WWW.FRONTIER-ECONOMICS.COM

frontier economics

RAPID EVIDENCE REVIEW OF COUNTERFACTUAL METHODOLOGIES TO EVALUATE RESEARCH & INNOVATION PROGRAMMES

16 APRIL 2024

CONTENTS

ACKNOWLEDGEMENTS

Frontier Economics would like to thank the academic experts and evaluation agencies who contributed to identifying relevant evidence as part of this review, as well as UK Research and Innovation (UKRI) staff who provided comments on the draft report.

We would also like to thank UKRI as the commissioner of this study and the team who we worked directly with for their input and support throughout the review.

EXECUTIVE SUMMARY

Impact evaluation is a critical part of the policy cycle, both in terms of **learning** (did policies work as intended, and what are the lessons for future policy design?) and **accountability** (is public money being spent well?) Evaluating the impact of an intervention robustly is difficult, as we need to consider impacts relative to the **counterfactual** of what would have happened in its absence. The counterfactual cannot be observed, and so must be estimated.

Objective of the report

UK Research and Innovation (UKRI) is committed to rigorous monitoring and evaluation of their investments.[1] UKRI commissioned Frontier Economics to conduct a rapid evidence assessment to examine different counterfactual methodologies used in robust evaluations of Research and Innovation (R&I) programmes relevant to its activities. The research questions that guided the evidence assessment are:

- What are the counterfactual methods, including qualitative and quantitative ones, used by funders in R&I evaluation, and what is the rationale for using them?
- What are the success factors and strengths associated with different methods?
- What are the challenges associated with different methods to measure counterfactuals in R&I evaluation, and how do they vary in terms of the likely reliability of the findings, the time the evaluation is likely to take, and the costs involved? How are some of these challenges overcome through the use of different approaches?
- In recent years, have any innovative techniques been used to establish counterfactuals in R&I evaluation?
- Based on the above analysis, what are potential lessons for UKRI that could enable us to conduct more robust impact evaluations (with supporting examples)?

Approach

The evidence assessment identified more than 140 studies using keyword searches and expert recommendations, following an agreed protocol. The review was focused on counterfactual evaluation methods, typically at level 3 or higher of the Maryland Scientific Methods (MSM) scale. We also included specific searches for studies using qualitative and mixed methods, common to R&I evaluations, to ensure we captured relevant examples of innovation in those approaches. Of the long list of studies, 30 shortlisted studies underwent a full manual review and synthesis. The studies were shortlisted based on their relevance to research questions, the spread of UKRI activity areas, and recent innovations to improve the robustness of counterfactual impact evaluations that have the potential to be applied to UKRI activities. The studies reviewed include 16 applied evaluations of R&I policies, six applied

[1] *UKRI strategy 2022 to 2027* (March 2022), https://www.ukri.org/publications/ukri-strategy-2022-to-2027/ukri-strategy-2022-to-2027/

evaluations of other policies with the potential for the methods to be applied to R&I, and eight theoretical studies focusing primarily on innovations in counterfactual methodologies.

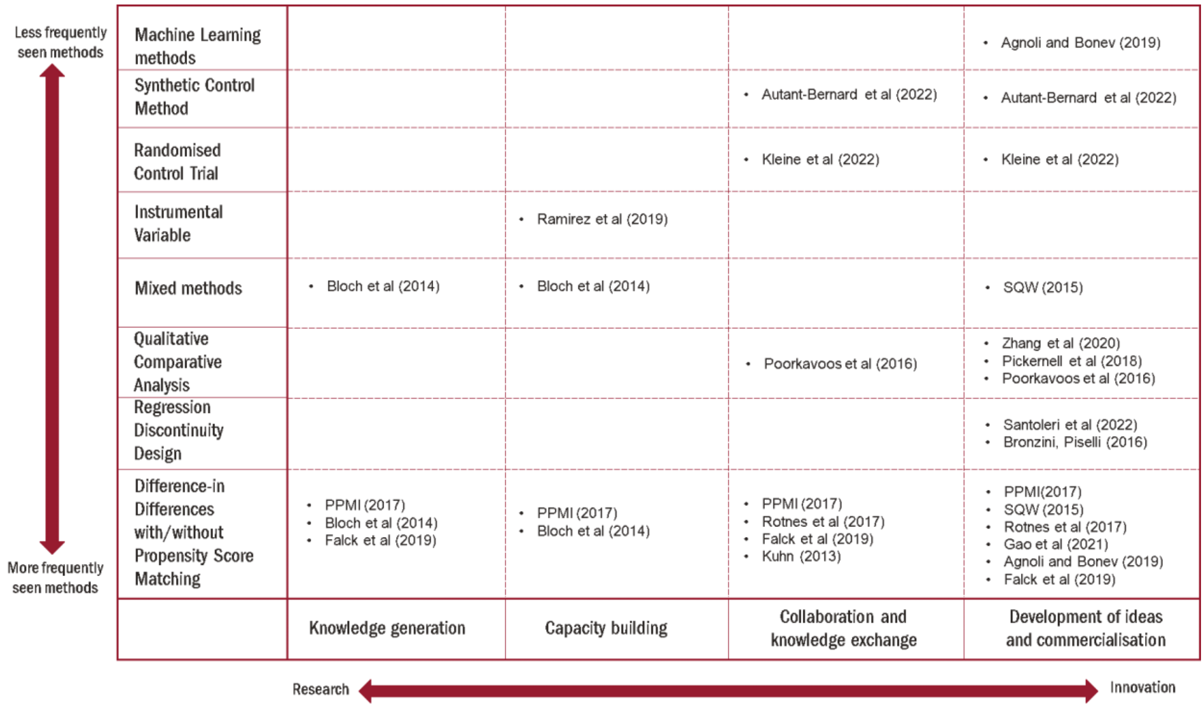
This study is a rapid evidence assessment rather than a systematic review, so the findings can only be considered representative of the studies we reviewed, rather than the entire potential evidence base.

Findings from the rapid evidence assessment

Counterfactual methodologies used in evaluating Research and Innovation programmes

Among the 16 applied R&I studies reviewed, a range of UKRI programme areas and counterfactual methodologies were covered. Figure 1 classifies the studies according to the broad UKRI programme areas and the evaluation methods used, the latter ordered by observed frequency in the studies. Some studies applied more than one methodology or examined multiple programme areas and so appear more than once in the diagram. Overall the review identified more studies on evaluating innovation- than research-related outcomes.

Figure 1 Range of counterfactual methods identified in applied Research and Innovation evaluation studies, by type of intervention and outcome areas



	Knowledge generation	Capacity building	Collaboration and knowledge exchange	Development of ideas and commercialisation
Machine Learning methods				• Agnoli and Bonev (2019)
Synthetic Control Method			• Autant-Bernard et al (2022)	• Autant-Bernard et al (2022)
Randomised Control Trial			• Kleine et al (2022)	• Kleine et al (2022)
Instrumental Variable		• Ramirez et al (2019)		
Mixed methods	• Bloch et al (2014)	• Bloch et al (2014)		• SQW (2015)
Qualitative Comparative Analysis			• Poorkavoos et al (2016)	• Zhang et al (2020) • Pickernell et al (2018) • Poorkavoos et al (2016)
Regression Discontinuity Design				• Santoleri et al (2022) • Bronzini, Piselli (2016)
Difference-in Differences with/without Propensity Score Matching	• PPMI (2017) • Bloch et al (2014) • Falck et al (2019)	• PPMI (2017) • Bloch et al (2014)	• PPMI (2017) • Rotnes et al (2017) • Falck et al (2019) • Kuhn (2013)	• PPMI(2017) • SQW (2015) • Rotnes et al (2017) • Gao et al (2021) • Agnoli and Bonev (2019) • Falck et al (2019)

Source: Frontier Economics

Note: These 16 papers are a subset of the papers reviewed in-depth as the figure shows the applied studies focussed on R&I programmes only. The rapid evidence assessment shortlisted 30 papers for in-depth review which consisted of 22 applied studies and 8 theoretical papers overall.

Key findings are that:

- Difference-in-Differences (DiD), both with and without the use of Propensity Score Matching (PSM), was the most commonly-used approach to R&I evaluation in our review, and the only methodology that had been applied to all of the broad UKRI programme areas among the papers reviewed.
- Other commonly-applied approaches were Regression Discontinuity Design (RDD), Qualitative Comparative Analysis (QCA) and mixed methods approaches.
- We identified fewer studies applying Synthetic Control Methods (SCM), Instrumental Variables (IV), Randomised Control Trials (RCTs) and Machine Learning (ML) methods to R&I evaluation.
- The less frequently used methods within the studies reviewed have been mostly applied to innovation-focussed programmes.

Apart from the applied studies on R&I-focussed interventions (shown in Figure 1 above) and theoretical studies, we also captured applied studies which explored innovations in counterfactual evaluation methodologies that could be applied to UKRI programme areas. They included, for example, innovations in the use of Propensity Score Matching to evaluate training programmes (Goller et al., 2020); and developments in the theory of Difference-in-Differences methods to increase confidence in the validity of the counterfactual (Roth et al., 2023).

Success factors and challenges

Based on the studies reviewed, the table below summarises the key strengths, success factors (that need to be met for the methodology to be applied successfully) and challenges identified in their application. Methods are ordered according to the Maryland Scientific Methods (MSM) scale, a five-level scale (where 5 is the highest) that ranks quantitative methodologies used in impact evaluation designs based on their robustness in identifying a causal impact.

Table 1 Counterfactual methodologies: strengths, success factors and challenges in application

Methodology	MSM scale	Strengths	Success factors	Challenges
Randomised Controlled Trial (RCT)	Level 5	■ Control group mimics counterfactual by design, meaning counterfactual estimates should be unbiased, and consistent in large samples.	■ Ethical issues in randomising support need to be considered. ■ Large and representative samples are needed to ensure statistical significance. ■ Detailed and reliable data on outcome variables across both groups are crucial for accurate analysis. ■ Strict adherence to random assignment is needed to ensure group comparability and minimise selection bias. ■ Consistent and standardised delivery of interventions across the experimental group.	■ Random allocation of support could diminish effectiveness. ■ Resource intensive to ensure pre-conditions of the methodology to produce valid results are met.
Regression Discontinuity Design (RDD)	Level 4	■ Uses cut-offs in treatment to imitate the experimental setting, mimicking the strengths of RCTs. ■ Does not necessarily require additional covariates if random	■ Treatment assignment needs to be deterministic. ■ There should be sufficient observations around the cut-off	■ Identifies local average treatment effects within the 'bandwidth' of the cut-off threshold only – treatment effects may not generalise.

Methodology	MSM scale	Strengths	Success factors	Challenges
		assignment at the cut-off is adhered to.	(i.e. the bandwidth) to estimate the treatment effect. ■ Observations near the cut-off need to be similar in all relevant aspects except for the treatment.	
Instrumental Variable (IV) design	Level 4	■ Can address the issue of endogeneity of treatment in observational studies, providing a more robust causal estimate. ■ IVs can be particularly powerful in the presence of natural experiments (where exogenous variation in the instrument occurs), which then mimics the experimental characteristic of an RCT.	■ Instrument must be valid: significant driver of treatment status, but no independent effect on the outcome of interest.	■ The main challenge of IVs is the difficulty of identifying a valid instrument. ■ Findings may not generalise if the relationship between the instrument and treatment differs across settings.
Difference-in-Differences (DiD) with and without Propensity Score Matching (PSM)	Level 3	■ Does not require random assignment to treatment or control group. ■ Can be used with both panel data and cross-sectional data. ■ Matching can create more comparable treatment and control groups based on observable characteristics.	■ Control group and treatment group must be sufficiently similar to have confidence in the validity of the control group as a counterfactual. ■ Data must be available to estimate propensity scores accounting for key drivers of treatment.	■ Testing the validity of the control group is difficult in practice, e.g. formal tests of the ‘parallel’ trends assumption. ■ Characteristics that may be affected by treatment cannot be included in the propensity score model.

Methodology	MSM scale	Strengths	Success factors	Challenges
				■ Unobserved factors influencing treatment cannot readily be accounted for.
Synthetic Control Method (SCM)	Level 3	■ Constructs a counterfactual that is designed to mimic the pre-treatment behaviour of a treated observation based on a weighting of multiple possible controls, helping directly address concerns around e.g. parallel trends. ■ Enables control-group analysis for small numbers of treated units.	■ Potential control units need to be shown not to be affected by the treatment. ■ Need to confirm that the synthetic control constructed does indeed mimic the treatment unit.	■ Requires careful consideration in determining the appropriate set of control units picked and the weights assigned to create the synthetic control. It may be hard to determine the optimal synthetic control. ■ Risk of 'overfitting' to the treatment and the results and, therefore, not being generalisable outside the specific context.
Mixed methods	N/A	■ Particularly suited to complex programmes where quantitative or qualitative approaches alone cannot capture the richness of impacts. ■ The use of both quantitative and qualitative data helps to triangulate results which improves the reliability of findings by corroborating results from different data sources.	■ Need to integrate approaches around a clear framework and agreed approach to analysis to avoid biases and subjectivity in findings.	■ It can be challenging to integrate qualitative and quantitative data to ensure coherence of findings or address conflicting findings. ■ Can be resource-intensive, both in terms of budget and time, to collect and analyse both quantitative and qualitative data ensuring their validity and reliability.

Methodology	MSM scale	Strengths	Success factors	Challenges
Qualitative Comparative Analysis (QCA)	N/A	■ Identifies how combinations of factors, instead of single factors in isolation, led to a given outcome. ■ It allows the evaluation of complex programmes with several combinations of causal pathways rather than a single case through which the outcome is observed.	■ Need to ensure that the truth table, which systematically represents all possible combinations of conditions and outcomes for the selected cases, reflects the logical relationships between conditions and outcomes in theory.	■ Can be context-specific and hence not generalisable to broader populations. ■ Findings can be influenced by biases of both participants and researchers, which can make the findings less objective.
Machine Learning (ML) methods	N/A	■ Ability to work with 'big data' with a high number of covariates. ■ Can help with covariate selection for the models by automatically learning from patterns in the data which can then guide the researcher to pick the right covariate set for modelling. ■ Can account for non-parametric and non-linear model specifications when such models are a better fit to model the observational data. ■ Can help to identify heterogenous treatment effects when treatment is expected to differ across segments of the treated group.	■ Need to ensure that assumptions such as control and treated units are comparable to avoid selection bias and that the treatment assigned to one unit does not affect the potential outcomes of other units are met.	■ Interpretability of causal estimates can be low, and results may lack a clear theoretical foundation. ■ Emerging field, so few applied studies to establish validity and effectiveness for wide-spread application.

Source: Frontier Economics

Apart from the specific methodological challenges outlined above, a general challenge associated with commonly-used counterfactual methodologies in the studies reviewed is the sample size available to conduct robust analysis and the ability of evaluations to produce statistically significant findings. The emergence of approaches such as Qualitative Comparative Analysis and Synthetic Control Method may be particularly suited to cases with small numbers of treated units.

The success factors and challenges relating to the methodologies in general also interact with the characteristics of R&I programmes which add to the challenges for robust impact evaluation. These characteristics include:

- Beneficiaries receiving multiple treatments over time, which makes it challenging to disentangle the impact of specific interventions or identify clear control groups.
- A complex and dynamic R&I ecosystem with multiple external factors influencing the success of the intervention and many beneficiaries of R&I programmes (universities, researchers and innovative businesses) interacting to generate overall impacts.
- Many R&I outputs are intangible (e.g. knowledge) and hard to measure.
- Sometimes a lack of clear control group (e.g. where research activity would not take place at all in the absence of an intervention, or where an intervention supports all eligible firms or researchers which can happen in small, emerging innovative areas).
- Where control groups are identified, permissions may be needed to collect data from them (e.g. those who are not successful in funding applications) in order to implement some counterfactual methodologies like Difference-in-Differences or Regression Discontinuity Design.
- The returns of R&I activities are inherently uncertain, with impacts being generated over a long-term horizon. This can entail many activities ‘failing’ while a small number are highly successful, which means a focus on ‘average’ impacts in the short run may be misleading about the impact of a given programme.

Innovations in recent years to establish counterfactuals in Research and Innovation evaluations

Our review highlighted three main types of methodological innovation relevant to the evaluation of R&I interventions:

- **Programme design enabling experimental evaluation design**: Randomised Controlled Trials are considered the most effective way to establish a counterfactual theoretically but require the right programme design which allows for randomisation of treatment. Building randomisation into the programme design while considering the programme objectives and costs associated with implementing the evaluation design provides evaluators an opportunity to consider Randomised Control Trials for evaluating an R&I programme. Despite a targeted search, our review only identified one use of

Randomised Controlled Trials in R&I evaluation where randomisation had been built into programme design (Kleine et al., 2022).

- **Improvement in existing counterfactual methodologies**: A range of papers have outlined theoretical developments that attempt to reduce the burden of satisfying the assumptions associated with quantitative counterfactual methods. These innovations attempt to improve the internal validity of findings by looking at ways to generate unbiased estimators of the causal impact using any given counterfactual methodology. For example, one study (Roth et al., 2023) discusses improved diagnostic tools to detect the violation of parallel trends.
- **Development of new methodologies**: The evidence reviewed highlights that new methodologies are being developed and applied to R&I recognising some of the limitations of more commonly-used approaches. For example, a Synthetic Control Method (Autant-Bernard et al. 2022) has been used where there are small numbers of treated units; and double / debiased machine learning models have been developed theoretically (Chernozhukov et al., 2018) and applied to R&I (Agnoli and Bonev, 2019) where causality needs to be demonstrated in complex settings with multiple possible covariates influencing treatment and outcomes.

Lessons learnt for UKRI

Based on the evidence reviewed, UKRI should consider evaluating R&I programmes using:

- Randomised Controlled Trials where randomisation of treatment is feasible ex-ante, where the ethical implications of random allocation of support are low, and where treatment can be standardised across all beneficiaries at the time of delivering the programme;
- Regression Discontinuity Design where R&I programmes include eligibility based on a clear scoring guide or other criteria, and the criteria to identify the beneficiaries are strictly adhered to during implementation of the intervention;
- Differences-in-Differences where treatment and control groups can be identified (through matching or other means) and where alternative estimation strategies with improved diagnostic tools are used if the programme and data characteristics threaten to violate the key assumptions underpinning its robust implementation;
- Differences-in-Differences with heterogenous treatment effect estimation strategies in programmes where treatment is rolled out in a staggered manner over time, i.e. different units can become treated at different points in time, rather than a simple one-off treatment;
- Mixed-methods for complex R&I programmes which are multi-faceted and will benefit from using data from different types of data and evidence sources to have a comprehensive evaluation;
- Qualitative counterfactual methodologies such as fuzzy-set Qualitative Comparative Analysis (an example of a theory-based approach) where sample sizes of treated and

control groups are smaller and the potential conditions leading to success are not straightforward binary factors (present or not); and

- Synthetic Control Method where one or a few units (such as universities, firms, sectors or areas) are exposed to an intervention, and where high quality secondary data over a period of time on both treatment and control units are available. Synthetic Control Method is also an approach to consider when significant heterogeneity is expected in the treatment effect across treated units, as could be the case for R&I programmes that target very large companies with tailored support.

1 Introduction

UK Research and Innovation (UKRI) commissioned Frontier Economics to undertake a rapid evidence assessment to understand the different counterfactual methodologies used in robust evaluations of Research and Innovation (R&I) programmes, including recent innovations in this area. This report sets out the findings and lessons learnt for UKRI.

This chapter provides the objectives for this rapid evidence assessment, and outlines the approach taken and the structure of this report.

1.1 Objectives of this study

Evaluations play a key role in policy design, helping to understand what changes have occurred as a result of a policy intervention, mechanisms of impact, and to what extent any changes can be attributed to the intervention. These insights can inform subsequent decision-making, and evidence the value for money of public spending. As such, the evaluation of R&I programmes is a core part of UKRI's strategy,[2] which commits to *"…monitoring and evaluating our investments rigorously to understand our impact and to learn from 'what works'."*

Robust impact evaluations can be challenging, as they need to evidence the difference that an intervention has made compared with *what would have happened in the absence of the policy intervention*: the **counterfactual**. The counterfactual cannot, by definition, be observed. There are different theory-based, experimental and quasi-experimental approaches that are used in impact evaluations which seek to estimate credibly what the counterfactual is.

The focus of this study is on evaluation approaches that seek to estimate the counterfactual, and which have been, or could be, applied to R&I programmes. We considered a wide range of UKRI areas of focus and activity types:

- **Knowledge generation** through basic and applied research funding to universities to create a diverse portfolio of high-quality research.
- **Capacity building** by investing in skills and career development of students, researchers and innovators, as well as on research infrastructure such as laboratory facilities, equipment and digital resources.
- **Collaboration and knowledge exchange** between academia, businesses and international partners to incentivise cross-sectoral partnerships and catalyse local and regional innovation capabilities.
- **Development and commercialisation of innovative business ideas** to support emerging concepts that have the potential to have a positive impact on the UK economy through increased productivity and sustainable growth.

[2] *UKRI strategy 2022 to 2027* (March 2022) (available at https://www.ukri.org/publications/ukri-strategy-2022-to-2027/ukri-strategy-2022-to-2027/)

Impact evaluations of R&I programmes need to take into account specific characteristics that can make the use of certain methodologies challenging or, indeed, unfeasible. In particular, key challenges of R&I programmes for robust impact evaluation include:

- Long lag times between public investment in R&I and observable socio-economic impacts;
- A complex R&I ecosystem with multiple external factors influencing success;
- Difficulties in attribution where many beneficiaries of R&I programmes (including universities, researchers and innovative businesses) receive multiple forms of public support, which all contribute to realised benefits;
- Many R&I outputs are intangible and hard to measure or place social and economic value on;
- Sometimes a lack of clear control group (e.g. where research activity would not take place at all in the absence of an intervention, or where an intervention supports all eligible firms or researchers which can happen in small, emerging innovative areas).
- Where control groups are identified, permissions may be needed to collect data from them (e.g. those who are not successful in funding applications) in order to implement some counterfactual methodologies like Difference-in-Differences or Regression Discontinuity Design.
- R&I is inherently uncertain, with many projects ‘failing’ while a small number are highly successful, which means a focus on ‘average’ impacts may be misleading.

UKRI is therefore interested in understanding the current evidence base on the use of different counterfactual methodologies in R&I evaluations.

1.2 Research questions

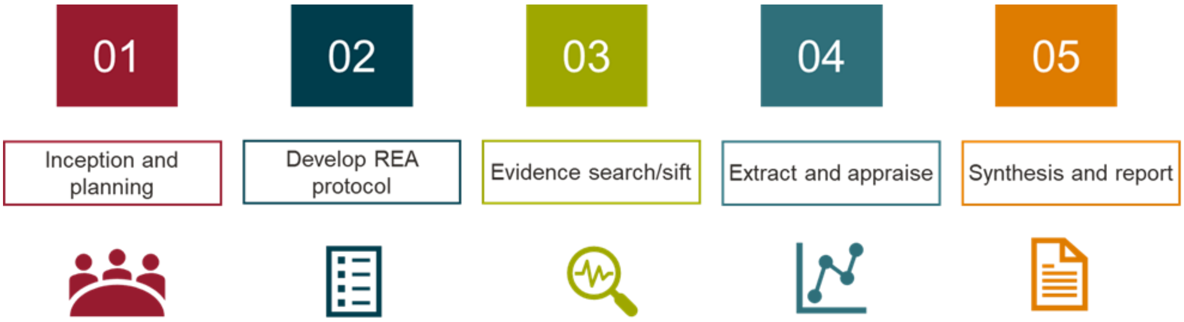
The following research questions have guided the review:

1. What are the counterfactual methods, including qualitative and quantitative ones, used by funders in R&I evaluation and the rationale for using them?
2. What are the success factors and strengths associated with different methods?
3. What are the challenges associated with different methods to measure counterfactuals in R&I evaluation and how they vary in terms of the likely reliability of the findings, the time the evaluation is likely to take and the costs involved? How are some of these challenges overcome through the use of different approaches?
4. In recent years, have there been any innovative techniques used to establish counterfactuals in R&I evaluation?
5. Based on the above analysis, what are potential lessons for UKRI that could enable us to conduct more robust impact evaluations (with supporting examples)?.

1.3 Approach

This rapid evidence assessment was undertaken over a period of four months. We drew on best practice for rapid evidence reviews (Collins et al., 2015) to implement a five-stage methodology summarised below:

Figure 2 The approach to the study



Source: Frontier Economics

Inception and planning

We held an inception meeting with UKRI to ensure a common understanding of the project aims, needs, approach, timeline and working arrangements. Given the breadth of available evidence across academic and grey literature that was agreed to be within the scope of the review, we carefully scoped out the research questions, working closely with UKRI to ensure the review was properly focused.

Developing the rapid evidence assessment protocol

We developed a protocol for the rapid evidence assessment which was agreed with UKRI. The full protocol can be found in Annex B .The protocol provided a structured approach to the evidence assessment, giving clarity on the process to be undertaken to conduct the review and allowing for transparency of methods and replicability of the work. The main elements of this protocol were:

- A set of **inclusion/exclusion criteria** for studies to include in the review.
- A **search strategy** to identify the relevant evidence. It included a mix of search keywords to use with databases such as Google Scholar, the Science and Innovation Policy Evaluation Repository and Science Direct; targeted searches on Google Search and evaluation agency websites to identify grey literature; and input from academic and global R&I funding bodies to ensure we did not exclude key unpublished studies.

- An **analytical framework** for how the evidence would be reviewed, assessed and captured to inform the research questions. Our framework ensured our approach focused on UKRI priorities for the review.

Conducting the search and sifting results

We implemented the search strategy agreed upon in the protocol. Details of the search outputs were recorded in an Excel format to ensure transparency and replicability. This step yielded a total of 147 papers to consider. The resulting list of studies were screened for relevance based on an abstract sift to ensure that the study informed at least one of the research questions. This generated a long list of 88 potential studies for review. Given the timelines for the project, it was agreed with UKRI to include 30 studies on the shortlist for full review. The shortlist attempted to cover the breadth of robust counterfactual methodologies, both in applied evaluation studies and theoretical papers.

Reviewing studies to appraise quality and extract insights

We manually conducted an in-depth expert review of the 30 papers, with key evidence extracted against the analytical framework. This generated a detailed summary table of each study along with key findings. Our review included a mix of theoretical studies (8 papers), and applied research or evaluation studies (22 papers).

Evidence synthesis and reporting

We drew on the summary table to draw key conclusions and insights related to each research question. We discussed initial findings and early conclusions with UKRI and revised the draft report based on detailed feedback.

1.4 Structure of the report

The rest of the report is structured as follows:

- Section 2 presents the findings from the evidence assessment, including short examples and vignettes from the studies reviewed;
- Section 3 presents lessons for UKRI drawing on the evidence reviewed;
- Section 4 describes the overall state of the evidence reviewed; and
- Section 5 draws overall conclusions.

We provide a bibliography of the papers reviewed in detail for this study, as well as annexes including a glossary of commonly-used terms (Annex A) and the full protocol (Annex B).

2 Findings from the rapid evidence assessment

This section summarises findings from the rapid evidence assessment based on a synthesis of 30 shortlisted studies that were reviewed in-depth.

In synthesising the studies, we were guided by the overarching research questions as set out in section 1.2 above.

This section is therefore organised to address these questions. We begin (section 2.1) by summarising the range of traditional counterfactual methods for Research and Innovation (R&I) evaluation that have been identified in our review and the success factors, strengths and challenges associated with them. We then (section 2.2) focus on the innovations in different approaches identified in our review, both in terms of incremental improvements on more commonly-used methods to evaluate R&I policies, and in terms of more novel approaches that are being developed but are less common in practice.

Throughout this section, we make use of boxes to call out specific illustrative examples of the points being raised from the papers that we reviewed.

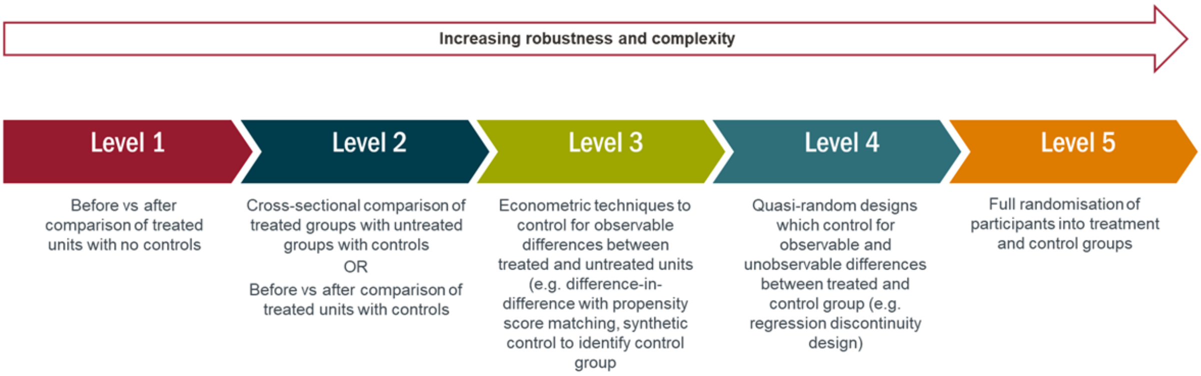
2.1 Counterfactual methods traditionally considered to evaluate Research and Innovation programmes

2.1.1 Quantitative methodologies

There are a range of potential quantitative approaches for impact evaluation. A commonly used framework for exploring the degree of robustness of evaluation approaches is the Maryland Scientific Methods (MSM) scale (Sherman et al. 1998),[3] shown in Figure 3. It is a five-level scale that reflects increased robustness in the approach taken and, therefore, confidence in the findings of an evaluation in terms of attribution of impact to the policy or programme being studied. In particular, increased levels of the scale reflect more sophisticated use of control groups where drivers of outcomes, beyond exposure to the policy, are more convincingly accounted for.

[3] https://www.policinginstitute.org/wp-content/uploads/2015/06/Sherman-1998-Evidence-Based-Policing.pdf

Figure 3 Maryland Scientific Methods Scale


Increasing robustness and complexity
Level 1
Before vs after comparison of treated units with no controls
Level 2
Cross-sectional comparison of treated groups with untreated groups with controls
OR
Before vs after comparison of treated units with controls
Level 3
Econometric techniques to control for observable differences between treated and untreated units (e.g. difference-in-difference with propensity score matching, synthetic control to identify control group
Level 4
Quasi-random designs which control for observable and unobservable differences between treated and control group (e.g. regression discontinuity design)
Level 5
Full randomisation of participants into treatment and control groups

Source: Frontier Economics based on Sherman et al. 1998

Level 1 is either (a) cross-sectional comparison between a treated group and an untreated group; or (b) before-and-after comparison of a treated group without any comparable untreated group. Using these methods, although we can say whether there is a difference, we cannot 'attribute' the change to the intervention.

Level 2 adds to Level 1 by using adequate control variables to adjust for (a) cross-sectional differences between treated and untreated groups or (b) before and after differences in other drivers of outcomes (beyond the policy intervention) for the treated group.

Level 3 involves methods in which changes in outcomes in the treated group after an intervention are compared with changes observed in a suitable control group (often summarised as 'difference in differences'). More sophisticated approaches in this Level attempt to demonstrate the comparability of the treatment and control group through techniques such as propensity score matching, or synthetic control approaches, helping account for issues such as selection bias[4] as fully as possible.

Level 4 methods exploit 'quasi-randomness' in treatment, using approaches such as Instrumental Variables (IV) or making use of rules in eligibility for treatment (Regression Discontinuity Design (RDD)) to justify that the only difference between treatment and control groups is their exposure to the policy. This gives rise to 'natural experiments' which can form the basis of evaluation.

Level 5 includes research methods which explicitly use randomisation while allocating treatment, such as Randomised Control Trials (RCTs). This is considered the 'gold standard'

[4] Selection bias is when participants in a programme (treatment group) are systematically different from non-participants (control group). Selection bias affects the validity of programme evaluations whenever selection of treatment and control groups is done non-randomly.

of attributing causality to the policy, as any differences between treatment and control group are, by definition, fully random.

The focus of the quantitative approaches in this review was on more robust approaches, identified as those Level 3 and above on the Maryland scale. Our review highlighted applications of different quantitative evaluation approaches to R&I across different levels of the scale.

Randomised Control Trials (RCTs)

Randomised Control Trials (RCTs), at Level 5 on the Maryland scale, entail explicit randomisation – that is, the random assignment of a policy or intervention to a group of individuals or firms, so that the outcome that the intervention is trying to influence will be compared between the group that received it (the so-called treatment group) and the group that did not receive it (control group). This approach originates from the health sciences and, in the last few decades, has gained prominence in economics and other social sciences (Ramachandran, 2020).

The strength of the RCT is that, by randomising the assignment, no significant differences between the treatment and control group could influence the outcome. Outcomes observed for the control group can, therefore, be taken as a measure of how the treatment group would behave had it not received the intervention – the counterfactual. If the randomisation is successful, the estimates obtained from an RCT converge to the true ‘treatment effect’ as the number of observations increases. An RCT produces valid results when the pre-conditions of the methodology are complied with.

Challenges of the methodology

The methodological requirements to implement an RCT, such as planning and designing the programme to enable an RCT evaluation, retaining participants of both treatment and control groups through the programme, and collecting data and monitoring their quality, can be resource intensive both in terms of budget and time.

Additionally, as with any quantitative research design, having a large enough sample is key to ensuring that an RCT can detect an impact. Research and Innovation (R&I) programmes may not lend themselves to large numbers of individuals or firms being treated by a specific intervention. Rather, individual policies may be complex interventions awarded to a single institution or consortium (e.g. a Research Group, or a Challenge Fund); or grant funding programmes where the number of treated units is small.

Further challenges of using RCTs for R&I programmes identified through our review (Dalziel, 2018), include:

- RCTs require standardised treatments which inhibit tailoring, experimentation and learning that may be needed given the uncertain nature of R&I.

- The narrow scope of an RCT will require additional research to help contextualise the findings and understand why impacts are observed.
- High demands placed on programme participants such that firms or individuals assigned a treatment or control group may migrate between groups (e.g., dropping out of treatment, control units later becoming treated, or those who are not treated changing behaviour as they are aware they are in the control group). This non-compliance will threaten the viability of the experimental approach when individual units can, to some extent, self-select into groups.
- Response attrition when outcome data from both treatment and control units needs to be collected through surveys, which risks low response rates and non-random response. This can lead to a biased estimation of the treatment effects.

In our review, we came across one evaluation study which applied an RCT counterfactual methodology to evaluate an innovation programme, which is described below.

Application of Randomised Controlled Trial (RCT) to identify casual effect of innovation vouchers awarded through a lottery on outcomes of UK-based small and medium businesses

Title of paper: Subsidised R&D collaboration: The causal effect of innovation vouchers on innovation outcomes (Kleine et al., 2022)

The paper analyses an innovation subsidy scheme in the UK, the “Innovation Vouchers” Programme, delivered between 2012 to 2016 by Innovate UK, prior to the formation of UKRI. The scheme provided UK-based Small and Medium Enterprises (SMEs) up to £5,000 to engage with the services of experts (that had not been previously engaged by the SME) from the public or private sector to pursue a particular innovation-related project within the firm. Given the relatively small monetary value of the support, the scheme was mainly targeted at small-scale projects, with intended outcomes such as Intellectual Property (IP), or the development of products, services and processes, rather than ‘breakthrough’ innovations.

In the 10 rounds studied, over 6,600 firms applied for a voucher, with 3,100 subsidies being awarded and 2,000 of them redeemed. The programme used a lottery to grant funding to ensure the random allocation of the innovation voucher and enable an RCT-led evaluation design. Allocation was random among SMEs who applied and met minimum criteria for receiving funding, enabling treatment and control groups to be established and minimise the risk of selection bias in estimating treatment effects.

The authors collected various outcome measures related to innovation outcomes and activities by means of three data sources: data held by Innovate UK, Companies House, and surveys conducted one- and two-years post-award. Achieved sample sizes were 1,107 firms in the treatment group, and 356 in the control group. No evidence of selection bias was found.

Receiving the voucher increased firms’ probability of receiving external funding for innovation activities and establishing new internal innovation management processes, but no long-term evidence was found of impact on the number of patents, design rights, trademark applications, or minimum viable products that were produced. The study did find a positive influence in the number of product and services developed for a subsample of companies that applied to the programme where this was the explicit stated aim.

These results were found comparing firms that were and were not awarded a voucher. However, one-third of those awarded did not redeem the voucher; the main estimates of the study therefore capture the "intention-to-treat", or the effect of being offered a voucher, rather than its use. The authors also study the effects of the programme for those firms that redeemed the voucher. The results are generally similar in sign and significance (or lack thereof) but with higher treatment effects. The authors also note that the small sample size in control group, especially in the survey in Year 2, could have an effect on the power of their statistical tests. They also point out the potential threat to the external validity of findings due to the non-random responses to the surveys. They include checks to test for this bias within their analysis.

Although the randomisation checks conducted do not show any significant difference between treated and untreated firms, they do find significant differences between those firms in the control group and those in the treatment group that redeemed the voucher, finding that the latter are smaller firms, more likely to have already chosen a supplier at the time of the application, and less likely to have received other sources of public funding.

The study suggests that this lack of impact on longer-term outcomes might be a consequence of the short time span of the study. In addition, the authors found that while there were no significant differences between the treated and untreated firms that answered the survey, there were observable differences between them and those firms that did *not* respond, potentially hindering the generalisability of their results.

Regression Discontinuity Design (RDD)

Rather than treatment being randomised, most evaluations of policy interventions are observational studies, where evaluators can observe the effect of an intervention but where the opportunity to control who is treated are limited. In these cases, studies have to deal with the issue that an unmeasured variable (a confounder) may unintentionally affect the outcome of an intervention and break the causal link between the intervention and the outcomes observed. Thus, counterfactual methodologies in these cases need to identify a counterfactual group comparable to the treatment group such that, as far as possible, confounding factors are accounted for.

The methodological approaches that most closely resemble the random allocation of treatment used for RCTs are those that achieve **quasi-randomisation** by exploiting a feature of the policy intervention or the environment that allows treatment and control groups to be identified where the only differences between them are exposure to the intervention. This creates a so-called ‘natural experiment’. Quasi-experimental methodologies are at level 4 on the Maryland scale.

In our review, the most prevalent example of a quasi-experimental approach applied to Research and Innovation (R&I) policy is **Regression Discontinuity Design (RDD).** RDD is used to evaluate the impact of policies where there is an assignment variable (e.g. a set of criteria or rules) with a cut-off above or below which treatment does/does not occur. It is assumed that units just below the cut-off (‘just unsuccessful’) provide a good control group for those just above the cut-off (‘just successful’), with the cut-off generating a discontinuity in the probability of treatment. Comparing the outcomes for these two groups can thus help estimate the intervention's causal impact without confounding variables.

We reviewed two papers using RDD to evaluate R&I policy (Bronzini and Piselli, 2016; Santoleri et al., 2022). In both cases, the studies exploited the programme design where funding decisions are based on scores determined by external expert panels that aim to measure the potential of the applicant. An example of how the programme design enabled the use of RDD is shown below.

Application of RDD: The effect of Research and Development (R&D) grants on patenting in Italian firms

Title: The impact of R&D subsidies on firm innovation (Bronzini and Piselli, 2016)

This paper evaluates the impact of an R&D subsidy programme implemented in 2004 by the regional government of the northern Italian region of Emilia-Romagna to foster innovation in beneficiary firms. The programme subsidised the innovative activities of eligible firms through grants that covered (for larger firms) up to 50% of the costs for research projects and 25% for pre-competitive development projects; larger shares of costs were covered for Small and Medium Enterprises.

The grants subsidised the cost of machinery, equipment and software, purchase and registration of patents and licenses, employment of researchers, use of laboratories, contracts with research centres, consulting and feasibility studies and, finally, the external costs for the creation of prototypes. In order to be eligible for a grant, the overall costs of a project had to be between €150,000 and €250,000.

The grants were assigned after an assessment of the proposals by a committee of independent experts appointed by the regional government, which granted each proposal a score. Since the recipients and non-recipients of the data are inherently different, one of the key challenges of the study was to find a comparable control group. The authors exploit this application process to compare the innovation outcomes of firms that obtained a score just above and below the granting threshold (75 out of 100). Their overall sample size was 612 firms.

The outcome variables studied, as a proxy for innovation, were the number of patents produced (the ‘intensive margin’) and a firm’s probability of patenting at all (the ‘extensive margin’) between 2005/06 and 2011, the post programme period. The authors acknowledge the pros and cons of using patents as a proxy for innovation. By exploring both intensive and extensive margins, the authors examine both the impact on the number of innovative firms, and the total level of innovation as proxied by patenting. By letting the outcome variable be a function of the score of the grant application, the average treatment effect of the programme is estimated by the value of the discontinuity at the threshold.

In order to ensure that the firms just above and below the cut-off threshold were similar, the study compares averages of key financial variables for the two groups. Comparing all treated and non-treated firms they find differences between groups, but these are ameliorated when the groups are restricted to a bandwidth around the cut-off. They also conduct tests to check the discontinuity requirement at the threshold.

The results show that the programme had positive effects on both the intensive and extensive margin. First, the programme positively impacted the number of patent applications of subsidised firms, the effect being significantly greater for smaller compared with larger companies. Second, the programme positively impacted a firm’s probability of applying for a patent, although the effect was weaker than the extensive margin and was only significant for smaller firms. The authors also carry out robustness checks to test the validity and sensitivities of their findings. They highlight that their findings may not be robust when it comes to external validity given the regional dimension of the programme.

There are two main types of RDD: sharp and fuzzy RDD. A sharp RDD is applied if all individuals on one side of the cut-off point are treated, and all individuals on the other side are not (“perfect compliance”). Fuzzy RDD relaxes the assumption of perfect compliance, and allows the possibility of some individuals above the threshold not receiving the treatment and some below receiving it. To account for imperfect compliance, fuzzy RDD often uses instrumental variables, an approach discussed in more detail below. The RDD applications identified in our review relied on sharp RDD approaches.

Application of Regression Discontinuity Design (RDD): The effect of Research and Development (R&D) grants on innovation outputs and business performance in Europe

Title: The Causal Effects of R&D Grants: Evidence from a Regression Discontinuity (Santoleri et al., 2022)

This paper investigates the EU’s “SMEs instrument”, a programme established in 2014 to fund innovation in European Small and Medium Enterprises (SMEs) through proof-of-concept and commercial development grants. The programme was developed by the Executive Agency for Small and Medium-Sized Enterprises an European Commission-funded agency aimed at providing business innovation support. The authors exploit the fact that firms received a score by the independent expert panel when they applied to the research funds, applying a ‘sharp’

RDD to compare the performance of subsidised and non-subsidised firms that have scores close to the threshold.

In principle, a sharp RDD should mean that no factors other than the panel's score influenced treatment. However, in their analysis, the authors included a number of additional covariates (such as sector, and rank within the competition) to increase the precision of the estimates and account for any potential imbalances between the treated and untreated firms that might correlate with the outcome variables.

The authors find that the grants positively affected firms' innovation outputs (measured as cite-weighted patents), revenue and employment growth, and private equity financing. The study also highlights heterogeneity between firms in these results, with that smaller and younger firms experiencing stronger treatment effects than larger and older firms.

Challenges of the methodology

Although quasi-random approaches help to provide 'as good as random' treatment in the absence of an experimental setting, the conditions under which quasi-randomisation can be obtained are quite limited.

In RDD evaluation approach, the key theoretical assumptions are that:

- Outcomes change 'smoothly' near the cut-off point, such that units just above and below the cut-off are comparable in all aspects except for their proximity to the threshold and therefore their treatment status.
- The 'cut off' value is essentially arbitrary rather than being used for other policy purposes which could in turn influence outcomes. For example, if those just below a quality threshold are then encouraged to apply for other programmes of support, this will limit the usefulness of RDD.

The RDD approach necessitates the determination of a 'bandwidth' around the cut-off point within which individual units can be shown to be statistically comparable in all aspects other than treatment status.[5] However, by definition, this means that any conclusions reached can *only* be relevant for the group within this bandwidth and may not be generalisable to those further away: the approach only gives a 'local' average treatment effect. This may, therefore, only give a partial picture of the impact of a programme or intervention.

The challenges associated with implementing an Regression Discontinuity Design (RDD) counterfactual methodology

Title: The Causal Effects of R&D Grants: Evidence from a Regression Discontinuity (Santoleri et al., 2022)

[5] https://www.betterevaluation.org/methods-approaches/methods/regression-discontinuity

In evaluating the “SME instrument” programme, the study seeks to ensure the validity of the methodology and the firms included in the analysis through a series of analyses such as:

Bandwidth Variation: Varied the ‘bandwidth’ – the range around the threshold used to define treatment and control groups – to test the sensitivity of findings to the choice. Narrow bandwidths offer precise comparisons near the cut-off; wider bandwidths enable broader comparisons and increase the sample size.

Automatic Bandwidth Selection: The authors implement a method suggested by Calonico et al. (2017) for selecting an optimal bandwidth which considers the trade-offs between bias and variance, and is useful where sample sizes around the threshold are limited.

Placebo Thresholds: The authors test whether ‘false’ thresholds yield zero impacts (as would be expected) to validate that the outcome effects can be attributed to the grants rather than other factors.

Instrumental Variable (IV) Design

An alternative quasi-experimental approach is the use of **Instrumental Variables (IVs)**. An IV is a third factor, separate from the intervention being studied, that influences whether a participant receives an intervention but (unlike a control variable) is not directly linked to the outcome of the study. A valid IV must satisfy two primary conditions:

1. it should not correlate with the characteristics of the treatment and control groups that are not accounted for in the model, a condition known as exogeneity; and
2. it must influence participation in the treatment, a condition known as relevance.

Put simply, for an IV design to be successful, it is essential to find an instrument that strongly affects selection into the programme but is not correlated with any characteristics affecting outcomes.[6] If a valid instrument can be identified, the IV approach offers in principle a powerful way to determine the causal impact of an intervention as part of an evaluation.

In our review, we came across one evaluation study which used an IV approach to identify the causal impact of human capital on innovation outcomes in Colombia. More details of why an IV was used in this study are described below.

Use of Instrumental Variables (IV) to address the issue of endogeneity in identifying causal impact of human capital on innovation outcomes

Title: Human capital, innovation and productivity in Colombian enterprises: A structural approach using instrumental variables (Ramirez et al, 2019)

[6] HM Treasury (2020), *Magenta Book Annex A Analytical Methods for use Within an Evaluation* (https://assets.publishing.service.gov.uk/media/5e96c41a86650c2dd9e792ea/Magenta_Book_Annex_A._Analytical_methods_for_use_within_an_evaluation.pdf)

The paper explores the Research and Development (R&D) –innovation–productivity linkage for the Colombian manufacturing industry, focusing on the role of human capital to affect the innovation behaviour and productivity of a firm. It uses firm-level data on 6,326 Colombian companies from the Development and Technological Innovation Survey and the Annual Manufacturing Survey.

The paper applies an IV study design to address the endogeneity issue when including human capital in a model to understand the relationship between R&D investments, innovation, human capital and productivity. The preferred instruments used are the lagged value of what the authors defined as "exogenous human capital" (a firm's proportion of workers with technical or higher studies, excluding those who carry out R&D tasks) and the lagged value of "endogenous human capital" (a firm's proportion of workers carrying out R&D tasks). The methodology assumes that the proposed IVs affect the decisions on the firm's human capital but are not related to any decision on R&D investment in the current and the previous year.

In all the models presented in the study, the effect of R&D expenditure on innovation, and the impact of innovation on productivity, are always positive and highly statistically significant. The importance of the IV approach is clear as the researchers estimate models with and without the instrument. Compared to the model without the instrument, the instrumented model shows that the effect of R&D on innovation was nearly 70% higher.

Challenges of the methodology

The main challenge when implementing an IV design is finding a valid instrument that satisfies the two conditions set out above. The challenge is that most factors that affect participation in a programme are also likely, in some way, to relate directly to the outcome variable.[7] As such, satisfying the assumption of a 'valid' instrument can be restrictive. This difficulty may explain why an IV approach is not frequently observed in the set of studies we reviewed for this rapid evidence assessment. The way the study using an IV design dealt with this challenge is described below.

Challenges associated with finding a valid Instrumental Variable (IV) for estimating causality

Title: Human capital, innovation and productivity in Colombian enterprises: A structural approach using instrumental variables (Ramirez et al, 2019)

As set out above, the instruments used are lagged values of firm-level exogenous human capital (the share of workers with technical or higher studies, excluding Research and Development (R&D) workers) and endogenous human capital (R&D workers as a share of the total workforce).

[7] Adapted from https://counteringdisinformation.org/topics/surveys/4-evaluative-research-counter-disinformation-programs

The authors carry out a number of different statistical tests to ensure first that there is a need for an IV approach (endogeneity) and that the instruments are valid in terms of their strength in determining the endogenous variable, and that they exclusively affect outcomes through this channel and not any others:

- **Strength**: Testing that the instruments are strong influencers of the outcomes. The strength of instruments is tested using the 'weak identification test' such as the Cragg-Donald Wald F statistic. A strong instrument is closely correlated with the endogenous explanatory variable but not with the error term.
- **Exclusivity**: Testing that the instruments affect the outcome only through the explanatory variables and not through some other unobserved pathway. Exclusivity is testing through over-identification tests, such as the Sargan or Hansen J test.
- **Endogeneity**: Testing whether explanatory variables are correlated with the error term, indicating potential bias and necessitating the use of IV. To evaluate endogeneity, the Durbin-Wu-Hausman test is commonly used.

Difference-in-differences (DiD)

While our review found examples of random and quasi-experimental counterfactual methods being applied to Research and Innovation (R&I) evaluations, the conditions for them to be valid can be restrictive. Much more prevalent in our review, therefore, were counterfactual methodologies where the control group is selected to be similar to the treatment group, but without being established through fully randomised or quasi-random approaches. These approaches then compare observed changes in outcomes for treated and control groups to show causal impacts of an intervention – the **Difference-in-Differences (DiD)** approach. Variations of this were common in our review: we reviewed 10 applied studies using DiD across a range of R&I policy areas.[8]

Fredriksson and de Oliveira (2019) provide an excellent, non-technical account of DiD as an evaluation methodology, which we draw on for this summary. The DiD estimate of a treatment effect in effect subtracts the change in outcomes for the control group from the before and after change for the treatment group. This 'double difference' can be calculated whenever pre- and post-treatment data is available for both groups. One example, applied to a UK Small and Medium Enterprise (SME) funding programme, is illustrated below.

Using Difference-in-Differences (DiD) to estimate causal impact of an Small and Medium Enterprise (SME) Research and Development (R&D) funding programme on financial and employment metrics, and the propensity to export

Title: Evaluation of Smart (SQW, 2015)

[8] (SQW, 2015; Kuhn, 2013; Røtnes, 2017; Bloch, 2014; Falck, 2019; Gao et al., 2021; Bonev & Agnoli, 2019; Goller, 2020; Zhang, 2019; PPMI, 2017)

The report evaluated ‘Smart’, a funding instrument developed by Innovate UK in 2011, for SMEs to engage in R&D projects that may lead to the development of new products, processes, and services. Smart encompassed three grants:

- **Proof of market:** offering a maximum grant of £25,000 to cover up to 60% of project costs to support activities such as market research, market testing, competitor analysis, analysis of IP position, and initial planning to take the project to commercialisation;
- **Proof of concept**: up to a maximum grant of £100,000 for up to 60% of costs to support activities such as initial feasibility studies, basic prototyping, specialist testing, IP protection, investigating production and assembly options; and
- **Development of prototype**: offering a maximum grant of £250,000 to cover up to 35% of project costs for medium-sized enterprises and 45% for small and micro enterprises for activities such as small demonstrators, Intellectual Property (IP) protection, trials and testing, market testing, marketing strategies, identifying routes to market, product design work, and pre-clinical studies.

To evaluate the impact of the programme, the paper employed a Difference-in-Diffferences (DiD) analysis with a focus on firm-level outcomes in terms of turnover, employment, R&D expenditure and propensity to export. The study employs both a two-period DiD approach (comparing the pre-Smart award period with the post-Smart period); and a pooled DiD analysis, which included pre-Smart data and two sets of post-Smart data: the observed period and the forecasted period. They also looked at segmented samples based on firm-level characteristics and award/project characteristics.

The authors exploit the fact that the characteristics of Smart helped to identify a control group: firms passing the grant scoring threshold but not receiving funding due to budget limitations rather than the quality of their application. This resulted in a sample of firms that applied unsuccessfully to the grant but do not differ significantly to those that were successful, allowing for the creation of a valid counterfactual group using applicant data.

While the approach mimics the random allocation of an Randomised Controlled Trial (RCT), with an exogenous factor – each funding round’s budget – affecting whether a firm receives the grant, there may still be some differences between the treatment and control groups. The treated firms still had a higher application score on average, which meant treated firms with very high scores had to be excluded to create comparable samples, which saw around 7% of firms dropped from the sample. The control group was also extended to cover firms just below the grant threshold and those just above but not funded on budgetary grounds. This results in a sample of 293 treated and 189 untreated firms.

The pre-treatment data on outcome indicators was obtained from the financial statements of applicant firms for the three years prior to applying for funding. The post-treatment data was obtained from surveys that were sent to treated and untreated firms that captured their financial metrics two years after receiving the grant and firms’ forecast of their own performance in three years.

Overall, their analysis showed limited evidence of the effect of Smart on all the key outcomes of interest when analysing the full sample even after including covariates as control variables. However, when looking at the pooled sample, the study finds strong statistical evidence for an impact on employment and turnover. There is less strong evidence when it comes to R&D expenditure and no evidence for longer-term propensity to export.

DiD was one of the most frequently used counterfactual methodologies in the studies reviewed as part of this rapid evidence assessment, and is intuitive to explain as a concept which may explain its attractiveness. It overcomes the challenges that come with biases that arise when using either a simple cross-sectional treatment-control comparison (where differences between the groups may not be accounted for) or a before-and-after study focusing just on the treatment group (where changes in outcomes may have materialised anyway but there is no comparison made to justify this). By combining insights from cross-sectional treatment-control comparisons and before-after studies, DiD provides more robust identification.

In its simplest application, DiD uses an interaction term in a regression model between time (pre- and post-intervention) and group (treatment and control) indicators which measures the treatment effect. One of the studies reviewed evaluated an innovation subsidy programme for SMEs in Germany using a DiD approach, where the treatment and control groups were defined by regions. This is described further in the box below.

Application of Difference-in-Differences (DiD) counterfactual methodology with interacted fixed effects

Title: Evaluating a place-based innovation policy: Evidence from the Innovative Regional Growth Cores Programme in East Germany (Falck et al., 2019)

The paper evaluates the Innovative Regional Growth Cores (IRGC) scheme, developed by the German Federal Ministry of Education and Research in 2001 as part of the broader programme “Entrepreneurial Regions”, aimed at promoting “regional innovation alliances in Eastern Germany.”[9] It subsidises “collaborative development and commercialization projects of firms and public research institutes co-located in regions in Eastern Germany, with the explicit goal of generating local spill-overs to promote regional economic development.” The programme subsidised participants, including private firms, universities, and public research institutes. Private businesses received 54% of the IRGC grants, with an average grant of €377,000.

To evaluate the programme, the authors conducted a Difference-in-Differences analysis. The counterfactual is obtained from firms that were located in eligible East German regions that were not awarded subsidies through the IRGC programme. The authors use “interacted fixed effects”, which combine region- and firm-specific effects to explore if regional impacts on firms vary based on firm characteristics. The sample consists of 231 directly treated firms, 3,069

[9] https://www.innovation-strukturwandel.de/strukturwandel/en/entreprenurial-regions/programme/programme_node.html.

indirectly treated firms (located in regions with at least one IRGC project), and 2,099 untreated firms (outside the targeted regions).

The paper finds that the programme did not lead to a measurable increase in overall Research and Development (R&D) spending for treated firms after the subsidies ceased, and there were no significant effects on R&D staff, overall employees, and turnover for directly treated firms in the observable post-treatment period. Additionally, there were no quantifiable local spill-over effects on non-subsidised innovative firms, and no economically meaningful effects on regional economic development were discovered.

Issues with DiD can arise when there are differences in observable characteristics between the two groups that also affect the outcome variable. These differences can lead to a biased estimation of the impact of the intervention since the changes over time might not be solely due to the treatment, but also due to these pre-existing differences between the control and treatment groups.

To address this issue, DiD in evaluation is often combined with **Propensity Score Matching (PSM)**, a methodology that isolates the intervention effect on the outcome variable by matching treated and control observations with similar characteristics. It does so in two stages. First, a *propensity score* is estimated for all observations (treatment and control) which predicts the probability of receiving treatment given the characteristics observed by the researcher. Second, the propensity score is used to match each treated observation to one or more control observations with similar scores. This approach aims to ensure that, when analysing the effect of a treatment or intervention, the researcher is comparing similar individuals (firms, researchers, etc.) to increase the similarity of the treatment and control groups and give further confidence in the reliability of the counterfactual analysis.[10]

The approach has different variations which we observed in our review. The most common approach used **is nearest neighbour matching** (Kuhn, 2013; Røtnes, et al., 2017; Gao et al., 2021) which pairs individuals in the treatment group with those in the control group who have the closest propensity scores. There are other approaches proposed by the literature, as **matching with stratification** (Abdia et al., 2017) which involves dividing participants into strata or groups based on their propensity scores, then comparing outcomes within these strata to estimate the treatment effect.

Most of the papers identified that applied DiD, with or without matching, studied innovation subsidy schemes for firms and had different firm and workers' performance metrics (sales,

[10] When the outcome variable of interest is a change in a variable such as quantity of research output or firm turnover comparing a pre- and post-intervention period, then the PSM approach in effect delivers a DiD analysis for a selected control group designed to look observably similar to the treatment group. If the common trends assumption holds, any residual differences in pre-treatment outcomes (e.g. level of turnover) can be 'differenced out' though in principle, if the matching procedure is effective and includes pre-treatment outcomes, there should be no residual statistical difference between the two groups.

patent activity, wages, etc.) as outcomes of the study.[11] However, we also found papers aimed at understanding the effect of a funding programme in academic research, which delved into outcomes on research quality and impact, and effects on researchers careers' and the prestige of their host institutions, summarised below.

Application of Difference-in-Differences (DiD) and Propensity Score Matching (PSM) to the evaluation of academic research funding programmes

Title: Assessment of the Union Added Value and the Economic Impact of the EU Framework Programmes (PPMI, 2017)

This paper evaluates the European Union's (EU's) Framework Programme 7 (FP7). FP7 was the main pan-European Research and Development (R&D) funding programme from 2007 to 2013, with a budget of over €50 billion, providing support ranging from academic research to firm innovation across a wide range of fields. The paper evaluates the effect of FP7 on scientific research, firm innovation, social impact, and the economy as a whole.

To analyse the impact of the programme on academic research outputs, the authors carry out surveys directed at successful and unsuccessful applicants, collecting data for 994 research groups that received the grant and 336 that did not. The authors observe some differences in terms of sectoral composition between their survey sample and those research groups that actually responded to the survey. To correct for this, they apply non-response weights, assigning a higher weight to those research groups with a lower probability of responding to the survey.[12]

This data was used to match comparable researchers by estimating the probability of receiving treatment. To do so, the authors used variables on the researchers' characteristics such as the sum of the evaluation criteria scores received in the grant application, whether the researcher had participated in a EU innovation funding programme before, the employment and budget size of the research team before receiving the grant, and the core scientific discipline of the research group.

After the matching, the effects of the funding were analysed through DiD regressions on a variety of outcomes on the research groups' size, scientific and commercial impact, and connectivity to other research groups. The results show that research units grew more in terms of employment and budget than non-funded units, and had higher impact publications.

[11] Røtnes (2017), Bloch (2014), Falck et al., (2019), Gao (2021), Agnoli, (2019), Goller et al. (2020), Zhang (2019), PPMI (2017).

[12] The probability was calculated through a logistic regression on the research groups' scientific field, country, and project characteristics as described in Valliant et al. (2013).

Challenges of the methodology

While the theoretical assumptions for the validity of DiD approaches may be less demanding, it is still important that studies can demonstrate they are met.

For example, DiD requires that factors outside the intervention (and other things that can be controlled for) affect the treatment and control groups in the same way, such that the control group represents a good approximation of what would have happened to the treatment group if it were not supported. This is difficult to validate in practice.

The method relies on ‘parallel trends’ assumption: in the absence of support, the treatment group would have followed the same post-treatment trend in the outcome(s) of interest as observed for the control group. This assumption is, by nature, untestable, and is mostly demonstrated by comparing trends pre-intervention for the two groups to assess whether they follow similar patterns.

A more formal approach to check the parallel trends assumption holds is to conduct ‘placebo regressions’, where the DiD method is applied to data in periods before the reform. These should not yield any significant effects. Verifying this assumption does, however, require pre-treatment data for the treatment and control groups, and therefore imposes additional resource and data requirements on any evaluation.

Another important assumption is the Stable Unit Treatment Value Assumption (SUTVA), which implies that there should be no spill-over effects between the treatment and control groups.

Implementing a Difference-in-Differences (DiD) approach - Determining whether the parallel trends assumption holds

Title: Evaluating a place-based innovation policy: Evidence from the Innovative Regional Growth Cores Programme in East Germany (Falk et al., 2019)

As discussed above, the paper evaluates the Regional Growth Cores (IRGC) programme using a Difference-in-Differences (DiD) approach. They use an event study design to assess whether the parallel trends assumption holds.

The analysis is detailed across three levels: directly treated firms, indirectly treated firms (those in regions where treated firms are located), and regional outcomes. For directly treated firms, the study observes significant pre-trends in key outcomes like Research and Development (R&D) expenditures, personnel, and turnover, which are not present for indirectly treated firms or at the regional level.

The event study model captures dynamic treatment effects and pre-trends, with treatment effects estimated for different post-treatment periods and adjustments made for potential linear pre-trends among treated firms. It also introduces ‘semi-dynamic’ DiD approaches to evaluate direct and indirect subsidies, including region-specific linear trends for indirect effects and regional analysis.

Combining DiD with statistical matching approaches like PSM helps to identify comparable control groups, which may give additional confidence that the common trends assumption holds, but comes with added challenges. PSM requires data on the drivers of participation and a clear conceptual understanding of them to justify the choice of matching model. This adds to the data and resource requirements of an evaluation. In practice, the model of participation may be limited to a set of factors that can be observed in data, but may exclude key drivers that cannot be observed. The approach also requires there to be substantial overlap between the propensity scores of those units which have benefited from the programme and those that have not (the 'common support' assumption). If either of these two conditions are not met, PSM is not a suitable methodology for estimating causal effects of an intervention.[13]

2.1.2 Mixed methods and qualitative methodologies

Our rapid evidence assessment also explicitly sought examples of qualitative and theory-based approaches to evaluation of Research and Innovation (R&I) policies, and innovations in those methods, recognising that these are commonly used reflecting the difficulties of evaluating R&I interventions (see section 1.1).

Judging the robustness of these approaches is less straightforward than the Maryland scale; in general, robustness will depend on the relevance of findings to the evaluation questions, whether views on additionality have been sought, whether alternative drivers of outcomes have been considered, the scope to triangulate across different sources of evidence, and the quality of the approach taken drawing on best practice guidelines.[14]

The evidence reviewed identified **mixed methods** that bring together quantitative and qualitative approaches, and methodologies that provide a standalone research approach emphasising qualitative research.

Mixed methods approaches

Our rapid evidence assessment identified a series of papers combining quantitative counterfactual analysis (in all cases we reviewed, Difference-in-Differences (DiD) with

[13] In practice, observations outside the common support can be excluded from analysis. This, though, limits the generalisability of findings only to observations within the common support.

[14] Supplementary guidance to the Magenta Book provides a framework for assessing qualitative evidence (https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/190986/Magenta_Book_quality_in_qualitative_evaluation__QQE_.pdf). The framework provides a 'checklist' of 18 questions to judge the quality of qualitative research. While a systematic review of each of these for a given evaluation study is unlikely to be proportionate, application of the four central principles outlined in the framework (whether the findings are contributory, defensible, rigorous in conduct and credible in claim) is a useful guide.

matching) with **a series of in-depth interviews and case studies of the intervention's recipients.**

The strength of the approach is that quantitative methods can determine *whether* there was an effect and the size of the effect, while qualitative methods help understand *mechanisms* through which effects were generated. Quantitative results can be triangulated with qualitative findings and vice versa. Triangulation, as a qualitative research strategy, is the use of multiple methods or data sources to develop a comprehensive understanding of a research problem or to test validity through the convergence of information from different sources (Dawadi et al., 2021).

Mixed methods are particularly valuable where an evaluation benefits from combining multiple data sources and types of data. For example, in complex contexts where identifying control groups is challenging or we lack sufficient observations for quantitative (counterfactual-based) evaluation, theory-based evaluations using mixed methods can be the best feasible approach. Mixed methods approaches are thus found in large and multi-faceted programs, such as the analysis of nationwide high-cost research and innovation support schemes in the EU (e.g. PPMI, 2017) and different countries, as the UK (e.g. SQW, 2015), Norway (e.g. Røtnes et al., 2017) and Denmark (Bloch et al., 2014). We summarise two R&I studies below, one focused on business innovation, and another on academic research.

Use of mixed methods approach to evaluate innovation clusters in Norway

Title: Evaluation of Norwegian Innovation Clusters (Røtnes et al., 2017)

The paper evaluates the Norwegian Innovation Clusters, a programme that funds the development of industrial clusters – a geographical concentration of interconnected businesses and institutions in a specific sector, fostering collaboration, innovation, and competitive advantage – to help generate innovation hubs. The programme is divided into three subprograms: ARENA, targeted at immature clusters, granting firms with annual funding of approximately £100,000–200,000 for a period of 3–5 years; Norwegian Centres of Expertise (NCE) for the development and internationalisation of existing clusters, providing funding of £300,000–450,000 for 10 years; and Global Centres of Expertise (GCE), for mature clusters with a global position, giving funding of £600,000–750,000 for 10 years.

The paper applies a difference-in-difference and a matching approach to evaluate the effect of the programme within a mixed methods framework. The authors collected public data from 16 public funding agencies, providing them with a rich dataset on firms' characteristics and funding received. This results in a sample of 1,068 participating companies – 47% having participated in ARENA, 37% in NCE, and 16% in GCE. Using data on firms' total assets and number of employees, the authors estimate their probability of receiving treatment (the propensity score). They then match each treated observation with up to five untreated observations that are most observably similar based on the propensity score, removing from the dataset those matches considered poor due to the units matched being outliers or significantly different in their covariates (nearest neighbour matching with trimming).

After matching observations, the authors conduct a Difference-in-Differences (DiD) regression, estimating the impact of the programme as the weighted average of the treatment effects of the matched observations. The authors analyse the effect of the innovation clusters, both for the individual subprograms, and in aggregate. They find that firms enrolled in the cluster program grow faster in employment and sales during their first three years of participation but that there is no significant effect afterward. They also find that the results are more positive for ARENA (immature clusters) than for NCE (mature clusters).

The authors mention that, although it would be desirable to estimate the effects per cluster, they lack sufficient data to carry out this analysis. To address this gap, the evaluation conducts in-depth interviews with members of seven cluster organisations to gauge the effectiveness of the cluster facilitation and the extent to which participation in the cluster had influenced the development of new collaborative relationships, Research and Development activities, and the overall performance and innovation capabilities of the firms involved.

The interviews highlighted the pivotal role of cluster facilitators in nurturing effective collaboration among cluster members, helping them navigate business, research, and political domains as a key driver of success of cluster initiatives. Active enterprise participation in cluster activities emerged as a vital determinant of innovation and collaboration outcomes, with engagement levels directly influencing benefits derived from the project.

Application of mixed methods to identify the causal impact of research grants on academic research

Title: Developing a methodology to assess the impact of research grant funding: A mixed methods approach (Bloch et al., 2014)

The paper evaluates the research grants of the Danish Council for Independent Research over the period 2001–2008, which entailed approximately 2,600 grants valued at $600 million being awarded to 1,600 principal investigators covering all main fields of science. The paper evaluated outcomes on the researchers' careers and the impact of their work. The authors conduct an analysis that combines Propensity Score Matching (PSM) and Difference-in-Differences (DiD) for the analysis of career and research outcomes, and a set of quantitative and qualitative self-reported information (through survey questionnaires and case study analysis) to gain a deeper understanding of the mechanisms through which the grant affected the researchers' production. The data used include:

- **Applicant data**: To study the effect on the scientific production of the recipient researchers, the authors use bibliometric data on publications and citations of successful and unsuccessful applications. To study the career progression of both groups, they use public administrative data on employment. This data is used to create the control group for the quantitative analysis and to match the observations.

- **Survey data**: The authors collect survey data from successful and unsuccessful grant applicants to supplement the bibliometric and employment data by collecting information on research projects as a whole and to examine the full range of effects listed above.
- **Semi-structured in-depth interviews** to 20 recipient researchers and 10 council members to 'validate survey results' for the former and to examine the main objectives and overall functioning of the funding programme for the latter.

The choice of the mixed-methods evaluation design was motivated the type and quality of quantitative data available for analysis. Although the authors did have data on all successful and rejected grant applicants, they did not have data on whether the grant recipients had received other grants and they had data on only the principal investigator of each grant. The combination of these data sources allows the authors to get a comprehensive view of the effect of the research programme using DiD with PSM (with successful applicants as the treated group and rejected applicants as the control group) and case study analysis. The study evaluated two broad sets of outcomes:

Career impact measures:

- Econometric analysis shows that grant recipients had a 7 percentage point higher probability of reaching Professor level compared with rejected applicants.
- Three out of four survey respondents who received grants said that the grants had led to other project applications or grants, around 60% said the grants had been 'essential' for their careers.

The interviews suggest that these positive effects might be even higher over the longer term. Several respondents pointed out that the new professional opportunities and prestige from receiving the grant had created a 'snowballing effect in their careers'.

Research impact measures:

The econometric analysis identifies limited evidence of a positive impact in publications and citations, though the authors caution that it was hard to validate whether researchers in the control group had received other forms of support. The bibliometric data was also heavily skewed, with a small number of researchers capturing most of the publications and citations, which complicated the matching procedure and generated small sample sizes.

- The survey found that more than nine in ten researchers said that the grant gave them the opportunity to produce very novel research results within their area, and almost two-thirds that the grant had resulted in unexpected results of great importance for their field.
- The interviews highlight that, while the grants are also perceived to have been beneficial to the development of the recipients' research projects, most interviewees focused their responses on the impact the grant had through career advancement.

Qualitative approaches

Our review identified **Qualitative Comparative Analysis (QCA)** as an applied approach to Research and Innovation (R&I) evaluation that draws explicitly on qualitative methods. QCA, an example of a theory-based evaluation approach, aims to understand the conditions in the intervention, or wider environment, which are critical to an outcome of interest. The approach involves identifying the combination of conditions which appear to be necessary for outcomes to be achieved.

One variant of QCA is **crisp-set Qualitative Comparative Analysis (csQCA)** (Palinkas et al., 2019). In csQCA, evaluators examine the presence or absence of specific conditions within a 'case' (a unit of analysis) in a binary way, whether the condition is either present or not. If cases were firms receiving grant funding for innovation, for example, this could include whether firms have a manager with previous innovation management experience. Combinations of conditions can then be examined to assess which are necessary for outcomes to occur, such as whether a patent was granted in the past year.

While csQCA uses qualitative data (presence/absence of conditions) and allows for the identification of patterns and configurations, it does involve some quantification in terms of analysing combinations and assessing consistency or coverage of cases. It therefore brings some of the features of quantitative analysis to qualitative study.

Overall, the strength of QCA in the context of evaluating R&I is that it allows the evaluation of complex programmes with several combinations of causal pathways rather than a single cause through which an outcome is observed. Additionally, QCA can be applied with relatively small and simple datasets.

Challenges of the methodology

Mixed-method and qualitative research designs are not free from limitations or challenges.

The decision of which mix of methods to deploy, for example, is often a difficult one to make; different mixes could impact the analysis and interpretation of findings, and therefore the conclusions drawn. The approach may often be based on pragmatism and feasibility, or methods that evaluators are familiar with, rather than strong *a priori* assumptions.

Integrating data from different methods also requires an approach to how the findings will be combined to draw overall conclusions ('triangulated'). This can often be challenging, in particular where different methods suggest different conclusions (Bloch et al., 2014).

Among the qualitative approaches reviewed as part of this evidence, QCA is a theory-based approach: the factors considered need to be based on a logic or theory of 'what matters' which therefore also needs to be developed for the approach to be deployed successfully. The approach also requires careful coding about whether factors are present (or not) in each case, which may sometimes be difficult to determine in an objective way. The results are also context-specific based on a selected set of cases and may not be generalisable (Zhang et al., 2019).

2.1.3 General challenges with implementing counterfactual methodologies

Apart from the methodology-specific challenges highlighted above, our review identified two broad themes of challenges that are applicable across many counterfactual methodologies used for evaluating impact evaluations:

Sample size

One of the major constraints to any quantitative impact evaluation is **sample size**, which is integral to determining statistical significance and statistical power of a quantitative evaluation methodology, or the probability that the results from the hypothesis we observe are not purely based on chance. A larger sample size generally increases the likelihood of achieving statistical significance because it improves the precision of estimates, reduces standard errors, and enhances the sensitivity of the hypothesis testing. However, including larger samples in an evaluation is likely to increase costs, in particular where data needs to be collected directly from programme participants.

Characteristics of Research and Innovation (R&I) programmes and their outcomes

As mentioned in section 1.1, Research and Innovation (R&I) programmes have characteristics which can contribute to certain counterfactual methodologies being more challenging to identify causal impacts. Here we briefly outline how some of these characteristics interact with the application of particular methods, with reference to the literature reviewed.

Given the need to demonstrate good use of public funds, R&I investments often use ‘merit-based’ or ‘excellence-led’ awards. This makes the use of Randomised Controlled Trials (RCTs) impossible, where the key requirement is the random assignment of treatment. There may also be ethical risks if support is denied to certain beneficiaries at random, in particular when awarding large amounts of funding. This suggests that randomisation may only be applied in limited circumstances relating to R&I policies. A recent study (Dalziel, 2018), summarised below, has highlighted the circumstances most relevant to the use of RCTs for R&I programmes.

When are Randomised Controlled Trials (RCTs) most appropriate for the evaluation of R&I programmes?

Title: Why are there (almost) no randomised controlled trial-based evaluations of business support programmes? (Dalziel, 2018)

The paper discusses the very limited use of RCTs to evaluate business support programmes, identifying three key reasons. They also identify, for some of them, potential ways to introduce their use:

- **Random allocation of support**. Randomly allocating support will greatly diminish the effectiveness of support as the capabilities of companies is a key determinant of outcomes, in addition to creating the ethical issues arising from randomly denying access to potentially beneficial business support. RCTs are more feasible where the potential impacts are more modest (microcredit and business training interventions are references as examples) compared with cases where interventions are more expensive or far-reaching where RCTs might be difficult to justify politically and socially.

- **Standardised treatments**. RCTs require standardised treatments. This can be problematic where business support interventions are designed and tailored to address specific challenges, or employ specialist knowledge and seek to build long-term relationships. RCTs may be more viable when support is purely financial, which is inherently 'standard'. It may also be possible to standardise some knowledge-based support in key intervention processes and functions without unduly compromising the support quality.

- **Reliability of results**. Business support programmes aim to produce outliers –firms whose performance is exceptional. When outliers are present, very large samples will be required to produce reliable results. RCTs often do not account for heterogeneity because they are designed to test the average effect of an intervention across all participants, assuming a relatively homogenous response.

R&I programmes are also in general complex in nature. Complexity can arise from different aspects, such as:

- High variability in the nature or the monetary amount of the treatment (even within a given programme) where R&I support is tailored to different researchers or firms, and/or builds on past interventions, the effects of which may still be unfolding. This can make attribution to specific interventions challenging.
- An ecosystem with multiple external factors influencing success such as beneficiaries receiving multiple forms of public and private support, or support being delivered in waves, which creates difficulties in establishing attribution or in clearly defining 'control' groups (Santoleri et al, 2022).
- Intangible outputs (e.g. knowledge created) which are hard to quantify, meaning evaluation may sometimes focus on what can be measured but not necessarily what 'matters' (Poorkavoos et al., 2016).
- Heterogeneity in treatment effects. Given the inherently uncertain nature of R&I policy, interventions often produce a few 'success' stories among many 'failures'. Hence, methodological approaches that estimate average treatment effects can fail to capture the factors that might explain success (Dalziel, 2018).
- Long-term nature of impacts which means establishing attribution is difficult given the influence of external factors over longer time horizons, and the data and resource

requirements to capture data from both treatment and control groups over longer time periods (Kleine et al., 2022).

These complexities can make the use of a single quantitative counterfactual approach unsuitable as no single counterfactual can apply across the range of evaluation questions or outcomes that may be of interest in evaluating a particular R&I programme. This is a key driver of the common use of qualitative, mixed-methods and theory-based approaches to R&I evaluation, as outlined earlier.

Some of these challenges discussed in this section have driven particular innovation in counterfactual methods when evaluating R&I programmes, which is the focus of the next section.

2.2 Recent innovation in the use of counterfactual methods to evaluate Research and Innovation programmes

Our review identified recent innovations in counterfactual methodologies applied to innovation policy evaluations, responding to some of the challenges set out in section 2.1 above. We summarise three broad types of innovation identified: designing the programs ex-ante to achieve randomisation, improving existing methodologies to respond to issues relating to Research and Innovation (R&I) evaluation, and in emerging methods that are, or could be, applied to R&I evaluations.

2.2.1 Research and Innovation programme design allowing randomisation

The main example of the first category is the randomisation of treatment ex-ante to be able to evaluate the programme ex-post through a Randomised Controlled Trial (RCT).

Despite the challenges associated with an RCT, the review found evidence of a recent application of RCTs in the context of innovation support for Small and Medium Enterprises (SMEs) (Kleine et al., 2022). This was an evaluation of a voucher programme for SMEs in the UK that granted them with up to £5,000 to engage in collaboration with academic experts that could boost their innovation know-how. The study was summarised in section 2.1.1.

This study focuses on a relatively inexpensive intervention where concerns about not focusing on an ‘excellence-led’ award in terms of value for money may have been low. The study was also able to focus on short-term outcomes that could reasonably be expected to appear within one or two years of the support being received (improvement of their internal innovation management processes), but still faced barriers around identifying longer-term outcomes which would necessitate additional data collection over a longer period (number of patents or commercialised products), and still faced hurdles in terms of being able to collect data from the treatment and control groups. As a result, the more general feasibility of using RCTs to evaluate R&I programmes appears to be relatively low (Dalziel, 2018).

2.2.2 Improvement of existing methodologies

Quantitative methods

As discussed in previous sections, the most common approach identified in our review to the evaluation of innovation policies is **Difference-in-Differences (DiD)** both with and without matching. Our review found examples of innovations in implementing these methods.

With regards to the **estimation of the propensity score** which often accompanies a DiD counterfactual estimation approach, the traditional approach in the existing literature has been through a logistic regression. This econometric approach has the advantage of being simple to explain and implement. However, it can lead to biased estimates in complex settings – i.e., in which there are many factors (or covariates) that could drive participation relative to the sample sizes; in which there are several treatments; or where there are non-linearities in the relationship between covariates and the treatment assignment (Gao et al., 2021). These can lead to a traditional logistic model 'overfitting' the data – that is, estimating a model that mistakenly captures noise or randomness as part of the underlying relationship – and mis-specifying the relationship between treatment and the factors driving it.

Recent theoretical evidence has proposed methodological alternatives that can improve the estimation of the propensity score in highly-complex settings. The methodologies identified rely on Machine Learning (ML) methods. One prevalent example is **Generalised Boosting Models (GBM)**, which have been found in the theoretical and applied evidence reviewed (Gao et al., 2021; Abdia et al. 2017). This approach combines multiple 'simple' models predicting treatment status into a 'strong' model through an iterative process. By integrating GBMs into the Propensity Score Matching (PSM) process, researchers can potentially improve the accuracy of propensity score estimation and achieve better balance between the characteristics of the treatment group and matched control group.

Using a Machine Learning (ML) methodology to improve Propensity Score Matching (PSM)

Title: Can public R&D subsidy facilitate firms' exploratory innovation? The heterogeneous effects between central and local subsidy programs (Gao et al., 2021)

The paper studies a public Research and Development (R&D) subsidy programme for firms in the Jiangsu Province, China. The programme, carried out between 2010 and 2014, aimed at fostering "exploratory innovation" in high-tech Small and Medium Enterprises (SMEs). Funding was delivered through national funding (1-2m CNY, approximately £100,000–£200,000, per supported firm) and the regional funding (0.3-0.5m CNY, approximately £30,000–£50,000, per supported firm). The study uses data from a survey conducted by the Jiangsu Provincial Department of Science and Technology, and obtains a final sample of 240 firms and 1,198 firm-year observations.

The authors study the impact of local and national R&D funding on firms' share of exploratory innovation patents. They use both PSM and DiD. The PSM procedure is conducted through a machine learning approach, generalised boosting model. Generalised Boosting Models (GBM) allowed for the estimation of the propensity score in the setting of the study, as this model is capable of handling multiple treatments (whereas traditional PSM methods only apply in the case of a single treatment), with non-linear relationships between the covariates and treatment assignment, by producing probabilities that represent the likelihood of being assigned to *each* of the treatments.

This methodology achieves this by building a strong prediction model through the combination of multiple simple models in an iterative process. In practice, it starts by estimating a simple decision tree that models the probability of treatment. A decision tree is a "supervised machine learning algorithm, used for both classification and regression tasks, which models decisions based on a set of conditions or rules and organises them in a tree-like structure" (Ciuffreda et al., 2023). Subsequently, more trees are estimated to provide the best fit possible to the residuals of the model of the previous iteration. Individual models may perform poorly in matching, but when models are aggregated they become stronger and more able to provide robust estimates.

The authors find that public R&D subsidies significantly increase firms' level of exploratory innovation, boosting it by almost 15%. The value of GBM allows the paper to explore different treatments: they find that local funding is more effective, increasing exploratory innovation in treated firms by 19% compared with 10% for national funding. The effect is also found to be stronger in industries with highly specialised agglomeration.

Other machine learning models discussed in the literature are **random forests** and **Least Absolute Shrinkage and Selection Operator (LASSO)** (Goller et al., 2020). Similar to GBM, random forest models estimate and combine multiple model to create more stable and accurate predictions. LASSO is a technique that aims to select a subset of covariates that are the most relevant for predicting outcomes. In the context of matching, this seeks to avoid the overfitting problem which can arise when there are many covariates that might drive selection into treatment relative to the number of observations.

Goeller et al. (2020) compare the performance of these two methodologies against a logistic regression to match observations in an evaluation of a training programme for the long-term unemployed in Germany. They conclude that LASSO might provide a benefit to the accuracy of matching procedures, as it provided the best estimations of the propensity score both in large and small sample size settings. The performance of random forest was not as strong – while it performed well in large sample sizes, they performed poorly in settings with either a low sample size, or low number of treated observations.

Overall, different machine learning techniques, such as LASSO or GBM, appear to be promising in applications of matching methods where there areh small sample sizes or many potential characteristics affecting the outcome of the policy intervention, broadening the number of settings in which these techniques can be reliably used. Recent evidence has also suggested the use of new econometric techniques for matching. Ziyu et al. (2022) proposed ‘propensity score differentiated matching’ to improve estimation in complex settings, although we did not identify practical application of this approach in evaluation studies.

The evidence reviewed also highlighted innovations on how to match individuals and estimate the treatment effect after the propensity score has been determined. A notable example is the **‘doubly robust’ estimator** (Abdia et al., 2017), useful when there are concerns about mis-specification of either the matching model, or the model used to estimate outcomes. This estimator relies on applying a procedure called Inverse Probability Weighting (IPW) (Kreif et al., 2013), by which the importance (weight) of observations in the outcome model is determined by their propensity score. In this approach, treated units receive lower weights if they have a high probability of being treated and control units receive lower weights if they have a high probability of not being treated. The rationale of this methodology is obtaining a more similar treatment and control group by focusing on those with a similar probability of being treated.

Our review also identified a recent study (Roth et al., 2023) focused on theoretical innovations to DiD, mostly in terms of how it can be applied in settings in which the pre-treatment parallel trend assumption needs to be relaxed. The paper considers improved diagnostic tools to detect the violation of parallel trends, recognising that even if pre-treatment trends in outcome(s) of interest for treatment and control groups are parallel, this does not ensure that post-treatment trends can be assumed to be parallel in the absence of support. The main take-away of this paper, summarised below, is that new tests and approaches are being developed that would allow for an application of DiD in a wider set of evaluations

Recent methodological advances in Difference-in-Differences (DiD)

Title: What’s trending in difference-in-differences? A synthesis of the recent econometrics literature (Roth et al., 2023)

The authors describe recent theoretical advances in Difference-in-Differences (DiD) and how they affect its application. The two key areas of these advances are allowing for multiple periods and variation in treatment timing across observations, and dealing with potential violations of the parallel trends assumption.

Allowing for multiple periods and variation in treatment timing

The coefficients of traditional DiD analyses are may not provide a good estimate of the average treatment effect when observations are treated at different times, a common situation in Research and Innovation (R&I) interventions (e.g. when funding is rolled out in phases). In particular, the traditional DiD approach would fail to distinguish those who are *never* treated from those who are *not yet* treated in looking at a treatment group observed at a particular time.

The study proposes the **Generalised Parallel Trend Assumption**, an extension of the parallel trends assumption for settings with multiple time periods and varying treatment timings. The assumption requires that the parallel trends assumption, typically applied in a two-group (treatment/control), two-period (before/after) scenario, holds across all combinations of periods and for all groups treated at different times. This means that, in the absence of treatment, the average outcomes for all groups (defined by their treatment timing) would have evolved in parallel. This assumption requires that the expected difference in outcomes between any two periods for any group should be equal to that of any other group, assuming no treatment had occurred.

Potential violations of parallel trends

The study also discusses approaches to relax the parallel trends assumption, a foundational assumption in DiD analysis that might not always hold in practice. Key points on relaxing this assumption include:

Conditional Parallel Trends: One approach to increase the credibility of parallel trends is to make them conditional on a set of covariates, which may vary across treatment and control groups and over time. Once these factors are accounted for, it may be more credible to assume that parallel trends holds. This adaptation assumes treatment is almost randomly assigned conditional on these covariates, adding a layer of robustness to the analysis against potential violations of parallel trends. The literature has proposed different approaches to obtaining conditional parallel trends, including regression-based and doubly-robust approaches.

Allowing for a certain degree of violation of the assumption: The authors discuss different ways in which the parallel trend assumption could be relaxed, while maintaining the validity of the comparison between treated and control group. They propose two approaches:

- Imposing that that the violation of parallel trends is not larger than the maximal pre-treatment violation of parallel trends.
- "Bracketing" the trend of the treatment group between the one of two control groups, allowing for the estimation of bounds on the average treatment effect of the treated.

Qualitative methods

In terms of **qualitative methods**, recent studies (Poorkavoos et al., 2016; Pickernell et al., 2019; Zhang et al., 2020) have applied an innovative approach to Qualitative Comparative Analysis (QCA) to build on some of the challenges faced with crisp set Qualitative Comparative Analysis (csQCA): **fuzzy-set qualitative comparative analysis (fsQCA)**. The main innovation of this methodology is to allow each case examined to have relevant factors affecting outcomes being present *to different degrees* (e.g. using a Likert scale, common in research), rather than in a binary sense of the factor being present or not as used in csQCA.

Like standard QCA, fsQCA is a methodology “that enables the systematic analysis of several cases to identify causal patterns that influence an outcome” (Castelló-Sirvent et al., 2020). Each condition in a case is assigned a value that represents its degree of presence or absence (for example, ”level of funding of technical training for employees“ in a study trying to understand what makes Small and Medium Enterprises (SMEs) innovate successfully), allowing for more nuanced categorisation than simple yes/no distinctions. The values that the firms, regions, or individuals in the study show for these variables are then assigned to different categories (e.g., full-, partial-, or non-membership to a category), based on their percentile with respect to the full sample. Systematic analysis (known as ‘configurational analysis’) is used to find patterns in these conditions associated with outcomes being realised. Finally, fsQCA synthesises the findings to highlight the most relevant combinations of conditions, offering insights into causal relationships in complex systems.

The outcome of these studies is the identification of a combination of conditions that are considered to be necessary and/or sufficient to lead to the outcome of the study. A necessary condition is one that must be present all (or almost all) of the time for the outcome to occur; the absence of the near-necessary condition prevents the outcome. A sufficient condition is one that, when present, causes the outcome to occur all or most of the time.

The application of this methodology requires detailed information on the subjects of the study. It does not require, though, a particularly large sample size, which makes it a fitting approach to study small sample outcomes. In the reviewed evidence, two papers (Pickernell et al., 2019; Zhang et al., 2020) applied fsQCA in the context of Research and Innovation (R&I) studies in England and China respectively. The box below summarises the key features of the study of English Local Enterprise Partnerships (LEPs).

Innovation in the use of Qualitative Comparative Analysis (QCA) approach to estimate causality

Title: Innovation performance and the role of clustering at the local enterprise level: A fuzzy-set qualitative comparative analysis approach (Pickernell et al., 2019)

The paper studies the impact of England's 39 Local Enterprise Partnerships (LEPs): collaborations between Local Authorities and businesses aimed at driving economic growth and job creation within specific regions by developing local economic priorities and leading economic projects. The paper aimed to study the level of radical and incremental innovation among Small and Medium Enterprises (SMEs) within the LEPs, and factors driving this. The paper focuses on UK firms in particular sectors, including software supply and consultancy, and manufacture of chemicals and chemical products.

The study applies fuzzy-set qualitative comparative analysis (fsQCA) to identify whether LEPs led to businesses increasing the share of their turnover arriving from innovative products or services. Information was gathered on variables that affected this. This included factors such as the share of firms within the LEP engaged in innovation, measures of interaction between

Higher Education Institutions and businesses/community, regional industrial specialisation, and measures of local-level clusters. To facilitate the fsQCA, values were assigned to three groups depending on whether the LEP was high, low or around the median value for each measure. The data used comes from the Department for Business Innovation and Skills 2015 report, 'Mapping Local Comparative Advantages in Innovation Framework and Indicators',[15] and includes a wide range of innovation-relevant economic data representative of the English LEP geographic areas between 2008 and 2012.

The fsQCA identified that it is a multi-faceted set of conditions, as opposed to single conditions, that are associated with English LEPs' SMEs innovation performance. The paper finds that the existence of clustering in conjunction with other factors, in particular prior levels of innovation, were key success factors. On the other hand, the impact of fostering local universities' third sector activities proves to be more peripheral in explaining the success of the programme overall.

2.2.3 Development of new methodologies

Our review identified two main methodologies that have been growing in prominence in policy evaluation in the last decade, with the potential for application to R&I interventions: synthetic controls, and causal machine learning approaches. It is important to note that applications of these techniques are more recent; our review found evidence of their application in evaluations that looked at other programmes (e.g. health-related interventions as discussed in Bouttell et al. (2017) or the application in Chernozhukov et al., 2018, to examine the role of institutions on economic growth based on Amemoglu et al., 2001) rather than for R&I programmes.

Synthetic control methods (SCM)

In the previous section, we identified a key limitation of Difference-in-Differences (DiD) is finding a control group that follows a pre-treatment parallel trend in the outcome variable with the treated group. We also identified fuzzy-set Qualitative Comparative Analysis (fsQCA) as a potential, yet limited, solution to the study of outcomes in settings with small sample sizes.

Recent literature (Bouttell et al., 2017; Degli Esposti et al., 2020; Kim et al., 2020; Autant-Bernard et al., 2022) has discussed the **Synthetic Control Method (SCM)** as a way to address these issues. As summarised in Abadie (2021), SCM is based on the idea that a combination of untreated observations often provides a more appropriate counterfactual for a treated observation than any single untreated unit. SCM seeks to formalise the selection of the control units using a data driven procedure. The methodology creates a control group that resembles the characteristics of the treatment group by weighting untreated observations. These weights are selected such that the resulting synthetic control resembles the treated unit

[15] https://assets.publishing.service.gov.uk/media/5a80ac8c40f0b62302694dce/bis-15-344-mapping-local-comparative-advantages-in-innovation-framework-and-indicators.pdf

before the intervention. Given this, the synthetic control group is developed to ensure the parallel trend assumption can be obtained.

This approach allows a DiD approach in settings with small sample sizes: a control group can be constructed specifically for even a single treated observation. Hence, SCM can be found in studies that have a very small sample of treated units, such as those that analyse outcomes at a regional level . It can also be a valuable approach when the researchers expect significant heterogeneity in the treatment effect across treated units, for example, where innovation support is targeted at very large companies. This is because SCM is able to provide an assessment of the impact of the intervention on individual beneficiaries (Autant-Bernard et al., 2022). SCM have also been applied in settings with larger sample sizes (summarised in Bouttell et al., 2017).

Application of a Synthetic Control Method (SCM) approach to identify impacts of Technological Research Institutes (TRI) programme in France

Title: Evaluating the impact of public policies on large firms: A synthetic control approach to science-industry transfer policies (Autant-Bernard et al., 2022)

The paper evaluates the French Technological Research Institutes (TRIs) programme. Implemented in 2010, it focuses on interdisciplinary research institutes that bring together companies and universities around a strategic research programme. The paper focuses on the impact of the programme on large firms, and studies a series of outcome variables such as Research and Development (R&D) investment, outputs, and their collaboration with smaller firms.

The dataset consists of 26 companies, 22 untreated and 4 treated. The sources of information that the authors used were the French annual survey of firms, a French employment database, and R&D survey on participants, and internal firm data from the innovation programme. Given the small sample size and the potential high heterogeneity in the effect of the programme between the treated firms, the authors study them individually, constructing for each a synthetic control from the untreated companies. However, the researchers only achieve a good synthetic control – i.e., a weighted combination of the untreated firms that satisfies the parallel trend assumption – for three of the four companies, thus being forced to discard the fourth one. Subsequently, they estimate the programme’s effect through an interrupted time series analysis.

The study finds different programme effects between the firms on different outcomes, suggestive of significant levels of impact heterogeneity, identified above as one of the key features of Research and Innovation (R&I) interventions which can be challenging for evaluation. The evaluation finds that the intervention had a positive significant effect on R&D for one firm, no effect for a second, and a negative effect for a third.

The methodology, nevertheless, includes several limitations.

- SCM require a good fit in the pre-intervention outcome of interest between treated units and the synthetic control. This can be difficult to achieve where there are outliers in characteristics that affect outcomes (Bouttell et al, 2017). If finding a suitable synthetic control is not possible, then the methodology cannot be implemented or will be limited (as seen in Autant-Bernard et al., 2022, where the approach could not be applied to one of the four treated firms).
- The implementation of SCM can also lead to 'over-fitting' the model, in which a close pre-intervention fit is achieved but where it may be harder to justify that the same synthetic control is a reasonable counterfactual post-intervention.
- The results can be highly dependent to the choice of covariates and the weighting in the control group, hindering the generalisability of the conclusions.

Machine Learning (ML) methods

Another avenue of recent innovations in evaluation research is the use of **double / debiased machine learning estimators** (Chernozhukov et al., 2018), **causal forests method** (Athey and Wager, 2018), and **Artificial Intelligence (AI)-based counterfactual reasoning models** (Xia et al., 2023). These are models that rely on weak theoretical requirements and allow for estimation in highly complex settings with a large number of covariates ('high-dimensional' settings). The main contribution of Chernozhukov et al. (2018) is to offer a novel procedure for estimating causality using ML and to perform statistical inference in high-dimensional settings. There are many methods that can be used to estimate double / debiased ML estimators – random forests, neural networks, or least absolute shrinkage and selection operator (LASSO).

A recent working paper (Agnoli and Bonev, 2019) was the only application of ML methods to R&I evaluation identified in our review, and the study has not yet been published. The authors evaluate the causal impact of technology standards, which are a set of requirements aimed to ensure interoperability between products, on innovation. The main challenge they face is of endogeneity: it may be that innovation leads to standardisation, rather than the other way round. They use ML methods – neural networks and random forests – to address this challenge. They find positive local treatment effects of technology standardisation activities on innovation (proxied by patenting applications by firms seen after the standardisation activities) which is significant five years after standardisation. The positive effect ranges from 113 to 158 more patent applications (depending on a parameter values in the model) after controlling for year, country and technology fixed effects.

3 Lessons on more robust impact evaluations of Research and Innovation programmes

In this section we offer some overall lessons learned for UKRI based on the findings in section 2 for the robust impact evaluation of Research and Innovation (R&I) policies.

Innovations in counterfactual methodologies for evaluating R&I programmes have come about in recent years to address some of the challenges associated with implementing evaluations of R&I programmes. We lay out the key lessons coming out of the review across the research questions below.

- Randomised Controlled Trials (RCTs) remain the most robust methodology to identify a counterfactual given its experimental nature. However, their use has been scarce when it comes to R&I evaluation.. Our review found one instance of an RCT being applied to evaluate an innovation programme. **UKRI could consider an RCT to evaluate programmes where randomisation of treatment is feasible ex-ante and ethical, and treatment can be standardised across all beneficiaries at the time of delivering the programme.** Such approaches are most likely to be useful for relatively small-scale, often financial, interventions that can be delivered to a large number of beneficiaries to ensure sufficient sample size, and where not receiving support on a random basis is unlikely to pose significant ethical challenges.
- Quasi-random counterfactual methodologies are the second most robust approaches, based on the Maryland Scientific Methods (MSM) scale, to identify a counterfactual. R&I programmes are sometimes geared towards identifying beneficiaries based on their capabilities or need. Such programme objectives can use an objective criteria such as a scoring to identify the beneficiary group. Regression Discontinuity Design (RDD) has been a common methodology used to exploit such programme designs. **Where R&I programmes include eligibility based on a clear scoring guide, or other objective criteria, UKRI could consider the use of robust RDD approaches to estimate the counterfactual.** However, it is important that the criteria are reasonably adhered to during implementation, otherwise the ability to use RDD is limited.
- The evidence highlighted that Differences-in-Differences (DiD) is a popular counterfactual methodology used in R&I evaluations, balancing robustness with less restrictive requirements compared with RCTs and quasi-random methods. The traditional DiD model and its implementation have seen recent innovations which allow ways to reduce the burden of satisfying some of its core assumptions such as parallel trends, and in improving how techniques such as matching can be used alongside DiD in selecting control groups. **UKRI could consider including these alternative estimation strategies in the DiD methodology if the programme and data characteristics threaten to violate the assumptions or lead to sub-optimal matching. Additionally, UKRI may also consider heterogenous treatment effect estimation in programmes where treatment is rolled out in a staggered manner rather than a simple one-off**

treatment which allows for differentiating between comparisons of treated and not-yet-treated observations, and already-treated observations. The coefficients of traditional DiD analyses are likely to not provide a good estimate of the average treatment effect when observations are treated at different times.

- Complexity is inherent in most R&I programmes, which can limit the ability to implement purely quantitative counterfactual evaluation methods robustly. A mixed-method approach allows evaluators to understand complex interactions and interdependencies between different activities and beneficiaries more comprehensively beyond the causal estimate measured through quantitative methodologies in isolation. **UKRI could consider using a mixed-method approach using quantitative methodologies and qualitative methodologies together to evaluate complex R&I programmes which are multi-faceted and will benefit from using data from different types of sources to generate comprehensive evaluation.**
- **UKRI could consider qualitative theory-based approaches such as Qualitative Comparative Analysis (QCA) and the more recent innovation of fuzzy-set Qualitative Comparative Analysis (fsQCA) when the focus is on the conditions that lead to a given outcome**. The particular value of fsQCA over QCA is allowing a more nuanced consideration of whether factors influencing success are present or not, and where the potential sample sizes of treated and control observations may be too small for robust quantitative counterfactual methods.
- Sample size is an important consideration for any counterfactual methodology. Certain characteristics of R&I programmes can make it challenging to obtain a large enough sample for quantitative counterfactual analysis. Synthetic Control Method (SCM) allows for counterfactuals to be constructed even when the numbers treated are small. **UKRI could consider SCM as a counterfactual methodology for evaluating R&I programmes where one aggregate unit, such as one university or limited number of businesses are exposed to an intervention, and where high quality secondary data pre- and post-intervention is available both for treatment and control units.**. **It should also be considered as an approach when the evaluators expect significant heterogeneity in the treatment effect across treated units due to multiple treatments, as could be the case for R&I programmes that target very large companies with tailored support.** However, it should be noted that this methodology is currently an active area of research, and successfully constructing a synthetic control is not guaranteed.

4 Reflections on the evidence

This section sets out our reflections on the evidence base reviewed which underpins this rapid evidence assessment, including where there are evidence gaps, and what this means for the conclusions drawn.

4.1 Assessment of the evidence reviewed

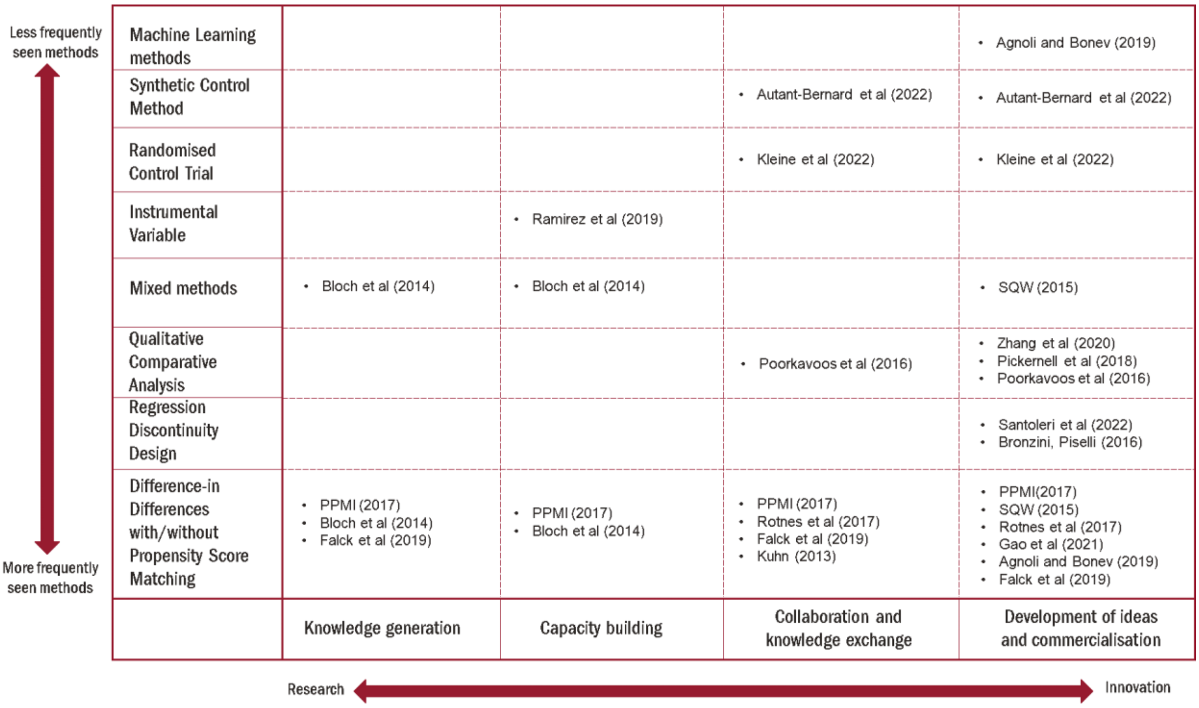
The findings presented in section 2 are based on the synthesis of existing evidence found as part of this review. This evidence has been identified with a targeted approach using the rapid evidence assessment protocol given the limited time horizon within which this study was conducted. While rapid evidence assessments do allow for a structured and rigorous search, and an assessment of the quality of the evidence identified, they cannot be exhaustive in the way that a systematic review is designed to be. Therefore, our conclusions, recommendations, and findings can only be drawn from the 30 studies reviewed, rather than being seen as fully representative of all the potentially relevant research.

The shortlist of studies we reviewed was chosen carefully, in agreement with UKRI, to provide broad coverage across a range of methodologies designed to understand counterfactuals (including experimental, quasi-experimental, theory-based, mixed methods and qualitative approaches) and across a broad range of UKRI areas of focus and activities. In terms of the applied papers which showcased how an evaluation study implemented a counterfactual methodology in the context of an Research and Innovation (R&I) intervention, we were able to capture a good spread across methodologies and areas of R&I as shown in Figure 4. The chart positions the evaluation studies focused on R&I programmes from our review in terms of the observed frequency of the methodology (on the vertical axis) and the area of UKRI focus (on the horizontal axis) ranging from areas more focused on early stage research to those more focused on nearer-to-market innovation.

The spread of studies across UKRI focus and evaluation areas demonstrates that different counterfactual methodologies can be applied to different R&I programme evaluations. The review found evaluation studies looking into academic research and capacity building (more research-oriented) as well as programmes which are geared towards turning research into innovative products and services. Overall the review had more applied studies focused on evaluating innovation-focused outcomes than research-focused.

Most of the applied studies we reviewed used Difference-in-Differences (DiD), Regression Discontinuity Design (RDD), Qualitative Comparative Analysis (QCA) and mixed method approaches to estimate the impact of R&I programmes. This is consistent with the findings in section 2 which highlights the challenges that come with evaluating R&I programmes that make it difficult to use certain quantitative counterfactual methodologies. However we also identified a growing (albeit still quite small) body of evaluations studies focussed on R&I programmes deploying the use of innovations in existing counterfactual methodologies such as fuzzy-set Qualitative Comparative Analysis (fsQCA), and more emerging methodologies such as Synthetic Control Method (SCM) and Machine Learning (ML) to evaluate R&I programmes. There is no one-size-fits-all approach or a one-to-one mapping between an R&I programme area and optimal counterfactual methodology to evaluate its impact. This implies that when considering approaches to evaluation, UKRI should not exclude particular approaches *ex ante* based on the programme area; rather, consideration will need to be given to the characteristics of the intervention, how it was delivered and the data available to inform evaluation design.

Figure 4 Range of counterfactual methods identified in applied Research and Innovation evaluation studies, by type of intervention



Machine Learning methods				• Agnoli and Bonev (2019)
Synthetic Control Method			• Autant-Bernard et al (2022)	• Autant-Bernard et al (2022)
Randomised Control Trial			• Kleine et al (2022)	• Kleine et al (2022)
Instrumental Variable		• Ramirez et al (2019)		
Mixed methods	• Bloch et al (2014)	• Bloch et al (2014)		• SQW (2015)
Qualitative Comparative Analysis			• Poorkavoos et al (2016)	• Zhang et al (2020) • Pickernell et al (2018) • Poorkavoos et al (2016)
Regression Discontinuity Design				• Santoleri et al (2022) • Bronzini, Piselli (2016)
Difference-in Differences with/without Propensity Score Matching	• PPMI (2017) • Bloch et al (2014) • Falck et al (2019)	• PPMI (2017) • Bloch et al (2014)	• PPMI (2017) • Rotnes et al (2017) • Falck et al (2019) • Kuhn (2013)	• PPMI(2017) • SQW (2015) • Rotnes et al (2017) • Gao et al (2021) • Agnoli and Bonev (2019) • Falck et al (2019)
	Knowledge generation	Capacity building	Collaboration and knowledge exchange	Development of ideas and commercialisation

Source: *Frontier Economics*

Note: *These 16 papers are a subset of the papers reviewed in-depth as the figure shows the applied studies focussed on R&I programmes only. The rapid evidence assessment shortlisted 30 papers for in-depth review which consisted of 22 applied studies and 8 theoretical papers overall..*

Apart from applied evaluation studies focussed on R&I programmes, the review also looked at six applied studies which evaluated programmes in other areas such as health which had similar outcomes to those from R&I programmes, and hence were considered to be relevant for UKRI. The rationale for including these studies in the review was to highlight the more innovative counterfactual methodologies being used for impact evaluations which may not necessarily feature in R&I programme evaluations in the current literature. In general, these counterfactual methodologies focussed on innovations in SCM (e.g. Bouttell et al., 2017) and ML methods for causal inference (Agnoli and Bonev, 2019).

The review also included eight theoretical papers focussed on innovations in counterfactual methodologies. The rationale for including these papers in the review was to identify recent theoretical innovations in identifying causal impact. Given the lag between research and its eventual application in empirical settings, the inclusion of these studies helped to identify the more recent developments that can potentially be applied to future R&I evaluations. Overall, the evidence from the theoretical papers reviewed as part of this study provided insights into the growing interest in applying more advanced ML approaches in the context of identifying causal impacts in evaluation studies (Goller, 2020).

4.2 Evidence gaps identified

In prioritising the studies we reviewed for this rapid evidence assessment, some possible gaps include:

- **A focus on English language papers**. Given the targeted nature of the review, the papers identified using the search strategy were all in the English language. Therefore the findings from this review are biased to the extent that they are based on evidence that has been reported in or translated to English. R&I policy interventions and theoretical developments in counterfactual methodologies are relevant globally. The full extent of insights therefore may not be clear by only looking at English language-based studies.
- **Bias towards published literature**. Although the review incorporated an approach to include unpublished literature to the extent possible, published literature, either in an academic journal, government website or working paper, still accounted for 29 out of the 30 papers that formed part of the evidence base for this review. Lags between completion of a study and their publication mean that there might well be recent developments in this topic which are not yet reflected in the published literature.
- **Limited number of search keywords and their combinations**. Given the time horizon within which this review was conducted, there was a trade-off between the breadth of literature that can be identified through different but relevant search keywords and the depth of literature to consider by exploring more papers within the same keywords. It was decided that a more pragmatic approach would be to give preference to using a wider set of search keywords and to focus on the breadth of studies, rather than looking at the depth within different approaches or areas of UKRI focus.

Further research could build on or validate the findings of this review by targeting non-English studies, further engagement with R&I agencies and researchers to assess whether there are unpublished studies or work in progress, or looking more in-depth at particular methodologies or areas of UKRI focus on providing a more systematic account of the literature.

5 Conclusions

This section summarises the core findings for each of the research questions based on the evidence reviewed. It lays out how the evidence reviewed helps to answer each of the research questions which guided the study.

Counterfactual methodologies used in Research and Innovation evaluation

The evidence reviewed as part of the shortlist of 30 studies included both theoretical and applied evaluation studies. These covered a broad range of robust counterfactual methodologies across quantitative, mixed methods and qualitative approaches. Looking at the applied Research and Innovation (R&I) evaluation papers specifically, the review found evidence across all of UKRI's relevant programme areas and these were spread across all the counterfactual methodologies.

Our review found that a few counterfactual methodologies – Difference-in-Differences (DiD), Regression Discontinuity Design (RDD), Qualitative Comparative Analysis (QCA) and mixed methods – are more commonly applied to evaluate R&I programmes. We also identified more applied studies on the innovation end of the R&I spectrum overall. DiD is one methodology found to be widely applied in evaluation studies across all the types of R&I programmes that are within UKRI's focus. Less frequently observed methodologies in our review included Instrumental Variables (IV) as well as more recent methods such as Synthetic Control Method (SCM) and Machine Learning (ML) methods in causal inference.

Outside of the R&I programmes, our review also identified applied studies which evaluated other programmes or outcomes that overlapped with UKRI's focus areas such as wider socio-economic growth. These provide relevant insights into potential use of counterfactual methodologies like SCM in future R&I evaluations under the right programme context.

Strengths and success factors of counterfactual methodologies used in evaluating Research and Innovation programmes

Theoretically, the strengths of quantitative approaches can be assessed by their ability to establish a robust counterfactual to estimate causal impact. Based on the Maryland Scientific Methods (MSM) scale, Randomised Control Trials (RCTs) are considered the most robust methodology in counterfactual evaluation, followed by quasi-experimental methods such as Regression Discontinuity Design (RDD) and Instrumental Variable (IV), and then methods which establish an observably-similar counterfactual group against which trends in outcomes can be compared, such as Difference-in-Differences (DiD) (with or without matching methodologies). However when it comes to implementing them to evaluate R&I programmes, their success in identifying a valid counterfactual depends on how well the programme design fits the methodological requirements to produce robust causal estimates and the available data to fulfil the assumptions that are demanded by the methodologies. More robust methods are more burdensome in terms of the programme design or data needed, which means the circumstances in which they can be used are somewhat more limited, or additional time and resource is needed to ensure the requirements are adhered to.

In the case of mixed methods approaches, they derive their strengths from the fact they are able to build upon a wider set of frameworks and available data to build evidence about the counterfactual. These approaches can be tailored to the very specific nature of individual interventions, and allow evaluators to explore mechanisms for change (i.e. *how* an intervention is affecting outcomes of interest). Mixed methods, in particular, bring together information from both quantitative and qualitative approaches. This allows for evaluating more complex programmes, common to Research and Innovation (R&I), and triangulation across a range of evidence and data sources to draw conclusions, which can produce more credible findings which could strengthen evaluation conclusions and implications for policy.

Challenges associated with different methods to measure counterfactuals in Research and Innovation evaluation

The main challenges with quantitative counterfactual methodologies to evaluate the impact of Research and Innovation (R&I) interventions are the availability of a good sample size to produce robust statistical inference, and satisfying the theoretical assumptions associated with the methodologies in complex settings. These assumptions can be difficult to meet, and not always possible to test convincingly.

Our evidence also looked at mixed methods and qualitative approaches, and these are not without limitations. For example, a mixed method which combines quantitative and qualitative approaches is a lengthy process as each of the research methods consumes time and resource. Additionally, the above challenges interact with specific characteristics of R&I programmes that add to the challenges for robust impact evaluation of these programmes.

Innovations in recent years to establish counterfactuals in Research and Innovation evaluations

Recent innovations in evaluating Research and Innovation (R&I) programmes aim to address some of the challenges identified. The innovations identified in the papers reviewed for this study focus on three main themes:

- **Programme design for experimental evaluation**: Randomised Control Trials (RCTs) are considered highly robust in theory when it comes to counterfactual development. R&I programmes where randomisation is feasible given the programme objectives, and where ethical concerns are limited, are well suited for this.

- **Improvements in existing methodologies**: Theoretical developments in the literature have sought to relax the assumptions related to a range of quantitative counterfactual methodologies, with potential implications for R&I evaluation. These innovations attempt to improve the internal validity of findings by looking at ways to generate unbiased estimators of the causal impact using any given counterfactual methodology. For example, improved diagnostic tools can help evaluators detect the violation of parallel trends, an essential assumption in a Difference-in-Differences (DiD) approach.

- **Development of new methodologies**: More recent methodologies beginning to be applied to R&I such as Synthetic Control Method (SCM) appear to offer particular advantages for evaluating impact where treatment is applied to a small samples such as particular sector, location or type of institution, and where trend data on key outcomes is available over time. Additionally, Machine Learning (ML) methods are also being employed to identify counterfactuals. These methods rely on weak theoretical requirements and accommodate more complex data features such as non-linearity and high number of covariates.

Lessons learnt for UKRI

The rapid evidence assessment identified the following key lessons for UKRI to consider for future Research and Innovation (R&I) evaluations, in particular to guide the choice of possible methodology.

Table 2 Lessons for consideration for UKRI across key counterfactual methodologies identified form the review of studies

Counterfactual methodology	Lessons for UKRI based on the review
Randomised Control Trial (RCT)	■ RCTs are theoretically strong in establishing a counterfactual, but difficult to implement for R&I interventions where there is an emphasis on quality-based award of support. ■ UKRI could consider the feasibility of RCTs in programme design based whether the methodological, data and resource requirements are proportionate and justified. The strongest case is where the ethical implications of random allocation of support are low, and where treatment can be standardised across beneficiaries. This could include small-scale financial interventions delivered to a large number of beneficiaries to ensure sufficient sample size, where not receiving support is unlikely to pose significant risk to those affected.
Regression Discontinuity Design (RDD)	■ Where R&I programmes allocate treatment based on a clear scoring guide, or where eligibility is restricted based on clearly-defined characteristics, this could enable the use of robust RDD approaches. ■ It is important that the cut-offs are adhered to, otherwise the ability to use RDD is limited.

Counterfactual methodology	Lessons for UKRI based on the review
	■ Programme design and data sharing should be set up to enable evaluators to collect data from those ineligible or just unsuccessful to enable RDD (or from broader groups such as all unsuccessful applicants for other methods using control groups) to be implemented in practice.
Difference-in-Differences (DiD) with Propensity Score Matching (PSM)	■ While commonly applied to a range of UKRI programme areas, recent innovations in the literature around DiD (e.g. the Generalised Parallel Trends approach set out in Roth et al., 2023) suggest ways to reduce the burden of satisfying some of its core assumptions such as parallel trends. UKRI could explore the application of these methods to a range of evaluations where DiD is implemented, in particular if assumptions appear hard to satisfy. ■ Other innovations have explored the theoretical application of DiD where treatment is rolled out in a staggered way (Roth et al., 2023), where traditional DiD may provide poor estimates of treatment effects. This allows for comparisons between treated, not-yet-treated, and never-treated observations. UKRI could consider these approaches where treatment is staggered.
Mixed methods and Qualitative Comparative Analysis (QCA)	■ Complexity is inherent in most R&I programmes. This may limit the feasibility or effectiveness of evaluation approaches that focus only on quantitative counterfactual methods. Supplementing them with qualitative approaches and other ways of establishing counterfactuals (such as self-reported participant views) in a mixed methods approach may be particularly suited to cases where R&I programmes have a diverse set of outcomes to evaluate, and where quantitative data may be limited or hard to collect for a control group. ■ The use of fuzzy-set QCA, an approach identified in our review (e.g. Pickernell et al., 2019), may be attractive for UKRI. It focuses on systematic analysis of conditions that lead to particular outcomes being realised, and allows for a more granular breakdown of factors (beyond a binary presence or not) than traditional QCA. ■ Consider budget and time requirements as part of the feasibility assessment of evaluation design as mixed methods can be demanding on both these fronts.

Counterfactual methodology	Lessons for UKRI based on the review
Synthetic Control Method (SCM)	■ Sample size is an important consideration for any quantitative counterfactual methodology. Many R&I interventions offer targeted support to small numbers of 'treated' researchers, institutions or firms. UKRI could consider SCM as a plausible counterfactual methodology for evaluating such programmes. In particular interventions which are very specific or unique, it may be difficult to identify comparable units that have not experienced such change. In such cases, SCM can be applied to generate a counterfactual constructed from a combination of control units to be similar to the treated unit. ■ In cases where there is heterogeneity in treatment (e.g. tailored support to large firms), SCM allows for individual assessment of the impact of the intervention. This can be helpful for UKRI to consider when evaluating beneficiaries with multiple treatments. ■ The approach relies on being able to observe outcomes for the treated unit(s) and a set of potential controls both before and after the intervention. Hence, it may be most useful where there is high quality and consistent secondary data at the relevant treatment unit level (e.g. universities, firms, sectors or areas) and over time.

Source: Frontier Economics

6 Bibliography

The bibliography provides full references for all papers and articles referenced in the report. We mark in **bold type** the studies that featured in the rapid evidence assessment.

Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59 (2), 391-425, https://doi.org/10.1257/jel.20191450

Abdia, Y., Kulasekera, K. B., Datta, S., Boakye, M., & Kong, M. (2017). Propensity scores based methods for estimating average treatment effect and average treatment effect among treated: A comparative study. *Biometrica Journal*, 59(5), 967-985 https://doi.org/10.1002/bimj.201600094

Acemoglu, D., Johnson, S., & Robinson, J. (2001). The Colonial Origins of Comparative Development: An Empirical Investigation. *American Economic Review*, 91 (5), 1369-1401. Available at: https://economics.mit.edu/sites/default/files/publications/colonial-origins-of-comparative-development.pdf

Agnoli, M. & Bonev, P. (2019). The effect of standardization on innovation: A machine learning approach. Available at: https://wwws.law.northwestern.edu/research-faculty/clbe/events/standardization/documents/agnoli_bonev_2019.pdf

Athey, S. & Wager, S.(2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, 1228-1242, https://doi.org/10.1080/01621459.2017.1319839

Autant-Bernard, C., Fotso, R., & Massard, N. (2022). Evaluating the impact of public policies on large firms: a synthetic control approach to science-industry transfer policies. *Review of industrial economics*, 4, 9-50, https://dx.doi.org/10.2139/ssrn.3619007

Bloch, C., Sørensen, M. P., Graversen, E. K., Schneider, J. W., Schmidt, E. K., Aagaard, K., & Mejlgaard, N. (2014). Developing a methodology to assess the impact of research grant funding: A mixed methods approach. *Evaluation and Program Planning*, 43, 105-117, https://doi.org/10.1016/j.evalprogplan.2013.12.005

Bouttell, J., Craig, P., Lewsey, J., Robinson, M., & Popham, F. (2017). Synthetic control methodology as a tool for evaluating population-level health interventions. *Jorunal of Epidemiology & Community Health*. 72(8), 673-678. https://doi.org/10.1136/jech-2017-210106

Bronzini, R., & Piselli, P. (2016). The impact of R&D subsidies on firm innovation. *Research Policy*, 45(2), 442-457. https://doi.org/10.1016/j.respol.2015.10.008

Calonico, S., Cattaneo, M., Farrell, M., & Titiunik, R. (2017). rdrobust: Software for regression-discontinuity designs. *The Stata Journal*, 17(2), 372-404. https://doi.org/10.1177/1536867X1701700208

Castelló-Sirvent, F., Roger-Monzó, V., & García-García, J. (2020). International economic policy: a fuzzy set qualitative comparative analysis on think tanks in the press. . *Economic Research-Ekonomska Istraživanja*, 34(1), 2609-2627. https://doi.org/10.1080/1331677x.2020.1835520

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21, 1, C1-C68. https://doi.org/10.1111/ectj.12097

Ciuffreda, I., Casaccia, S., & Revel, G. (2023). A Multi-Sensor Fusion Approach Based on PIR and Ultrasonic Sensors Installed on a Robot to Localise People in Indoor Environments. *Sensors*, 23(15), 6963. https://doi.org/10.3390/s23156963

Collins, A., Coughlin, J., Miller, J., & Kirk, S. (2015). *The Production of Quick Scoping Reviews and Rapid Evidence Assessments: a How-to Guide.* Defra and NERC. Available at: https://connect.innovateuk.org/documents/3058188/3918930/JWEG%20HtG%20Dec2015

Dalziel, M. (2018). Why are there (almost) no randomised controlled trial-based evaluations of business support programmes? *Palgrave Communications*, 4(1), 1-9. https://doi.org/10.1057/s41599-018-0069-9

Dawadi, D., Shrestha, S., & Giri, R. (2021). Mixed-methods research: A discussion on its types, challenges, and criticisms. *Journal of Practical Studies in Education*, 2(2), 25-36. https://doi.org/10.46809/jpse.v2i2.20

Degli Esposti, M., Spreckelsen, T., Gasparrini, A., Wiebe, D. J., Bonander, C., Yakubovich, A. R., & Humphreys, D. K. (2020). Can synthetic controls improve causal inference in interrupted time series evaluations of public health interventions. *International Journal of Epidemiology*, 49(6), 2010-2020. https://doi.org/10.1093/ije/dyaa152

Falck, O., Koenen, J., & Lohse, T. (2019). Evaluating a place-based innovation policy: Evidence from the innovative Regional Growth Cores Program in East Germany. *Regional Science and Urban Economics*, 79, 103480. https://doi.org/10.1016/j.regsciurbeco.2019.103480

Fredriksson, A., & de Oliveira, G. (2019). Impact evaluation using difference-in-differences. *RAUSP Management Journal*, 54 (4), 519–32. https://doi.org/10.1108/RAUSP-05-2019-0112

Gao, Y., Hu, Y., Liu, X., & Zhang, H. (2021). Can public R&D subsidy facilitate firms' exploratory innovation? The heterogeneous effects between central and local subsidy programs. *Research Policy*, 50(4), 104221. https://doi.org/10.1016/j.respol.2021.104221

Goller, D., Lechner, M., Moczall, A., & Wolff, J. (2020). Does the Estimation of the Propensity Score by Machine Learning Improve Matching Estimation? The Case of Germany's Programmes for Long Term Unemployed. *Labour Economics*, 65, 101855. https://doi.org/10.1016/j.labeco.2020.101855Kim, S., Lee, C., & Gupta, S. (2020). Bayesian Synthetic Control Methods. *Journal of Marketing Research*, 57 (5), 831-852. https://doi.org/10.1177/0022243720936230

Kleine, M., Heite, J., & Huber, L. R. (2022). Subsidized R&D collaboration: The causal effect of innovation vouchers on innovation outcomes. *Research Policy*, *51*(6), 104515. https://doi.org/10.1016/j.respol.2022.104515

Kreif, N., Grieve, R., Radice, R., & Sekhon, J. S. (2013). Regression-adjusted matching and double-robust methods for estimating average treatment effects in health economic evaluation. *Health Services and Outcomes Research Methodology*, 13, 174–202. https://doi.org/10.1007/s10742-013-0109-2

Kuhn, J. M. (2013). *An evaluation of the Danish Innovation Assistant Programme.* Danish Agency for Science, Technology and Innovation. Available at: https://ufm.dk/en/publications/2013/an-evaluation-of-the-danish-innovation-assistant-programme-en-effektmaling-af-videnpilotordningen

Palinkas, L. A., Mendon, S. J., & Hamilton, A. B. (2019). Innovations in Mixed Methods Evaluations. *Annual Review of Public Health*, 40, 423-442. https://doi.org/10.1146/annurev-publhealth-040218-044215

Pickernell, D., Jones, P., & Beynon, M. J. (2019). Innovation performance and the role of clustering at the local enterprise level: a fuzzy-set qualitative comparative analysis approach. *Entrepreneurship & Regional Development*, 31(1-2), 82-103. https://doi.org/10.1080/08985626.2018.1537149

Poorkavoos, M., Duan, Y., Edwards, J. S., & Ramanathan, R. (2016). Identifying the configurational paths to innovation in SMEs: A fuzzy-set qualitative comparative analysis. *Journal of Business Research*, 69(12), 5843-5854. https://doi.org/10.1016/j.jbusres.2016.04.067

PPMI (2017). *Assessment of the Union Added Value and the Economic Impact of the EU Framework Programmes .* European Comission. Available at: https://publications.europa.eu/resource/cellar/af103c38-250d-11e9-8d04-01aa75ed71a1.0001.01/DOC_1

Ramachandran, M. (2020). The Logic of Randomised Controlled Trials in the Social Sciences. *Social Scientist*, 48 (1/2), 41–52. Available at: https://www.jstor.org/stable/26899498

Ramírez, S., Gallego, J., & Tamayo, M. (2019). Human capital, innovation and productivity in Colombian enterprises: a structural approach using instrumental variables. *Economics of Innovation and New Technology*, 29(6), 625-642. https://doi.org/10.1080/10438599.2019.1664700

Roth, J., Sant'Anna, P. H., Bilinski, A., & Poe, J. (2023). What's trending in difference-in differences? A synthesis of recent economterics literature. *Journal of Econometrics*, 235 (2), 2218-2244. https://doi.org/10.1016/j.jeconom.2023.03.008

Røtnes, R. A., Rybalka, M., Norberg-Schulz, M., Walbækken, M., & Håkansson, A. (2017). *Evaluation of Norwegian Innovation Clusters.* Innovation Norway. Available at: https://www.regjeringen.no/contentassets/377067362c2d4f5e87cd87063bce7a62/evaluation-of-norwegian-innovation-clusters_.pdf

Santoleri, P., Mina, A., Di Minin, A., & Martelli, I. (2022). The Causal Effects of R&D Grants: Evidence from a Regression Discontinuity. *The Review of Economics and Statistics*, 1-42. https://doi.org/10.1162/rest_a_01233

SQW (2015). *Evaluation of Smart.* Available at: https://www.sqw.co.uk/expertise/innovation/evaluation-of-smart-randd-grants-impact-and-process-evaluation

Valliant, R., Denver, J. & Kreuter, F. (2013). *Practical Tools for Designing and Weighting Survey Samples*. NY: Springer-Verlag, http://dx.doi.org/10.1007/978-1-4614-6449-5

Xia, H., Muskat, B., Li, G., & Prayag, G. (2023). AI-based counterfactual reasoning for tourism research. *Annals of Tourism Research*, 101, 103617. http://dx.doi.org/10.1016/j.annals.2023.103617

Zhang, M., Li, B., & Yin, S. (2020). Configurational paths to regional innovation performance: the interplay of innovation elements based on a fuzzy-set qualitative comparative analysis approach. *Technology Analysis & Strategic Management*, 32(12), 1422-1435.. https://doi.org/10.1080/09537325.2020.1773423

Zhang, Y., Xu, B., & Rashid, H. (2019). Air Treatment Effect Assessment for Improving Vehicle Emission Standards: Counterfactual Analysis Based on Machine Learning. *Nature Environment & Pollution Technology*, 18(5). Available at: https://neptjournal.com/upload-images/NL-73-41-39-Final.pdfZiyu, Z., Kuang, K., Li, B., Cui, P., Wu, R., Xiao, J., & Wu, F. (2023). Differentiated matching for individual and average treatment effect estimation. *Data Mining and Knowledge Discovery*, 37(1), 205-227. https://doi.org/10.1007/s10618-022-00886-5

Annex A Glossary

Table 3 Glossary of key terms and acronyms

Term	Explanation
AI	Artificial Intelligence
Average treatment effects (ATE)	In an experimental study, the post-treatment difference between the means of the outcome of the study of the treatment and control groups
Bias	Systematic difference between the true expected value of a parameter and the value it is estimated with in the sample of the study
Common support	In the context of matching in a counterfactual study, the overlap in value ranges of the propensity score between the treatment and control groups
Covariates	Observable characteristics of the participants or their environment that can potentially affect the values of the dependent variable of the study
csQCA	Crisp-set Qualitative Comparative Analysis
DiD	Difference-in-Differences
EU	European Union
fsQCA	Fuzzy-set Qualitative Comparative Analysis
GBM	Generalised Boosting Models
IP	Intellectual Property
IPW	Inverse Probability Weighting
IV	Instrumental Variable / Instrumental Variables
LASSO	Least Absolute Shrinkage and Selection Operator
LEP	Local Enterprise Partnership
ML	Machine Learning: a subfield of artificial intelligence that encompasses models and algorithms that mimic a human-like learning process to improve their accuracy

Term	Explanation
Matching	In the context of statistics, the process of linking observations that show similarity in covariates relevant to predicting the outcome of the study with the aim of performing like-for-like comparisons
MSM Scale	Maryland Scientific Methods Scale, a five-point scale (5 being the highest) reflecting increased confidence in the robustness of quantitative evaluation methodologies.
Overfitting	The parametrisation of a model that follows the data points in the sample of study too closely, mistakenly capturing noisy variation as a feature of the true systematic relationship between independent and dependent variables, and, thus, hindering its predictive power.
PSM	Propensity Score Matching
QCA	Qualitative Comparative Analysis
R&D	Research and Development
R&I	Research and Innovation
RCT	Randomised Control Trial
RDD	Regression Discontinuity Design
Regression	Statistical model that aims to determine how a dependent variable is related to a series of independent variables
SCM	Synthetic Control Method
SME	Small and Medium enterprise
Statistical significance	The claim that there is a high probability that the difference in values we observe in a hypothesis test is not due to random variation. The academic convention is that significance levels above 95% are considered ‘significant’.
SUTVA	Stable Unit Treatment Value Assumption
UKRI	UK Research and Innovation

Annex B Details of rapid evidence assessment protocol

The evidence assessment followed the principles of a Systematic Review to objectively assess a body of evidence, and was undertaken over three months. It did not attempt to capture all of the relevant evidence, but to prioritise what to review based on relevance to the study.

The protocol provides transparency about how the review was designed and conducted, and how the evidence was analysed. It includes the inclusion and exclusion criteria that guided the scope of the review, the strategy for the search, refinement and extraction of evidence, and the subsequent synthesis of that evidence. We discuss each part of the protocol below.

Scope of the review: criteria for inclusion

The criteria to determine which studies were eligible for inclusion are shown below.

Table 4 Inclusion and exclusion criteria

Topic	Inclusion criteria	Exclusion criteria
Type of literature	• Academic literature such as theoretical/conceptual and applied studies • Grey literature including evaluation reports and analysis documents (internal or public) by UKRI, think tanks, government agencies and other evaluation agencies • Review studies, synthesis reports or meta-studies which focus on the methods used	• Blog posts • Webpages • Newspaper articles • Review studies, synthesis reports or meta-studies which do not detail methods used
Type of publication	• Peer-reviewed published literature in journals • Non peer-reviewed literature including working papers • Unpublished studies identified through targeted requests to Frontier and UKRI networks	• Books • Videos • Audios
Language of publication	• Focus on English-based literature but include any that are pointed out by experts and agencies approached by UKRI and Frontier networks	-

Topic	Inclusion criteria	Exclusion criteria
Geography	- No geographic exclusion, studies will be drawn from global literature.	-
Time-period	- Review will focus on studies published since 2003. Prominent older academic (mostly theoretical) literature captured via snowballing where findings continue to be relevant and where the methodology is robust	-
Research topic / methodology	- Outcomes and R&I programmes relevant for UKRI (see below) - Conceptual qualitative and quantitative methodologies for causal inference	- R&I programmes or outcomes that are not relevant for UKRI - Methodologies based on anecdotal observations

Source: Frontier Economics

The criteria around research topic ensured the review focused on literature relevant to UKRI and its stakeholders. Our understanding of the relevant programmes and outcomes is summarised below.

Types of Research and Innovation (R&I) programmes that UKRI funds

UKRI invests across the R&I ecosystem to support academic research, capacity-building, knowledge exchange, business innovation and commercialisation of solutions to national and global challenges.[16] It does so by bringing together seven Research Councils, Innovate UK and Research England.[17]

The types of R&I programmes supported by UKRI through its investments are:

- **Knowledge generation** through **basic and applied research funding** to universities to create a diverse portfolio of high-quality research. Types of funding include:
  - 'Place-based' investments that provide a portfolio of R&I investments across regions of the UK with an aim to spread the reach of investment geographically in support of wider 'levelling-up' objectives, and foster R&I clusters across the country (such as the Strength in Places Fund).
  - Investment in international R&I programmes such as the International Science Partnerships Fund.

[16] https://www.ukri.org/what-we-do/

[17] https://www.ukri.org/councils/

  - A focus on multi- and inter-disciplinary R&I programmes and those which support wider government strategic objectives through larger-scale investments in programmes such as the Strategic Priorities Fund.
  - Investment not just in projects but also in individual researchers with promising, high-potential research agendas through e.g. Future Leaders Fellowships or other fellowship programmes.
  - A wide range of open and guided calls for funding for specific research programmes across the different UKRI councils.
- **Capacity building** by investing in skills and career development of students, researchers and innovators as well as on research infrastructure such as laboratory facilities, equipment and digital resources. This includes:
  - Studentships and funding for PhDs and post-doctoral researchers.
  - Training schemes run by different Councils.
  - Support which is focused on increasing diversity in the research base.
- **Collaboration and knowledge exchange** between academia, businesses and international partners to incentivise cross-sectoral partnerships and catalyse local and regional innovation capabilities. This includes:
  - Funding to support business-academia and knowledge transfer partnerships which aim to ensure research helps meet industry needs.
  - Joint training initiatives across academia and businesses.
  - Setting up institutions that promote knowledge exchange such as Innovation and Knowledge Centres.
  - Block funding to Higher Education Institutions to carry out knowledge exchange and commercialisation activities.
  - Project funding and training specifically for knowledge exchange and commercialisation including follow-on funding and fellowships.
  - Support to develop networks across a range of sectors, technologies and areas of research interest.
- **Development and commercialisation of innovative business ideas** to support emerging concepts that have the potential to have a positive impact on the UK economy through increased productivity and sustainable growth.
  - Public seed funding for early stage businesses.
  - A wide range of open and guided calls for funding for specific innovation programmes.
  - ‘Challenge-led’ investments seeking to provide holistic R&I and wider support to key societal missions, such as the UKRI Challenge Fund investments.
  - Large scale initiatives focused on innovation such as Catapults and Catalysts.
  - Support for business development to help address innovation challenges through programmes like Innovate UK EDGE.

- Facilitating engagement with policy and regulatory leads to help address barriers to innovation and commercialisation.
- Funding and facilitating access to capital facilities which help to test, scale and validate innovate ideas to help bring them to market.

We focussed on literature that captures R&I programmes across these broad areas as part of this study.

Outcomes of relevance

We took a broad approach to identifying evidence relating to outcomes of particular interest to UKRI. Our starting assumption was that studies which evaluate interventions that map into the types of policies and programmes set out above will by their nature consider outcomes that are of potential interest to UKRI.

Where we identify studies that do not themselves evaluate 'UKRI policies' but where the counterfactual methods could in principle be adapted to them, we will consider whether the outcomes covered in the evaluation would also be relevant to UKRI. This takes a broad perspective as in principle UKRI is interested in a wide set of outcomes. For example, UKRI's strategic themes include broad policy areas such as net zero, security and resilience, opportunities and outcomes, health and wellbeing, and infectious disease.[18] UKRI also spotlight a wide range of potential outcomes and impacts of interest across the Councils.[19]

Based on the most recent UKRI Annual Report, we note that outcomes around the following areas would be particularly relevant, but as discussed we will take a broad view:

- **People and careers around R&I** for example outcomes around:
  - Skills
  - Equity, Diversity and Inclusion in terms of the R&I workforce and landscape
  - Public engagement with and acceptance of innovative ideas
- **Place and regional impact** for example outcomes around:
  - The location of R&I activity, investments and facilities
  - International collaboration and R&I activity
- **Research, knowledge and ideas** for example outcomes around:
  - Quantity and quality of research outputs
  - Collaboration and inter/multi-disciplinarity of R&I
  - Knowledge transfer and dissemination
- **Innovation and commercialisation** for example outcomes around:

[18] https://www.ukri.org/who-we-are/our-vision-and-strategy/ukri-strategic-themes/

[19] https://www.ukri.org/who-we-are/how-we-are-doing/research-outcomes-and-impact/

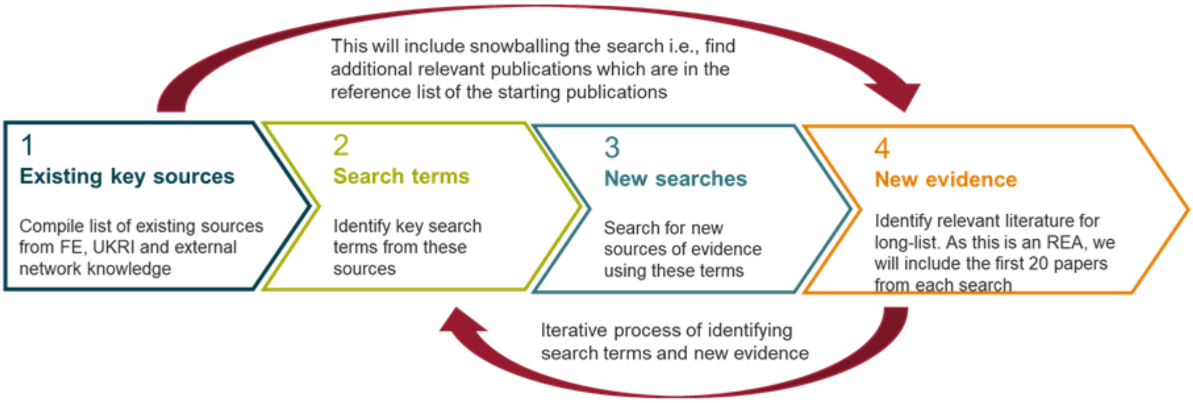
  - □ Technological development of new ideas towards commercial readiness
  - □ Co-investment and follow-on investment
  - □ Collaboration (domestic and international) to foster innovation
  - □ Spin-out of research ideas into viable commercial opportunities
  - □ Intellectual property
  - □ Diffusion of innovative ideas and technologies
  - □ Development of innovative clusters and spillovers of innovation
- **Wider socio-economic benefits** for example outcomes around:
  - □ Growth of innovative businesses (revenue, GVA)
  - □ High-skill, high-quality employment
  - □ Wider productivity and growth impacts
  - □ Social impacts including those on environment, health and wellbeing

Search strategy

Overall strategy

The overall search strategy for identifying the longlist of literature is set out in Figure 5 below. It was an iterative process drawing on existing knowledge, searches of key databases and refinement/snowballing of the search based on the results generated. The results were used to create a longlist of papers for possible inclusion in the rapid evidence assessment.

Figure 5 Summary of overall search strategy


This will include snowballing the search i.e., find additional relevant publications which are in the reference list of the starting publications
1
Existing key sources
Compile list of existing sources from FE, UKRI and external network knowledge
2
Search terms
Identify key search terms from these sources
3
New searches
Search for new sources of evidence using these terms
4
New evidence
Identify relevant literature for long-list. As this is an REA, we will include the first 20 papers from each search
Iterative process of identifying search terms and new evidence

Source: Frontier Economics

The search strategy included the following steps:

- **Step 1** – Compile a list of existing published and unpublished literature based on references from internal Frontier Economics evaluation experts, Frontier academic network, wider UKRI and government agencies (e.g. Evaluation Task Force) working on evaluations, and national and international evaluation and R&I agencies.[20] It was important that the search strategy covered these different types of evidence in order to minimise publication bias. For evaluations, grey literature and unpublished evidence were particularly important. This is because it has been demonstrated that often there is a publication bias which results in studies that do not find effects or impacts being less likely to be published (Gough et al., 2013[21]). The list below sets out examples of external institutions and evaluation portals or report links that we included in our search -

  - **Austria**: Austrian Science Fund
  - **Germany**: Fraunhofer Group for Innovation Research; German Research Foundation; IZA
  - **United Kingdom**: UKRI Evaluation database; Evaluation Task Force
  - **United Nations**: Inspection and Evaluation Division (IED)
  - This initial literature formed the basis for identifying the first set of keywords for the search process.

- **Step 2 and 3** - The topic (title, keywords and abstract) of articles in online databases were be searched using keywords (including wildcards[22]) identified and agreed with UKRI based on the review of publications from Step 1. The use of synonyms and antonyms of key words were explored where necessary. Where possible, search terms within group of similar terms were combined using Boolean Operators.
- **Step 4**: Only the first 10 hits from each iteration of the search process were selected and no links were followed from the original hit. For each search, a record of the year, the database, source location, type of paper and search terms used were recorded.

The development of keywords were an iterative process and steps 2 to 4 were repeated. Additional targeted searches were undertaken with sets of keywords suggested by UKRI to fill gaps in certain methodologies which were of interest to their work and to cover the breadth of known methods. To keep the approach agile, we did not set limits on the number of iterations of steps 2 to 4 but the search process was constrained by timelines and scope of this work, and ended when no further relevant studies are identified through additional searches.

Additionally, we were also guided by the initial list of literature identified in Step 1 to identify new evidence through snowballing which involved using the reference list in the publications. This ensured that key literature is included which could potentially be missed if only using search strings. We used free open source Natural Language Processing (NLP) tools available

[20] We included studies recommended by academic experts and an initial search of publication records of any key academics or institutes at the cutting-edge of evaluation methods (particularly with application to R&I) recommended.

[21] https://apo.org.au/sites/default/files/resource-files/2013-12/apo-nid71119.pdf

[22] Wildcards are symbols that can optionally replace a single letter in a word. For example, Web of Science wildcard is the asterix symbol (*) or question mark (?)

for literature review such as connectedpapers.com at this stage to assist the snowballing process[23]. However, this did not help identify any additional literature beyond the databases.

Once the records of each individual search were completed they were combined to give a full longlist of the evidence found, removing any duplicates. We skimmed through the title of the articles to identify any literature that was picked up by the search strategy that are obviously not relevant to the study based on the research questions and scope of review and were removed from the longlist. This list was shared with UKRI for their review.

Databases

In steps 2 and 3, the search drew on the following sources of grey and academic literature:

- Google Scholar;
- Online database Science Direct and evaluation studies databased Si-per and World Bank;
- References to unpublished literature from internal Frontier evaluation experts, Frontier academic network, wider UKRI and evaluation agencies as set out in Step 1.

Evidence sift: selecting the shortlist of studies for in-depth review

The search strategy outlined above yielded a large number of studies. Hence, the next step was to screen them to check whether they meet the inclusion and exclusion criteria. This was intended to remove articles captured by the search strings that are not relevant to the study based on the agreed research questions and scope of the review.

The screening of the long-list of studies identified was then done manually to make sure that the shortlist -

- covered a broad range of robust counterfactual methodologies;
- had a good spread of applied studies across the types of R&I programmes that UKRI focusses on; and
- captured recent innovations in counterfactual methodologies across both theoretical and applied studies

Full review of targeted text in the documents to derive short-list

After narrowing down the long-list of papers, we examined the abstract, introduction, and possibly conclusions based on a speed-reading of the articles. Anything that again didn't meet the scope of the study based on the inclusion and exclusion criteria but had passed through the first stage screening were excluded.

[23] See https://libguides.princeton.edu/c.php?g=1171670&p=8559773

This process yielded a shortlist of 30 relevant papers, 25 which were included in the short-list directly and 5 which were kept as a reserve. We reviewed the list against other similar review studies being conducted by UKRI to minimise the extent of overlap.

We selected a priority list of the 25 studies to include based on:

- Relevance of the article given the type of R&I programme that is the focus of the article. Articles where the type of R&I programme is closer to the UKRI focus were preferred.
- Robustness of the counterfactual identification methodology (as defined by the Maryland Scientific Methods Scale (SMS)[24]) that is described/used in the article. Methodologies which are at the higher end of the scale were preferred.
- Prominence of any innovative counterfactual identification methodology that have been used in recent years.
- Geography – while studies from all over the world were relevant for this study, we gave additional weight to UK studies if decisions needed to be taken about which to shortlist to give greater external validity.

We clearly documented the selection process used to narrow down the list of literature from the longlist identified through the search strategy to the shortlist of 25 studies. We then implemented snowballing to identify three additional papers and two more from the reserve list to get the list of 30 papers for in-depth review. We shared this list with UKRI and agreed on it before analysing the evidence.

Analytical framework – how will the evidence be synthesised?

The analytical framework laid out how evidence was reviewed, assessed and captured to inform the research questions. Our proposed framework ensured our approach was focused on UKRI's priorities for improving its understanding of counterfactual identification for R&I evaluations.

Developing a template for information extraction helped to ensure that the extraction was done in a way that was consistent for each piece of evidence. For each of the 30 papers, we extracted the following information relevant to the research questions through an in-depth review-

- Year of publication
- Journal (if relevant)
- Evaluation agency or organisation which published the study (if relevant)
- Authors and affiliations
- Geographical context
- Type of R&I programme studied

[24] https://whatworksgrowth.org/resource-library/the-maryland-scientific-methods-scale-sms

- Counterfactual identification methodology
- Types of outcomes measured
- Context associated with identifying counterfactual
- Key findings (high-level)
- Innovativeness of paper based on the methodology being used is a recent development
- Internal validity of findings – overall assessment of the study quality and robustness of counterfactual methodology used
- External validity of findings – overall assessment of relevance to UKRI

The information was captured in a table of studies which formed one of the deliverables of the project.

An essential part of a rapid evidence assessment is to critically appraise the evidence found by the search. This ensures more relevant and reliable evidence is given greater consideration at the synthesis stage. Critically appraising the evidence involves evaluating each piece of evidence to consider both the relevance of the evidence to the research questions and the robustness of the methodological quality utilised.

Frontier Economics Ltd is a member of the Frontier Economics network, which consists of two separate companies based in Europe (Frontier Economics Ltd) and Australia (Frontier Economics Pty Ltd). Both companies are independently owned, and legal commitments entered into by one company do not impose any obligations on the other company in the network. All views expressed in this document are the views of Frontier Economics Ltd.

WWW.FRONTIER-ECONOMICS.COM