European
Commission

$\sigma^2 = \frac{1}{n}\Sigma(x_i-\bar{x})^2$
$S_x^2 = \frac{1}{n-1}\Sigma(x_i-$
$\sigma = \sqrt{\frac{1}{n}\Sigma(x_i-\bar{x})^2}$

# How to use administrative data for European Social Funds counterfactual impact evaluations

## A step-by-step guide for managing authorities

Social
Europe

# How to use administrative data for European Social Funds counterfactual impact evaluations

## A step-by-step guide for managing authorities

**European Commission**
Directorate-General for Employment Social Affairs and Inclusion
Directorate G

**This guidance material has been produced based on the work by the following experts: Gianluca Mazzarella and Paolo Paruolo, Centre for Research on Impact Evaluation (CRIE), Joint Research Centre (JRC) and the European Commission**

## LEGAL NOTICE

Manuscript completed in May 2020

Neither the European Commission nor any person acting on behalf of the European Commission is responsible for the use that might be made of the following information. More information on the European Union is available on the Internet (http://www.europa.eu).

Luxembourg: Publications Office of the European Union, 2020

PDF ISBN 978-92-76-17908-5 doi:10.2767/721497 KE-02-20-253-EN-N

## Contents

# 1. Introduction

This document is a step-by-step guide for managing authorities and other evaluators of European Social Fund (ESF) interventions to help them build capacity on the use of administrative data for counterfactual impact evaluations.

Administrative data are systematically collected by governments or other organisations for administrative purposes. Due to their accuracy and low cost, administrative data can be very useful for evaluations. The advantages of using administrative data also include the possibilities to link the data to survey data and to update them. In some cases, administrative registries alone contain the information needed for the entire evaluation exercise. However, most of the time the preferred option would be to link administrative data to survey data as the two types of information are complementary. Survey data, for instance, give insights into areas such as perceptions and detailed household expenditures. This type of information cannot be measured by administrative data. This document will focus only on the use of administrative registries for evaluation purposes.

Administrative data (possibly linked to survey data) are useful for recording both policy outcomes and contextual information (i.e. other factors that affect the outcome in addition to the policy). Data on outcomes and contextual factors need to be linked to programme implementation data concerning who is treated and who is not treated.

In a nutshell, the entire process could be summarised with the following steps.

KEY STEPS FOR USING ADMINISTRATIVE DATA FOR COUNTERFACTUAL IMPACT EVALUATION

Step 1: Analyse the intervention (especially its objectives and the mechanism used to s elect participants) to decide on the data needs and the CIE method to be used.

Step 2: Identify the data sources for (1) monitoring data (data on participants, details and timing of the intervention); (2) pre-intervention information (demographic characteristics, pre-intervention outcomes, territorial information); (3) post-intervention information (outcomes related to the intervention's objectives).

Step 3: Identify the institutions collecting the data identified in the second step

Step 4: Learn about the data access rules of each data holder and reach a new agreement if necessary

Step 5: Link and anonymise or pseudonymise the data

Step 6: Aggregate the data in bigger categories

Step 7: Distribute the data to evaluators, choosing one of three options: physical transfer, secure access or secure labs

The rest of the document will explain the steps summarised above in more details.

## 2. ADVANTAGES OF USING ADMINISTRATIVE DATA

The main advantages of using administrative data for evaluation are set out below.

1. **Cost and time savings:** Administrative data are already collected all the time, so they can eliminate the need to design a survey or a sampling plan, run a survey, correct non-responses, etc.
2. **Completeness:** Administrative data usually include information about the entire population and are very detailed. This level of detail and completeness would be hard to collect through surveys.
3. **Flow vs stock sampling**: Survey data typically take a picture of the population of interest at a certain moment in time. This is called *stock sampling*. Administrative data, in contrast, come from a continuous flow of information, which can be used to reconstruct the required information at any point in time.
4. **Accuracy:** Surveys contain *self-reported* information that may limit data quality and create a risk of *recall bias*[1] or other social biases e.g. due to fear of the tax authorities, people may underreport their income. Administrative data, in contrast, provide more accurate information.
5. **Administrative data can be used to design the sample of a survey:** Because administrative data relate to the whole population, they can serve as the backbone for carrying out surveys. Finally surveys and administrative data can be linked to complement the information contained in both sources.

[1] *Recall bias* (or *response bias*) is the possibility that a respondent reports wrong (or simply inaccurate) information in a survey. This can be the case when the interviewer asks the respondent to recall work periods.

## 3. Challenges of using administrative data

While administrative data have many advantages (as explained above), they also present some disadvantages.

1. **Bureaucratic burden:** It is very unlikely that all the data sources needed for the evaluation will be collected by the same institution; often it is necessary to link different sources in order to have a comprehensive database. This means that the institutions holding the data (ministries, regions, agencies, etc.) have to coordinate their efforts to link their data and to make them available to the evaluator (possibly in an anonymised form). This involves specific agreements with all of these institutions, which may require time and constitute additional bureaucratic burden.
2. **Need for harmonisation:** Administrative data are typically collected by governments for registration, transaction and record keeping and not for the purpose of evaluation. Therefore, it is likely that different institutions collect the same information in different ways, triggering a need for harmonisation. A typical example of this is the collection of information about qualifications classified in different categories. In addition, it is important to pay attention to the timeframe over which the data are collected: some institutions may have a time lag in providing data. Finally, data may refer to different populations. For example, some registries cover the national population, whereas others could be at regional level. This is something that has to be considered when constructing a harmonised archive.
3. **Need for anonymisation and pseudonymisation:** Administrative data can contain personal information which cannot be disclosed. Later in this document we will explain some necessary steps to properly handle personal data.

## 4. USING ADMINISTRATIVE DATA FOR EVALUATION: KEY STEPS

The next part of the document guides managing authorities through the different steps of using administrative data for evaluation, starting from **identifying the key characteristics** of the intervention to be evaluated, and ending with **delivery of the data** to the evaluator.

### 4.1 Characteristics of the intervention

On the one hand, there is a direct connection between the intervention to be evaluated and the data necessary for the evaluation. In particular, the selection of participants largely influences the **counterfactual impact evaluation** (CIE) method that could be used. On the other hand, the CIE method determines both: (i) the 'pre-intervention information' that has to be contained in the data (e.g. propensity score matching requires a lot of 'pre-treatment information' in order to be reliable); and (ii) the subsample required for the evaluation. In addition, the objectives of the intervention determine the evaluation's outcomes (these are what is known as 'post-intervention information').

*Selection of participants*

First of all, all the relevant information determining whether an individual is eligible (or not) for the intervention has to be present in the data. Some examples are provided below.

1. If the intervention is focused on individuals in a certain **age range**, the day (or at least the month) of birth has to be provided.
2. If the intervention targets people who live in specific **areas**, residential data must be available. The granularity of the data must be at least at the level to which the intervention is assigned. For example, if the intervention is assigned at NUTS2 level, this has to be the minimum granularity level available in the data.
3. If the intervention supports workers who belong to a specific **sector**, this has to be available in the data.

Moreover, the CIE evaluation process tries to compare individuals who are as similar as possible. For that reason, it might not be necessary to extract all individuals' records from the administrative registers, but only a subgroup of them. In the evaluation process, there is a need for an appropriate <u>control group</u> and for a <u>treated group</u> (i.e. those who participated in the intervention). For example, if the aim is to evaluate a *Youth Employment Initiative* intervention focused on people below 25 years old, the evaluator would not need information on people older than 35, irrespective of the method used.

This sample selection must be done very accurately; otherwise, the evaluator would not be able to find all the necessary information. It is recommendable to try to be 'generous' in the selection (e.g. for the Youth Employment Initiative do not select individuals that are below, say, 27/28 years old, but all the individuals that are below 35 years old).

*Objectives of the intervention*

It is important to be sure that evaluators find all the 'post-intervention information' they need in the data. For this reason, one needs to pay particular attention to the issues below.

1. ***Which*** outcomes are directly connected to the intervention's objective?

   For example, if the intervention is aimed at supporting youth employment, the evaluator has to find all outcomes about job situation (employment status, wages, etc.); if the intervention is about early school leavers, the evaluator needs data about school attendance. Furthermore, some outcomes may not be among the intervention's primary outcomes but may be indirectly affected by the policy and, as such, help better understand the overall effectiveness of the policy. Even if such outcomes are not compulsory for the evaluation process, collecting them could provide significant added value.

2. ***When*** should the outcomes be collected?

   It is important to pay particular attention to the *timeframe* for the outcomes. For example, if the intervention is aimed at 'long-term results', we must be sure the data covers a sufficient period after the end of the intervention.

## 4.2 From data needs to data sources

Unfortunately, it is very unlikely that all the information the evaluator needs will be available in the same data source. Most likely, it will be necessary to find complementary sources, collected probably by different institutions. When carrying out this task, it is important to: (i) keep in mind that the three main categories of data should be complementary; and (ii) identify which data sources have been included. In the latter case, there is not necessarily a one-to-one relation between the categories of data and data sources. For example, some additional information may be contained in monitoring data. There are three main categories of data.

1. **Monitoring data:** the evaluator will definitively need the monitoring registries to know who the participants in the intervention are, when they started/concluded the intervention, and all possible information concerning different activities and their intensity. The data have to be as detailed and reliable as possible to avoid misclassification and differing treatment of the same intervention.
   a. *Misclassifications:* If the data are not reliable, there is a risk that the evaluator will consider some individuals as a 'control' who have actually been treated.
   b. *Differing treatment of the same intervention:* The same ESF-financed intervention may, for example, include different training programmes, with some of them effective and others not. Evaluators will only be able to make this distinction if they have enough detailed information.

2. **Pre-intervention information:** the evaluation's quality also depends on the richness of the 'pre-treatment information' provided. For this reason, it is very important to collect enough data on the individuals' situation before the intervention starts. Possible pre-intervention information falls into three broad categories.
   *a.* *Basic information:* the starting point is to make sure that all the basic demographic information is available, e.g. date (or at least month) of birth, gender, level of education, marital status, citizenship.
   *b.* *Work histories/school attainment:* if the evaluated intervention is supporting employment, it is fundamental that the evaluator has access to individuals' previous work histories. If the intervention is supporting an early school leaver, or targeting schools in general, the evaluator will probably need information about school attendance and the marks of the children. It is also important to make sure all the outcomes being evaluated in the intervention have been measured before the intervention. In this way, differences between the post- and the pre-intervention situation can be measured.
   *c.* *Territorial data:* spatial information is a relevant aspect to take into account. It can help evaluators understand if the individuals exposed to the intervention are concentrated in some areas more than in others, and if these areas are different with respect to certain other dimensions (higher unemployment, school dropout rate, etc.) that can influence the intervention's success. Such data becomes fundamental if the intervention is designed to provide support only to individuals that live in certain areas (such as poor regions).

3. **Post-intervention information:** Finally, one needs to carefully identify all outcomes related to the intervention's specific objective.

As it is very unlikely to find all this information in the same data source, it is important to understand in which registries the relevant information is held.

In order to link datasets, *unique identifiers* are useful. For example, in general, a name and surname are not enough on their own to serve as the unique identifier of an individual, as different individuals may share the same name. Good examples of unique identifiers are ***social security codes*** as these are specifically generated to be unique in the population.

## 4.3 Data owner(s)

After identifying all the different data sources necessary for the evaluation, we have to identify the institutions that collect the data. Having done so, it is important to understand if the data holders have already given access to data for evaluation/scientific purposes. If so, they probably have their own policy for anonymisation and safe access to the data (to be discussed in the next two sections). Otherwise, these rules will have to be set up jointly with the institution that owns the data. If the rules are already in place, it is important to carefully check two things.

1. Do the rules fit with the evaluator's needs, in particular those concerning anonymisation? Usually the first step of an evaluation study is to estimate the 'take-up rate', i.e. the fraction of eligible individuals that participated in the intervention. This involves the evaluator reconstructing the eligibility from a data source that includes participants and non-participants and then merging this source with the monitoring data. For this reason, the data have to be anonymised in a consistent way.
2. Is use of the data allowed for evaluation purposes?

If all requirements are met and the data fits the evaluation needs, the process is finished. You can skip the remaining part of this document!

If the institution has not yet distributed its data or if its requirement does not fit with the evaluation needs, a new agreement will have to be negotiated.

The next two sections cover main aspects of (a) the anonymisation of personal information in the data; and (b) policies concerning distribution as required by the data owner. It may be the case that not all the data sources are directly managed by the managing authority. However, it is important that the MAs know about these aspects in order to reassure the other data owner and to drive the step-by-step creation of the statistical database.

## 4.4 Privacy issues

Administrative data contain personal information that can be sensitive to share. For this reason, all the steps described in this document (anonymisation, transfer of the data, data storage, etc.) have to comply with the national regulations and with the *General Data Protection Regulation*. For this reason, some safeguards need to be put in place before the data are made accessible to evaluators.

One option is to have the entire dataset delivered to the national statistical office, which can take charge of all the steps described in this section. Other options exist in Member States that have a specific agency that can link personal data. Finally, the different institution/agencies can cooperate, pseudonymise the data in a consistent way (as will be explained in the next section) and release the different pseudonymised data sources.

The two main actions that have to be taken are:

1. **data anonymisation**
2. **aggregation of detailed categories into bigger classes.**

As we have seen before, the data acquisition process is not easy. As a result, it seems reasonable to construct a *harmonised* version of the data that can be used not only for a single evaluation, but also for a series of evaluations the country has to provide and, if necessary, for research purposes. Many countries have created an infrastructure in which all the available administrative data and the available survey data are contained. For this reason, in principle, if the aim is to construct a harmonised infrastructure, no information needs to be aggregated, even if it is not relevant for the specific evaluation, as it may be important for further studies. As such, aggregation only needs to be considered if the data has to be physically delivered to an external evaluator for a single study.

## 4.5 Anonymisation of data

Information such as names, surnames and social security codes cannot be left in the data given to the final evaluator. However, it is essential that the data contain an anonymous identifier of the individuals so that it is clear whether two records refer to the different individuals or to the same person (e.g. two periods of work by the same person). In addition, the evaluator will need to be able to link data from different anonymised sources.

The anonymisation process requires some basic knowledge of ‘unique identifiers’. A unique identifier is a characteristic or set of characteristics that cannot be the same for two different individuals in the population. A social security code is a typical example of a unique identifier. An example of a set of characteristics that can uniquely identify an individual when taken jointly is that person’s name, surname, place of birth and date of birth.

One possibility would be to substitute these characteristics with progressive numbers, which would apply a simple method of anonymisation to the data. However, the evaluator will no longer be able to link different sources, and the entire anonymisation process will have to start again from scratch every time the dataset has to be updated.

For all these reasons, the best option is to **encrypt** this information in a safe way that would still enable the ***unique identification of*** the individuals in the data to allow further linkage. This process is usually called *pseudonymisation*.

Pseudonymisation has to be done using appropriate encrypting algorithms and not by simply substituting these variables with progressive numbers. These algorithms will ensure that the same individual listed in two (or more) different sources is anonymised consistently. In addition, if a data source has to be added subsequently to the list or simply updated, there will be no need to anonymise all the other sources again, just the one that has been added/updated.

For example, let us assume that the name is sufficient to identify an individual. Set out below is a list of people listed in the dataset containing the periods of employment (left-hand table) and of people who received the training programme (right-hand table):

| |
|---|
| Amelia |
| Oliver |
| Margaret |
| Jack |
| Emma |
| Harry |
| Mary |
| Jacob |
| Samantha |
| Charlie |

| |
|---|
| Amelia |
| Oliver |
| Margaret |
| Jack |
| Emma |
| Mary |
| Jacob |
| Samantha |

The (2 panels of the) next table report anonymised individuals using one of the most common pseudonymisation algorithms (the 'secure hash algorithm' with 256 bit encryption):

| Name | Anonymised identifier |
|---|---|
| Amelia | 99c9739507c22853bc4c94dd1f32128f7166535ddb40223342a7eceaace75542 |
| Oliver | fe43748968757a8b155064e53c7e485492944af270639a64c2df7e356420647a |
| Margaret | 120db209fa630e4d62c48bdf80e1bc8d15dcf679b5c7acb2b149a8418af71d8d |
| Jack | 0cdb3938b8c06c35c9cdc170d04838897974c4068e90376196212cec44d84648 |
| Emma | 4902c6142cdf355d71f08ec7bfcfcf3e2d0eb27dd29a2645b0e84c5118042360 |
| Harry | f1c592491d592fb044782f273a273c6b5b8ad162d30cfe348d32ee098bbfafcc |
| Mary | 343252b076100987e6f581f408a7f48db74bfcc91d37a9b02a4505b7b329cbeb |
| Jacob | a5641fc5e195539d68981f92c7d4162a3196938cb706440bf47a792bc0c90e07 |
| Samantha | b8facd2633acd807d66f5e1a02227cb896aa1bf3ef23f71117fc57f75bcccc0a |
| Charlie | e88ec278eee7099ece595084a72af8406b4a0df2c3fe2849f92eca9b24228d81 |

| Name | Anonymised identifier |
|---|---|
| Amelia | 99c9739507c22853bc4c94dd1f32128f7166535ddb40223342a7eceaace75542 |
| Oliver | fe43748968757a8b155064e53c7e485492944af270639a64c2df7e356420647a |
| Margaret | 120db209fa630e4d62c48bdf80e1bc8d15dcf679b5c7acb2b149a8418af71d8d |
| Jack | 0cdb3938b8c06c35c9cdc170d04838897974c4068e90376196212cec44d84648 |
| Emma | 4902c6142cdf355d71f08ec7bfcfcf3e2d0eb27dd29a2645b0e84c5118042360 |
| Mary | 343252b076100987e6f581f408a7f48db74bfcc91d37a9b02a4505b7b329cbeb |
| Jacob | a5641fc5e195539d68981f92c7d4162a3196938cb706440bf47a792bc0c90e07 |
| Samantha | b8facd2633acd807d66f5e1a02227cb896aa1bf3ef23f71117fc57f75bcccc0a |

The same individuals can be recognised in two different datasets without the need for their name to be known. In other words, pseudonymisation algorithms preserve the possibility to carry out data linkage.

## 4.6 Aggregation of detailed categories in bigger classes

Some information contained in the data may be too detailed, causing an individual's identity to be revealed. However, this information may be difficult to delete because of its relevance for the evaluation.

For example, while data recording an individual's exact address are not relevant for evaluators, they still need to know the area where the people participating in the intervention reside.

To ensure that evaluators can perform all the 'spatial analysis' they need, the address can be recorded in less detailed categories i.e. the *city* or, better still, *districts* (if available). Other examples come from:

1. information concerning the employer, which can be substituted with information about the sector of industry
2. different qualifications, which can be recorded using the highest achievement.

## 4.7 Distribution of the data

Data delivered to the evaluator should be covered by some kind of confidentiality disclosure clause. The evaluator cannot use these data for other purposes, nor share the data with people not directly involved in the evaluation.

Additional measures can be taken to guarantee confidentiality. This section summarises the three main ways of sharing the data with the evaluator.

The first solution (physical transfer of the data) is the most common, but it is also the least secure. The second and third solutions (secure access and secure labs respectively) require time and money so are best only used if there are a certain number of planned evaluations or if the data have to be distributed for other purposes as well (e.g. monitoring, scientific research). In the latter two solutions, the data owner would have control of all individuals who can access the data and would be able to check all outputs produced using it. This is not feasible if the data are physically transferred to the evaluator.

### *Physical transfer of the data*

The first possibility is that the data are physically delivered to the evaluator. If so, one needs to remember to ensure that the evaluator provides secure storage that guarantees the security of the data provided. It would be preferable for the data to be saved using a standard format (such as.csv,.dat,.txt, etc.). In this way, evaluators will take charge of the conversion of the data in their preferred format.

While this solution is obviously the least secure option, it is the cheapest. Under this approach, evaluators have to provide the statistical software necessary for the data analysis and, as stated, the secure storage that will hold the data.

### *Secure access*

Another option is to share the data on a dedicated server, providing remote access only to the evaluator.

Under this approach, the evaluator connects to a remote server featuring a *virtual desktop* with the data to be analysed. In this way, not only is the data stored in a secure environment, but the institution responsible for the data can also check the evaluator's calculations and prevent disclosure of a table with too few observations, which would enable individuals to be identified.

On the other hand, in this setup, the institution has to provide the evaluator with all the statistical software necessary for the evaluation. This can increase the cost of the evaluation process.

### *Secure labs*

The safest option is to dedicate a specific room for analysis of the data, known as a *secure lab*. Secure labs are rooms which only authorised people can enter and access the computer on which the data are stored (physically or remotely) and analysed. Typically, these rooms have darkened windows so that people outside the room cannot see the monitors inside, and/or are kept under camera surveillance to guarantee that users do not break lab rules. In addition, the computers in the room do not have an internet connection to prevent the information from being shared online.

When the secure labs approach is used, the institution responsible for the data has to provide the evaluator with all the facilities necessary for the evaluation (such as statistical software) and will have direct control of the outputs produced by the evaluator.

## Getting in touch with the EU

### In person

All over the European Union there are hundreds of Europe Direct Information Centres. You can find the address of the centre nearest you at: http://europa.eu/contact

### On the phone or by e-mail

Europe Direct is a service that answers your questions about the European Union. You can contact this service

– by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),

– at the following standard number: +32 22999696 or

– by electronic mail via: http://europa.eu/contact

## Finding information about the EU

### Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: http://europa.eu

### EU Publications

You can download or order free and priced EU publications from EU Bookshop at: http://bookshop.europa.eu. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see http://europa.eu/contact)

### EU law and related documents

For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex at: http://eur-lex.europa.eu

### Open data from the EU

The EU Open Data Portal (http://data.europa.eu/euodp/en/data) provides access to datasets from the EU. Data can be downloaded and reused for free, both for commercial and non-commercial purposes.

Publications Office
of the European Union