systems such as Hospital Episode Statistics (HES), or collection can be prospectively arranged. Such comparator areas can be local, regional, national, or a synthetic comparator created by combining a number of areas. To be an informative comparator the area must represent a good match to the intervention area in all relevant characteristics and not be impacted by the launch of the new service being evaluated.[5] The goodness of the match can be determined qualitatively or quantitatively by comparing the known features of the two areas.

The most common use of such control data is to directly compare the post intervention outcomes in the two areas, using the same approach as scenario 3, but with the contemporary control data are used instead of the historic intervention area data. As before, there are two categories:

A. Only aggregate data are available post intervention launch for the two areas. As in previous scenarios, an example of this would be analyses based on audit data alone but now across multiple areas.

B. Disaggregated data are available post intervention, allowing a disaggregated matched comparison can be made which again, results in the same total estimated impact as part A but gives us an understanding of the respective trends. This situation would occur where an intervention is only launched in one part of a larger geographic area or patient group where the decision makers has access to the data of the full set prospectively, for example one GP practice in a CCG area.

Under this scenario, comparator area data are used either instead of, or due to a lack of, historic evidence as used in scenario 3. Using simple analytical techniques there is no way to incorporate both, which we will explore in scenario 6. There is no definitive rule for whether historic or contemporary comparator evidence is more appropriate, it is situation dependent. For example, if the intervention of interest was not the only change at the point of launch of the intervention, the control area data would likely be most appropriate if the second new service was launched in both areas, but not if it were only in the control area. A number of other factors must be considered, for example, what if comparator data are available but is not a good match, how does one define a suitable match, and what if there are multiple comparators potentially telling different stories?

Scenario 5 – all pre and post-launch data are available for the intervention area

In this scenario and scenario 6 we explore the addition of more advanced analytical approaches to the analysis of the data, specifically the use of interrupted time series (ITS) or 'segmented regression' analysis. This approach has been well presented in the literature,[6-8] but in brief, the method considers the trend in an outcome of interest over time, segmenting it into the period before the intervention was launched, and after it. The example of using pre- and post-launch data for the intervention area is shown in the explanatory **Error! Reference source not found.**, where the pre-launch data are used to infer a post-launch counterfactual case, with the nature of the change in the outcome define a-priori. Using the framework described by Bernal[8] it is possible to define the regression model using the equation detailed below, where Y is the aggregated outcome, β represents the relevant coefficients, T the time since the start of the study, t the specific time-point, X is a dummy variable indicating when the new intervention is active, and ε the error term.

$$Y_t = \beta_0 + \beta_1 T_t + \beta_2 X_t + \beta_3 X_t T_t + \varepsilon_t$$

The application of such a regression model allows for the formal estimation of whether any change in the outcome of interest is statistically significant under a frequentist framework, and for any change to be quantified by estimating the area between the two regression lines, shown in **Error! Reference source not found.**, over the analysis period.

The use of such method requires time series data both before and after the launch in the intervention area, as in scenario 3B.

[Figure 1 here]

Scenario 6 - data are available on both control and intervention areas pre- and post-launch

We demonstrated in scenario 4 that the addition of control area data typically implied the exclusion of historic intervention area data in informing the counterfactual. Using ITS methodology it is possible to formally incorporate comparator data, potentially from multiple areas or a synthetic area, alongside the full set of intervention area data. The method uses the pre-intervention data to formally test whether the comparator areas can be considered a good match. If so, the post-launch comparator data are then used to infer the post-launch counterfactual of the intervention area. Therefore, this approach assumes that the control area is indicative of what would have happened to the outcome in the intervention area had the launch not occurred, much as we assumed in scenario 4 but with a formal assessment of the trend and reliability of the comparator. The equation detailed in scenario 5 can be extended by incorporating a Z term as a dummy for assignment to the treatment or control population, as detailed by Linden[7]:

$$Y_t = \beta_0 + \beta_1 T_t + \beta_2 X_t + \beta_3 X_t T_t + \beta_4 Z + \beta_5 Z T_t + \beta_6 Z X_t + \beta_7 Z X_t T_t + \varepsilon_t$$

Comparing the Scenarios

Each of the scenarios outlined above is characterised by a set of core assumptions, made implicitly or explicitly, if used to evaluate the impact of a new intervention on some outcome of interest. Similarly, the variability in the ease of implementation, and data and analytical requirements of each scenario implies a range of pros and cons associated with each. These are presented in Table 1, which highlights that the more analytically simple and data light the scenario the stronger the core assumption required about the nature of the interaction with the outcome and time trends in the data.

Table 1: Summary of the different analytical methods

Method	Core assumptions	Pros	Cons
Scenario 1, only data after launch in the intervention area	Only the change in the data after the launch is relevant to the evaluation.	Requires little data or technical knowledge.	Unable to comment on the change in the outcome of interest because of the intervention, only its trend after launch.
Scenario 2A, first and last time point of intervention period	The two data points are fully indicative of the change.	Requires little data or technical knowledge.	Highly dependent on small array of data. Risks loss of important details of data, intervention effect, or trends.
Scenario 2B, disaggregated change from starting period	Last pre-intervention period fully represents the counterfactual.	Only requires one pre-intervention data point. Analytically simple.	Highly dependent on small array of control data. No consideration of trend in counterfactual.
Scenario 3A, simple average of historic intervention area data	Simple averaging of before and after data incorporates all factors, there is no value in an assessment of the trends.	Only requires small amount of pre and post data. Analytically simple.	Fails to explore trends in data.
Scenario 3B, matched pre and post intervention	There is a repeating periodic fluctuation, e.g. seasonality, that impacts the outcome of interest and the trend over time is informative.	Simple means of adjusting for periodic fluctuations.	Result varies given matching approach. Blunt means of adjusting for periodic fluctuations that can result in incorrect estimates.
Scenario 4A, comparison of averages post intervention in control and intervention areas	The selected control area fully represents the counterfactual of the intervention area.	Allows for use of control area data. Only requires post-launch data.	Fails to explore trends in data. Makes no use of historic data. Difficult to determine if the control area represents a reasonable comparator.
Scenario 4B, matched post intervention control and intervention area	The selected control area fully represents the counterfactual of the intervention area and the trend over time is informative.	Allows for use of control area data. Explores trends in data without having to define a cycle length. Only requires post-launch data.	Makes no use of historic data. Difficult to determine if the control area represents a reasonable comparator.
Scenario 5, ITS analysis of intervention area	Regression of pre-intervention data fully represents post-intervention counterfactual and the trend over time is informative.	Allows for use of historic control data. Explores the trends.	Reliant on historic intervention area data being predictive of counterfactual.
Scenario 6, ITS analysis of control and intervention area	Control area fully represents counterfactual of intervention area but the match can be tested by exploring the pre-intervention data. The trend over time is informative.	Allows for use of control area and exploration as to the closeness of the control and intervention areas.	Assumption that the control area continues to represent a good match after the intervention period.

Case study

To explore the practical implications of the different scenarios, and demonstrate the potential for varied conclusions, we have created a case study to which each is applied. To inform the case study a time series dataset of an outcome unit of interest (e.g. bed days, hospital admissions, or indicators of quality and care outcome) has been simulated. The data values and number of time points has been selected to best inform the characteristics of each of the scenarios described in Table 1 while representing the uncertain nature of real world data relevant to this setting.

This data relates to two distinct groups (intervention and control) and a maximum of 30 observations are available over some defined time period at regular intervals (e.g. every week, month, or year). The data are structured such that in both areas the outcome was increasing for the first 15 observations at a rate of 4/3 per time period from a mean value of 20 units at time 1, after which point the intervention is implemented in the intervention area but not the control. From time point 15 onwards in the intervention area the outcome decreases at the same rate of 4/3 units per period, while in the control area the outcome levels off, assumed to be due to factors unrelated to the intervention. All time points are subject to some level of variation to mimic what is observed in real world data, simulated using a normal distribution (mean 45 and standard deviation 5). We assume that after launch (t=15) the new service becomes fully operational, with no run in period. The last time point in the intervention area (t=30) was set as an extreme outlier (estimated as occurring with a probability of 0.99999 on the simulated distribution) to explore its impact on the results, for example if an exogenous factor affected the intervention such as failure of a key piece of machinery. **Error! Reference source not found.** shows the fabricated data in full, with each data point representing the time period before, such that data point 1 being the total outcome over times 0 to 1.

[Figure 2 here]

Using the informative structure of the simulated case study it is possible to estimate two possible underlying effect values. If the control area is the best indicator of the counterfactual the intervention resulted in a reduction of 151 units over the period, if the historic intervention area is best, a reduction of 324 units. While these values can help us to understand the results of the different scenarios they must be interpreted with caution; as while they inform the underlying trend used to simulate the data the case study time points were sampled independently.

In the next part we explore what the data availability would look like under each of the scenarios outlined in the previous section, estimating what the impact and conclusions would be regarding the effectiveness of the intervention. As outlined earlier, in many of the cases only a limited set of the data are available, indeed it is only scenarios 4 and 6 where the full dataset is available to the decision maker. **Error! Reference source not found.** and 4 provides an overview of the data availability across all of the scenarios.

[Figure 3 here]

[Figure 4 here]

Table 2 gives an overview of the results of the different possible scenarios and possible interpretations.

Table 2: Summary of the different scenarios results

Scenario	Possible interpretation of the result	Estimated change[1]
Scenario 1, only data after launch in the intervention area	The outcome of interest appears to have decreased over the post-launch time-period	not possible to estimate a change in the outcome
Scenario 2A, first and last time point of intervention period	There appears to have been an increase in the outcome from the pre-launch to post-launch period. Extrapolating the observed values over the entire 15 months of intervention suggests that the new intervention had increased the outcome by 37.6 units ((44.9-42.4)x15)	37.6
Scenario 2B, disaggregated change from starting period	The outcome of interest appears to have decreased over time from the pre-launch time-period, with an estimated change of -120.1 units over the period ((34.4-42.4)x15)	-120.1
Scenario 3A, simple average of historic intervention area data	There appears to have been little change from the pre- to post-launch periods in the outcome, with the average value going from 35.1 to 35.4 ((35.4-35.1)x15)	4.9
Scenario 3B, matched pre and post intervention	There appears to have been little change from the pre- to post-launch periods in the outcome, with the average value going from 35.1 to 35.4. However, it appears from the data that there was an increasing trend in the outcome before the intervention and a decreasing trend afterwards ((35.4-35.1)x15)	4.9
Scenario 4A, comparison of averages post intervention in control and intervention areas	Compared to the control area the intervention area had a lower average level of the outcome after the launch of the intervention	-146.0
Scenario 4B, matched post intervention control and intervention area	Compared to the control area the intervention area had a lower average level of the outcome after the launch of the intervention. The control area appeared to have a flat trend in the outcome over the post-launch period compared to a decreasing trend in the intervention area ((35.4-45.1)x15)	-146.0
Scenario 5, ITS analysis of intervention area	Compared to the pre-launch intervention area the post-launch saw a decrease in the trend over time in the outcome, from positive to negative, which was statistically significant See the Supplementary Appendix for regression	-258.8
Scenario 6, ITS analysis of control and intervention area	Both control and intervention areas saw a shallowing of the trend over time. The intervention area saw a greater decrease in the trend, being negative compared to the relatively flat trend in the control. This different was	-146.0

	statistically significant. The control area was found to be a match to the intervention area in the pre-launch period (the regressions lines are aligned), See the Supplementary Appendix for regression	

[1]negative values indicate that the new service reduced the outcome

Error! Reference source not found., 4, and Table 2 demonstrate the large potential for variation in the estimated impact of the intervention, and the overall conclusions that could be drawn given the different scenarios. Estimations of the change in the outcome vary from predicting the intervention increased the outcome by 37.6 units over the post-intervention period (scenario 2A), to decreasing it by 258.8 (scenario 5). Similarly the interpretations differ in their ability to identify the trends in the different areas and time periods, as well as the overall impact of the intervention.

In the case study presented here, with full access to the data and knowledge of the underlying trends in the simulated data, it is clear that several of these scenarios result is a very incorrect conclusion. However, the appropriateness of the scenarios and accuracy of their conclusions compared to any 'true' effects are clearly much harder to determine in the real world.

Discussion

In this paper we have explored a range of possible scenarios and analytical approaches available to a decision maker when evaluating the impact of a new intervention on an outcome of interest, highlighting the implicit assumptions made in each. Through our simulated case study we have demonstrated how these scenarios can yield very different estimates of effectiveness.

Comparison of the methods explored here suggests that it is intuitively appealing to conclude that the approach outlined in scenario 6, using the ITS methodology including the control area comparison, is the most accurate as it incorporates the most complete set of data whilst taking the most complete approach to statistical analysis. However, the most appropriate methodology may be driven by other factors, primarily the availability of informative data and the validity of the core assumptions detailed in Table 1.

Furthermore, the use of ITS analysis (scenarios 5 and 6) is not without assumptions, primarily relating to the suitability of the historic and control area data to inform the counterfactual, and the functional form of the trends modelled. It also requires a significant level of data and analytical ability to implement. However, the inability to observe exactly what would happen in the intervention area without the new service, necessitates such assumptions in order to estimate the impact of its launch. Fears about the robustness of such assumptions are likely to be best addressed by the identification of additional relevant evidence to either adjust the existing data or inform a

new comparator. For example, methods are available to overcome concerns over additional service changes in the time period covered by the data,[7] to incorporate multiple control areas,[7] and to conduct a more rigorous selection of control area through matching.[9]

As with all such analyses, the ITS methodology can be extended to consider the significance of the findings beyond pure chance. This can be achieved through a frequentist framework, considering the statistical significance of the regression estimates, as discussed in Linden et al.,[7] or through a Bayesian framework.[10] Such considerations should play an important role in the decision making process, as a single estimate of the impact on an intervention can be misleading. Specifically, it fails to take account of the uncertainty, of the informative data or the consequences of making an incorrect funding decision. However, it is important to reflect that even if there is substantial uncertainty it is the expected impact of the intervention that should be most informative to the commissioning decision, rather than the significance of the impact, [11].

An intrinsic element to any analyses explored in this paper is an understanding of the data under interrogation: the application of robust methods is only helpful if the data being used is consistent and relevant to the question being addressed. Prior to any analysis it is important to understand the data, answering questions such as: how was it generated; is an estimate of the rate of an event more relevant than its frequency; is it consistent over the time period of interest; what is the route of causality between the intervention of interest and the data; and when plotted do there appear to be any unexplainable outliers?

Analyses such as those presented here are most robust when combined with qualitative methodologies through a mixed-method approach, with the qualitative findings ideally facilitating a more detailed understanding of the trends seen in the data and informing the suitability of the different counterfactual scenarios. Such a mixed-methods approach may extend the quantitative incorporate health economic considerations, such that the generalisable cost-effectiveness of the intervention is considered.

Furthermore, the use of robust methodologies, such as ITS analysis, does not replace the need for the robust selection of outcomes and data collection, as any analysis can only be as robust as the data that informs it. Failure to prospectively design the launch of an intervention and associated evaluation to ensure , the required level of data collection, and sufficient consideration of a contemporaneous control, will likely lead to an erroneous result whatever evaluative method is used.

Patient and public involvement

As the informative dataset was simulated there was no patient nor public involvement in this study, nor was consent required for access to patient data.

Author contribution

SH devised the idea for the paper, generated the informative data, conducted the analysis, and led the drafting of the paper. LB and GR provided recommendations on the generation of the data and the analysis, in addition to contributing to the drafting of the paper.

Competing interests

The authors have no competing interest to declare

Funding

This article presents independent research by the National Institute for Health Research Collaboration for Leadership in Applied Health Research and Care Yorkshire and Humber (NIHR CLAHRC YH) and the NIHR Applied Research Collaboration Yorkshire and Humber (ARC YH). The views and opinions expressed are those of the authors, and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Data availability statement

The simulated case study data is available as supplementary appendices. Please contact the corresponding author sebastian.hinde@york.ac.uk for full access.

References

1. Health and Social Care Act 2012, avaiable at: http://www.legislation.gov.uk/ukpga/2012/7/contents/enacted (accessed 31/05/2018). **c.7.**
2. Bath Research & Development, *Research, evaluation and evidence: a guide for commissioners.* NHS Research and Development Forum, accessed: 31/-5/2018. **http://www.rdforum.nhs.uk/content/wp-content/uploads/2016/04/GUIDE-Evaluation-and-Research-for-CCGs-Final-version-April-2016.pdf**.
3. The Better Care Fund, *How to... understand and measure impact.* NHS England, 2015. **https://www.england.nhs.uk/wp-content/uploads/2015/06/bcf-user-guide-04.pdf.pdf accessed: 31/05/2018**(4).
4. NHS England, *Evaluation strategy for new care model vanguards.* NHS England, 2016. **https://www.england.nhs.uk/wp-content/uploads/2015/07/ncm-evaluation-strategy-may-2016.pdf accessed: 31/05/2018**.
5. Linden, A., *Conducting interrupted time-series analysis for single- and multiple-group comparisons.* The Stata Journal, 2015. **15, Number 2, pp. 480–500**.
6. Cruz, M., M. Bender, and H. Ombao, *A robust interrupted time series model for analyzing complex health care intervention data.* Statistics in Medicine, 2017. **36**(29): p. 4660-4676.
7. Linden, A., *Conducting interrupted time-series analysis for single- and multiple-group comparisons.* Stata Journal, 2015. **15**(2): p. 480-500.
8. Bernal, J.L., A. Gasparrini, and S. Cummins, *Interrupted time series regression for the evaluation of public health interventions: a tutorial.* International Journal of Epidemiology, 2016. **46**(1): p. 348-355.
9. Linden, A., *Improving causal inference with a doubly robust estimator that combines propensity score stratification and weighting.* J Eval Clin Pract, 2017. **23**(4): p. 697-702.
10. Spiers, G., et al., *Transforming community health services for children and young people who are ill: a quasi-experimental evaluation.* 2016. **4**(25): p. 1-222.
11. Claxton, K., *The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies.* J Health Econ, 1999. **18**(3): p. 341-64.

Figure legends and captions

Figure 1: ITS analytical method

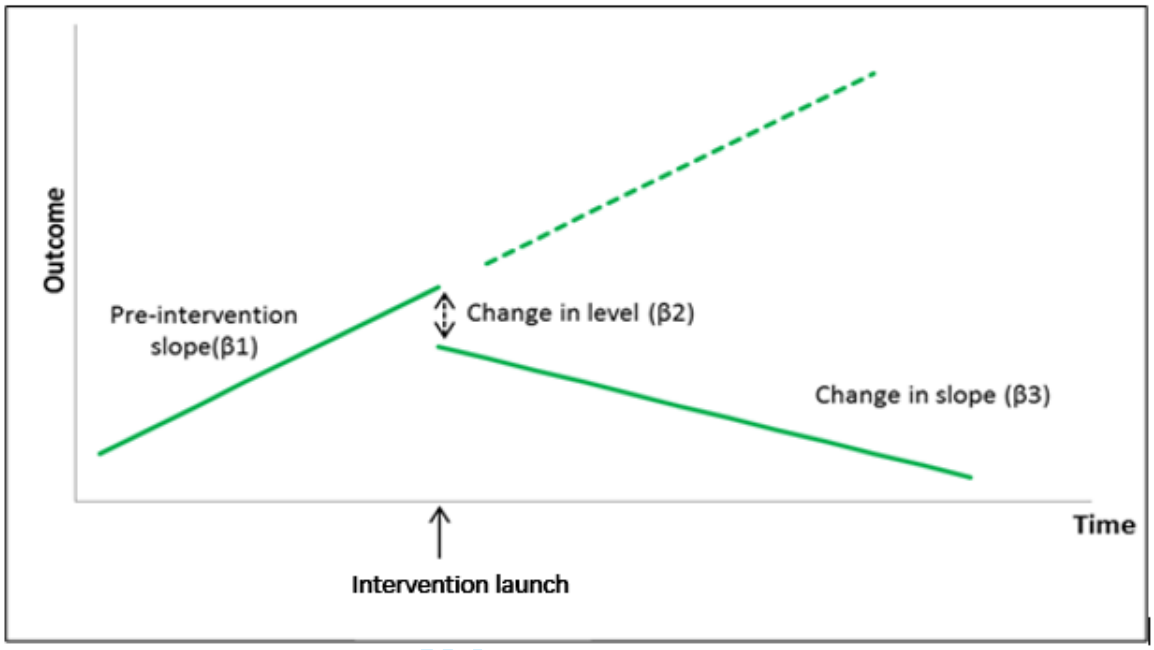


Figure 2: Fabricated time series data

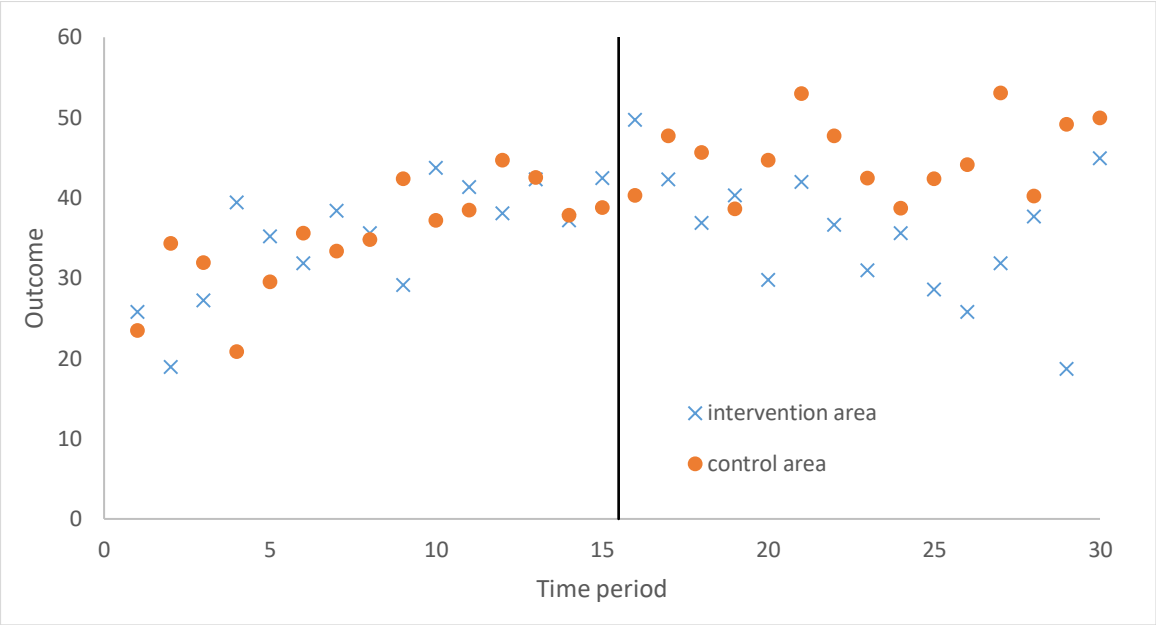


Figure 3: Data availability across the different scenarios of the case study, scenarios 1-3

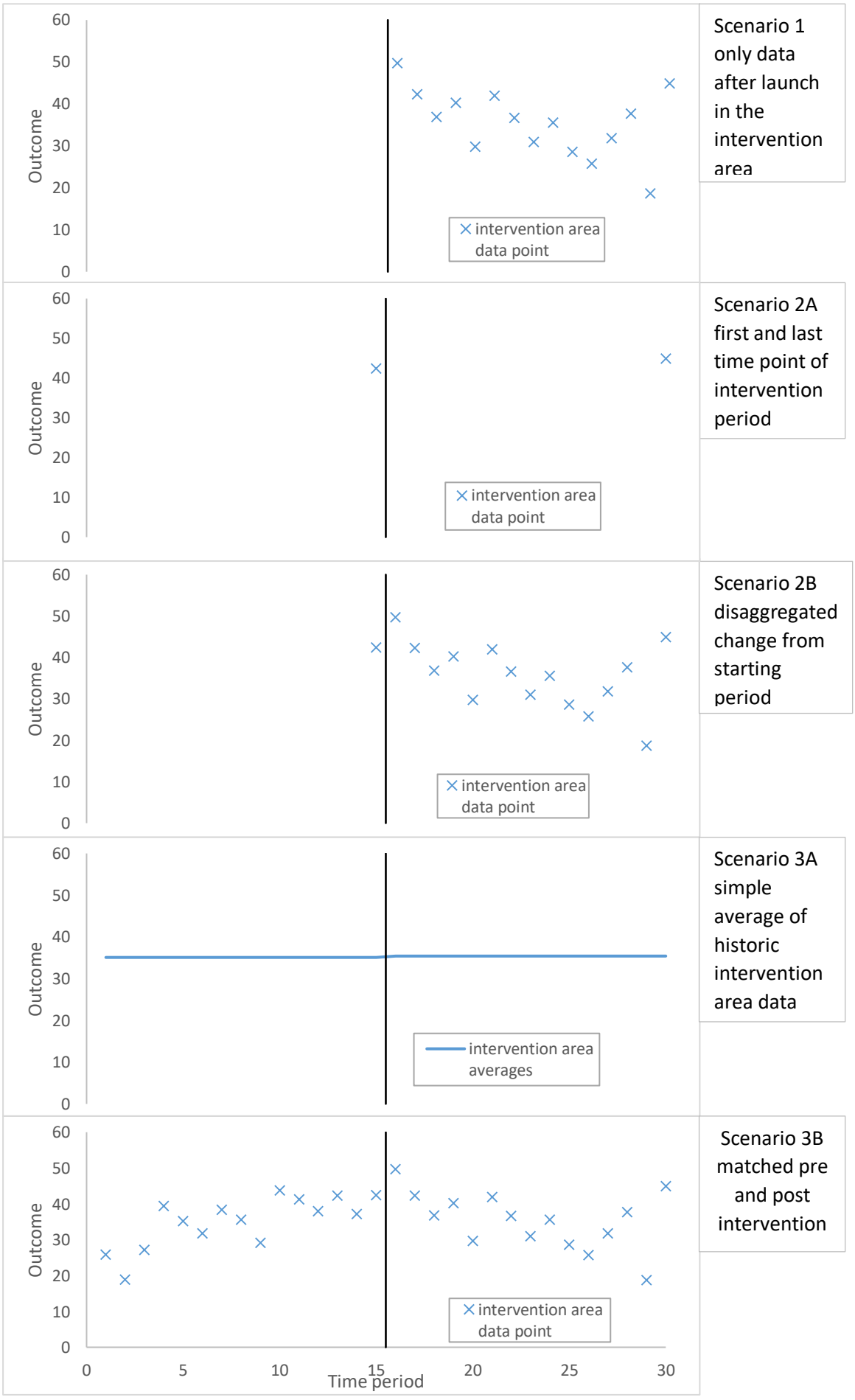
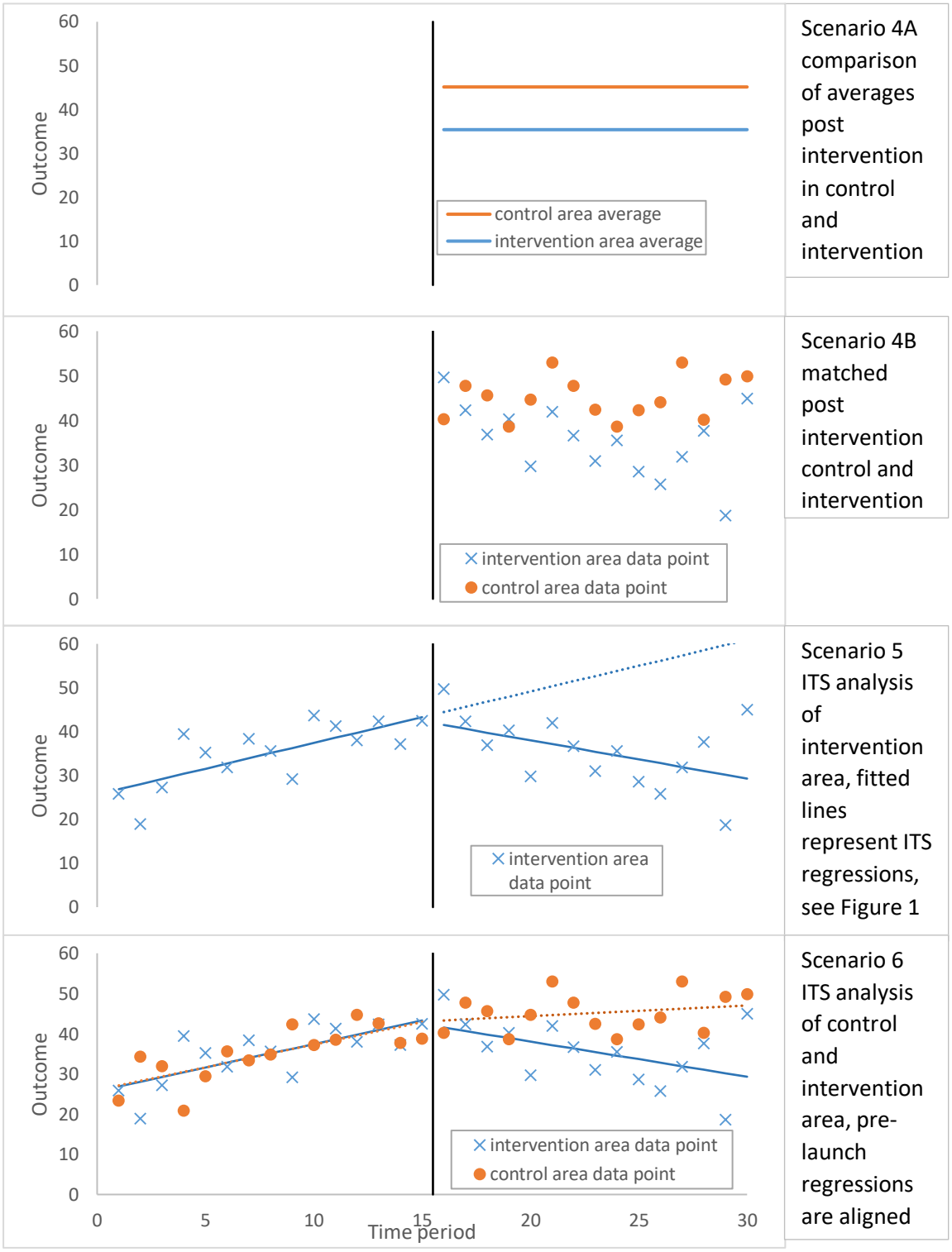


Figure 4: Data availability across the different scenarios of the case study, scenarios 4-6



Supplementary Appendix: Regression output for ITS analysis (scenarios 5 and 6)

This appendix reports the regression outputs for the ITS analysis presented in scenarios 5 and 6 using the ITSA program in Stata.

Regression output for scenario 5

Outcome	Coef.	Newey-West Std. Err.	t	P>t	[95% Conf.	Interval]
β1	1.172109	0.2951369	3.97	0.001	0.565446	1.778772
β2	-2.93635	4.039829	-0.73	0.474	-11.2403	5.367642
β3	-2.04554	0.6483852	-3.15	0.004	-3.37832	-0.71276
β0	26.88192	2.872154	9.36	0	20.97813	32.78572
Treated (β1[_t]+β3[_x_t16])	-0.8734	0.5773	-1.5129	0.1424	-2.0601	0.3133

Regression output for scenario 6

outcome	Coef.	Newey-West Std. Err.	t	P>t	[95% Conf.	Interval]
β1	1.128589	0.2891484	3.9	0	0.548371	1.708808
β4	-0.23341	3.888218	-0.06	0.952	-8.03569	7.568873
β5	0.04352	0.4131738	0.11	0.917	-0.78557	0.872614
β2	-0.76556	3.07017	-0.25	0.804	-6.92631	5.395184
β3	-0.86161	0.3851036	-2.24	0.03	-1.63438	-0.08885
β6	-2.17078	5.074068	-0.43	0.671	-12.3527	8.011078
β7	-1.18393	0.7541274	-1.57	0.122	-2.6972	0.32934
β0	27.11533	2.620871	10.35	0	21.85617	32.37449
Treated (β1[_t]+β5[_z_t] +β3[_x_t16] +β7[_z_x_t16])	-0.8734	0.5773	-1.5129	0.1364	-2.0319	0.285
Controls (β1[_t]+β3[_x_t16])	0.267	0.2544	1.0496	0.2988	-0.2434	0.7774
Difference (β5[_z_t] +β7[_z_x_t16])	-1.1404	0.6309	-1.8077	0.0764	-2.4063	0.1255