The Strategy Unit.

Evaluating innovations

Why, what and how? A menu of potential options…

NHS
Midlands and Lancashire
Commissioning Support Unit

Why do you evaluate?

Evaluation is a systematic process

Evaluation critically examines the programme/intervention/innovation of interest. It is needed to understand the difference the innovation or intervention makes (and how), so that informed decisions can be made about improvements to the innovation or the future of the programme.

Robust evaluation tells us not only *whether* an innovation worked, but also *why* and *how* – allowing us to learn lessons for spreading successful interventions and developing new ones. Therefore, evaluations should be developed to:

- Assess whether an innovation has achieved its intended outcomes
- Understand how the innovation delivered its intended outcomes, or why it may not have done so
- Identify how cost effective the innovation has been, based on a comparison of the input costs and any savings delivered by implementing the innovation
- Assess the impact of the innovation on the outcomes and experience of the patients, carers and staff involved
- Inform decision makers about how to build on or improve an innovation

Evaluation is not just about demonstrating success, it is also about learning why things don't work. As such, identifying and learning from mistakes is one of the key parts of evaluation.

What do you evaluate?

Not all evaluations are created equal

The size and cost of any innovation will have an impact on the size and cost of the evaluation. Large scale programmes with complex interactions between multiple innovations require more robust (and therefore more time consuming and costlier) approaches to evaluation.

Different points in the life cycle of an intervention also require different levels of evaluation. For example, an evaluation of an early prototype would be less comprehensive than one for an 'embedded' innovation that is 12 months post implementation.

- **Randomised Control Trials (RCTs)** are the gold standard for evaluation, but are often impractical to run for innovations, as they need to be set up prospectively, are complex to manage and expensive to run.
- **Quasi-experimental study designs**, such as matched cohort studies and Interrupted Time Series Analysis, provide a more practical alternative, if developed and delivered robustly.
- **A mixed methods approach with quantitative and qualitative research** will provide insight not only on whether the innovation was successful, but the elements that were key to any success and learning on what can be improved.

Overall, the scale of the evaluation should provide enough evidence for you to decide whether to continue, stop, roll-out or change the innovation.

There are some essentials for evaluation

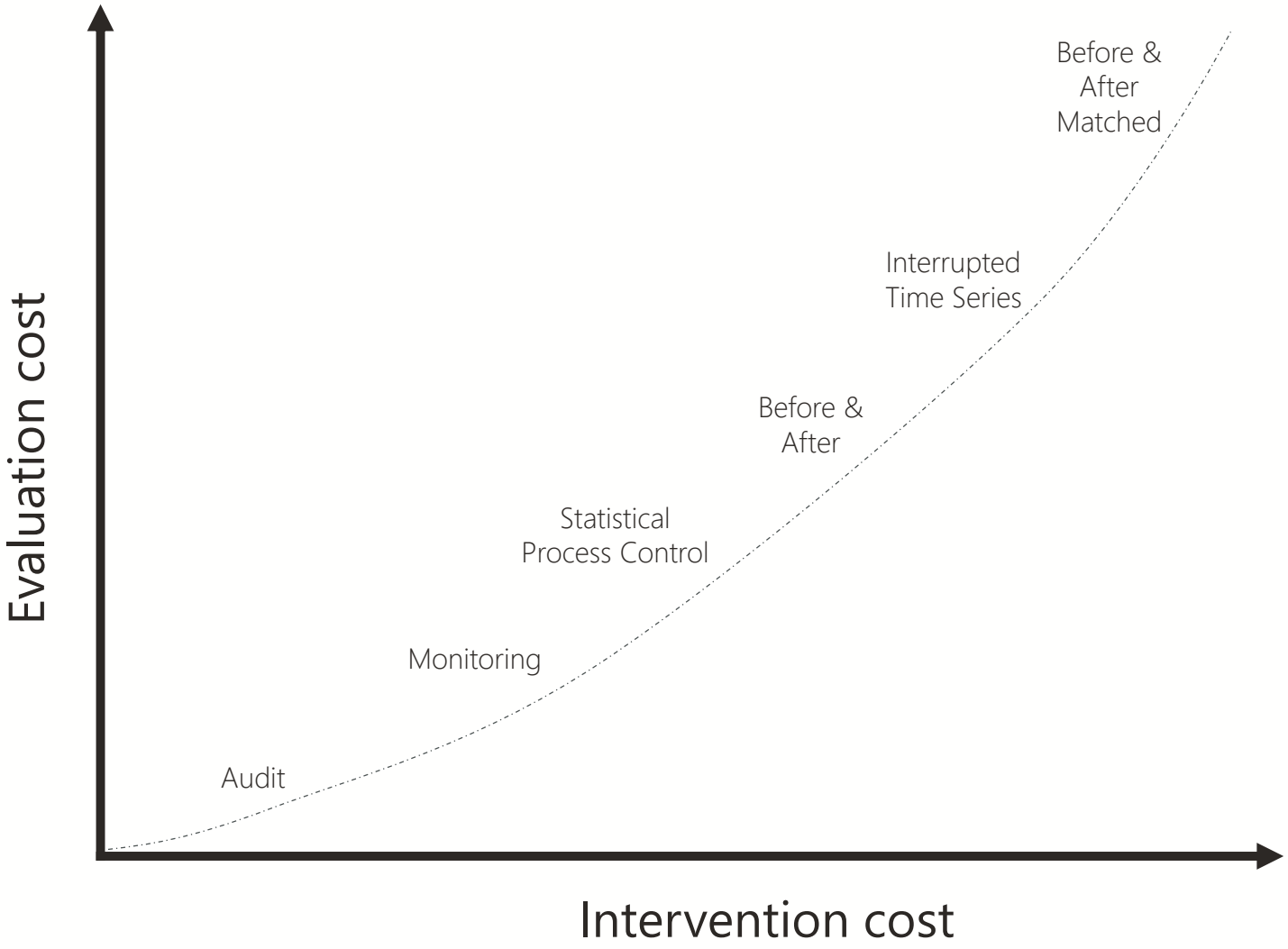
- **A theory of change** on why and how the innovation will achieve the expected outcomes – this can be captured in a Logic Model.
- **An agreed set of outcome measures,** linked to the theory of change
- **An evaluation framework,** which includes:
  - An agreed description of the innovation being evaluated, including the theory of change
  - An agreed set of evaluation principles and the core questions the evaluation will examine
  - A description of the planned evaluation method
  - A description of the counterfactual (an estimate of what would have happened in the absence of the innovation)
  - A detailed plan of how the evaluation be delivered
- **A baseline position for each of the outcome measures** (including historic trends) to quantify the issue the innovation has been developed to address.

Timing of the evaluation is crucial. Establishing a qualitative baseline retrospectively is very difficult.

How do you evaluate?

Quantitative Methods

Evaluation can be scaled to the innovation…



- **The cost and complexity of the evaluation should be proportionate to the cost of the intervention**
- Different approaches can be taken at different stages of the innovation life (e.g. audit data may be sufficient at the pilot stage, but would not be sufficient to scale up an expensive innovation across an STP footprint)
- The approach should always provide enough evidence to move to the next stage of development

Audit

An audit is an investigation into whether a service or activity is in line with agreed standards, to see if and where improvements can be made. The standards used can be determined by the service providers themselves or externally. The knowledge produced by an audit usually applies only to the particular context in which it was conducted, meaning that it is not possible to conclude that the same effects would be seen elsewhere in different circumstances.

Pros

- Provides early feedback on potential impact and implementation issues
- Easy to implement and relatively inexpensive to set up

Cons

- Does not provide generalisable results about impact
- Does not account for regression to the mean
- Does not demonstrate causality

When to use

Can be used at any stage of the implementation and running of the innovation but most appropriate in the early implementation stage before monitoring can show any trends

Monitoring

This uses data to try to determine the progress of a particular intervention or service against a set of targets or objectives. Data are usually collected and used at regular intervals to report progress to management teams, for example, number of people accessing services, number of contacts/therapies delivered. Data from performance measurement might be used to identify areas of concern in a service.

Pros

- Provides early and regular feedback on potential impact and implementation issues
- Easy to implement and relatively inexpensive to set up

Cons

- Does not account for regression to the mean
- Does not demonstrate causality

When to use

Once the innovation has been implemented, but before it has had time to demonstrate statistically significant results – to provide early signs of whether the implementation has been successful.

Statistical Process Control (SPC) Monitoring

Statistical Process Control (SPC) is a set of statistical methods based on the theory of variation that can be used to make sense of any process or outcome measured over time, usually with the intention of detecting improvement or maintaining a high level of performance. Control limits are set to distinguish between 'common cause' and 'special cause' variation. 'Common cause' variation could suggest early signs of success or failure of the innovation that warrant additional analysis.

Pros

- Provides early and regular feedback on potential impact and implementation issues
- Easy to implement and relatively inexpensive to set up
- More robust than monitoring alone

Cons

- Does not account for regression to the mean
- Does not demonstrate causality

When to use

Once the innovation has been implemented but before it has had time to demonstrate statistically significant results to provide early signs of whether the implementation has been successful. The control limits can be set using historic data

Quasi-experimental designs are a practical alternative to RCTs

Examples of quasi-experimental designs include:

- Before and after studies
- Interrupted time series analysis
- Before and after matched studies

These are explored in more detail in the following slides. They have several benefits and risks in common, shown below:

Benefits

- Reduces the risk of encountering ethical issues compared to RCTs
- Less time and resources required compared to RCTs
- Choice of methods that can be tailored to the problem
- Can be applied retrospectively as well as prospectively

Risks

- Selection bias in control and comparator group
- Large sample sizes required to prove statistical significance
- Unobserved differences can't be accounted for
- Dependency on readily available data that is of good quality

Before and After

A simple 'before and after' study consists of the comparison in the trend for an outcome(s) amongst the intervention group for a period of time before and after the intervention has come into place. The average for the outcome(s) measure for the group is compared between the before and after time periods to determine if a significant change has occurred.

Pros

- Relatively easy to implement
- Requires less data in comparison to more sophisticated methods

Cons

- Prone to regression to the mean (see appendix)
- Neglects the trend in outcome(s) across the wider population

When to use

Applied retrospectively, once the intervention has been implemented for a sufficient amount of time for the impacts to be realised

Interrupted Time Series

Interrupted time series (ITS) analysis is a quasi-experimental design with which to evaluate the longitudinal effects of interventions, through regression modelling. The trend for the outcome measure in question is analysed before the intervention had been introduced, at the point of intervention and after the intervention has been introduced.

Pros

- Accounts for trends in the outcome measure that are consistent over time
- Ability to analyse intended and unintended outcomes

Cons

- Trend in outcome measure needs to be linear
- Requires sufficient amount of historical data to generate robust trend estimates

When to use

Applied retrospectively, once the intervention has been implemented for a sufficient amount of time for the impacts to be realised

Before & After Matched

Matching is a non-experimental evaluation method that uses large data sets and statistical techniques to construct the best possible comparison group for a given treatment group. Matching relies on the assumption that participants and non participants are similar in terms of any unobserved and observed characteristics variables that could influence both the probability of being part of a intervention and potential to benefit from the intervention. Trends before and after the intervention has been implemented would be compared between the two groups.

Pros

- Method can be strengthened by combining with difference in difference analysis
- More robust than a simple before and after analysis

Cons

- Matching quality can vary depending on data availability

When to use

Applied retrospectively, once the intervention has been implemented for a sufficient amount of time for the impacts to be realised

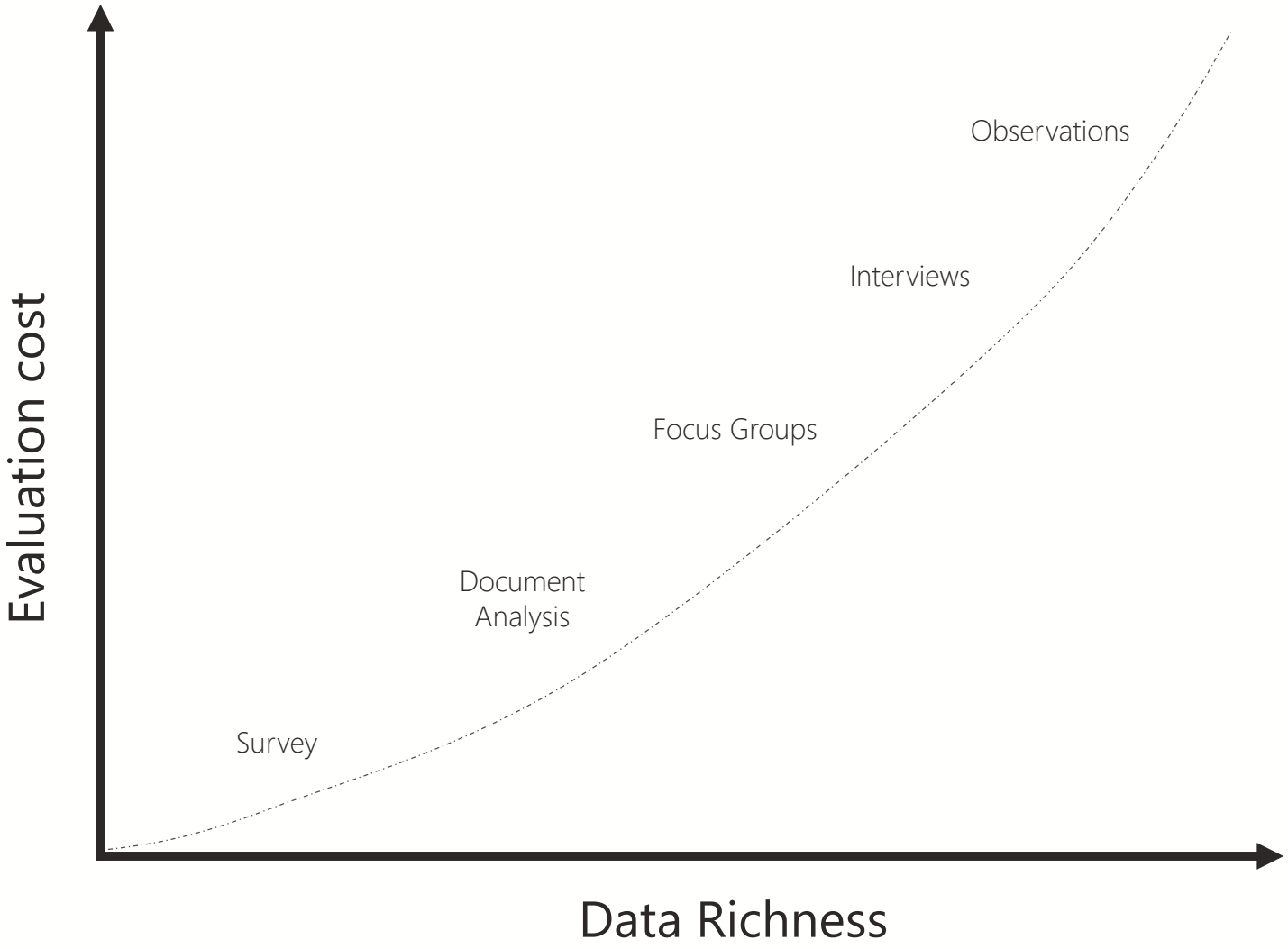
How do you evaluate?

Qualitative Methods

Qualitative Studies

- ‘Qualitative’ studies are not about measuring quality, nor are they the collation of anecdotes!
- **They explore the complexity of social phenomena through narrative means**, using abstraction to look for meaning.
- **They are used to answer questions about experience, meaning and social reality** (from the perspective of the participant).
- They involve the systematic collection, organisation, description, synthesis and interpretation of **textual, verbal or visual data**.
- **In the health and care sectors, the participants are wide ranging** and could include the patient, the clinician, the commissioning manager, the hospital executive, the STP chair. Participants are increasingly from beyond the NHS as well and include those working within the Public Health, Local Authority and Voluntary sectors.

Options in qualitative methods



- **Sample size in qualitative evaluations may typically be small, but the depth of data collected can be 'rich'.**
- Different approaches can be taken at different stages of the innovation life.
- The method used should collect sufficient data required for decision making.
- The method used should also be mindful of the impact on participants' time.

Document Analysis

Document Analysis, also known as 'content analysis' is an analytical method which relies on data that has already been collected and reported. As an initial step in an evaluation process, it can provide evaluators with an understanding of the innovation's origins and intentions, progress that has been made and the context within which it operates. Patterns and themes in existing data can be identified. It could also provide a thematic framework for subsequent stages of a qualitative evaluation. Findings can be used to develop theory of change and/or triangulated with other evaluation methods.

Pros

- Lower costs, as does not require new data collection.
- Provides qualitative understanding of baseline, which might otherwise have been lost (e.g. with staff changes).

Cons

- Potential for poor quality data in documents or information that does not lend itself to analysis
- Delay in access to documents can make this stage redundant.

When to use

It is usually a precursor step at the beginning of an evaluation to inform subsequent primary data collection steps. Documents could include: business cases, previous analysis/evaluation and their reports, case studies, website information.

Qualitative Surveys

Surveys are a systematic method for gathering information from a sample of people. A qualitative survey differs from a quantitative survey, in that it looks for the characteristics of diversity in the sample rather than counting the variation or distribution. Sampling is purposive (targeted to relevant stakeholders, e.g. those experiencing an innovative service) in a qualitative survey, rather than by chance (randomised).

Pros

- Participation 'burden' is low
- Reaches a wider sample size; not constrained by availability or geographical context
- Can be administered via online methods

Cons

- Requires a promotional strategy to reach relevant respondents
- Survey has to be designed carefully, minimising any ambiguity in questions.

When to use

When breadth of perspectives is required rather than depth. Qualitative surveys may be the first step of qualitative data collection, with findings from the survey informing the approach of subsequent steps.

Interviews

Interviews allow for the exploration of individual perspectives on the innovation. Analysis of interviews allows identification of themes that recur across the participant group. Interviews with those who have designed the innovation help to understand the aim and ambition. Interviews with those delivering the innovation addresses the 'implementation fidelity', that is, 'is the innovation being delivered in the intended manner?' Service user interviews explore how the innovation impacts on their experience.

Pros

- Structured analysis of individual (anonymised) reflections.
- Identifies themes as well as unique views
- The reflective nature of the process can galvanise interviewees to take further action

Cons

- Relies on interview recall – this is 'what they say, rather than what they do'
- Data collection and analysis is time intensive
- Some dependency on rapport between interviewer and interviewee

When to use

There is a requirement for individual perspectives because the topic is sensitive, contentious or detailed. Can be used at all stages of an evaluation with the same or different stakeholder groups. Interviews can also be repeated with the same individuals over time as part of a longitudinal evaluation. Findings can be used to improve, stop or roll-out the delivery of the innovation.

Focus Groups

In qualitative studies, focus groups are essentially group interviews. This approach allows a number of perspectives to be gathered at once and group dynamics can add to the 'richness' of the data being collected. In a focus group setting, an evaluator takes on more of a facilitator role rather than being an interviewer. In practice, with engaged participants, focus groups can be more rapid than individual interviews.

Pros

- Helps the group generate and share ideas
- The group dynamic can identify emerging consensus and facilitate resolution of potential issues. It can also draw out previously unrecognised issues.

Cons

- More difficult to manage a group and allow everyone to participate
- Keeping to the topic of interest and gathering sufficient depth of data

When to use

If the perspective required is not likely to be sensitive and also if the innovation is still at an early stage and require broader exploratory discussions. Focus groups might also be preferred when views of representatives of a stakeholder group are required, rather than specific individuals.

Observations

In observational studies, the evaluator can play a passive role, systematically observing and then describing events, behaviours, and artefacts within the social or natural setting. In a meeting with healthcare decision-makers discussing the future rollout of an innovation, the evaluator may take on a participant-observer role, possibly contributing to discussions, as well as observing the meeting. In other instances, the evaluator may act as a non-participant, for example when observing professionals interacting or using a technological innovation.

Pros

- Does not depend on willingness to participate - there is no burden of time for participants.
- Collects 'live' data in the natural setting – 'what people do rather than what they say'

Cons

- Risk of observer bias is high
- Observer 'paradox' – people behave differently when they know they are being observed
- Very resource intensive

When to use

In innovation projects, observations can be used when it is important to know how the innovation is being used or experienced in its natural setting. It also provides a more objective assessment than interview or survey data, which can be regarded as 'self-reported'. Observational data can also indicate the fidelity of an intervention across settings.

How do you evaluate?

Mixed Methods

Different combinations of methods are possible

- **A mixed-method evaluation 'triangulates' two or more evaluation methods, drawing on both quantitative and qualitative data.** The triangulation is conducted systematically, that is, there is intentional or planned use of diverse methods, potentially at every stage of the evaluation process.
- There are dual purposes of mixed methods:
  - **Clarification:** broaden and deepen the understanding of the processes through which the innovation achieves its outcomes **and** how these might differ by the context within which the innovation is implemented.
  - **Corroboration:** strengthens the reliability of data, and the validity of the findings and recommendations especially when one method alone would still leave "gaps" in the evaluation knowledge.
- The mixed methodologies used are tailored to the evaluation questions and can be run in parallel or sequentially, as relevant to the innovation being tested. For example, quantitative outcomes can be measured at appropriate timepoints, and qualitative fieldwork can occur in between the quantitative to understand stakeholder perspectives of the quantitative findings.

Testing emerging findings with stakeholders and including their views in the triangulation process helps to make sense of the findings. It can also help ensure the work is relevant and useful for decision-making purposes.

Appendix

Supporting materials

Regression to the mean (1)

The issue of 'regression to the mean' can occur whenever something which varies over time is measured once and is then measured again at a later point in time. Observations made at the extreme the first time round will tend to come back to the population average the second time round.

Regression to the mean is a particular challenge when an intervention is focused on particular types of patients (for example patients with high emergency care use). Say we look at people with frequent hospital admissions at present. On average, these individuals will have lower rates of unplanned hospital admissions in the future, even without intervention. So, if a therapist is working with patients who are currently having frequent A&E attendances, they may notice how the patients have fewer admissions over time. However, this reduction might well have occurred anyway due to regression to the mean, and it cannot necessarily be attributed to the input of the therapist. Regression to the mean occurs simply because after one extreme period, the next period is statistically likely to be less extreme.

The way to control for regression to the mean is to create a matched control group and look for differences between the intervention group and a similar control group. If you are using a before and after approach you can reflect that this may be an issue when writing up the caveats in the evaluation report.

Regression to the mean (2)