Department for Work & Pensions

GSR
GOVERNMENT SOCIAL RESEARCH
Social Science in Government

Fit for Work: Scoping the feasibility of an impact evaluation

Francisco J. Gonzalez Carreras,
Stefan Speckesser, Jim Hillage

June 2018

Research Report 961

A report of research carried out by the Institute for Employment Studies on behalf of the Department for Work and Pensions

This document/publication is also available on our website at: https://www.gov.uk/government/organisations/department-for-work-pensions/about/research#research-publications

If you would like to know more about DWP research, please email: Socialresearch@dwp.gsi.gov.uk

First published June 2018.

ISBN 978-1-5286-0512-0

Executive summary

Fit for Work was an occupational health assessment and advice service launched in December 2014 and funded by the Department for Work and Pensions (DWP) to which General Practitioners (GPs) and employers could refer employees on (or at risk of) long-term sickness absence.

- Following very low referrals, it was announced that the Fit for Work assessment service would come to an end in England and Wales on 31st March 2018 and 31st May 2018 in Scotland. However, employers, employees and GPs will continue to have access to the same Fit for Work helpline, website and web chat, which offer general health and work advice, as well as support on sickness absence.
- The impact feasibility study was completed before the decision was taken to close the service, therefore the options considered were not influenced by this decision. The report also refers to Fit for Work as an ongoing service as that was the status at the time it was written.

This report assesses the possible approaches that could be adopted to estimate the impact of the assessment element and/or the assessment and Return to Work Plan (RtWP) elements by comparing a suitable outcome measure (e.g. length of sickness absence, proportion of individuals returning to work after a certain period of time) for both a group of participants and a control group of similar individuals who did not receive the intervention.

Three options are considered for potential designs:

- A Randomised Control Trial (RCT) at individual level would be the most robust potential design. An alternative might be to organise an RCT at GP level. However there could be a range of potential problems with such an approach, including recruiting GPs to take part and ensuring any experiment was conducted correctly as the service has national coverage and voluntary participation.
- An alternative experimental approach would be a Randomised Encouragement Design (RED) conducted at employer level. This would involve contacting a randomised sample of non-referring employers, and providing them with information about Fit for Work and encouragement to refer. This intervention group would then be compared with a matched sample of organisations who did not receive any encouragement to refer. This method relies on the 'encouragement' of the employer being correlated with making a referral and therefore participation by individuals, and there are significant risks with this being successful.
- The final option would be to construct a control group from early leavers from the service and compare their experiences and outcomes with a sample of those completing their engagement with the service. Econometric techniques could be used to minimise selection bias. This would be an easier approach to set up than either an RCT or RED. However there may be significant differences between early leavers and completers that cannot be observed and allowed for in the analysis, which could raise significant concerns about the validity of such an approach.

In our view, all three approaches outlined above have potential risks which are likely to severely limit their operation or the validity of the findings of a study, not least from low levels of participation. Therefore none is recommended as a robust and reliable way forward.

Content

List of figures and tables

Author profiles

Francisco Gonzalez Carreras is Biostatistician at Queen Mary University of London Women's Health Research Unit, the Blizard Institute and undertook feasibility scoping for an impact evaluation. He is an economist with experience in quantitative evaluations and cost-benefit analysis. He has worked for the Social Care Institute for Excellence (SCiE) and the Department for Business, Energy and Industrial Strategy (BEIS).

Jim Hillage is Director, Employment Policy Research at the Institute for Employment Studies and designed and managed the evaluation. He has led national evaluations including the Employer Training Pilots, Activity and Learning Agreements, and the Fit Note pilot. He has worked for the Department for Work and Pensions (DWP), Health and Safety Executive (HSE), and Department for Business, Energy and Industrial Strategy (BEIS).

Stefan Speckesser is Associate Research Director, Education and Labour at National Institute of Economic and Social Research (NIESR). He managed the feasibility scoping for an impact evaluation. He is an economist with experience regarding evaluation methodology and policy impact. He has worked for the Department for Work and Pensions (DWP), and the Department for Business, Energy and Industrial Strategy (BEIS).

Abbreviations

Abbreviations

APS – Annual Population Survey
ATE – Average Treatment Effect
ATT – Average Treatment on the Treated
CDiD – Conditional Difference-in-Difference
DiD – Difference-in-Difference
ESA – Employment and Support Allowance
FfW – Fit for Work
FN – Fit note
GP – General Practitioner
HML – Health Management Ltd
ITT – Intention-to-treat
IV – Instrumental Variable
JSA – Jobseeker's Allowance
LFS – Labour Force Study
MI – Management Information
NINOs – National Insurance Numbers
PSM – Propensity Score Matching
RCT – Randomised Controlled Trial
RED – Randomised Encouragement Design
RtWP – Return to Work Plan
WPLS – Work and Pensions Longitudinal Study

1 Introduction

Fit for Work is an occupational health assessment and advice service to which General Practitioners (GPs) and employers can refer employees on long-term sickness absence (off work, or at risk of being off work, for four weeks or more). The key aim of the service is to reduce long-term sickness absence levels by supporting employees who are off sick to return to work sooner than they would have otherwise done. This feasibility study aims to consider how to evaluate this part of the intervention. The Fit for Work programme also includes an advice service which is more wide ranging, including support to keep people in work and help them back after shorter absences.

The service is funded through the Department for Work and Pensions (DWP) and was rolled out across England, Wales and Scotland during 2015. The service is delivered by Health Management Ltd (HML) in England and Wales and through the Scottish Government in Scotland.

In England and Wales, it was initially planned that the programme would be rolled out across both countries between mid-March and the end of September in nine consecutive waves. However, the period was shortened and the roll-out to GPs was completed by the end of July, meaning that all GPs in over 8,000 practices in England and Wales could refer eligible patients from that date. In Scotland the roll-out to GPs started in February and consisted of three phases covering three different areas. The third and last phase finished at the end of June, when 988 Scottish practices could refer patients to Fit for Work Scotland. Employers have also been able to refer employees in England, Wales and Scotland since September 2015.

Employees can be referred to the assessment service either by their GP or their employer. Employees' eligibility is initially assessed by the referrer and their participation is voluntary. On referral patients/employees are contacted via telephone by the assessment service and their eligibility and consent to take part in an assessment are confirmed. Almost all assessments take place by telephone (and the rest are conducted face-to-face). In England and Wales, the assessment starts with collecting some background information (about the participant's workplace and job) before asking about health conditions and other barriers (social and attitudes/beliefs) to returning to work. In Scotland, this is a two-stage process with enrolment covering eligibility and consent confirmation and the collection of background information, separate from the assessment of health and other barriers to returning to work. The result of the assessment in all locations is a tailor made Return to Work Plan (RtWP). For the purposes of this exercise, individuals who receive this plan are considered to be 'completers'. Eligible individuals who leave the programme before being assessed or receiving a plan could be considered to be drop-outs or early leavers.

The aim of this report is to consider possible options for an impact evaluation of Fit for Work and their feasibility, based on assumptions about the available data sources that it might be possible to access. It is based on a review of the relevant literature and discussions with researchers in DWP and elsewhere with experience of conducting impact evaluations.

2 Identifying impact

In this chapter we review the main approaches for estimating the impact of a social intervention such as the Fit for Work service and how some of the inherent problems can be at least partially overcome by both experimental and non-experimental techniques.

Impact evaluation aims to establish what difference a policy intervention has made and whether it has achieved its objectives. It therefore can be defined as: measuring the net change in outcomes amongst a particular group, or groups, of people that can be attributed to a specific intervention using the best methodology available, feasible and appropriate to the evaluation question(s) being investigated and to the specific context.[1]

2.1 Outcomes and counterfactuals

Impact evaluation does this by estimating the effects of the particular intervention on a set of defined outcome variables. Outcomes are those measurable achievements which either are themselves the objectives of the policy – or at least contribute to them – and the benefits they generate.[2] Given the key aim of Fit for Work is to reduce long-term sickness absence, relevant outcomes are likely to relate to the length or frequency of absences from work due to ill-health.

The main difficulty in this task is to isolate the intervention effects from other factors that can also affect the outcomes but cannot be attributable to the intervention. Therefore a key question to be addressed by any impact evaluation is to identify what would have happened had the intervention not taken place, commonly referred to as the 'counterfactual'. The perfect counterfactual would be obtained if we could observe the same individual (household, country, etc.) in two different states: having and not having participated in the particular intervention. However, it is impossible to compare one observation with itself under two different circumstances: participation and non-participation.

The counterfactual for those who participate in an intervention is therefore made up from people who did not take part in it, but are very similar to those who did. The closer the counterfactual or comparison group is to the intervention group, the more valid the comparison. Ideally the two should be identical except that one group receives the intervention and the other does not. In the case of Fit for Work an ideal counterfactual would be a group of individuals who were eligible for the assessment service but were randomly assigned to a control group or to whom it was not offered due to external reasons, independent of the characteristics of the individuals (e.g. because the rollout of the programme was random and progressively reaching to different geographical areas), so they carried on as normal, receiving whatever support was or was not available. This would result in two groups that would be similar in all terms but in their participation status. Therefore if differences were found in their sickness absence variables, they could be attributed to the participation in an effective programme.[3]

1 Duflo, E. et al, 2007, pp.3895-3962.

2 HM Treasury, 2011.

3 Even though there are small differences in local and employer services, it can be assumed in principle that all individuals have access to the same services. Further specifications of the models and techniques applied could also allow for these differences.

2.2 Different types of impacts

Impact evaluation theory normally distinguishes between two different types of impact. The first is the **average treatment effect**, which corresponds to the expected effect on the relevant population as a whole. The second is the **average treatment effect on the treated**, which focuses on the impact on the individuals who took part in the intervention.

In this instance the first refers to the effect of Fit for Work on the eligible population as a whole, i.e. all employees off work for at least four weeks due to sickness absence, whether they take part in the intervention or not. It takes into account the proportion of the population who take part in the service as well as the effect on those who do not. This is also sometimes referred to as an 'intention to treat' estimate. As the Magenta Book observes, where participation is voluntary, trying to undertake an estimate where the proportion participating is small, the impact may also be small and can be very hard to detect.[4]

The average treatment effect on the treated approach measures the effect on those who took part in the service and while it may be easier to detect an effect with small participation rates, depending on how participants are selected it may be difficult to account for bias. For example it could be possible to compare the outcomes for eligible employees who use the service and those who do not. However, using a group of non-participants, who would be otherwise eligible for the service, to estimate the counterfactual is not straightforward because the referral process means the decision to take part is not random.

In making a referral, GPs may feel the service is more beneficial to some groups of eligible patients, than others. Some patients might also be more likely to consent to being referred to the service than others as they may be more proactive or more willing to return to work, than those who did not consent. These reasons make participants' and non-participants' groups systematically different. Therefore a simple comparison of the average return to work time between them would be potentially affected by selection bias. The difference in outcomes between the two groups would capture the effect of the treatment but also other effects caused by the intrinsic differences between members of the two groups, such as personal characteristics, greater willingness to return to work, etc. It would therefore be difficult to isolate the impact of the service from the other factors to estimate a robust effect.

2.3 Controlling for selection bias

There are a number of ways in which any selection bias can be controlled either by the design of the evaluation or through approaches to analysing evaluation data.

2.3.1 Experimental research designs

The problem of selection bias can be largely or completely eliminated by using an experimental design for the evaluation and allocating participation to the intervention at random.

Randomised Controlled Trials (RCTs) can remove selection bias entirely. First, a sample of N individuals is drawn from the relevant population. This sample does not need to be a random sample of the total population, although if it is not then

4 HM Treasury, 2011.

the conclusions are restricted. For example we could pick a sample from the total population according to criteria such as gender or age; however, if we do so, the conclusions of the experiment could be extrapolated only to that particular sub-population (i.e. age group or gender).

In the case of Fit for Work, the randomisation could theoretically take place at the point of referral from a GP. Some eligible individuals could be randomly referred to the service (and become the treatment group) while others (the control group) would remain in the 'business-as-usual' situation. In this case, that would mean they would receive a fit note and any other treatment to which they were referred by the GP. A random assignment between the two groups should guarantee that they were homogeneous in all of their characteristics other than their participation status. That is, it makes selection bias disappear. There are other advantages attached to a randomisation approach. The statistical techniques required to analyse the data are well-known and straightforward, and the model assumptions involved in RCTs are not as difficult to hold strong as some of those needed in non-experimental techniques.

For this reason, randomisation is considered the 'Gold Standard' in impact evaluation. The pillars of the theoretical background for RCTs are drawn from clinical trials in medical statistics. The application in social experiments, however, sometimes involves different challenges. Interventions like Fit for Work are complex with multiple components and can therefore be difficult to implement systematically and evaluate. One of the main challenges is the ethical concern about randomly assigning people to a programme when it is widely accepted that one situation (typically being on the programme) may appear more beneficial than the other, and therefore those involved in the randomisation may not be willing to take part in such an experiment. Another frequent issue in this type of experiment is partial compliance, where some of the treated group do not finish the treatment, or where the randomisation process does not operate robustly. As the Fit for Work programme is voluntary, neither the referrers nor researchers have any control over this compliance problem. Other issues might involve spillover or contagion effects, when non-participants take advantage of the programme via their employer, although not formally part of it and therefore potentially confound the results.

Thus, RCTs are not without difficulties. An alternative to randomisation involves using quasi-experimental techniques to control for selection bias so that we can find an accurate estimate of the policy impact, net of the effect that some participant characteristics might have on the outcome variable. This is discussed further in Section 2.3.2.

2.3.2 Quasi-experimental research designs

Most commonly, programme delivery is not randomised and the estimation of causal effects cannot be carried out, benefiting from experimental designs, which are robust against selection bias. Treatment effects can then only be estimated using comparison groups of non-participants, whose characteristics – both observed and unobserved – are likely to differ from those of the participants. Most commonly, econometric designs can be used, which are conditional on a large number of observable characteristics in order to correct for selection bias. When unobservable differences in the gains from Fit for Work influence an individual's decision of whether to participate in the programme or not, such designs need to be augmented by further addressing differences in unobservable characteristics.

It is also worth noting that *all* impact evaluations rely on the assumption that there are not general equilibrium effects from the intervention affecting both participants *and* non-participants. While interventions of small scale – as is usually the case for RCTs – may well justify this assumption, a non-experimental evaluation of large scale programmes may have to aim for e.g. an early period of the programme, when sufficient numbers of non-participants can be observed, who have not been affected in their behaviour by the programme.

Under such conditions, the following econometric approaches – individually or in combination – could be used for an empirical impact evaluation of Fit for Work:

- **Propensity Score Matching (PSM):** PSM is a widely used method in impact evaluation and is supported by literature.[5] The main assumption behind PSM is that a score (normally the probability of participation) can be created using the observed characteristics of participants and non-participants (socio-demographic, etc.). Individuals with similar scores in both groups can be compared and the impact would be the difference in the outcome variable between these groups. There are different paths to follow when carrying out PSM and trying to minimise possible bias.[6] Its main drawback is that the technique matches according to observed characteristics. Unobserved ones are not controlled for and this is potentially a source of bias that makes the conclusions less robust than in RCTs.
- **Difference-in-Differences (DiD):** This approach would involve comparing the average change over time in the outcome variable for the treatment group, with the average change over time for the control group. The outcomes are measured both before and after the intervention in the treatment group (p) and in the comparison group and the 'effect' is the difference estimated from the change (difference) in the two, measured in the following way:

$$DID = (After_P - Before_P) - (After_C - Before_C)$$

- The DiD approach assumes that the key characteristics of the regions, practices or individuals involved that are not included in the model are constant over time. Subtracting 'after' situation from the 'before' situation takes account of these constant characteristics as they should be the same at both times. Due to the panel data structure, effect estimates obtained from DiD are robust in the sense that bias resulting from differences in observable and unobservable characteristics would be cancelled out, as long as these differences are constant over time. By definition, the DiD approach requires pre-programme and post-programme measurements of variables for both the participant and control groups.
- **Conditional Difference-in-Differences (CDiD):** An alternative that includes elements of the two above techniques is the CDiD approach as suggested by Heckman et al[7] which would further address time-constant differences in unobservable characteristics. Our previous work on employment outcomes of labour market policy interventions[8] demonstrated that extending DiD along these lines is an effective mechanism to account for further differences between participants and matched non-participants, and can address both level differences in outcomes (i.e. time-constant bias) or dynamic differences (e.g. conditional transition probabilities). It contains elements of both matching and

5 Heckman et al., 1998b; Rosenbaum and Rubin, 1983; Rosenbaum and Rubin, 1985; Dehejia and Wahba, 1999; Dehejia and Wahba, 2002; Smith and Todd, 2005.

6 Different PSM approaches have been discussed in the literature, including Nearest Neighbours in propensity scores and local linear regressions. PSM requires the specification of particular functionals (modelling the participation process in Probit models, kernel weights, bandwidths, etc.), which have been benchmarked in the relevant literature to achieve bias-minimising properties, see for example Galdo, J. *et al*, 2008, pp.189-216.

7 Heckman, J. *et al,* 1998a.

8 Bergemann, A. *et al*, 2009, pp.797–823.

DiD approaches, as the difference in the differences is only measured for panels of matched participants and controls. As a combination of matching and DiD, criticisms tend to be similar to those attached to the individual techniques. First, matching is only done over observed characteristics. Second, the DiD elements indeed avoid bias from time-invariant non-observed variables but not from time variant variables not accounted for in the models.

In any case, identification and estimation of programme impacts require consistent data to be collected both from people participating in the service as well as non-participants about their characteristics, experience of the programme (and in the case of the controls, their experience outside the programme) and their outcomes.

3 Outcomes and data sources

In this chapter we examine the potential outcome measures that could be used to estimate the impact of Fit for Work and the potentially available data sources.

3.1 Outcome and impact measures

A key aim of the Fit for Work service is to get people back to work from a period of long-term sickness absence sooner than would otherwise have been the case. Therefore a primary outcome measure that would be of interest to an impact evaluation would be the length of time for which individuals are absent from work. This could be measured in two ways:

- Directly by collecting absence start and finish dates from individuals (e.g. through surveys or from data collected by the service) or from their employers.
- Indirectly by collecting fit note start and finish dates either from individuals, their employers or General Practitioners (GPs).

The former would be the most accurate as the period covered by fit notes do not correspond exactly to a sickness absence period. Fit notes are not issued for the first seven days of an absence and individuals may return to work before the end date written on the fit note.

The length of absence may not be the most reliable impact measure for all cases. Although the service operates on fairly quick turnaround times individuals may theoretically postpone their return to work that they would have otherwise made to complete their participation and, for instance, receive a Return to Work Plan (RtWP). However given the orientation of the service, length of absence would appear to be an important outcome measure.

Another variable of interest is the sustainability of return to work, both in terms of any subsequent sickness absence after a return to work following the Fit for Work intervention and/or continued periods of employment (e.g. for six months post-intervention). Such an outcome could be used in addition to length of absence to, for example, investigate whether Fit for Work participants do not return to work sooner than individuals in the control group, but that their return to work was more secure.

However, measuring such an outcome variable would involve additional data collection, for instance through follow-up surveys or administrative data. Access to linked administrative data would open a much wider range of possibilities regarding outcome variables. Individual and social benefits could be estimated, for example if a net impact on employment rates was found in the longer-term. In addition, Fit for Work also aims to avoid individual transitions out of employment and an impact on such outcomes could be directly evidenced, e.g. in Jobseeker's Allowance (JSA) or Employment and Support Allowance (ESA) registers.

Closely related to long-term impacts on employment rates, the analysis could also investigate whether Fit for Work focus reduced the probability for people losing employment to claim out-of-work benefits. Although this might only affect a small proportion of the programme participants, it could be an important measure of the social (and fiscal) benefit of the programme.

Administrative data could also be useful to account for variables that greatly influence the length of sickness outcomes, for instance, and this would help to reduce the size of the sample needed for the study.[9] Any use of administrative data that involves bundling information from several administrative sources needs a linkage variable that can uniquely identify every individual. National Insurance Numbers (NINOs) would be the best option for this purpose but they are not currently being collected by the Fit for Work service. This gap makes linking data difficult to achieve. Other less optimal alternatives can be considered, however they are less effective at creating matches. These include using an individual's date of birth, postcode and gender together for linkage purposes. In this case there might be further issues with the availability of variables in the different sources, consistency of measurements, and changes in some variables, such as address over time which serve to further limit its effectiveness.

An additional outcome variable could be the number of visits to a GP. If participants were using services significantly less than the control group we could conclude that the programme was saving health costs. This data is available, but is not linked to non-health data at present, and would therefore require significant data security, ethical and legal considerations. Estimates of health service usage could instead be collected via employee surveys.

3.1.1 Effect size

A further consideration in terms of outcome is the likely size of the impact of the service, for example what would be the expected reduction in the length of absence among participants compared with the controls. The smaller the minimum detectable effect we are seeking to measure, the larger the samples of participants and controls required to enable it to be measured. The sample size is also related to the statistical power and significance level required. If we increase the power or lower the significance level we will need to increase the sample size.

3.2 Data sources

There is limited information available about the eligible population, i.e. long-term sickness absentees. Estimates from the Labour Force Survey indicate that around 1.8m employees had a long-term sickness of 4 weeks or more in a period of 12 months.[10]

Identifying a sample for an impact evaluation is complicated by the two points of referral as employees can be referred to the service either by their GP or by their employer and there may be systematic differences between the two groups.

A further complication is the consensual nature of the service. Referrals by either route are only made with the employee's consent and the continued participation is voluntary, which means that employees need not complete their assessment or receive an RtWP. This could introduce significant logistic complications to the membership of the control and intervention groups.

9 Duflo, E. *et al*, 2007, pp.3895-3962.

[10] DWP (Oct 2016 Work), Health and disability green paper: data pack. https://www.gov.uk/government/statistics/work-health-and-disability-green-paper-data-pack

3.2.1 Management Information

Data are collected on all individuals referred to Fit for Work. The data collection process varies between the service in Scotland and the service in England and Wales, but essentially the same data are collected. The amount of data collected depends on an individuals' interaction with the service and for instance there is more limited information about those referred but who do not take part in an assessment compared with those who both receive an assessment and an RtWP and are discharged from the service in the normal way.

Data collected includes:

- Socio-demographic health and employment characteristics for all referred individuals who consent to take part in the service.
- Participants' engagement with the project (start and finish date attendance at assessment, receipt of RtWP etc.).
- Some information about early leavers (e.g. people who do not complete an assessment or decline to receive an RtWP), including basic characteristics and contact details collected on referral.

Currently the length of sickness absence is not collected for service participants. While absence start date is collected, absence end date is not. This gap would have to be filled for the management data to be used to estimate impact.

3.2.2 Survey data

Surveys of participants could be conducted to collect additional data, such as what happened to people once they had finished their interaction with the service. In addition, data would also need to be collected from samples of non-participants in order to make comparisons.

As the eligible population is such a small proportion of the working population it would be difficult to identify a comparison sample of non-participants from a general population survey. However, GPs and employers potentially have the capacity to identify individuals for a comparison group. GPs could identify patients that present themselves in order to receive fit notes. These could be included in a sample of non-participants who are otherwise similar to Fit for Work participants. In the same fashion, employers could identify employees with a health condition who are absent from work, but who are not taking part in the service, to form a comparison sample. Any comparison sample would need to be identified at the point of potential referral to the service regardless of whether this is done by a GP or employer. The size and scope of the surveys would depend on the analytical design adopted.

3.2.3 Linked administrative data

A range of outcome variables could be obtained from administrative data available from the Department for Work and Pensions (DWP) or other sources and could be used for the evaluation if they could be linked to data on participation and eligibility. Such data could allow for estimating the impact of Fit for Work on long-term employment levels or benefit dependency, for example based on the Work and Pensions Longitudinal Study (WPLS). This census-level dataset combines:

- A register of all claimants of DWP benefits.

- Data on individual earnings and employment obtained from HMRC records P14 and P45, i.e. a census level dataset of all dependent employment subject to taxation/national insurance payments.
- If JSA/ESA register data could be used, participants could be identified in these datasets. This would also make it possible to look for potential controls in these data. These controls would be composed of people who did not take part in Fit for Work but who were otherwise comparable. They could be identified through matching methods, i.e. finding a control group that, according to other observable characteristics, would be comparable to Fit for Work participants.[11] Once participants and controls were identified, it would be possible to compare trajectories and find out whether Fit for Work participants were less likely to move onto out-of-work benefits.

Administrative data would also be crucial for employing a CDiD approach, as such a design requires outcome variables for both groups of participants and drop-outs before and after participation in Fit for Work.

The use of administrative data is an inexpensive strategy for acquiring some (albeit incomplete) measures of outcomes and impacts of the programme as no data collection would have to take place. Using linked administrative data, we would not be able to observe some critical variables, e.g. if people formally retained employment, but did not actually return to their workplace. For instance, outcome measures such as the days lost due to sickness absence for employers would be difficult to obtain.

In practical terms, the data could easily be merged on the basis of references such as postcode, date of birth and gender. Ideally, it would be most easily facilitated if the service collected National Insurance Numbers (NINOs) of participants. The latter is not likely as NINOs are not currently being collected. No matter what linking variables are used, the feasibility of this approach would be subject to a rigorous assessment of the option in terms of lawfulness of such data linkage. In our view, it would particularly require:

- A full understanding of which data could be shared to carry out the research.
- A mechanism to merge data, which is both feasible and robust, and to supply it to independent researchers.
- The involvement of a trusted 'third party' to ensure that both DWP and the independent researchers only obtain the minimum data required to carry out the impact assessment.

If legal constraints on this data linkage prevented this option, informed consent of Fit for Work participants could be sought at the start of the 'customer journey' allowing their records to be linked to other sources of administrative data, though it is possible that seeking this consent could affect participation rates. We further recommend exploration of alternative data sources, which could be linked to management information, such as data from NHS Digital. An accurate account of these data is not currently available to the independent research community and would therefore require further scoping, stakeholder engagement and descriptive work than could have been carried out in this feasibility study. Consent for linking to such sources would also be required.

Other potential sources of data include fit notes issued by GPs. These are administered electronically and could form a valuable database containing data on certified sickness absence. Fit note data at the aggregate level will be made available

11 See in section 2.3.2 for the description of PSM. Briefly, a scoring is first created using observed characteristics of FfW participants and non-participants. Then FfW and non-participants with similar scores are compared in their absence variables and differences, if any, are attributed to participation.

via NHS Digital. These data would have to be supplemented by further characteristics of the Fit for Work programme management information e.g. by an evaluation of the programme at practice level as the issuing of RtWPs results in a termination of fit notes. Therefore, GP practices with referrals to Fit for Work – other things being equal – should show trivially reduced average sickness absence based on fit notes. Average durations of the RtWPs, as well as socio-economic information about the geographic area and the population characteristics covered by practices, would need to be added to obtain consistent measures of sickness absence across groups of practices with or without referrals to Fit for Work, but these data could be a very valuable research resource as long as there are marked differences in referral patterns. Note however that this data source would also only offer an outcome measure for sickness absence, which may differ from the true absence from the workplace.

4 Possible designs for an impact evaluation

In this section we consider a range of possible research designs for measuring the impact of Fit for Work. The critical first step is to identify a way of defining and measuring outcomes for a comparison group and we have looked at potential experimental or non-experimental design at various levels:

- Area – comparing areas of the country.
- Referrer – making comparisons at the level of General Practitioner (GP) practices or employers.
- Person – comparing individual employees who participate with those who do not.

4.1 Geographical level

Matched area comparisons between different geographical areas where the programme is operating and areas where it is not could be used to create a comparison group. Comparing participant and non-participant areas could potentially overcome the problem of the two referral routes. This would constitute an 'intention to treat' approach (see section 2.2) as it would examine the effect on the eligible population as a whole regardless of whether they use the service.

Such an approach using analytical methods such as Difference-in-Differences (DiD) (see section 2.3.2) would be possible if the programme was implemented in only some regions or gradually introduced across the country e.g. starting in some areas and moving out to the rest over time. This was potentially possible as although Fit for Work is a national programme the initial plan was for a progressive roll-out.

However, the low volume of participation in the early stages of the roll-out suggested that it was quite unlikely to find an estimate for regional differences in sickness absence or the proportion of individuals above an absence length threshold. In addition, the roll-out timetable was shortened and the service was offered nationwide before comparison areas could be established.

We could regress area-level outcomes on area-level referral rates. We could not attribute causality unless we used an instrumental variable. However, instrumental variables are quite difficult to find (see section 4.2). An approach would be to use the Randomised Encouragement Design (RED) described in section 4.2.

4.2 Referrer level

We have also considered the possibility of conducting the impact evaluation at **practice level** and examined the possibility of comparing absence levels in referring and non-referring practices. While it appears that not all GP practices are participating, all have been contacted and are theoretically aware of the programme. Therefore there is likely to be some systematic difference between referring and non-referring practices. However these differences could potentially be addressed, for example using DiD designs, which would eliminate time-constant differences between referring and non-referring practices. This would be an intention to treat design because not all eligible individuals referred from a practice would take it up.

If it was possible to link fit note data to programme management information at practice level, this could be used to establish the average duration of certificated sickness absence in GP practices both participating and not participating in Fit for Work before and after referral to the service. This would enable analysis of the population of sickness absentees receiving a fit note rather than a sample of referred individuals.

However, there are some important drawbacks. First, we do not know whether it would be possible to gain access to a database of fit notes. Secondly, it may not be possible to link fit note data to Return to Work Plan (RtWP) data to maintain a continuity of certified absence information for Fit for Work participants or flag up Fit for Work participants when they come back to the fit note system. In addition to this, the fit note provides information on certified absence and this variable is different from actual absence time. Considerations around consent for data linking apply.

We also considered an option at **employer level** involving a RED. All employers are theoretically aware of the potential to refer their long-term sickness absentees to Fit for Work. However currently only relatively few have actually made referrals. It would be possible to have a strategy of contacting a random subsample of selected employers from those who have previously been informed about the programme. Those contacted would be given further information about the programme, more support and, if possible, incentives to encourage them to refer employees to Fit for Work. Not all employers would do this and thus these targeted or incentivised firms could be used as a proxy for participation. This method is called an 'instrument variable' approach, where the 'instrument' would be the encouragement.

Instruments[12] are variables that are included in the analytical models instead of other variables (in this case referring or non-referring firm). The problem of selection bias which would occur if we were just comparing referring and non-referring firms can be mitigated because the participation is proxied/instrumented through this 'encouragement' variable (i.e. encouraged/not encouraged organisation) which is random. This randomness is what allows selection bias to be avoided. However the effectiveness of the approach relies on the instrument being an effective proxy for participation and there being a close correlation between those firms encouraged and those subsequently participating. If that was not the case it would not be clear whether any study was measuring the effect of the encouragement or the effect of taking part.

Operationally at a first stage it would be necessary to define the population of firms among which to study the effect of Fit for Work, e.g. by size, sector or location, bearing in mind that the conclusions of the analysis will only be valid for firms which have these characteristics. After this it would be necessary to draw a representative sample of this population of firms. Finally, a sub-sample of this representative sample would need to be randomly encouraged to participate (i.e. refer employees to Fit for Work when needed).

Data would have to be collected from both sets of firms through surveys to estimate the average length or frequency of long-term sickness absence time at firm level. In terms of calculations, the method would need to subtract the variable of interest of the encouraged from the non-encouraged group and the referring proportion, controlling for the proportion of referring firms in each group.

12 Instrumental variables can be used when programmes are open to any potential participant. In this case comparing participants with non-participants would not be accurate as participation would be driven by personal characteristics and so would the outcome measured, e.g. length of sickness absence in Fit for Work. An Instrumental Variable (IV) is a variable that (1) is correlated with participation and (2) is completely independent of the outcome of interest. Substituting participation by this instrument in the quantitative analysis allows researchers to find an unbiased estimate of the programme effect.

A RED could also potentially be used with GP practices based on the same principle and could additionally test whether methods for increasing GP involvement with Fit for Work were effective as well as the impact of an increased level of involvement on sickness absence.

There will be a random sample of GPs that would be encouraged to refer to Fit for Work and this would be the instrument, provided that we could find a significant correlation between referrals and encouragement. The same issues around measuring the effect of the encouragement as opposed to the effect of the proxy would apply here too.

4.3 Individual level

We also considered the potential options for a study undertaken with individuals as a unit, using either an experimental or non-experimental design, although we also considered a possible randomisation at employer level.

Randomised Control Trial (RCT)

As identified in Chapter 2, the ideal design would be to randomly allocate eligible employees to the service and to an alternative 'business as usual' route at the point of referral. Organising such an RCT through employers was considered infeasible because:

- it would be difficult to identify potentially referring employers;
- if they could be found, it was considered unlikely that employers would agree to take part in such a trial; and
- even if they did it was difficult to envisage how such a trial could be organised.

However, RCTs are a common method in health research and therefore the possibility of using randomised allocation within GP practices is at least theoretically feasible.

When faced with an eligible patient, GPs could use a randomisation tool to indicate whether a patient would be referred to the service or receive usual care.

Information about the patient would be collected during this first visit and follow up data collection would be necessary to measure the length of sickness absence after the treatment, whether the patient dropped out of the programme or switched to the non-prescribed programme, etc.

As outlined in section 2.3.1, RCTs have a number of attractive advantages particularly as they robustly control for selection bias.

However there are a number of fundamental problems that would have to be overcome in order for such a design to be implemented:

- First a sufficient number of GPs would have to agree to take part. They may have ethical concerns about denying eligible patients randomised to the control group access to the service. They may also have concerns about the practicality of an experiment, given for example the infrequency with which most GPs would see an eligible patient. Although GPs are likely to understand the principles of an RCT and their appropriateness in testing the effectiveness of particular interventions, they may have concerns about randomisation of a service which has been rolled out and should arguably constitute usual care.
- Secondly, participation in the service is voluntary and employees need to give their consent before being referred to the service which would complicate the randomisation at the point of referral and could result in attrition. Even after being referred, employees can withdraw consent and not proceed to an assessment

or not receive an RtWP. Therefore there could still be a 'selection effect' if those agreeing to enter the programme and see it through are different to those who decline to be referred or drop out.

- Thirdly, organising an efficient randomisation process could be problematic.
- Fourthly, patents referred to the usual care option could be subsequently referred by their employer, further contaminating the randomisation.
- Fifthly, we would need to collect survey data from patients at the point of referral which would require further co-operation by both GPs and the individuals concerned.
- Finally, even if it was conducted perfectly, such a study would only produce findings of restricted external validity as they would only refer to employees referred through GPs and not those of the whole population, i.e. including those referred by employers.

For these reasons (practical, ethical and technical considerations), randomisation of individuals is not deemed feasible.

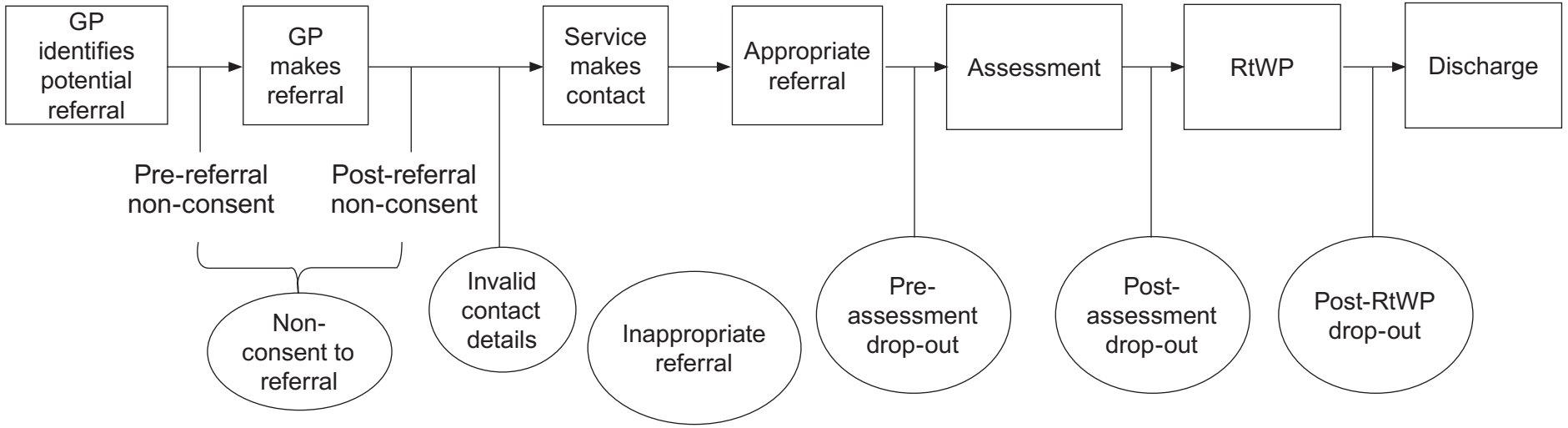
Non-experimental designs

Another option is to attempt to identify a control group from people who start the Fit for Work assessment process, but do not complete it.

Figure 3.1 depicts the flow of individuals through the process (based on referral from a GP). The GP or employer in discussion with the employee identifies a potential referral to Fit for Work. However, the approach would not be possible in practice because at this early stage in the process the employee may not consent to be referred. Even if a referral is made the employee may reconsider and refuse to answer further questions from Fit for Work, or the service may find that the contact details provided by the GP or employer are invalid.

This group of people who either do not consent to referral or who are not referred would make a potentially strong comparison group as they are, on the face of it, very similar to those who are referred. However contact details for this group are unavailable, they have not been informed about the evaluation and therefore may not be able to be contacted even if their details were available. For these reasons, a quasi-experimental design has too many practical constraints to make it feasible.

Figure 3.1: Flow of individuals through the Fit for Work process



In the next step of the Fit for Work process, the service makes contact with the individual and checks their eligibility for the service. At this stage some referrals are found to be inappropriate (e.g. the referral is for someone who is unemployed). Once eligibility is confirmed, the service takes initial details from the referred employee and a date is made for the assessment (the process of taking this initial information is slightly different in Scotland from the process in England and Wales due to the different structures of the services – see Chapter 1). The assessment takes place and subsequently an RtWP is generated.

For the purposes of this exercise the key point is that the data from the process evaluation of the service indicate that some individuals:

- Withdraw consent and/or decline to continue with the service before their assessment.
- Withdraw consent or decline to continue with the service between the assessment and receiving the RtWP.
- Return to work either before the assessment, before receiving an RtWP or after receiving an RtWP, but before their official discharge.

The drop-outs considered as controls for the study would be those classified in Figure 3.1 as 'pre-assessment drop-outs' (people that were appropriately referred but who dropped out before the assessment). These also include individuals who returned to work. Once the individuals have passed through the assessment stage they are not considered appropriate drop-outs as they would have received at least part of the intervention.

Data collection

A range of data on participants is captured by the management information. However this will need to be supplemented by surveys to collect:

- More information from early leavers about their post-referral sickness and employment experience and why they left the service early.
- Data on the longer-term sickness and employment experiences of participants which are not captured on the management information.

One issue with this approach is ensuring that there is a large enough number of referrals to allow for a sufficient number of drop-outs. Assuming a conservative response rate to the survey of 33 per cent, then a sample frame of around 1,800 drop-outs will be required and this could be achieved from around 36,000 appropriate referrals.[13]

The crucial issue when tackling this drop-out option is that the proportion of the drop-outs that can be considered controls has to be estimated to some extent. Calculations are done assuming that five per cent of referrals can be used as controls. If this percentage dropped to 2.5 per cent, for example, the total number of referrals would rise to around 72,000. Thus, the feasibility of this option relies heavily on both the overall level of participation and the proportion who drop out at the appropriate point.

Further details on sample sizes that would potentially be required are presented in Appendix 2.

Propensity Score Matching (PSM) will be an essential element of this approach in order to address differences in observable characteristics between participants and non-participants in terms of socio-demographic and health characteristics, as both are likely to affect the outcome. In addition, a Conditional Difference-in-Difference

13 Assumptions and calculations in Appendix 2.

(CDiD) approach – for example exploiting further linked administrative data if that is possible – could be employed to address further differences in matched samples due to unobservable characteristics which are constant over time.

In order to underpin the validity of PSM and CDiD, a series of tests on the validity of the underlying assumptions will be carried out. Matching techniques such as PSM use balancing tests which are aimed at supporting the assumption that, after the matching, treated and control groups only differ in their participation status. The usual tests in relation to the assumptions underpinning DiD will also be carried out.

The CDiD technique contains elements of both PSM and DiD. Therefore, both types of test are needed so that the differences in the outcome variables in CDiD samples can be interpreted as a causal impact, for example in terms of average length of sickness absence.

The feasibility of this approach would depend on several factors, mainly:

- Whether the number of drop-outs would be sufficient to estimate counterfactual outcomes (relevant to PSM and CDiD).
- Whether these drop-outs would be willing to provide information as part of a survey and supply informative data on conditions and outcomes relevant to referral and impact (relevant to PSM and CDiD).
- Whether administrative data could be made available to model CDiD. This would require pre-programme information on both Fit for Work completers and drop-outs, for example the differences in pre-programme employment and benefit rates from DWP administrative records.

A comparison between completers and drop-outs would have the following benefits:

- There is an easy way to find a counterfactual or comparison group.
- It would work for both participants referred by GPs and those referred by employers.
- The personal information is collected at the point of referral and therefore the people to be surveyed are already known.
- At a simple level, data on participants are already available in the Fit for Work management information and it would only be necessary to survey the drop-outs. More complex analytical approaches (e.g. DiD) and/or more comprehensive outcome data (e.g. sustainable employment) would require more extensive data collection (and analysis) and would increase the costs of the study.
- PSM and CDiD work well when there is a rich source of information, which can be obtained from programme management information in combination with a survey and a sufficiently large number of observations available when the programme is approaching steady-state provision.

However, this design might have significant pitfalls, particularly about whether early leavers from Fit for Work could be viewed as an effective control group. Although information on their characteristics and why they left the programme could be collected, there are likely to be systematic differences between early leavers (drop-outs) and completers that cannot be observed and allowed for in the analysis which would limit the external validity of such a study. In other words there may be unobserved variables (therefore not included in the model) that are affecting the outcome but which we cannot account for. Therefore the obtained outcomes omitting these variables could be biased. Furthermore, this approach relies on there being a sufficient number of individuals dropping out and that sufficient data can be obtained from the control group to allow for the application of PSM.

5 Conclusions and recommendations

This study has considered potential options for measuring the impact of Fit for Work. The options available are constrained by certain aspects of the service model and because it has been rolled out across the country. We have considered a range of experimental and non-experimental approaches which were scoped in detail. Our approach has been primarily desk-based, and further consultation with policy-makers and key stakeholders such as General Practitioners (GPs), employers and the service providers would be needed to assess the practical considerations and feasibility of such a study.

The main aim of the policy is to reduce the incidence of long-term sickness absence by supporting employees who are off sick to return to work sooner than they would have otherwise done. It also includes an advice service which is more wide-ranging including keeping people in work and helping them back after shorter absences. Given the main aim of the programme we recommend that the most appropriate effect measure is length of absence, ideally supplement by the sustainability of a return to work after sickness absence.

5.1 Evaluation alternatives

We have considered a range of potential approaches to estimating the impact of Fit for Work. However, **all options have significant barriers to implementation** and therefore none is recommended as a robust and reliable way forward.

Experimental design

The most effective study design would be to use an experimental research design with randomised programme participation. Such an approach with randomised programme participation would have technical benefits were it possible and it could theoretically be possible to organise a Randomised Control Trial (RCT) at GP level, but there are significant practical limitations with such an approach in terms of recruiting GPs to take part and ensuring any experiment was conducted correctly given for example the national coverage of the service and its voluntary participation. In our view, these pitfalls are likely to prove insurmountable and rule out the feasibility of such an approach altogether.

An alternative experimental approach at the level of referral would involve a Randomised Encouragement Design (RED) among employers. Such a design could offer a powerful variable to mitigate selection bias because it would introduce variation in referral patterns, which is independent of the characteristics of patients, employers or GPs. Random Encouragement could be conducted at employer level (or possibly GP level) and would involve contacting a randomised sample of currently non-referring employers (or GPs), providing them with lots of information about Fit for Work and possibly incentives to take part (which taken together would constitute the ‘encouragement’), and comparing absence levels of the encouraged sample with a matched sample of non-encouraged organisations. While this would not measure impact directly, impact could be estimated if it could be shown that the ‘encouragement’ instrument was correlated with making a referral and therefore participation. However this association could only be established at the end of the study so there are significant risks attached. Given the current level of referral to Fit

for Work and that there are large numbers of employers and GP practices, we believe that this approach to measuring the impact of the programme is theoretically feasible but might be difficult to operate in practice and therefore not recommended.

Non-experimental design

A further option we have considered in detail is constructing a control group from early leavers from the service and carrying out a survey of the drop-outs/early leavers in order to collect information about their health condition, socio-demographic characteristics and sickness absence experience. The counterfactual information would be compared with completers applying a Propensity Score Match (PSM) approach in order to minimise selection bias, which should be complemented by Conditional Difference-in-Difference (CDiD), potentially exploiting available administrative data sources to capture potential differences in unobservable characteristics before the programme.

The main advantage of this approach is that the contact details of early leavers are available, so that costly (and potentially long-term) surveys aiming to capture the universe of the eligible population could be avoided. It would also be easier to set up than either an RCT or RED. However there may be significant differences between early leavers and completers that cannot be observed and allowed for in the analysis and which could raise significant concerns about the validity of such an approach. This approach also relies on data on length of absence being collected but given that the absence finish date is not included in the current suite of management information, this is currently not possible, rendering this option unfeasible in practice too. Furthermore, the number of service users may be insufficiently large to generate the required sample sizes.

Finally, data linkage could have potentially provided a cost effective way of generating more comprehensive and robust estimates of impact. However currently this is not possible because National Insurance Numbers (NINOs) are not collected by the service.

5.2 Conclusion

Therefore we conclude that all approaches outlined above have significant potential risks which are likely to severely limit their operation or the validity of the findings of a study. Limited levels of participation also add significantly to the risks attached to the approaches outlined in section 5.1 as they all require sufficient samples of the treated population to detect an effect and affect their feasibility.

Appendix 1: Summary of options

A summary with the different options, their requirements, pros, cons can be seen in Table A1 on next page. The column ‘Robustness’ ranks the robustness of each approach within the group of options, with 1 being the most robust and 6 the least. The costs for both Propensity Security Matching (PSM) options and Conditional Difference-in-Difference (CDiD) with linked data include the survey costs plus data management and analysis work. The RCT option includes the design, implementation and monitoring work on the experiment plus data management and analysis. Finally, both linked data with Jobseeker’s Allowance (JSA)/Employment and Support Allowance (ESA) and fit note options would involve only data management and analysis work.

Table A1 Summary of different approaches

Approach	Dependent variable	Description	Needs	Pros	Cons	Robustness
PSM	Length of sickness	Compare completers vs. drop-outs.	One survey of completers and drop-outs after programme. Needs consent from individuals. **Timeline**: Around 12 months, including survey.	Straightforward and relatively quick and cheap rough approach.	Effect of unobservables cannot be tested.	5-6
PSM with further follow ups	Length of sickness and sustainability of employment	Compare completers vs. drop-outs.	Survey completers and drop-outs after programme and two follow ups on each. Needs consent from individuals. **Timeline**: Depending on surveys and follow ups, around 20-22 months.	It will add to the former an evaluation on sustainability of employment.	Effect of unobservables cannot be tested.	5-6
CDiD with linked data	Length of sickness absence	Compare completers vs. drop-outs in sickness absence but also in pre-programme records.	Survey completers and drop-outs after programme and access to pre-programme records, such as earnings and benefits. Needs consent from individuals and access to pre-programme data. **Timeline**: Dependent on the period to be studied and time needed to guarantee access to data for research. Around 14-15 months including the survey.	It will be more robust than PSM, less affected by possible bias due to non-observed characteristics.	It needs these pre-intervention records.	4

Approach	Dependent variable	Description	Needs	Pros	Cons	Robustness
Linked data with access to JSA/ESA	Several: length of work absence, benefit claims, etc.	Compare FfW participants with potential controls in JSA/ESA and compare their trajectories with respect to benefit claims, for instance.	NINOs of FfW participants in order to identify them in other databases such as JSA/ESA. Needs access to datasets. **Timeline**: Depends on when we receive NINOs, how long it takes to receive JSA/ESA data for research and how long the period to be studied is. From start to end of analysis: 3-4 months.	Ideally data will include all participants, records could be traced over a long period of time thus allowing for study on a longer-term basis. Preferable to CDiD or PSM if data is available.	Will need access to JSA/ESA records. FfW would need to collect NINOs in order to identify the treated.	2
RCT[14]	Length of sickness absence, potentially others depending on the follow up (sustainability of employment)	Individuals visiting GPs will be randomly assigned to FfW or FN approach. Both groups would need to be surveyed at the end of the sickness absence.	A proper randomisation design and at least a follow up of both groups. Needs GPs' collaboration and consent to use data for research. **Timeline**: Design and implementation of RCT plus analysis of data: around 6-7 months.	If properly designed and executed this approach will provide the best impact evaluation.	Difficult to set up properly, will need collaboration from GPs. If FfW is considered to be better than FN it would raise ethical issues.	1

14 Costing referred only to the GP strand of the RCT. The cost of the RED at firm level would need to be scoped out separately.

Approach	Dependent variable	Description	Needs	Pros	Cons	Robustness
Fit note compared with FfW records	Length of sickness absence, sustainability of employment	Individuals within a FN framework could be compared with records in FfW.	Data on FNs should be available and we should be able to identify FfW users if they show up in FN datasets. It depends on how long it takes to get access to FN data for research and how long the period studied is. **Timeline**: From start to end of analysis: 3-4 months.	Will compare the total FN users with the total FfW users. We could compare averages on absence, etc. before and after the programme. Preferable to PSM or CDiD approach.	We need access to FN records. We should be able to flag up FfW users if they show up in FN data. FN reports certified absence rather than actual absence time.	3
RED	Average length of sickness absence in the firm, average number of cases, average duration of sickness absence. Potentially others depending on the information available.	A set of randomly picked employers will be encouraged to participate. They will be compared with non-participating employers.	We might need to do a pre-definition of an evaluation sample of employers to be surveyed. A random subset of them will be picked to be encouraged/informed in more depth. Needs employers' collaboration and consent to use data for research. At least one follow up would be needed for both groups. **Timeline**: Design and implementation of RED, information, collection plus analysis of data: around 6-7 months.	A quite robust method to estimate the impact of the programme at employer level.	Implementation problems as collaboration is needed. The estimated impact is valid for companies with characteristics of the evaluation sample. Estimated impact for firms that would participate when encouraged to do so.	1

Appendix 2: Survey assumptions

Assuming a conservative response rate to the survey of 33 per cent, then a sample frame of around 1,800 drop-outs will be required and this could be achieved from around 36,000 appropriate referrals. Provided that the key categorical information from early leavers is already captured, then a 15-minute telephone survey should be sufficient to collect data on their post-referral sickness and employment experience and why they left the service early.[15]

The greater the sample, the more likely it is to find the impact, if this exists. More particularly, small sample sizes would be less likely to estimate robust impacts if the effects were small to moderate. Given the nature of the intervention, these effects sizes could still yield social net benefits so investment in larger sample surveys could well be justified. The crucial issue when tackling this drop-out option is that the proportion of the drop-outs that can be considered controls has to be determined to some extent. Calculations are done assuming that 5 per cent of referrals can be used as controls. If this percentage dropped to 2.5 per cent, for example, the total number of referrals would rise to around 72,000. Thus, the feasibility of this option relies heavily on this proportion.

Assumptions

We need to find the minimum samples size necessary in order to find an established minimum detectable effect, given a significance level and test power.

Sample size with continuous variables

Assuming that a decrease in the length of sickness absence by three days would be a meaningful reduction from a policymaker's point of view, we could define the effect size (ES):

$$ES = \frac{|\mu_p - \mu_c|}{\sigma}$$

where is the mean of Fit for Work participants' length of sickness absence; is the mean length of sickness absence in the comparison group. Greek sigma (σ) is the assumed standard deviation of the sickness absence. Its value is taken from Department for Work and Pensions (DWP) Research report No 896[16], although this will need to be changed according to the actual data collected once it is available. The assumed standard deviation adds up to 18 days.

15 The main concern is to have a minimum of drop-outs to create the control group. There will be more than enough treated individuals to be compared to these controls.

16 Using table 2.6, page 34, the mean in weeks for English regions (excluding Scotland) are transformed into days and then the standard deviation is discovered. This standard deviation was calculated as 13.4 days. This figure is increased by 35 per cent to allow for a possible increase in variability in the future. With a smaller standard deviation the minimum sample size would be also smaller and costs would be lower. Source: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/438234/rr896-fit-for-work-service-pilots.pdf. Accessed 01/09/2015

Thus the ES would be:

$$ES = \frac{|\mu_p - \mu_c|}{\sigma} = \frac{3}{18} = 0.17$$

We will use a two sided test with a five per cent level of significance and we want to get the minimum number of people needed to ensure that the power of the test is 80 per cent and detect this difference.

For that purpose we apply the equation:

$$n_1 = 2\left(\frac{Z_1\text{-}\alpha/2 + Z_1\text{-}\beta}{ES}\right)^2 = 2\left(\frac{1.96+0.84}{0.17}\right)^2 \approx 560$$

The minimum sample size needed to detect this difference would be approximately 560 people in each of the control and treatment groups (rounded up to 600). Making the assumption that only one in every three individuals[17] approached would be willing to answer the questionnaire, we would estimate needing a sample of up to 600 x 3 = 1,800 people. If proportions between referrals and drop-outs were constant (five per cent), the total number of referred people needed to carry out this analysis would be 1,800/0.05 ≈ 36,000 referred individuals. This number would change, downwards if the proportion of ineligible referrals fell or upwards if the proportion of people who leave early was lower. The numbers also depend on the validity of the data used to estimate the standard deviation and the desired power of the eventual study.

However, even if the sample size doubled and the rate of drop-out halved the required flows through the programme would still look feasible.

Sample size with proportions

It might **also** be interesting to test the proportion of individuals that went back to work within a certain period of time, for instance one month (‘successes’**).** This is modelled with binary variables and what interests the researcher is the proportion of successes in the two groups compared: Fit for Work participants and the ‘business as usual’ group, i.e. those receiving fit notes.

It will be necessary to test whether the difference between these proportions is statistically significant or not. This test is slightly different to the one shown above and needs the proportions of successes in each group.

We can name the proportion of successes or positives in the Fit for Work group as and the proportion of positives in the fit note groups as π_{FN}. We need to fix the proportions that we want to test in our study. For instance, if we wanted to test whether Fit for Work gives a 60 per cent success rate versus the 50 per cent provided by the fit note intervention, we would need to fix:

$\pi_{FfW} = 0.6$

$\pi_{FN} = 0.5$

The general formula to find out the sample size for the test would be:

17 This ‘rule of thumb’ is normally used when calculating response rates. If we were quite conservative about this figure we could use a 20 per cent response rate. The sample size needed under these assumptions for this response rate would be 600 x 5 = 3,000 individuals. Based on the above reported proportion of drop-outs, the number of referrals needed for 3,000 drop-outs would be 60,000.

$$n \simeq \frac{\pi_{FfW}\,(1-\pi_{FfW}) + \pi_{FN}\,(1-\pi_{FN})}{(\pi_{FfW}-\pi_{FN})^2} \times [Z_1-{}^{\alpha}/_{2} + Z_1-\beta]^2$$

In the example above for a 60 per cent success rate in Fit for Work and 50 per cent in the fit note group, keeping the five per cent significance level and the 80 per cent power, the sample needed would be found as follows:

$$n \simeq \frac{0.6\,(1 - 0.6) + 0.5\,(1- 0.5)}{(0.6 - 0.5)^2} \times [1.96 + 0.84]^2 \simeq 385$$

So we should have 385 participants per arm/group, a total of 770.

The main drawback of this approach is that we need to set and and thus it has to be recalculated depending on the different proportions to be tested.

Appendix 3: References

Bergemann, A., Fitzberger, B., Speckesser, S. (2009) 'Evaluating the Dynamic Employment Effects of Training Programs in East Germany Using Conditional Difference-in-Differences', *Journal of Applied Econometrics*, 24, 797–823.

Dehejia, R. H. and Wahba, S. (1999) 'Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs', *Journal of the American statistical Association,* 94**,** 1053-1062.

Dehejia, R. H. and Wahba, S. (2002) 'Propensity score-matching methods for nonexperimental causal studies', *Review of Economics and statistics,* 84**,** 151-161.

Duflo, E., Glennerster, R., Kremer, M. (2007) 'Using randomization in development economics research: A toolkit', *Handbook of development economics,* 4**,** 3895-3962.

DWP (Oct 2016 Work), *Health and disability green paper: data pack*. https://www.gov.uk/government/statistics/work-health-and-disability-green-paper-data-pack

Galdo, J., Smith, J. And Black, D. (2008) 'Bandwidth Selection and the Estimation of Treatment Effects with Unbalanced Data', *Annals of Economics and Statistics*, GENES, 91-92, 189-216.

Heckman, J., Ichimura, H., Smith, J., Todd, P. (1998a) *Characterizing selection bias using experimental data*, National bureau of economic research.

Heckman, J. J., Ichimura, H., Todd, P. (1998b) 'Matching as an econometric evaluation estimator', *The Review of Economic Studies,* 65**,** 261-294.

HM Treasury (2011) *The Magenta Book: guide to evaluation*.

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

Rosenbaum, P. R. and Rubin, D. B. (1985) 'Constructing a control group using multivariate matched sampling methods that incorporate the propensity score', *The American Statistician,* 39**,** 33-38.

Smith, J. A. and Todd, P. E. (2005) 'Does matching overcome LaLonde's critique of nonexperimental estimators?' *Journal of econometrics,* 125**,** 305-353.