ecdc
EUROPEAN CENTRE FOR
DISEASE PREVENTION
AND CONTROL

TECHNICAL GUIDANCE

Trend analysis guidance for surveillance data

www.ecdc.europa.eu

ECDC TECHNICAL GUIDANCE

Trend analysis guidance for surveillance data

EUROPEAN CENTRE FOR
DISEASE PREVENTION
AND CONTROL

This report was commissioned by the European Centre for Disease Prevention and Control (ECDC), coordinated by Joana Gomes Dias, and produced by Epiconcept (Paris, France).

Contributing authors

Epiconcept: David Prieto-Merino, Esther Kissling and Lore Merdrignac

ECDC: Joana Gomes Dias, Chantal Quinten, Chiara Entradi and Gaetano Marrone

Suggested citation: European Centre for Disease Prevention and Control. Trend analysis guidance for surveillance data. Stockholm: ECDC; 2024.

Stockholm, 2024

ISBN 978-92-9498-616-0
doi: 10.2900/294038
Catalogue number TQ-05-23-037-EN-N

Contents

Figures

Tables

Abbreviations

AIC	Akaike information criterion
BIC	Bayesian information criterion
EU/EEA	European Union/European Economic Area
MAR	Missing at random
MCAR	Missing completely at random
MNAR	Missing not at random
TESSy	The European Surveillance System

Executive Summary

Monitoring trends in communicable diseases is one of the six objectives of ECDC's long-term surveillance strategy. To facilitate this activity, ECDC collects information on communicable diseases and related health events from European Union and European Economic Area (EU/EEA) countries. This communicable disease surveillance at the European level allows ECDC to analyse the data and disseminate information that supports the prioritisation of resources, identification of outbreaks and evaluation of the impact of control measures. Collectively, these efforts aid the effective prevention and control of infectious diseases, while also minimising the burden on EU/EEA countries.

Trend analysis is an important methodological tool used to identify patterns in communicable disease surveillance data. The data that are analysed are a collection of observations made sequentially through time. In epidemiological surveillance, counts of disease cases or health events are organised sequentially for a specified time unit and over a set period of time.

A trend analysis can provide:

- **A depiction of patterns of change** in an indicator over time (e.g. information on increases or decreases and the magnitude thereof). This is the main focus of this document.
- **A prediction of future measures** of indicators by projecting the pattern into the future.
- **A baseline to compare time periods**, by comparing the patterns before and after a public health intervention.
- **A baseline to help detect outbreaks or events** when the number of cases exceeds what is expected according to the pattern.

This document supports those undertaking trend analyses using regression models. It provides guidance so that experts can carry out a trend analysis independently and understand when extra statistical support is required.

For the purposes of this document, it is assumed that trend analysis is done with surveillance data. Therefore, all references to 'data' refer to surveillance data. Additionally, it is assumed that all trend analyses are done by time. Thus, references to 'trend analysis' refer to time trend analysis unless otherwise indicated.

Trend analysis requires careful consideration and execution to make the most correct inferences around the data. This document outlines the objectives, strengths and limitations of each of the methodological approaches, offering a holistic perspective of the effects of each potential choice and how it may affect the analysis.

Objectives

The specific objectives of this document are to:

- describe the approach needed for trend analysis for a given dataset with specific characteristics (e.g. a given type of trend, a particular geographical unit or set of units, or a data completeness issue) and
- illustrate the approach by providing concrete examples that use ECDC datasets.

Code that can be used for trend analysis (using the statistical software packages R and Stata)—including describing and exploring data, performing trend analysis and selecting a model—has also been provided as supplementary material.

Structure of this document

The guidance in this document is divided into four parts, as follows:

- **Part A** – **Considerations before embarking on a trend analysis:** This section covers the preliminary questions one should consider to determine when it is appropriate to do trend analysis, as well as the underlying assumptions.
- **Part B** – **Trend analysis fundamentals:** This section covers the steps required to carry out a trend analysis in a dataset covering one geographical unit.
- **Part C** – **Trend analysis across multiple geographical areas:** This section covers how to take several geographical units into account in a trend analysis.
- **Part D** – **Trend analysis with missing data:** This section covers how to deal with missing data when conducting a trend analysis.

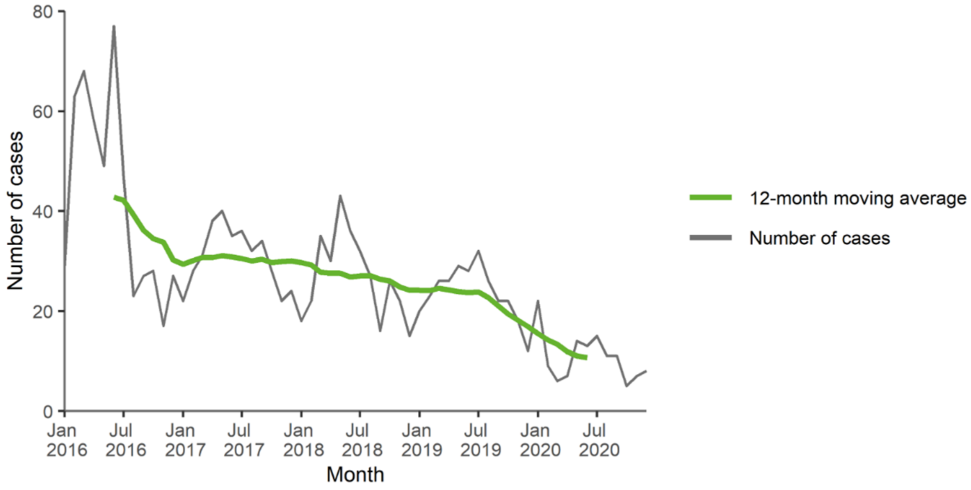
Introduction

ECDC collects information on communicable diseases and related health events from European Union and European Economic Area (EU/EEA) countries. This communicable disease surveillance at the European level allows ECDC to analyse the data and disseminate information for the effective prevention and control of infectious diseases, while also minimising the burden on EU/EEA countries [1]. Monitoring trends in communicable diseases is one of the six objectives of ECDC's long-term surveillance strategy.

What is trend analysis?

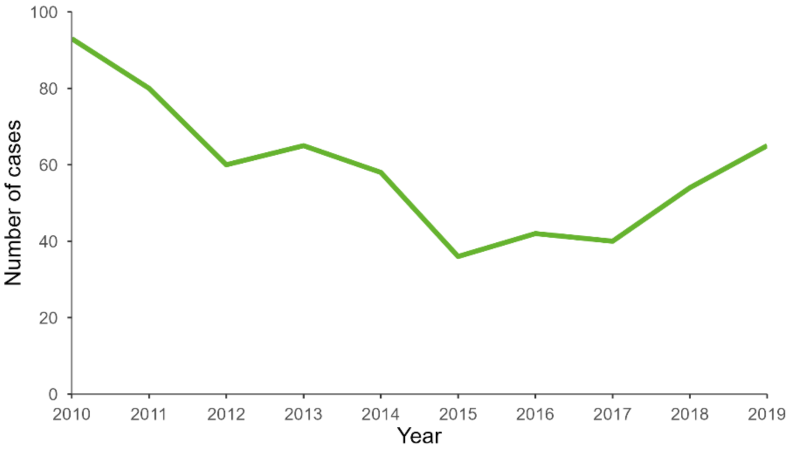
Trend analysis of surveillance data is carried out to identify patterns in the data. It is analysed longitudinally and may also be called 'time series analysis'. The data used are a collection of observations made sequentially through time [2]. In epidemiological surveillance, counts of disease cases or health events are organised sequentially for a specified time unit. Examples of tables and figures of surveillance data can be found in the ECDC annual epidemiological reports [3] (Figure 1, Figure 2).

Figure 1. Number of confirmed brucellosis cases by month, EU/EEA, 2016–2020



Source: Country reports from Austria, Cyprus, Czechia, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Malta, the Netherlands, Norway, Poland, Portugal, Romania, Slovakia, Slovenia, Spain and Sweden

Figure 2. Number of confirmed congenital syphilis cases by year, EU/EEA countries reporting consistently, 2010–2019



Source: Country reports from Bulgaria, Cyprus, Czechia, Denmark, Estonia, Germany, Hungary, Iceland, Ireland, Latvia, Lithuania, Luxembourg, Malta, Norway, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden and the United Kingdom.
The United Kingdom (UK) was a Member State of the European Union (EU) at the time of collating the data for this report. The UK withdrew from the EU on 31 January 2020. Only countries reporting consistently have been included and countries that only report acute cases have been excluded.

Objectives of a trend analysis

A trend analysis can provide:

- **A depiction of patterns of change** in an indicator over time (e.g. information on increases or decreases and the magnitude thereof). This is the main focus of this document.
- **A prediction of future measures** of indicators by projecting the pattern into the future.
- **A baseline to compare time periods**, by comparing the patterns before and after a public health intervention.
- **A baseline to help detect outbreaks or events** when the number of cases exceeds what is expected according to the pattern.

The above information can help contribute to prioritisation of resources, identification of outbreaks and evaluation of the impact of control measures. Note that not all types of trend analyses may be possible or valid, given the data source.

Aim and objectives of this document

This is an operational document for those undertaking trend analyses using regression models. It provides guidance so that experts can carry out a trend analysis independently and understand when extra statistical support is needed.

The specific objectives of this document are to:

- describe the approach needed for trend analysis for a given dataset with specific characteristics (e.g. a given type of trend, a particular geographical unit or set of units, or a data completeness issue) and
- illustrate the approach by providing concrete examples that use ECDC datasets.

Code that can be used for trend analysis (using the statistical software packages R and Stata)—including describing and exploring data, performing trend analysis and selecting a model—has also been provided as supplementary material.

For the purposes of this document, it is assumed that trend analysis is done with surveillance data. Therefore, 'data' in this report refers to surveillance data. Additionally, it is assumed that all trend analyses are done by time, and thus references to 'trend analysis' in this document should, unless otherwise indicated, be assumed to be synonymous with time trend analysis.

Structure of this document

The guidance in this document is divided into four parts, which are as follows:

- **Part A – Considerations before embarking on a trend analysis:** This section covers the minimum requirements to do a trend analysis and the preliminary questions one should consider to determine the appropriateness of a trend analysis, as well as the underlying assumptions informing the choices.
- **Part B – Trend analysis fundamentals:** This section addresses the theoretical underpinnings of trend analysis. It covers aspects such as the type of trends, the concept of seasonality, the choice of statistical models and the algorithmic steps required to carry out a trend analysis in a dataset covering one geographical unit.
- **Part C – Trend analysis across multiple regional areas:** This section addresses considerations for the inclusion of multiple geographical units in a trend analysis, including the theoretical framework and the steps required to carry out a trend analysis in a dataset that covers several geographical units.
- **Part D – Trend analysis with missing data:** This section discusses the various patterns and mechanisms of missing data, as well as the methods to deal with them when conducting a trend analysis.

Datasets used as examples

The datasets used as examples in this document are freely available from ECDC's 'Surveillance Atlas of Infectious Diseases' [4]. These are some of the datasets used:

- **Reported cases of confirmed salmonellosis:** Monthly reported totals of confirmed salmonellosis cases at the EU/EEA level from 2010 to 2019 are used in two examples in Part B. In the first example, these totals are aggregated by year. In the second example, monthly totals are used. Yearly reported totals of confirmed salmonellosis cases from 2007 to 2016 are used in two examples in Part D. Some missing data have been artificially introduced in order to meet the example objectives.
- **HIV notification rates:** Yearly notification rates of confirmed cases of HIV infection from 2007 to 2016 are used in an example in Part C.
- **Hepatitis B notification rates:** Yearly notification rates of all cases of hepatitis B infection (regardless of disease phase) from 2008 to 2017 are used in an example in Part C. Two countries were excluded due to incomplete data.

All datasets can be found in the supporting material.

Using this document outside of an ECDC context

While this document was commissioned by ECDC and written primarily with ECDC disease experts in mind, it can be used by anyone carrying out a trend analysis.

Part B focuses on a pooled trend analysis of countries to provide an estimate of the trend in the EU/EEA. However, this type of analysis could also be applied to other geographical units, such as regions within a country.

Part A. Considerations before embarking on a trend analysis

Before carrying out a trend analysis, there are the two fundamental questions to answer:

- Are the data suitable to be analysed for trend (for their structure and their relationship with time)?
- Is trend analysis the appropriate technique to answer the question that needs to be addressed?

1. Minimum requirements of data to do trend analysis

1.1 Stable data source

Prior to trend analysis, it is important to understand the data source(s) well. If the systems collecting the data are not stable over the study period, then trend analysis may be biased [5]. This could include changes to the case definitions or the measurement procedures. For some changes in data sources (e.g. an increase in the number of sentinel general practitioners reporting a particular syndrome), statistical methods can be used to account for these changes (e.g. collecting information about catchment areas to calculate rates, rather than using absolute numbers). However, not all changes can be easily accounted for.

Before carrying out a trend analysis, qualitatively evaluate the data source and investigate possible changes. This can be done by consulting with experts who are familiar with the underlying surveillance systems, for example. Questions to ask include:

- Were there any known changes to clinical or laboratory case definitions over time?
- Were there any known changes to measurement procedures over time?
- Were there any known data submission/collection issues over time?
- Were there any known changes to who did the reporting over time?

Obtaining a clear and thorough understanding of the data is the most important part of performing a trend analysis.

1.2 Data availability

The choice of time or study period for analysis may be limited by the start date of the surveillance. For example, the study period will be limited for a new disease or pathogen, such as COVID-19.

2. What type of question is trend analysis appropriate for?

Trend analysis can be applied to historical data to describe trends. These trends can also be used to project data for future time periods, which is known as forecasting. However, forecasting comes with several caveats.

In the context of this document, trend analysis takes two variables into account: the outcome (e.g. the measure of the health topic) and the time unit. It is important to note that while time may be correlated with the outcome, time—by itself—does not cause the disease or health event. There are generally many other factors changing over time that are associated with changes in the outcome. When forecasting, the assumption is that the time-dependency of these factors is stable. If it is not stable, then a forecast using trend analysis will be wrong. For example, a disease may be related to weather. The weather conditions may have remained stable during the time period where the trend is estimated. However, if the weather changes dramatically in a future time period, the real disease pattern can be very different from what was forecast.

Additionally, the further forward in time the forecast or projection, the more uncertainty around the forecast. Any projection of time series trends should be applied to the short-term only.

Nevertheless, trend analysis is a pragmatic approach to prioritising resources based on trends. But it is important to remember that uncertainty increases after the short-term.

Part B. Trend analysis fundamentals

1. Theory

1.1 Study period

The study period (the number of years or months) to include in the trend analysis will depend on various factors, particularly the availability of stable surveillance data. Additionally, factors external to the surveillance procedure may influence the choice of study period, including changes to policy (e.g. the introduction of a vaccination campaign or hepatitis C screening). These factors may have a strong effect on the disease under study. Depending on the study objectives, the study period could exclude when this external event occurred (if it is not relevant to the surveillance question being asked) or this time period could be included and the factor modelled explicitly in the analysis. Note, explicit modelling is not covered in this guidance document.

Interrupted time series analysis

When evaluating trends in surveillance data before and after an intervention (e.g. introduction of a vaccination campaign) a quasi-experimental study design called interrupted time series analysis can be used. In brief, the underlying pre-intervention trend is estimated in an attempt to understand if there are any changes after the intervention.

This is a powerful study design that is reasonably easy to carry out and generates a message that is easy to understand. However, it cannot make inferences about individual-level outcomes. Factors other than the intervention in question may affect the differences in trends.

For further reading on interrupted time series analysis, see Kontopantelis E, et al. (2015) [6] and Lopez Bernal J, et al. (2016) [7].

The length of the study period will also depend on the previously mentioned factors. While, in theory, a longer time period will increase the number of data points and may improve the certainty around a trend, often there can be issues of stability of the surveillance system over long periods of time. Assessing trends in a dataset with a long time period may also introduce an additional element of complexity. Secular trends (e.g. seasonality of five years, 10 years, etc.) may occur. ECDC often uses 10 years as a study period when doing trend analysis.

1.2 Unit of aggregation

Surveillance data can have different units of aggregation by time. The source data may be, for example, daily, weekly, monthly or yearly data. The unit of aggregation to choose for a trend analysis depends on the objectives of the analysis.

If the seasonality (patterns of change within a natural year) is not of interest, then aggregate the data by year and carry out a trend analysis by year. Note that if the data are in units at a higher resolution than a year and one chooses to ignore seasonality and not include any seasonal terms in the model, then the model will be misspecified if seasonality does exist. The inferences from the model can be biased.

If seasonality is of interest, then assess the time unit by which to aggregate the data. The choice of data aggregation can depend on a variety of factors, including:

- **Missing data**: For a proportion of the data, the precision in the time unit variable may be lower. For example, only the week of onset may be available, rather than the day of onset. If many records are missing the day of onset, then the day as a unit of aggregation in the trend analysis may not be appropriate.
- **Noisy data**: The higher the resolution of unit of aggregation by time is, the more 'noise' there will be in the data. Daily data may vary considerably from day to day and produce more variance in the final model. The data will be difficult to visualise. When the same data are aggregated by week of reporting, the time series may be smoother. On the other hand, when data analyses are carried out by the day of reporting, bias due to weekend effects may occur.
- **Weeks versus months:** If the objective is to describe annual seasonality in surveillance data, then aggregation by weeks or months are often good choices. A good understanding of the health topic under study (its expected seasonality, its transmission dynamics), as well as how the data are collected, is important to facilitate choosing between these units of aggregation. Note that each of these choices comes with challenges:
  - **Weeks** can be standardised into International Organization for Standardization (ISO) weeks. ISO weeks always include seven days and begin on a Monday and end on a Sunday. While most years include 52 ISO weeks, this method means that sometimes years can include 53 weeks, which can be a nuisance.

- **Months** do not always have the same number of days, though a year has 12 months. If this affects the results, months may not be the best choice. Additionally, months are a time unit of lower resolution than weeks, so aggregating by months may mask an important seasonality trend.

The choice of time unit of aggregation may depend predominantly on the data available (e.g. if only monthly data are available and no data of a higher resolution). If a choice is possible between units of aggregation (e.g. weeks versus months), data visualisation (plots) and statistical methods assessing model fit can help determine which unit is best to use.

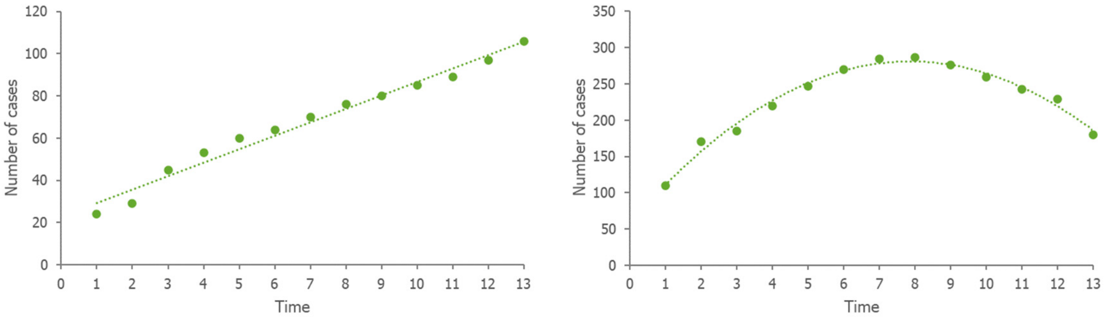
1.3 Definitions

A time series of surveillance data consists of several components. For the purposes of this document, the following nomenclature will be used:

1.3.1 Trend

A **trend** is a pattern that appears in the data over time and does not repeat. There are linear and non-linear trends (Figure 3).

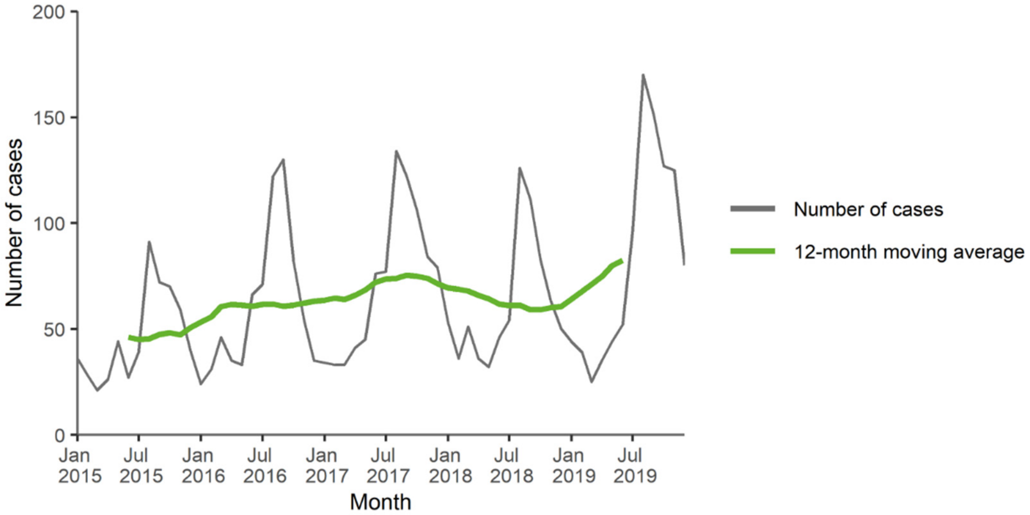
Figure 3. Examples of a linear trend (left) and a non-linear trend (right)



1.3.2 Cyclical variation

Cyclical variation in a time series is a component that changes over some units of calendar time and then repeats. Cyclical variations can happen at different time units. For example, circadian rhythms are cyclical variations repeating themselves over a 24-hour period. This is very common in some physiological parameters. Some time series show yearly seasonality related to the change in seasons. An example of this is a disease that is associated with summer months, like leptospirosis (Figure 4).

Figure 4. Distribution of confirmed leptospirosis cases by month, EU/EEA, 2015–2019



Source: European Centre for Disease Prevention and Control (ECDC). Leptospirosis. Annual Epidemiological Report for 2017. Stockholm: ECDC; 2022 [8].
Data source: Country reports from Austria, Cyprus, Czechia, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, the Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden and the United Kingdom (UK).
The UK was a Member State of the European Union (EU) at the time of collating the data for this report. The UK withdrew from the EU on 31 January 2020.

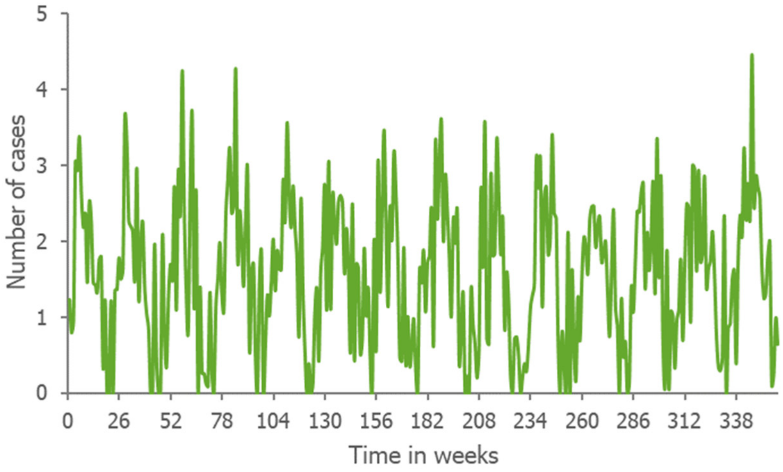
Another example of cyclical variation is a secular trend, where a cycle exists without a fixed frequency, usually over a number of years. An example of this is the two- to three-year cycle in pertussis data. Several cyclical variations might appear combined together within the pattern of a time series.

Trend and seasonality are time-dependent patterns in the data. Their value at time t will be highly correlated to their value at time $t-1$. However, autocorrelation will not be included in the trend analyses for purposes of simplicity.

1.3.3 Noise

The **noise** in the data is a component that is neither a trend nor seasonality. Noise can be random fluctuations (Figure 5), outbreaks or unexpected events underlying the data.

Figure 5. **Example of noisy data by week**



When using yearly data, yearly seasonality will not need to be taken into account. However, if data with a higher resolution than year (i.e. monthly, weekly or daily data) are used, one will often need to take seasonality into account to obtain an accurate estimation of the trend.

2. Types of trends

Different types of trends can exist within a dataset. Some common ones include:

- a polynomial trend of the order 1 (i.e. a **linear** trend),
- a polynomial trend of higher orders (e.g. a **quadratic** or **cubic** trend) and
- an exponential trend.

2.1 Linear trend (polynomial trend of the order 1)

There is a linear trend when the outcome increases or decreases by a constant amount for each unit increase in time (an example is given in Figure 6). This type of trend is easiest to quantify and convey to a layperson. The shape of a linear trend is a straight line.

2.1.1 Generic equation for a linear trend

The equation is a polynomial of order 1:

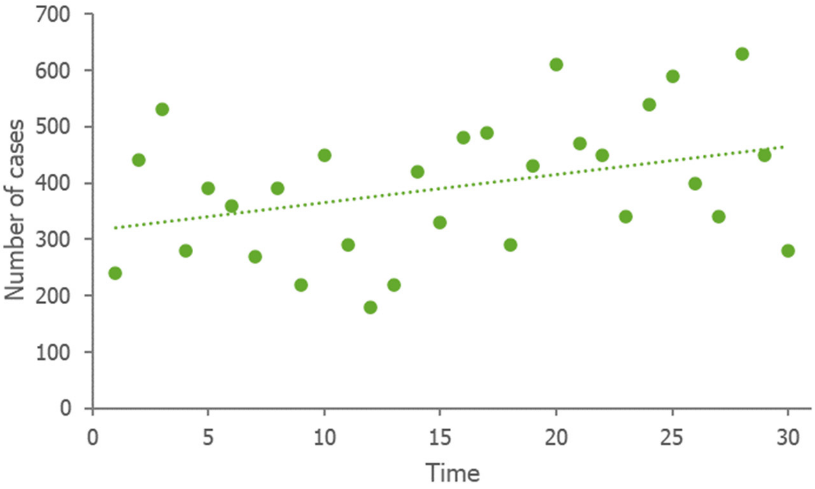
$$y_t = \beta_0 + \beta_1 t$$

where y_t is the outcome at time t, β_1 is the coefficient and β_0 is the intercept. The intercept is where the trendline crosses the y-axis (i.e. when $t = 0$) and the coefficient is the slope of the trendline (i.e. the expected change in the outcome variable for every unit increase in time).

2.1.2 Example of a linear trendline

Figure 6 provides an example of a linear trendline with a polynomial trend of the order 1.

Figure 6. Example of a linear trendline (polynomial order 1), with the number of cases over time



Equation

The equation for the trendline in the linear example is:

$$y_t = 317 + 5t$$

Interpretation

In the example, the coefficient is 5 (i.e. with each increase in time unit, the number of cases is expected to increase by 5). As the intercept is 317 at time $t = 0$, one would expect to see 317 cases.

Note that the time equal to zero $(t = 0)$ (the time origin) will refer to a unit of calendar time depending on how the time variable is coded. For example, if the time unit is years and the time variable is defined as $t = (year - 2000)$, then its origin, $t = 0$ is in the year 2000. So, according to this model, around 317 cases should be seen in the year 2000.

The equation for the trend draws the straight line in Figure 6, so a point on the line is the expected number of cases at that time point predicted by the trend equation. But the actual data points do not lie on the line. Every data point has a deviation from the line (above or below, larger or smaller). These deviations are assumed to be random, meaning that they are unpredictable because there is no obvious pattern in them. The model for the observed data can be written: Y_t (note the capital Y) at any time t in an equation as the expected value according to the trend (y_t) plus the random deviation (z_t) :

$$Y_t = y_t + z_t \text{ and therefore } z_t = Y_t - y_t$$

A large variability of the observed data around the trend (i.e. large values of z_t) means that there is a large potential for error when a future value of Y_t is predicted using the trend y_t. It also means that the trend will be estimated with less precision (with a larger confidence interval).

2.2 Polynomial trends of higher orders

A polynomial trend of an order higher than 1 is a non-linear function that can model changes in the trend. Here, time is added to the model with different powers (e.g. $y = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \cdots$). The highest power in the equation is called the 'order' of the polynomial. Note that a polynomial trend of order 1 is in fact a linear trend. The interpretation of polynomial trends of orders higher than 1 may be more difficult to convey to a layperson.

Polynomial trends of orders higher than 1 have the following shapes:

- The shape of a **quadratic trend** is a parabola (if the number of observations is enough) that has one bend. This can be a U-shaped curve or an umbrella-shaped curve. Within the time period under analysis, only part of this curve may be visible, but if the model is projected into the future and goes far enough into the past, the whole shape can be revealed.
- The shape of a **cubic trend** has two curves (if the number of observations is enough).
- The general shape of a **polynomial trend** of the order k has $(k - 1)$ curves (to appreciate the whole shape, a projection of the trend outside the period of analysis might be needed).

2.2.1 Generic equations for polynomial trends

The equation of a polynomial trend of the order 2, a **quadratic** trend, is:

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2$$

where y_t is the outcome at time t, β_0 is the intercept and β_1 and β_2 are the coefficients.

The equation of a polynomial trend of the order 3, a **cubic** trend, is:

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3$$

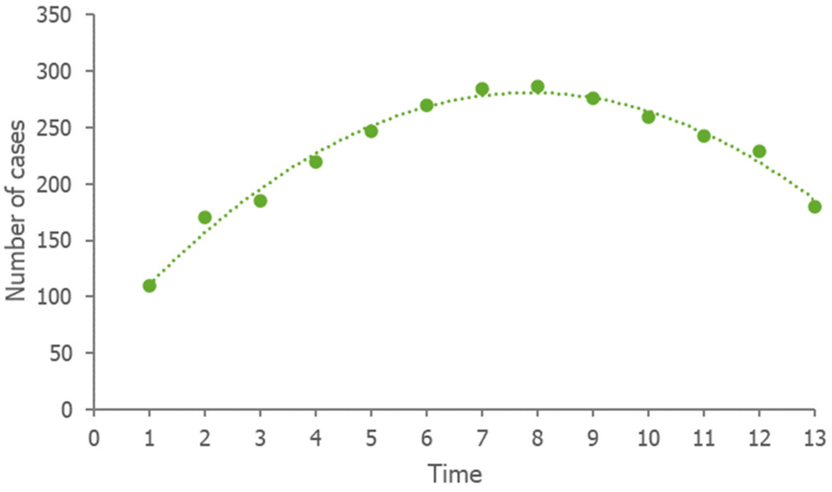
The general equation of a **polynomial** trend of the order k is:

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_k t^k$$

2.2.2 Example of a quadratic trendline

Figure 7 provides an example of a quadratic trendline with a polynomial trend of the order 2.

Figure 7. Example of a quadratic trendline (polynomial order 2), with the number of cases over time



Equation

The equation for the trendline in the quadratic example (Figure 7) is:

$$y_t = 57.5 + 56.8t - 3.6t^2$$

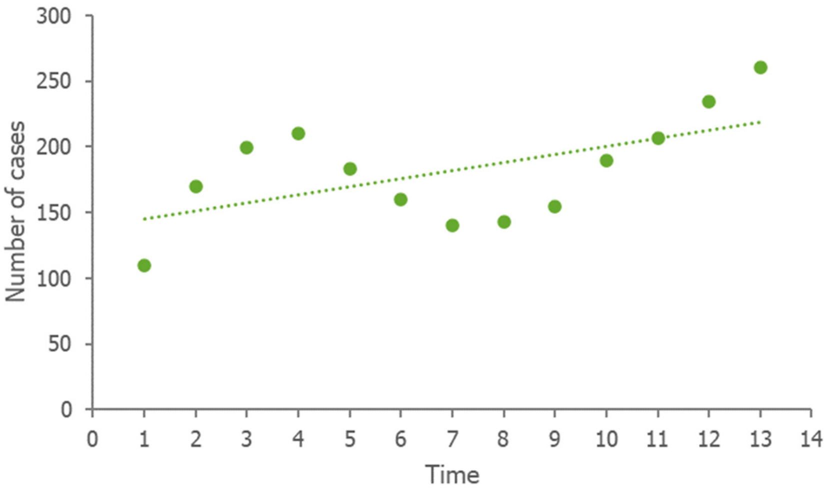
Interpretation

If the slope were to remain unchanged (e.g. if there was no effect of t^2), then for each increase in time unit, the number of cases would increase by 56.8 (but, in fact, the slope does not remain unchanged). The slope is not a constant number, but itself a (linear) function of t (its value changes with t). In the example, this means that although initially cases increase with time, at some point (around time $t = 8$) cases will stop increasing and will start decreasing with time. A negative coefficient of a quadratic term indicates an umbrella-shaped curve and a positive coefficient indicates a U-shaped curve.

2.2.3 Example of a cubic trendline

Figure 8 provides an example of a cubic trendline with a polynomial trend of the order 3.

Figure 8. Example of a cubic trendline (polynomial order 3), with the number of cases over time



Equation

The equation for the trendline in the cubic example (Figure 8) is:

$$y_t = 76.4 + 63.4t - 11.3t^2 + 0.6t^3$$

Interpretation

If the slope were to remain unchanged (e.g. if there was no effect of t^2 or t^3) then for each increase in time unit, the number of cases would increase by 63.4 (but, in fact, the slope does not remain unchanged). The slope is itself a (quadratic) function of t.

2.2.4 Patterns in the slopes of polynomials

A trend following a polynomial of degree k means that the slope of that trend is itself a polynomial of degree $k - 1$. When $k = 1$, the slope is a polynomial of degree 0 (a constant) and the trend is linear. For polynomials of orders higher than 1, it is difficult to explain the effect of time (the slope) and it is much easier to refer to the graphical representation.

When carrying out an analysis using polynomial trends, remember to keep the lower order terms in the model. For example, when creating a quadratic trend, do not forget to keep the $\beta_1 t$ term in the model. When creating a cubic trend, do not forget to keep the $\beta_1 t + \beta_2 t^2$ terms in the model.

2.3 Exponential trend

An exponential trend is where there is a percentage of growth with each time unit. The trend is multiplicative rather than additive (i.e. for every added time unit the trend is multiplied by a certain number, rather than having a certain number added to it). Many infectious diseases with predominant underlying person-to-person transmission are multiplicative.

The exponential trend is a curve that increases or decreases exponentially. The trend curve will increase if the coefficient is positive and decrease if the coefficient is negative. Note that values following an exponential trend will never reach zero.

2.3.1 Generic equation for an exponential trend

The equation of an exponential trend is:

$$y_t = \beta_0 \beta_1^t$$

where y_t is the outcome at time t, β_1 is the time coefficient and β_0 is the intercept. At time $t = 0$ the expected number of cases is $y_0 = \beta_0 \beta_1^0 = \beta_0$, at time $t = 1$ $y_1 = \beta_0 \beta_1^1 = \beta_0 \beta_1$, at time $t = 2$ $y_2 = \beta_0 \beta_1^2 = \beta_0 \beta_1 \beta_1$, and so on. So, for every time unit increase, the expected number of cases is multiplied by β_1.

Note that if the natural logarithm of the data is taken, there will be a linear equation (see Part B, Section 2.1 'Linear trend (polynomial trend of the order 1)'):

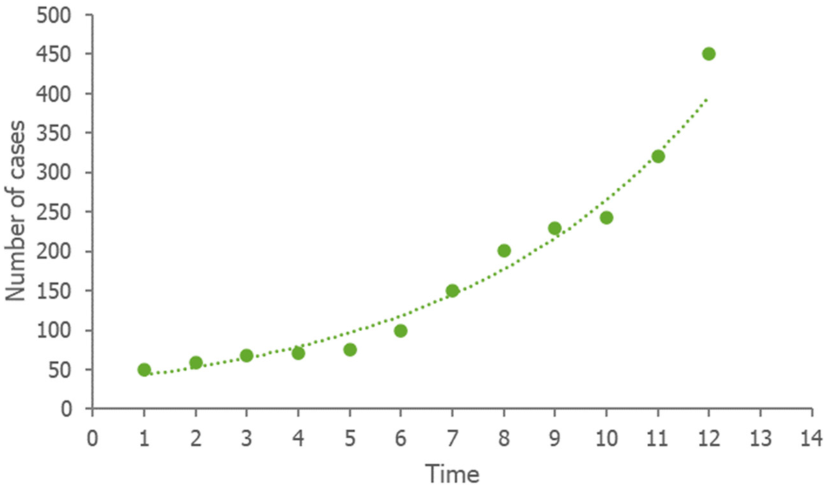
$$\ln(y_t) = \ln(\beta_0 \beta_1^t) = \ln(\beta_0) + \ln(\beta_1)t = A_0 + A_1 t$$

The exponent of A_0 is the intercept β_0 and the exponent of A_1 is the coefficient β_1 of the original data: $\beta_0 = e^{A_0}$ and $\beta_1 = e^{A_1}$.

2.3.2 Example of an exponential trendline

Figure 9 provides an example of an exponential trendline.

Figure 9. **Example of an exponential trendline, with the number of cases over time**



Equation

The equation for this trendline is:

$$y_t = 35.5 * 1.2^t$$

Interpretation

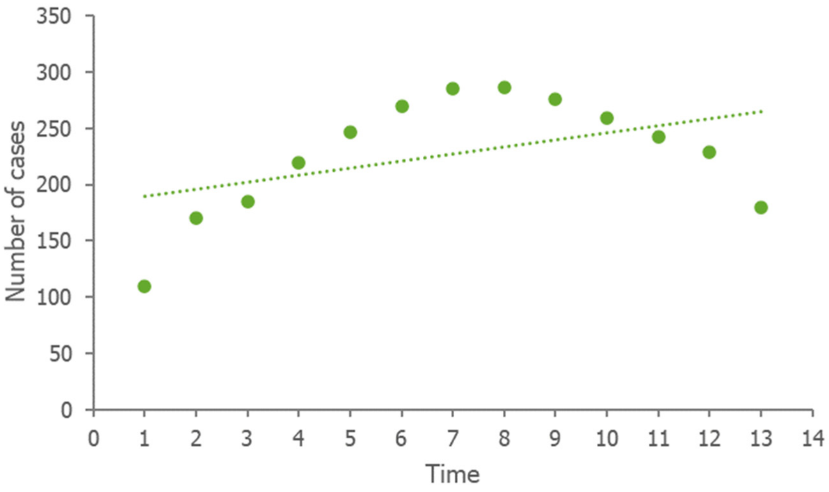
For each increase in time unit, the number of cases is multiplied by a factor of 1.2 (i.e. cases are increasing by 20%).

The interpretation of the exponential trend is somewhat easier to understand and convey than that of polynomial trends and can sometimes be preferred.

2.4 Other non-linear trends

It is important to fit the correct trend to the data. If the wrong trend is modelled, biased inferences and predictions could be made. For example, if a linear trend is fitted to data that display a quadratic trend, there can be a wrong interpretation of the data (Figure 10).

Figure 10. Example of a wrong trend line, with number of cases over time



The linear trend indicates that there is a steady increase of around six cases per time unit. But, looking at the data, there is an initial increase followed by a decrease. This is an example of why visualising the data and the trendline is an important step in trend analysis.

2.5 Other trends

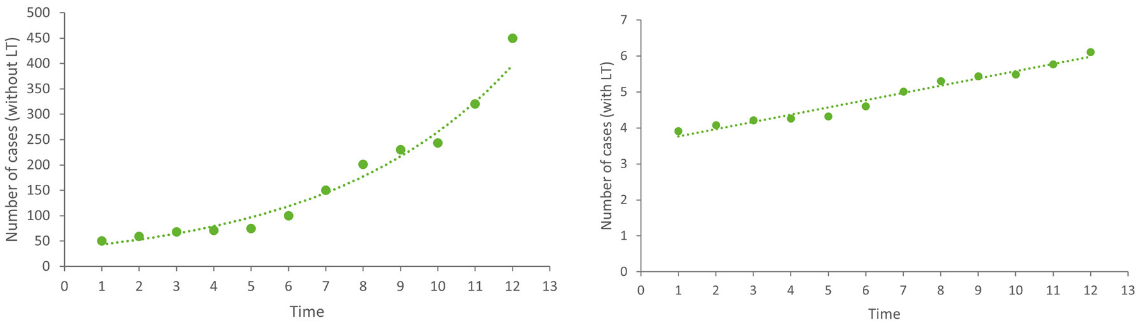
Data can have other trends as well. Logarithmic trends are inverse to exponential trends: these trends start with a very steep increase or decline, followed by a flattening. Another type of trend is the so-called 'power trend', which is similar to the exponential trend, but has a more symmetrical curve. The power trend models a steady increase or decrease in the data. These types of trends are not discussed further in this guidance document.

Non-linear trends can be modelled with polynomials; however, polynomials have symmetrical features and will not fit non-symmetrically distributed data well. In such instances, piecewise polynomials (known as splines) can be used to fit the distribution.

2.6 Transforming data

If the data follow an exponential trend (see Part B, Section 2.3 'Exponential trend'), then log transforming the outcome variable may produce a straighter trend that can be estimated with a linear trend. Log transforming data with an exponential trend can also be useful to stabilise the variance. If the data follow an exponential trend, then often the variance over time increases with the mean. When the variance of the data around the trend is not constant over time (known as heteroscedasticity), it can be more difficult to make correct inferences. A log transformation can help reduce heteroscedasticity. Furthermore, in some data with an exponential trend and changing variance, data are not normally distributed around the trend and this again makes the inference and the prediction more difficult. A log transformation may help the residuals to become normally distributed.

Figure 11. Example of number of cases over time without and with log transformation



LT: log transformation

Transformations of data should be carried out with caution, as the transformed model may be difficult to interpret. The conclusion needs to be back-transformed to the original data. For example, one could log transform a time series of the yearly number of cases of a disease, then fit a linear trend and, by projecting it, predict that in 2022 the series should reach a value of 5.7 (with a 95% prediction interval of $5.0–6.4$). Then, after back transforming the logarithm by exponentiating these numbers, the expected number of cases for 2022 should be: $e^{5.7} = 299$ with a 95% prediction interval ranging from $e^5 = 148$ to $e^{6.4} = 602$.

3. Seasonality (cyclical variation)

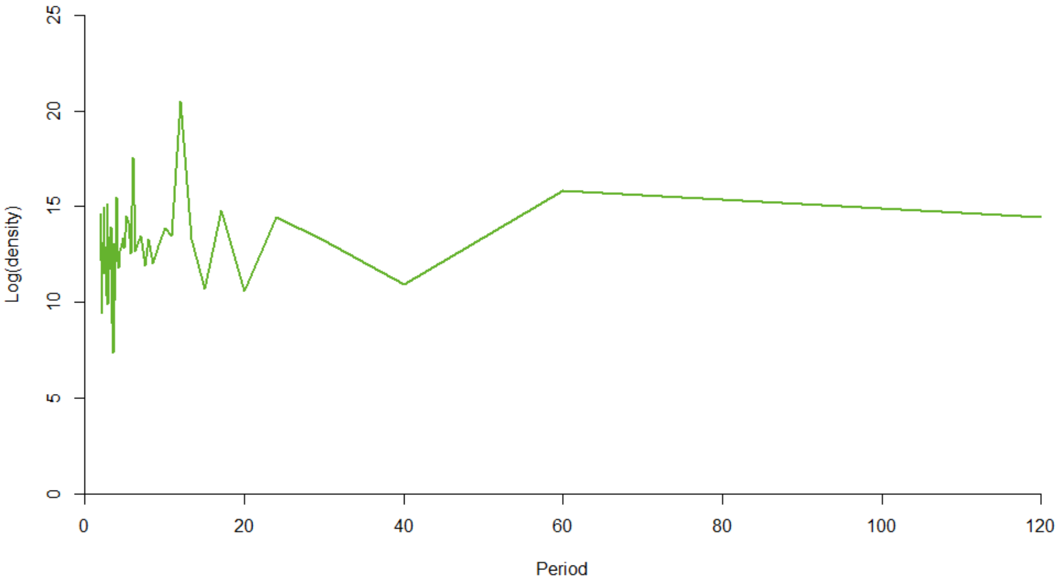
Cyclical variation can have an annual frequency (seasonality), but can also have a weekly frequency, daily frequency (circadian rhythms) or any frequency related to calendar time. In surveillance analysis, most of the time seasonality is the relevant cyclical variation. Therefore, such cyclical variations are referred to as 'seasonality' in this document.

3.1 Detecting seasonality

To detect seasonality in the data, the first step is to assess the data visually: is any seasonality visible? Ask disease experts if the disease or health event has seasonal tendencies. Also, after removing the main trend in the data, create a periodogram, which is a discrete Fourier transformation of the autocovariance function of detrended data [2]. This may sound complicated, but most statistical software packages provide an easy way to create a periodogram. The interpretation of the periodogram, which displays power over frequency, is straight-forward: if there are strong spikes at certain frequencies, then there is seasonality of that frequency in the data.

The example outputs (Figure 12) from the periodogram of detrended monthly salmonellosis data indicate that there is a strong spike at 12, which corresponds to a yearly frequency. This is very clear looking at the detrended data (Figure 12). The periodogram indicates that there may be other frequencies at other periods (six months, four months, etc.) as well.

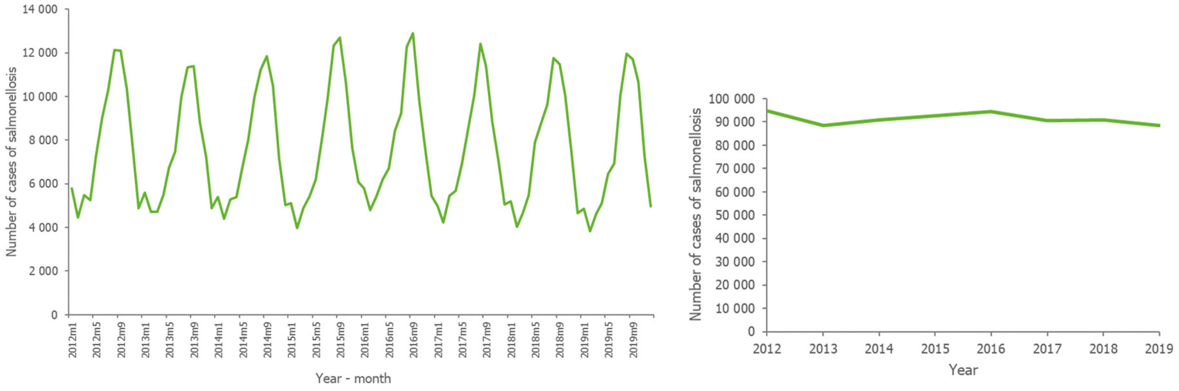
Figure 12. Example periodogram of detrended monthly salmonellosis data



3.2 Addressing seasonality with data aggregation

If seasonality is not of interest (measuring seasonality is not part of the analysis objectives), then data can be aggregated into a wider time unit to address seasonal variations within that unit. For example, if a time series has monthly seasonality, but only an estimation of a long-term trend over several years is of interest, data from the months may be collated to obtain one yearly data point and then a linear regression can be fitted (Figure 13).

Figure 13. Examples of the number of cases of salmonellosis by year-month and by year



3.3 Addressing seasonality with differenced data

In some fields of time series analysis, data are 'differenced' (taking the difference between a current and a previous data point). Data can also be 'seasonally differenced', where a difference is taken from a current data point to a previous data point at a seasonal lag (e.g. 12 for monthly data). This differencing is often used in autoregressive integrated moving average (ARIMA) and seasonal autoregressive integrated moving average (SARIMA) models.

For further reading on ARIMA, SARIMA and other time series analysis methods (e.g. exponential smoothing), see Chatfield, C (2004) [2] and Hyndman, RJ and Athanasopoulos, G (2021) [9].

3.4 Why take seasonality into account, if it is of no interest?

If seasonality is of no interest and data are not to be modelled with time units that are at a higher resolution than yearly, then data can be aggregated by year (if it is not already) and a trend analysis performed, ignoring seasonality entirely.

However, if data are to be modelled at time units that are at a lower level than yearly and there is seasonality in the data, then this must be considered in the model even if there is no interest in seasonality (e.g. to model weekly influenza data in order to plan vaccination campaigns and hospital preparedness). The reason for this is that seasonality is a deviation of the data around a general trend of the year (more cases will be expected in some months than those predicted by the trend, and in some months fewer cases will be expected). If seasonal deviations from the linear trend are not captured by specific seasonal coefficients as another systematic pattern of the time series, then the model will include them in the random variations (i.e. z_t) of the model. This will increase the standard deviation of the errors (i.e. S_z) and, as explained above, will reduce the precision of the estimation of the linear trend and increase the number of errors in the predictions.

3.5 Methods to include seasonality in trend analysis

If there is seasonality in the data, consider using sine and cosine terms in the model to take seasonality into account in the analysis. Any curve can be described by a linear combination of sine and cosine terms. After determining the main periods in the data from the periodogram, the sine and cosine terms can be created by taking the sine and cosine function and multiplying its argument by 2π divided by the period. These terms can then be included in the regression model.

Alternatively, seasonal periods can be included as dummy variables in the model. For example, if the data have a monthly seasonality, then a variable can be created with the month of the year for each data point. Including this variable in the model as a factor will create 11 categories (leaving out one month as the reference category) and will capture the average increase in the outcome of each month compared to the reference month. Month or week (depending on the level of aggregation) can be added as a factor variable in the model.

4. Analysis

4.1 Descriptive analysis of the data

After deciding on the study period for the trend analysis and data cleaning, the next step is to visualise the data. The case counts of a disease (the number of disease cases reported) can be plotted by each time unit.

Visualising the data will provide many clues as to what type of trend is in the data and what type of model should be used. It might also be possible to observe whether there are any sudden changes in the trend, which could be associated with external factors. Scripts for data visualisation are included in the supplementary material.

> After understanding the data (see Part A, Section 1.1, 'Stable data source'), visualising the data is the next most important step in a trend analysis.

4.2 Assessment of linearity

As mentioned in Part B, Section 2.4, 'Other non-linear trends', ignoring a non-linear data structure will lead to a biased inference around the data. However, there might often be very mild non-linearity of a trend in the data. Deciding whether to model a linear or non-linear trend will not only depend on the statistical outputs when comparing trends, but also on the study objectives and the magnitude of difference between linear and non-linear trends. Note that understanding the data and the study objectives, as well as undertaking a visual assessment of the data, are as important as the statistical outputs in trend and non-linearity assessment.

In this guidance document, the main non-linear types of trends presented are polynomial trends of the orders 2 and 3, and exponential trends. Exponential trends can often be approximated, in the short term, by polynomial trends of the orders 2 and 3. Here, the assessment of non-linearity begins by comparing polynomial regression types before moving on to assessing exponential trends.

Polynomial regression of the orders 2 and 3 can be compared to a linear model. Using polynomial trends of an order higher than 3 may not be needed in shorter term surveillance data and are harder to interpret.

Using polynomials of higher orders may improve data fit. Eventually, the data can fit perfectly by allowing enough curves (polynomials of high orders). This is what is called 'overfitting'. The objective is not to fit the data perfectly, but to obtain a useful model that can indicate something about reality and be extrapolated onto future data.

There are various methods for assessing which model fits best, including:

- calculating the R^2/adjusted R^2 (if using ordinary linear regression),
- conducting a likelihood ratio test (models must be nested) or
- using an information criterion (models do not have to be nested, but the same data must be used).

4.3 Calculating the R^2/adjusted R^2

When using ordinary linear regression, the R^2 (which is the percentage of variance explained by the model) can be calculated. It is a value between 0% and 100% (or 0 and 1 if using proportions), and the higher the number the more the variance of the outcome is explained by the model (i.e. the better the model fits the data). The R^2 is a standard output after an ordinary linear regression command in most statistical software (note that a polynomial of an order greater than 1 is also a linear equation of powers of the explanatory variable X, X^2, X^3, etc.).

Often, statistical software will also provide the adjusted R^2, which takes into account the number of parameters in the model and the sample size. If the adjusted R^2 statistic is available, then using this one is recommended.

The adjusted R^2 of a model with a linear trend can be compared with the adjusted R^2 of a model with a quadratic trend and the adjusted R^2 of a model with a cubic trend. The model with the highest adjusted R^2 is the model that best explains the data. However, note that this does not mean that it explains the data well, but that it fits better than the other models. As observed in the previous section, depending on the understanding of the data and study objectives, very small differences between R^2 can be ignored and the simplest model can be chosen.

> Some statistical software provide R^2 or $pseudo - R^2$ for regression models other than ordinary linear regression, but it is not recommended to use these R^2 to assess non-linearity.

4.4 Conducting likelihood ratio tests

The likelihood ratio test is a method to compare the model fit of one model to another nested model. If there is no statistically significant difference between the models, then using the simpler model (the model with fewer terms) is recommended. If there is a statistically significant difference between the models, then the model with more terms is preferable.

In Table 1, there is a statistically significant difference at the 5% level between the linear and the quadratic models. This indicates that the quadratic trend fits the data better. There is no statistically significant difference between the cubic and the quadratic trend models, indicating that a quadratic trend can be used.

Table 1. Likelihood ratio test

Type of polynomial trend in the model	Likelihood ratio test p-value
Linear (order 1)	Reference
Quadratic (order 2)	0.03 (compared to the linear model)
Cubic (order 3)	0.12 (compared to the quadratic model)

Note that the likelihood ratio test can be used with various regression types. However, it does not take overfitting (too many parameters in the model) into account.

4.5 Using information criterion

Another way of comparing models is by using an information criterion. The two most popular information criterion are, the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). For the purposes of this document, the AIC will be used.

The information criterion is calculated using maximum likelihood estimates from the model and the number of parameters in the model. The AIC penalises models with more parameters, so to avoid choosing more complex models.

In general, the lower the AIC, the better the model fit. In general, models that differ by 0–2 points from the model with the lowest AIC are credible alternative models, while models that differ by 4–7 points have moderate credibility and models that differ by more than 10 points have almost no credibility [10].

To compare models using the AIC, it is important that they are calculated on the same dataset. Otherwise, the AIC will not be comparable. This means that the number of records in the data (and the actual records) must be the same in both models. If, for example, a log transformation of the data is performed on one model, the AIC will not be comparable to another AIC on the non-transformed data. There is a workaround to compare AIC in such an instance, but it is not given in all statistical software.

Note that absolute values of AIC may differ depending on the software used. It is important to look at the values of AIC within outputs from the same software.

4.6 Comparing an exponential trend to a polynomial trend

As described previously, if it is established that the data have a non-linear trend, tests should be conducted to see if quadratic or cubic trends fit better. Then one can check whether an exponential trend fits the data best. It is important to take the nature of the data into account, as well as the audience for the trend analysis. An exponential trend may be easier to convey than a quadratic or cubic trend, while the data may fit well to either a polynomial or an exponential trend.

With exponential trends, it is possible to log transform the data (see Part B, Section 2.6, 'Transforming data') and apply a linear model. This means that it is no longer possible to compare the models using the likelihood ratio test, as they are no longer nested. Comparing the AIC between models is also no longer straightforward, as the data are no longer the same.

Also, if linear regression is being used, the adjusted R^2 of a model on log-transformed data and the R^2 of a model on non-transformed data can be compared.

Limitations

While checking for non-linearity using polynomial models is simple to carry out and easy to understand and convey, not all forms of non-linearity can be modelled with polynomial terms. Therefore, this method may not be suitable in all instances (see Part B, Sections 2.3, 'Exponential trend'and 2.4, 'Other non-linear trends'). Furthermore, linear and exponential trends are impossible to maintain in the long term (linear trends can go below zero, but with exponential trends, at some point, the total population will have had the disease or health event). They can be extrapolated for a while, but not forever. However, for a short-term time series, without any long-term forecasting, these trends are likely to be adequate.

4.7 Other ways to assess non-linearity

There are other methods for model selection, such as the use of a training and validation dataset, which are not covered in this guidance document.

5. Choice of statistical model for the outcome

Choosing between linear regression (which has a continuous variable as an outcome), Poisson regression (which is often specifically used for count data) or other regression analysis models with count data as an outcome will depend on:

- the nature of the data,
- how the model meets the assumptions of each regression type and
- how simple the model can be to explain the trend (i.e. can it be easily communicated and is it useful for decision-making?).

It will also depend on the properties of the outcome variable and its relationship to the predictor variables (in this context: time). With surveillance data, there are generally counts or rates, although there might sometimes be proportions.

Information criterion is an example of a statistical method that can be used to choose between regression types. Models with different regression types can be compared (e.g. a linear versus a Poisson model) using information criterion (see Part B, Section 4.5, 'Using information criterion').

As long as the outcome variables are in the same scale, the observed values and predicted values of two models can also be compared to determine the root mean squared error. A lower mean squared error means a better fit. This cannot be done to compare models with different outcome scales (e.g. to compare the logistic model of the proportion of the infected population with a linear model of the number of infected cases). However, one should endeavour to find the simplest method to adequately explain the data.

5.1 Count data

If there are counts of disease cases in the surveillance data, linear regression or Poisson regression can be used.

Each regression type comes with assumptions that must be checked prior to embarking on the analysis. The interpretation of the Poisson regression model is more complex than that of the linear regression model. In the context of simple trend analysis of surveillance data, linear regression models should be preferred unless, by the nature of the series to model, the Poisson regression model is strongly indicated.

5.2 Other types of data

Not all data for trend analysis are count data. Rates or proportions over time may also be outcomes of interest.

Traditionally, rate data are modelled including rates with Poisson or negative binomial regression (if overdispersion is present). In such cases, the denominator (time at risk) is an exposure in the model. As with counts, the assumptions of Poisson regression (see Part B, Section 5.4, 'Poisson regression') must be met. If these assumptions are not met, using a scale factor within Poisson regression or using negative binomial regression should be considered.

Rates can also be modelled with linear regression. As the expected number of events (λ_t) become larger, the Poisson distribution approximates a normal distribution. Therefore, a linear regression could be used to model this data.

5.3 Linear regression

The equation of the linear regression model, as shown earlier, is:

$$y_t = \beta_0 + \beta_1 t + z_t$$

where y_t is the observed outcome at time t, β_1 is the slope, β_0 is the intercept and z_t is the error term or residual (the difference between the actual observed value and the predicted value at time t):

$$z_t = y_t - (\beta_0 + \beta_1 t)$$

5.3.1 Assumptions of linear regression

While relying on many assumptions, linear regression is a commonly used and easy to understand regression type. According to the central limit theorem, if a data series is sufficiently large, the sampling distribution of the estimated parameters (intercept and slope) will be normal even if the data are not normally distributed. This means that the inferences (i.e. confidence intervals and p-values) of the linear regression will be valid. The assumptions of linear regression include linearity, homoscedasticity, independence of observations and normality.

Linearity

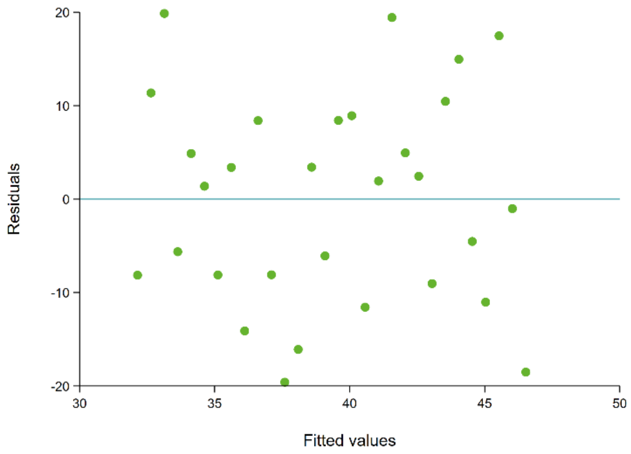
The relationship between the outcome and independent variables must be linear.

To check the assumption of linearity, plot the data against the model. The data points should be more or less symmetrically distributed around the model line.

It is also possible to plot the values of the model against the residuals (Figure 14). The residuals in the context of linear regression are what are left when the data points are subtracted from the model. The points should be more or less symmetrically distributed around the horizontal.

If the linearity assumption is violated, the predicted model and any forecast are likely to be biased. If the model does not meet the assumption of linearity, check to see if adding a polynomial of time (quadratic or cubic) improves the model. Alternatively, the data may have an exponential trend and a log transformation of the data may be suitable.

Figure 14. Plot with fitted values versus residuals to check the linearity assumption



Homoscedasticity/heteroscedasticity

The residuals are assumed to have a constant variance. If the scatter of residuals is not stable over time, then there is heteroscedasticity in the model.

To check for homoscedasticity, inspect the plot of residuals in Figure 14 against the values of the model over time. If the mean or the variability of the residuals increases or decreases substantially, then heteroscedasticity may be present.

If the homoscedasticity assumption is violated, confidence intervals around the model may not be reliable and inferences around the trend will not be reliable.

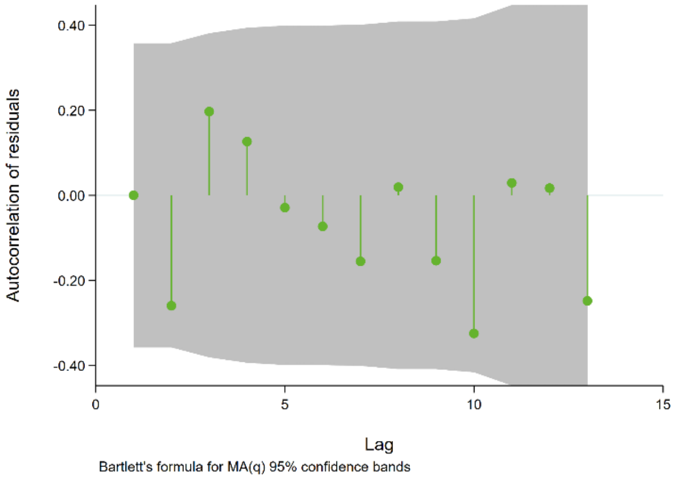
If the model does not meet the assumption of homoscedasticity:

- Check to see if all seasonality has been taken into account.
- Consider using a shorter study period, depending on the nature of the homoscedasticity.
- Consider a log transformation of the data, as an exponential model may be a better fit (see Part B, Section 2.6, ‘Transforming data’).

Independence of observations
The residuals of the model should be independent and not autocorrelated.

To check the independence of the residuals, plot the residuals over time or create an autocorrelation graph. If the residuals are not correlated, each correlation should be close to zero; however, some random variation is expected. Therefore, the values should be within the confidence interval (grey area in the graph), as they are in Figure 15.

Figure 15. Autocorrelation graph to check the independence of observations



If the independence of observations assumption is violated, it is possible that a wrong regression type or model has been selected. This means that the trend analysis outputs will not reflect the data.

If the model does not meet the assumption of independence of residuals:

- Check that all seasonality has been taken into account. If an existing seasonality has not been taken into account, observe this seasonality in the residuals, as in Figure 15 (see Part B, Section 3.5, 'Methods to include seasonality in trend analysis').
- If, after adjusting for seasonality, there is still autocorrelation of observations, a lag factor may need to be added to the model.

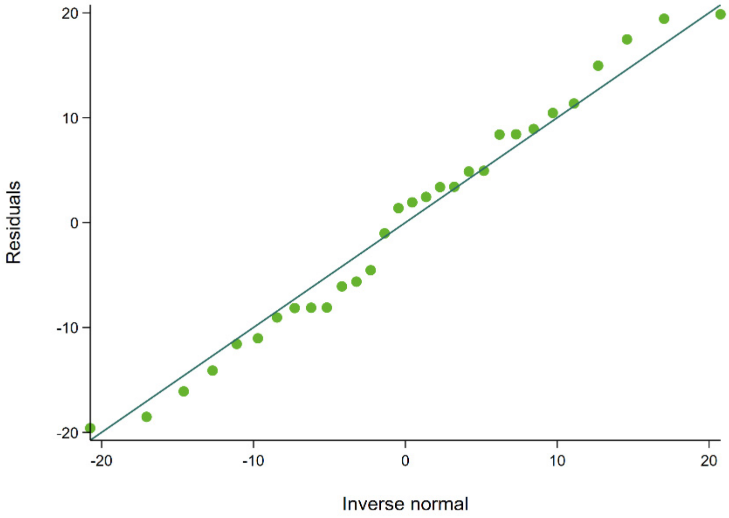
Violation of independence assumption and time series

In a time series, observations are most often dependent on previous observations. For example, the cases of disease X observed in the present month depend on the number of cases observed in the previous month in the same population. Note that the independence assumption relates to the residuals of the data. After removing the trend (and perhaps also the seasonality) from the data, a lot of the dependence of the data can be removed. However, very often, there is a remaining dependence of residuals. If this is the case, then there is a violation of the independence assumption and any regression model will have biased inferences. Frequently, trend analysis is nevertheless carried out, and some violation of the independence assumption is tolerated by the analyst. If there is uncertainty regarding how to interpret the extent of the dependence of residuals, a statistician should be consulted.

Normality
The errors of the model (the residuals) are assumed to be normally distributed.

To check whether the residuals are normally distributed, create a histogram of the residuals and a specific diagnostic plot called the normal probability plot (also called a 'quantile plot'; Figure 16).

Figure 16. Normal probability plot (quantile plot) to check whether the residuals are normally distributed



In Figure 16, the residuals are plotted against the theoretical percentiles of the normal distribution. If the assumption of normality of residuals holds, then the fit should be linear. In the above example, the fit is linear and the assumption of normality of the residuals holds.

If the normality assumption is violated, confidence intervals and significance tests are no longer reliable.

If the model does not meet the assumption of normality of residuals:

- Check whether a different statistical model (e.g. Poisson) fits the data better.
- Consider a log transformation, as the data may have an exponential trend.
- Check to see if the data have one or several outliers that cause the assumption of normality to be violated.
- Check that these are true data points (i.e. if they are genuine or if there was perhaps a data entry error).
- Check if it is it possible to adapt the study period.

Note that outliers may also violate some of the assumptions above. Outliers are data points that may be particularly high or particularly low compared to other values. 'True outliers' may occur if there is a large outbreak of the disease under study in a particular year or if there is an abrupt change in data collection. Outliers may also occur due to data entry errors or other problems in the data. These are not 'true outliers', but outliers due to erroneous values.

Outliers can violate the assumptions of many regression types.

5.3.2 Interpreting linear regression output

In a linear regression output, for each time unit increase, the expected value of the outcome changes by β_1, with all other variables held constant.

In Table 2, the coefficient β_1 for time indicates that, for each time unit increase, there will be around 50 additional cases. The standard error of the coefficient indicates the model's precision around the coefficient's unknown value. It is used to calculate the confidence intervals and significance tests.

Table 2. Example of output from statistical software after a linear regression of count data on time

	Coefficient	Standard error	$P > \|t\|$	Lower 95% CI	Upper 95% CI
β_1 (Time)	49.6	24.1	0.049	0.20	98.93
β_0 (Constant)	3 165	428	<0.001	2 289	4 041

CI: confidence interval

In this example, $P > |t|$ is the p-value (usually two-tailed) that tests the hypothesis that the coefficient equals zero. In the case of the time coefficient $\beta_1 = 0$, the model is a horizontal line (i.e. there is no expected change in outcome over time). Here, the p-value is 0.049, indicating that there is only a 4.9% chance of obtaining this estimate for the coefficient, or one even further away than the null hypothesis tested $(\beta_1 = 0)$ using a random sample if no true change in outcome occurred over time. The lower and upper CIs present the lower and upper confidence intervals around the coefficient (usually 95% confidence). The confidence intervals are related to the p-value; as can be seen, the lower 95% CI is close to zero, which is the case where the p-value is close to 0.05 (as in this example). The pragmatic interpretation of a confidence interval is that there is 95% confidence that the true coefficient is within the range of 0.20 and 98.93.

The coefficient of the constant indicates which count to expect at $time = 0$, and the p-value indicates that it is statistically significant and different from zero. The interpretation and importance of the constant depends on how $time = 0$ has been defined, which is not often pertinent in trend analysis.

5.4 Poisson regression

Poisson regression is often seen as a good choice for count data. However, with large sample sizes, the central limit theorem holds and linear regression may be preferred, as it is easier to interpret.

5.4.1 Poisson regression equation

The equation of the Poisson regression model is:

$$y_t \sim Poisson(\lambda_t)$$

Where λ_t is the expected number of events at time t (and also the variance), it is assumed that λ_t follows an exponential trend. Therefore, its logarithm would follow a linear trend:

$$log(\lambda_t) = \beta_0 + \beta_1 t + z_t$$

5.4.2 Assumptions of Poisson regression

The assumptions of Poisson regression include: Poisson response, independence of observations, linearity and $mean = variance$.

Poisson Response
The response variable is a count, described by a Poisson distribution.

Independence of observations
The residuals of the model must be independent and should not be autocorrelated. This can be checked by plotting the residuals against the model. If there are patterns in the residuals, then there is likely to be some autocorrelation in the data.

If the residuals are not independent:

- Check to see whether all seasonality has been taken into account. If all seasonality has not been taken into account, there will be no independence of observations (see Part B, Section 3.5, ‘Methods to include seasonality in trend analysis’).
- Consider adding a lag factor to the model if, after adjusting for seasonality, there is still autocorrelation of observations.

As mentioned in Part B, Section 3.5, observations in a time series are most often dependent on each other. Removing trend and seasonality from the data may not result in entirely independent residuals. If there is uncertainty about the effect of dependence of residuals on the analysis, consult a statistician.

Linearity
The log of the mean rate, $log(\lambda_t)$, must be a linear function of t.

To assess the linearity assumption in Poisson regression, plot the residuals against the fitted values of the model and check whether there is a trend in addition to the patterns that are expected to be seen from a Poisson response variable.

If the linearity assumption is violated, then the predicted model is likely to be untrue. In particular, any forecasts will contain errors.

If the model does not meet the assumption of linearity:

- Check to see if adding a polynomial of time (quadratic or cubic) improves the model.
- Consider a log transformation of the data if the data have an exponential trend.

Mean = Variance

This is an assumption that is very specific to Poisson regression. By definition, the mean of a Poisson random variable must be equal to its variance. If this assumption is not met, this is known as overdispersion.

Check for overdispersion by using statistical tests and plot the residuals against the fitted values. If the data are overdispersed, the $mean = variance$ assumption is violated, and the inferences around the data will not be correct.

If the data are overdispersed:

- Use Poisson regression with a scale factor.
- Use negative binomial regression.
- Use the AIC and graphical visualisation to choose between a model with Poisson regression with scale factor and a negative binomial model.

Diagnostics of the assumptions of Poisson regression are more complex than those of linear regression and rely on a good familiarity with the properties of Poisson-distributed response variables. You may need to consult a statistician when performing these diagnostics.

5.4.3 Interpreting Poisson regression output

The time coefficient, β_1, means that for each increase in time unit, the expected number of events is multiplied by a constant, $exp(\beta_1)$, when all other variables are held constant. Where $\beta_1 > 0$, then $exp(\beta_1) > 1$ and the expected number of events increases exponentially over time. But where $\beta_1 < 0$, then $exp(\beta_1) < 1$ and the expected number of events decreases exponentially with time.

In Table 3, the time coefficient indicates that, for each time unit increase, the expected number of cases should decrease by a factor of $exp(-0.127) = 0.88$. Therefore, for each time unit increase, the outcome will decrease by 12% $(\lambda_{t+1} = 0.88\lambda_t)$.

Table 3. Example output from statistical software after a Poisson regression of count data on time

	Coefficient	Standard error	$P > \|z\|$	Lower 95% CI	Upper 95% CI
β_1 (Time)	-0.127	0.008	<0.001	-0.141	-0.112
β_0 (Constant)	4.907	0.049	<0.001	4.810	5.004

CI: confidence interval

In this example, the standard error of the coefficient indicates the uncertainty around the coefficient's unknown value. It is used to calculate the confidence intervals and the p-value used in significance tests.

$P > |z|$ is the p-value (usually two-tailed) that tests the hypothesis that the coefficient equals zero, the exponential of the coefficient is 1 (if $\beta_1 = 0$ then $exp(\beta_1) = 1$) and there would be no change in the expected number of cases over time. Here, the p-value is < 0.001 , indicating that it would be almost impossible to see this data if the true value of the coefficient were $\beta_1 = 0$. The lower and upper CIs present the lower and upper bound of the confidence intervals around the coefficient (usually 95% confidence intervals). These can be exponentiated to estimate a reasonable range where the true effect might be: $(exp(-0.141)\,; exp(-0.112)) = (0.87; 0.89)$. The confidence intervals are related to the p-value; as can be seen, the confidence intervals clearly exclude zero (no change over time). This can also be seen from the p-value.

The constant indicates that, when $time = 0$, the expected count is around $exp(4.9) = 134$. This is statistically significant and different from zero, which is not always useful information in trend analysis.

6. Algorithm for performing the trend analysis

Step 1: Define the study objectives

Before embarking on a trend analysis, define the study objectives. These should include whether there is interest in describing seasonal aspects of the data and what kind of outputs are needed for communications and decision-making.

Step 2: Understand the data

- Check the data. Is the data clean? Is the data collection stable over time? Are there any external factors affecting the data?
- Check for any outliers. Are they 'true outliers' (see Part B, Section 5.3, 'Linear regression'), or a result of data entry error?
- Decide on the study period.

Is the data stable over the study period?

If yes: Go to step 3.

If no: Go back to disease experts and discuss the data. Is a trend analysis possible/useful?

Step 3: Decide on the time unit of analysis

Decide what time level to draw conclusions about and aggregate the data to minimise the effect of fluctuations at lower time levels. For example:

- To see long-term trends over many years and possibly some long-term cycles, aggregate the data by year to remove the nuisance of potential monthly fluctuations.
- To see long-term trends but also seasonal variations, aggregate the data at month level.
- To do weekly predictions, aggregate the data at week level.

Whatever level of aggregation is chosen, bear in mind the data availability at that level, including:

- There must be sufficient data at that level. For example, if there is a lot of information missing over many weeks, it is not reasonable to aggregate at week level, but better to use month level.
- Ideally, there should be a good number of cases at each observation point (low or zero counts are not good for modelling time series).
- The data should not be too noisy. For example, data are often in batches in health information systems. So some days there will be no reports and other days a lot of cases will have been reported together, despite having occurred over the previous days. This makes the data very noisy and aggregating days into weeks might help to cancel out this noise.

Step 4: Plot the data

Plot the data over the selected time unit. Is there a clear picture of the data? Visualising the data helps with decision-making for subsequent steps, notably deciding on regression type and determining linearity/non-linearity of trends. Consider these questions:

- Is there a long-term trend in the data?
- Does the trend look linear?
- Are there any clear breaks in the trend, such as sudden changes in the level or the slope of the trend?
- Is there any cyclical pattern that repeats itself?

Step 5: Decide on the preliminary regression type

What properties does the outcome variable have? Is it a count, a rate or a proportion? Is the model simple enough to be easily communicated and useful for decision-making?

Is the outcome variable count data?

If the data for trend analysis is count data, are the counts large (≥ 30)?

If yes: Consider choosing linear regression as the preliminary regression type and go to step 6.

If no: Consider choosing Poisson regression as the preliminary regression type and go to step 6.

Is the outcome variable rate data?

If the data for trend analysis is rate data, how do the rates change over time?

If the rates **change linearly** and the variance is stable over time, consider a linear regression and go to step 6.

If the rates **change exponentially** and variance around the trend seems to be increasing or decreasing, consider a Poisson regression. If overdispersion is present, consider a negative binomial regression and go to step 6.

Is the outcome variable proportion data?

If the data for trend analysis is proportion data, consider using a logistic model and go to step 6.

Is the type of outcome variable unclear?

If the type of outcome variable is unclear, the data are not understood well enough. The data must be well understood before they are worked with.

Step 6: Consider inclusion of seasonal terms

Is the data aggregated by a unit that is less than one year (e.g. monthly, weekly)?

If no: Go to step 7.

If yes: Choose how seasonal terms will be included: sine and cosine terms, or seasonal indicator variables?

If **seasonal indicators** are chosen, go to step 7.

If **sine and cosine terms** are chosen, carry out a spectral analysis and use the results of the periodogram to determine possible periodicities in the data. Select initial periodicities for the sine and cosine terms and go to step 7 (this will be further explored in step 8).

Step 7: Assess the non-linearity of the trend

Is the data aggregated by a unit less than one year (e.g. monthly data)?

Carry out a simple regression of the count variable on the time variable, taking into account seasonality if the time units are at a higher resolution than year (and assuming there are no other predictor variables). Plot the resulting model over the observed data points.

Does it look like a linear trend adequately models the observed data?

Calculate the AIC and, if using linear regression, the adjusted R^2. Add a quadratic trend. Plot the resulting model and the data.

Does it look like the linear trend adequately models the data?

Calculate the AIC and, if using linear regression, the adjusted R^2. Carry out a likelihood ratio test between the linear and the quadratic model. Add a cubic trend. Plot the resulting model and the data.

Does it look like the trend is polynomial of order >1?

Calculate the AIC and, if using linear regression, the adjusted R^2. Carry out a likelihood ratio test between the quadratic and the cubic model.

If a non-linear trend (quadratic or cubic) is detected, check whether an exponential trend might fit better. Exponential trends may be easier to communicate in a simple way, compared to quadratic or cubic trends, if the fit is similar (some quadratic and cubic trends cannot be approximated by an exponential trend at all).

Does the trend look like it might be exponential?

Log transform the data and carry out a linear regression, then calculate the adjusted R^2. It is also possible to calculate the AIC if the workaround is used.

Once it is determined what type of trend the data has (linear, quadratic, cubic, exponential), go to step 8.

Step 8: Build the model

Are sine and cosine or seasonal terms used?

If no: After step 7, the model is nearly pre-final. Decide whether a log transformation of the data is needed (for exponential trends) or not (for linear, quadratic or cubic trends). Run the regression and go to step 9.

If yes: In step 6, initial periodicities for the seasonality were selected. Calculate the adjusted R^2 if using linear regression and calculate the AIC. Add sine and cosine terms of a further periodicity of importance to the model. Compare the adjusted R^2 if using linear regression; compare the AIC and calculate the likelihood ratio test. Continue this step with other periodicities of importance until the model is pre-final. Run the regression and go to step 9.

Step 9: Validate the model

Are the regression assumptions met?

Using diagnostic plots and tests, check that the assumptions of the chosen regression type (e.g. linear regression or Poisson regression) are met (See Part B, Section 5.1, 'Count data' for more information).

If they are met, see the discussion about overfitting below.

If they are not met, it may be necessary to go back to steps 2, 3, 5 and 6, depending on which regression assumption was not met.

Is there overfitting?

Check the coefficients and standard errors in the model. Do the standard errors exceed the absolute values of the coefficients?

If no: Go to step 10.

If yes: It is possible that the model is overfitted. Consider using a simpler model. This may be particularly pertinent to models with a weekly seasonal term. Go back to steps 2 and 3.

Step 10: Interpret the model

Now that the model is final, the regression outputs can be used to interpret the model. What type of trend has been found? What is its magnitude? Is it a statistically significant trend (to the 5% level)? Is the trend easy to communicate? What more information might be needed to interpret the trend?

7. Examples

All examples are carried out in just one geographical area (e.g. region, country or the EU/EEA). Do not combine several locations. Please see Part C for trend analysis across multiple geographical areas.

7.1 Example: Yearly trend of counts of salmonellosis cases in the EU/EEA

This example uses the algorithm in Part B, Section 6, 'Algorithm for performing the trend analysis' to demonstrate the process of obtaining a yearly trend of counts of salmonellosis cases in the EU/EEA. The dataset is freely available on ECDC's 'Surveillance Atlas of Infectious Diseases' [4].

Step 1: Define the study objectives

The objective of this study is to carry out a trend analysis of counts of salmonellosis cases in the EU/EEA in order to determine whether salmonellosis cases have been increasing or decreasing in the past 10 years, on average, and assess the magnitude of change.

Step 2: Understand the data

The data have been discussed with disease experts. There are no issues of missing data (non-reporting), and the surveillance systems have remained stable. It is possible to proceed with a trend analysis.

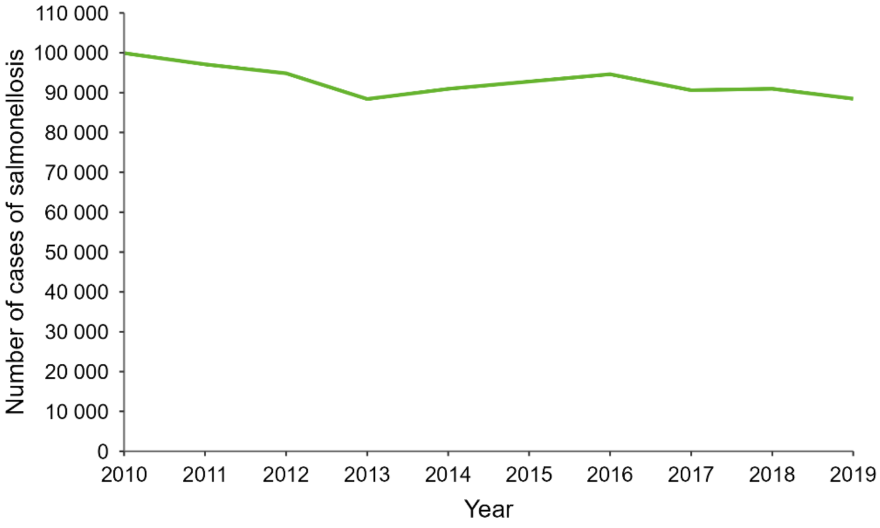
Step 3: Decide on the time unit of analysis

For the purposes of this analysis, the salmonellosis dataset aggregated by year will be used.

Step 4: Plot the data

The count variable has been plotted against the year variable (Figure 17), ensuring that the y-axis starts at zero.

Figure 17. **Number of confirmed cases of salmonellosis by year, EU/EEA, 2010–2019**



As seen in Figure 17, there seems to be a downward trend, but it is not entirely clear from the plot. No strong non-linearities are observed.

Step 5: Decide on the preliminary regression type

There are count data and very high numbers. Linear regression will be used for the trend analysis.

Step 6: Consider inclusion of seasonal terms

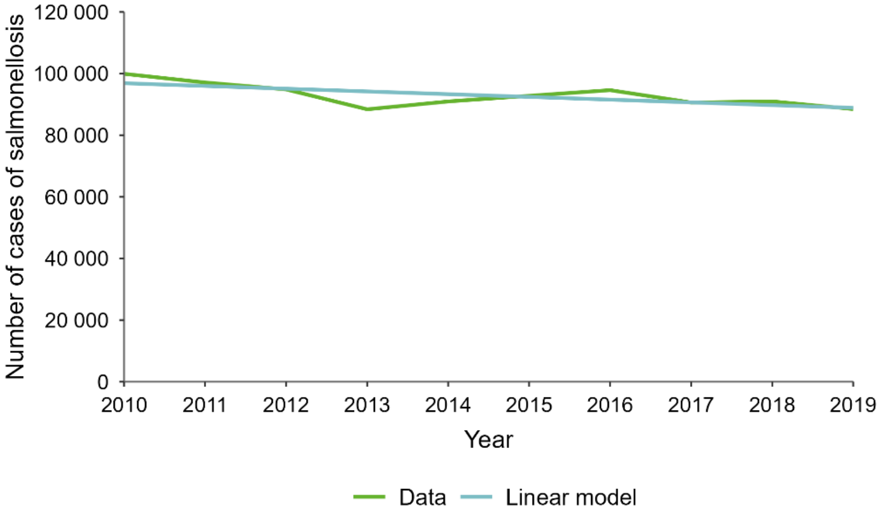
For the purposes of this analysis, no seasonal terms will be included (only yearly trends are of interest).

Step 7: Assess the non-linearity of the trend

The linear trend has been observed and the regression model has been plotted over the data. Note that for the purposes of this analysis, the intercept is not of interest. Therefore, it may be easiest to work with a time unit in integers that begins with one, rather than the year (2010). This could be particularly helpful when creating the quadratic and cubic terms.

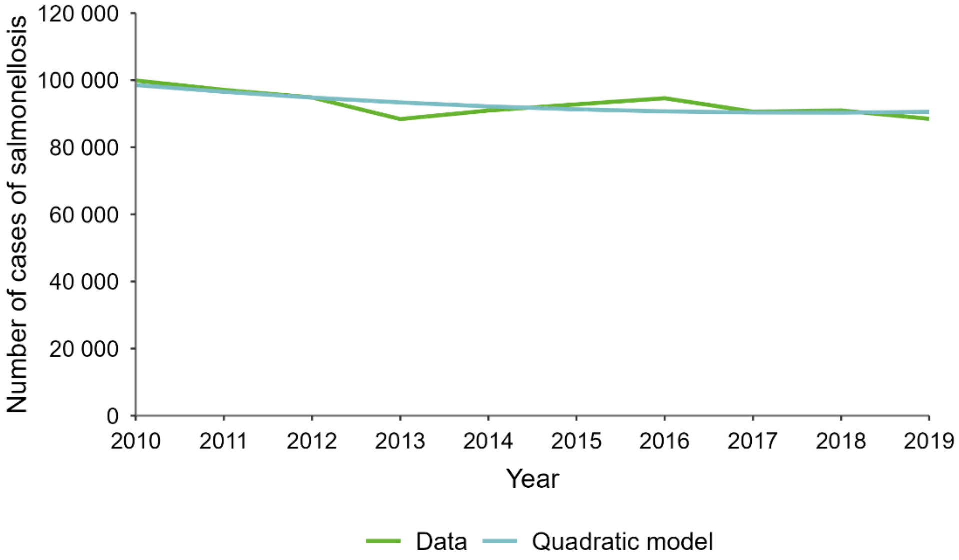
While the model does not fit the data perfectly (Figure 18), this trend looks acceptable for the purposes of the study objectives. The adjusted R^2 is 45.5% and the AIC is 188.6.

Figure 18. Number of confirmed cases of salmonellosis by year and linear model for the trend, EU/EEA, 2010–2019



Next, a quadratic trend is run, and the regression model is plotted over the data (Figure 19). The adjusted R^2 is 48.3% and the AIC is 188.8

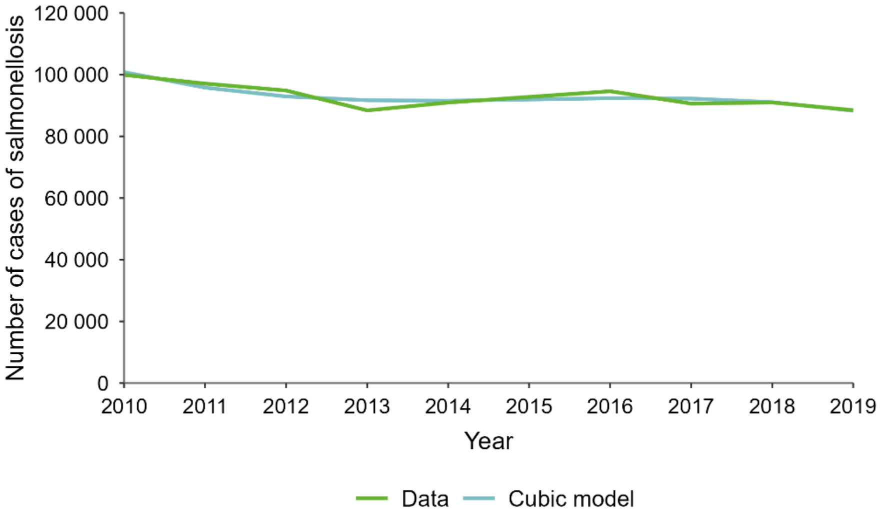
Figure 19. Number of confirmed cases of salmonellosis by year and quadratic model for the trend, EU/EEA, 2010–2019



As seen in Figure 19, the quadratic trendline does not look very different from the linear trendline. The adjusted R^2 is slightly better and the AIC is slightly higher. The p-value of the likelihood ratio test comparing the linear and quadratic trend is 0.1722, indicating that there is no significant advantage of the quadratic model over the linear model.

Next, a cubic trend is run, and the regression model is plotted over the data (Figure 20). The adjusted R^2 is 69.1% and the AIC is 184.1.

Figure 20. Number of confirmed cases of salmonellosis by year and cubic model for the trend, EU/EEA, 2010–2019

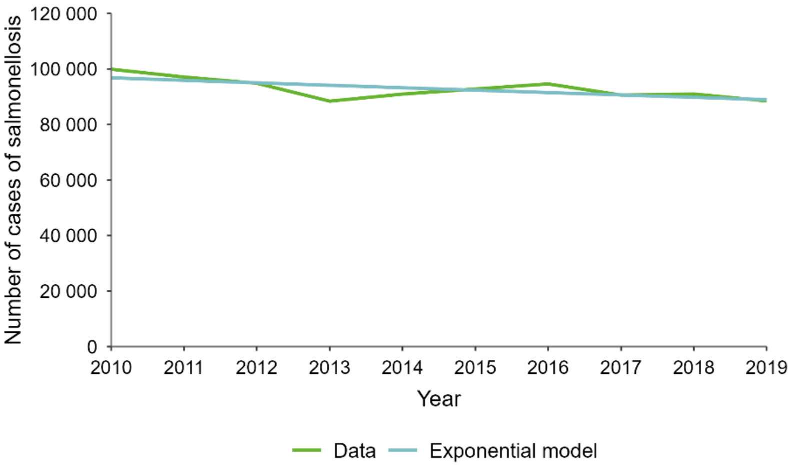


As seen in Figure 20, two curves are allowed in the model, which models the data slightly better. The R^2 is substantially better and the AIC is better, but only by two and four points for the quadratic and linear models, respectively. The likelihood ratio test indicates that the cubic model performs better than the quadratic model $(p = 0.010)$ and the linear model $(p = 0.014)$.

At this point, the objectives of the study should be considered. Modelling a trend with a higher number of degrees in polynomials will always result in a better fit, but it could run the risk of overfitting. Within the study, is it useful to model the ups and downs of the salmonellosis cases trend and comment on its magnitude, even if this curvature is unlikely to continue the same way going forward (i.e. the cubic model is likely not good for predicting)? Is it preferable to have a simple message to convey (i.e. use a linear trend model) or is it important to convey the complexities of this trend over the past 10 years?

Before this question is answered, the data are checked for an exponential trend, are log transformed, and a linear regression is run on them and plotted with the observed data (Figure 21). The adjusted R^2 is 44.9%. The AIC is -40.2 and is not comparable to those on the non-transformed data. The plot does not look very different from that of the linear model.

Figure 21. Number of confirmed cases of salmonellosis by year and exponential model for the trend, EU/EEA, 2010–2019



As seen in Figure 21, judging from the R^2 and the plot, the exponential model does not provide any benefit over any of the polynomial models (including the polynomial model of order 1, the linear model).

Based on the above information and with the study objectives in mind, while the cubic model is interesting, it is preferable to have a simple message to convey a general trend over the 10-year period. Therefore, the model with the linear trend is chosen as the pre-final model.

Step 8: Build the model

The model was built in the previous step. It is a simple linear regression model with time modelled as a linear trend.

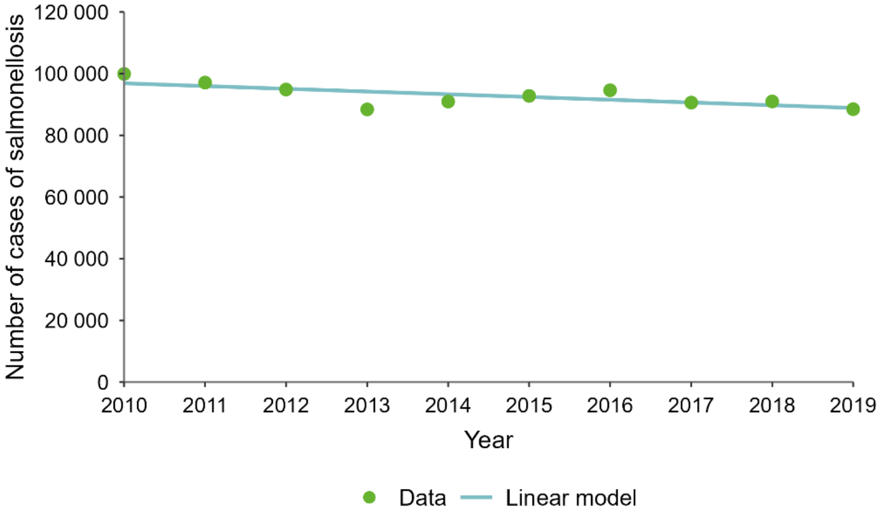
Step 9: Validate the model

Next, regression diagnostics are used to validate the model. Linear regression is being used, so the model must be assessed against assumptions of linearity, homoscedasticity, independence, and normality of error terms.

Linearity

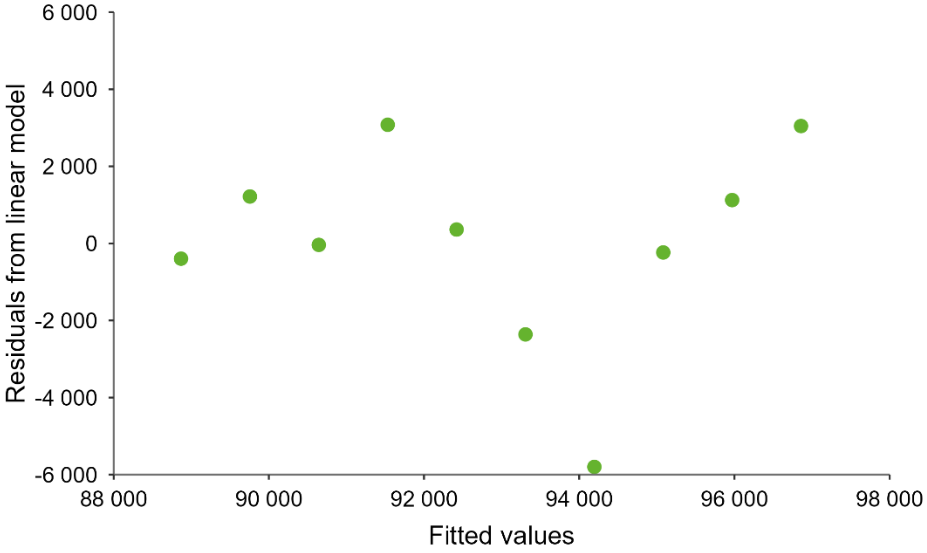
To validate the linearity assumption, it is checked whether the data points are symmetrically distributed around the model (Figure 22). This is more or less the case, but there is a wave pattern in the data.

Figure 22. Number of confirmed cases of salmonellosis by year and linear model for the trend, EU/EEA, 2010–2019



Then, one must look at the residuals against the model (Figure 23).

Figure 23. Plot with fitted values versus residuals to check the linearity assumption

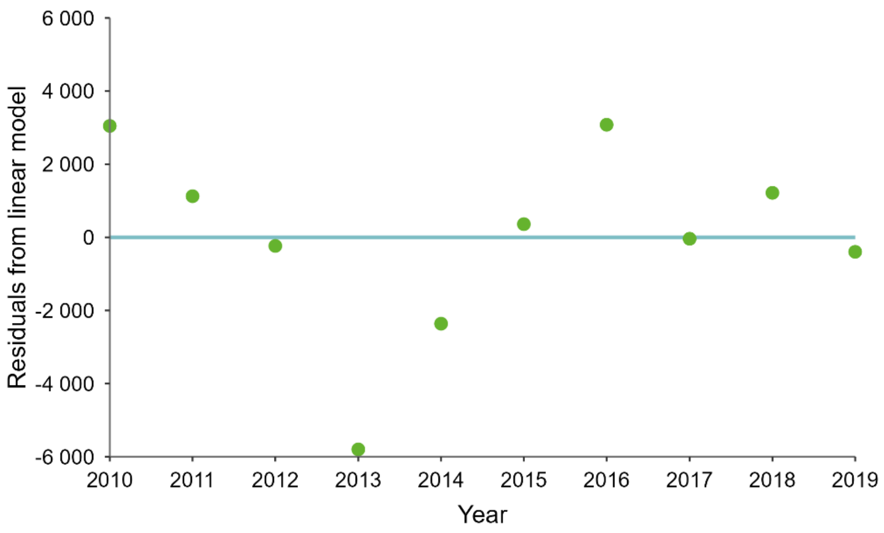


As seen in Figure 23, the data points should be symmetrically distributed around the horizontal axis. With few data points, this is hard to assess adequately. The wave structure is visible again.

Homoscedasticity

To validate the homoscedasticity assumption, the plot of residuals in Figure 23 is inspected against the model over time (Figure 24).

Figure 24. Residuals from linear model over time to check the homoscedasticity assumption

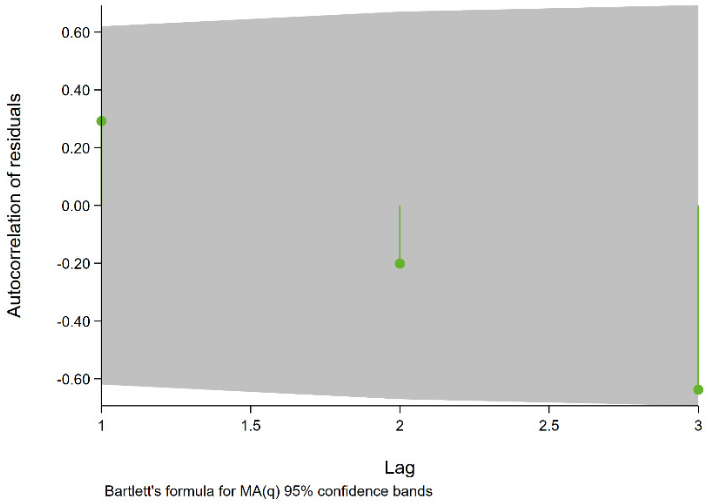


Again, with few data points it is hard to assess this adequately. The data points should ideally be symmetrically distributed over time.

Independence

To validate the independence assumption, the residuals over time are assessed and an autocorrelation graph of the residuals is created (Figure 25). With so few data points, it is hard to determine a specific pattern over time, but there does not appear to be any strong signal.

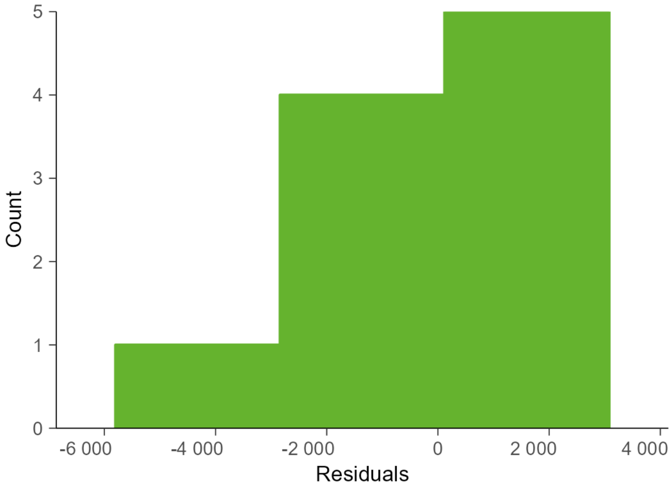
Figure 25. Autocorrelation graph to check the independence of the residuals



Normality

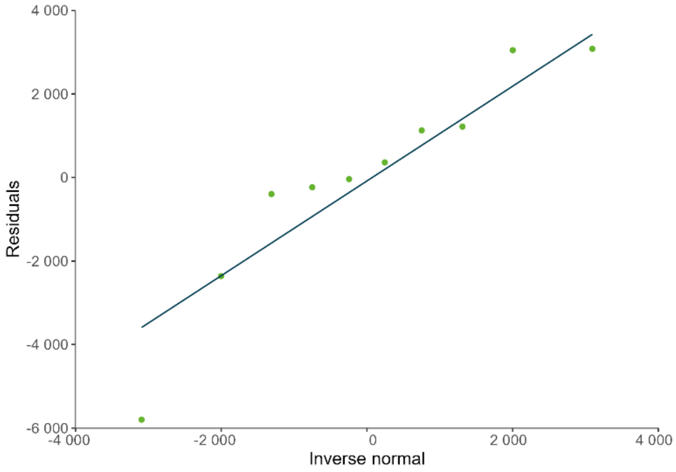
To validate the normality assumption, a histogram of the residuals is created (Figure 26).

Figure 26. Histogram to check whether the residuals are normally distributed



As seen in Figure 26, the residuals do not look normally distributed, but with so few data points it is hard to interpret a histogram. The quantile plot is also looked at (Figure 27).

Figure 27. Normal probability plot to check whether the residuals are normally distributed



As seen in Figure 27, the residuals follow the normal distribution to a certain extent, but there is one particular outlier that corresponds to the data point for 2013 (Figure 24). There is a substantial dip here. It would be interesting to discuss the reasons for this dip with the disease experts (i.e. is it possible that there was underreporting or that some countries did not report in that year?).

The verdict

The data meet the linear regression assumptions to a certain extent. It is hard to evaluate this assumption adequately in light of so few data points. When drawing and communicating inferences from the model, it must be made clear that the trend in the data has been simplified for communication purposes and that this trend is not to be extrapolated to further years.

As the absolute values of the coefficients do not exceed the standard errors and there is only one parameter in the model, it does not appear that the model has been overfitted.

Step 10: Interpret the model

The final model indicates that, for each increase in year between 2010 and 2019, there is an average decrease in salmonellosis cases in the EU/EEA of around 888 units (Table 4). This trend is statistically significant.

Table 4. Model summary

	Coefficient	Standard error	p-value	Lower 95% CI	Upper 95% CI
β_1 (Time)	-887.8	304.2	0.019	-1 589.2	-186.4
β_0 (Constant)	97 746.3	1 887.2	<0.001	93 394.3	102 098.4

CI: confidence interval

7.2 Example: Yearly and monthly trends of counts of salmonellosis cases in the EU/EEA

This example uses the algorithm in Part B, Section 6, 'Algorithm for performing the trend analysis' to demonstrate the process of obtaining yearly and monthly trends of counts of salmonellosis cases in the EU/EEA. The dataset is freely available on ECDC's 'Surveillance Atlas of Infectious Diseases' [4].

Step 1: Define the study objectives

The objectives of the study are:

- To carry out a trend analysis of monthly counts of salmonellosis cases in the EU/EEA in order to determine whether salmonellosis cases have been increasing or decreasing in the past 10 years on average, and to assess the magnitude of change.
- To understand the seasonality of reported salmonellosis cases in the EU/EEA.

Step 2: Understand the data

The data were discussed with disease experts. There are no issues of missing data (non-reporting) and the surveillance systems have remained stable. It is possible to proceed with a trend analysis.

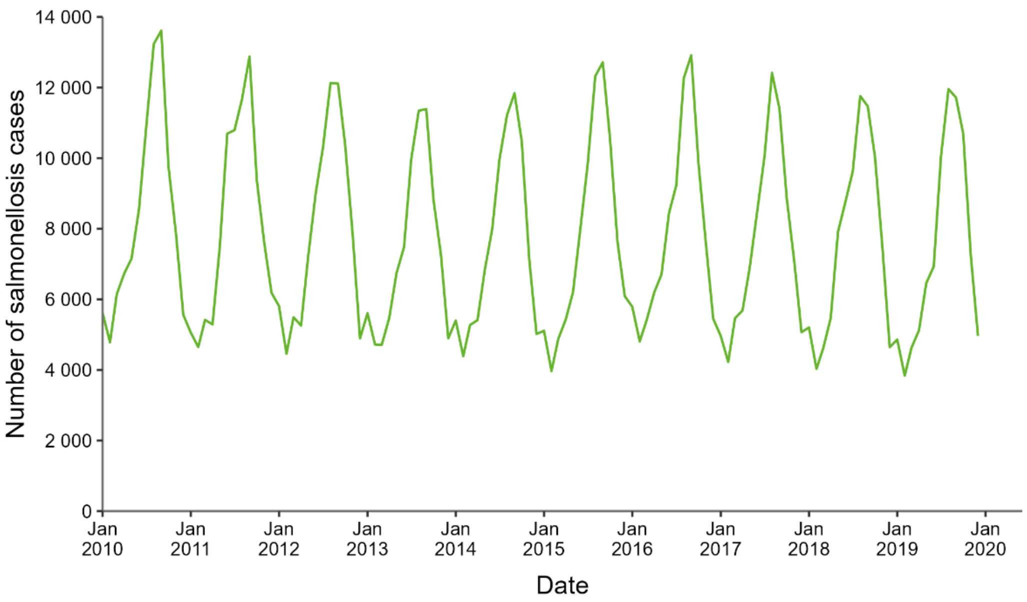
Step 3: Decide on the time unit of analysis

For the purposes of this analysis, the salmonella dataset aggregated by month-year will be used.

Step 4: Plot the data

The count variable has been plotted against the month-year variable, ensuring that the y-axis starts at zero (Figure 28).

Figure 28. Number of confirmed cases of salmonellosis by month, EU/EEA, 2010–2019



As seen in Figure 28, there is very strong seasonality with peaks in salmonellosis cases in the summer and troughs in the winter. There is sometimes a strange peak and trough around the new year, which could possibly be an effect of the holiday season. There may be a slightly downward yearly trend as well, but this is difficult to determine from the graph.

Step 5: Decide on the preliminary regression type

There are count data and very high numbers. Linear regression will be used for the trend analysis.

Step 6: Consider inclusion of seasonal terms

Seasonal terms are included for a better understanding of salmonellosis seasonality. Weekly data would probably be interesting for this study, but as they are not available, monthly data will be used instead. The term 'month' is included in the model for an understanding of how much the numbers increase or decrease on average per month compared to the reference month, after adjusting for the yearly trend.

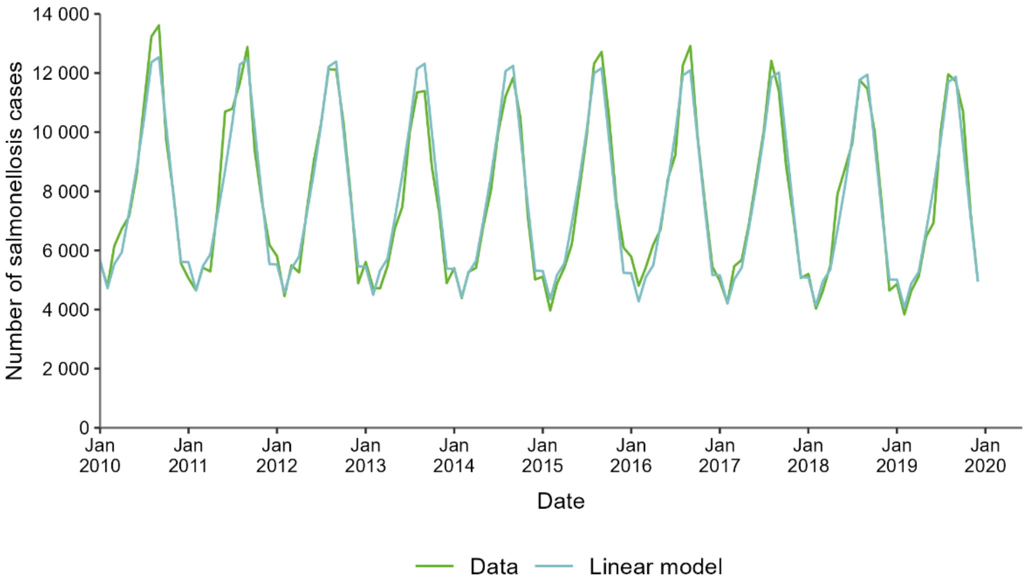
February, the month with the lowest numbers of salmonellosis cases, is used as the reference.

Step 7: Assess the nonlinearity of the trend

First the linear trend for month-year is checked, taking month into account as a seasonal indicator term, and the regression model is plotted over the data (Figure 29). Note that for the purposes of this analysis, the intercept is of no interest. Therefore, it may be easier to work with a time unit in integers that begins with one, rather than the year (2010). This could be particularly helpful when creating the quadratic and cubic terms.

Another important point to note, is that no assumptions are made about linearity for the seasonal term. By including it in the model as a categorical variable, the increases and decreases can vary in any way by month, compared to the reference month. However, the assumption is made that the increases and decreases are similar, on average, for each month over the years. If it looks like the seasonality varies over the years, then perhaps it might make sense to describe the seasonality separately for years where there is common seasonality.

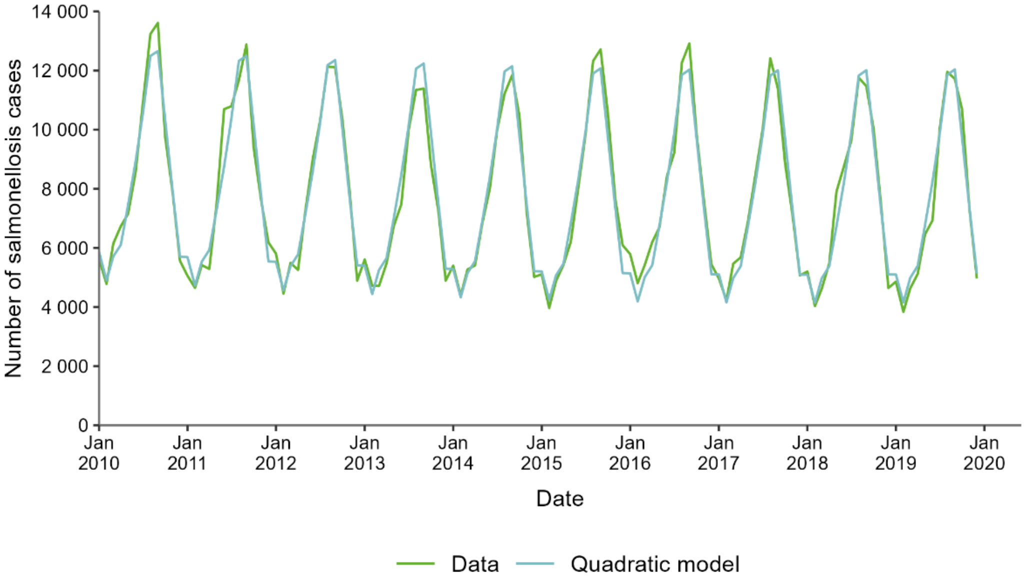
Figure 29. Number of confirmed salmonellosis cases by month and linear model for the trend, taking month into account as a seasonal indicator term, EU/EEA, 2010–2019



As seen in Figure 29, it seems that the model fits the data very well. The adjusted R^2 of the model containing time in month-years and a seasonal indicator for month is 96.1%. The adjusted R^2 of the model, without the seasonal indicator, is 44.9%. The AIC is 1860.4.

Next, a quadratic trend is run and the regression model is plotted over the data (Figure 30). The adjusted R^2 is 96.2% and the AIC is 1858.6.

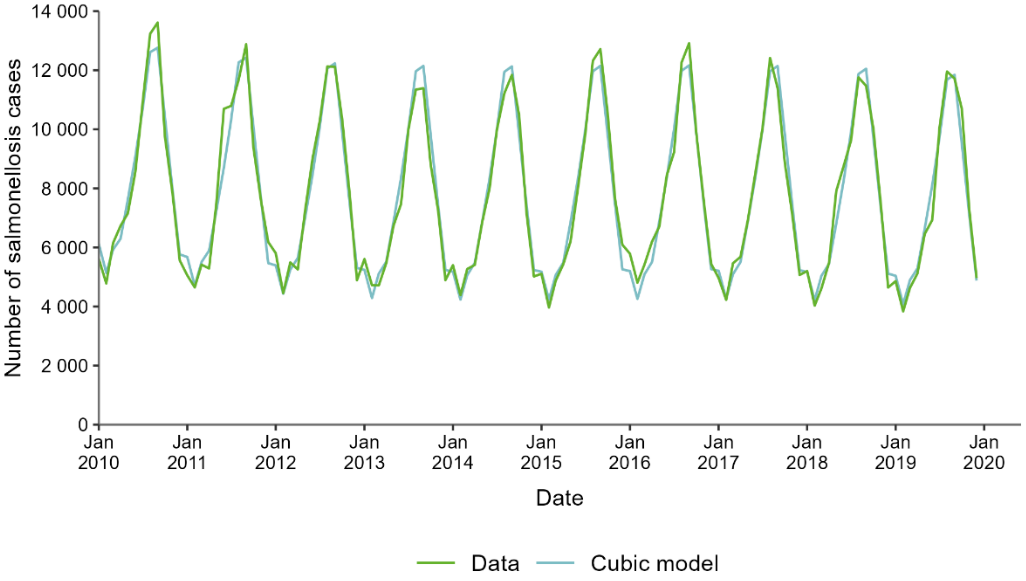
Figure 30. Number of confirmed salmonellosis cases by month and quadratic model for the trend, taking month into account as a seasonal indicator term, EU/EEA, 2010–2019



As seen in Figure 30, the trendline does not look very different from the linear trendline and the adjusted R^2 is virtually the same. The AIC is slightly lower in the quadratic model, but by less than two points, meaning that there is little difference between the models. The p-value of the likelihood ratio test comparing the linear and quadratic trend is 0.051, indicating that there is no significant advantage of the quadratic model over the linear model at the 5% level. However, this may suggest that a cubic model could be considered.

Next, a cubic trend is run and the regression model is plotted over the data (Figure 31).

Figure 31. Number of confirmed salmonellosis cases by month and cubic model for the trend, taking month into account as a seasonal indicator term, EU/EEA, 2010–2019

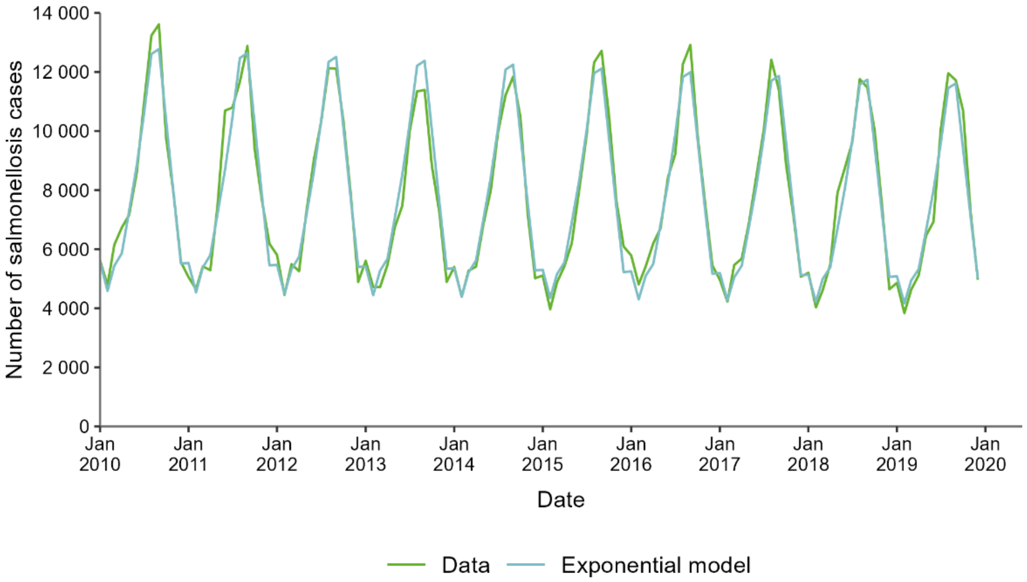


As seen in Figure 31, the cubic model allows for two curves, which can almost be seen in the plot. This models the data slightly better. The adjusted R^2 is 96.3% and the AIC is 1854.3. This is around six points lower than the linear model. The likelihood ratio test indicates that the cubic model performs better than the quadratic model $(p = 0.013)$ and the linear model $(p = 0.007)$.

At this point, the objectives of the study must be considered. Modelling a trend with a higher number of degrees of polynomials will always result in a better fit, but runs the risk of overfitting. Within the study, is it useful to model the ups and downs of the salmonellosis cases trend and comment on its magnitude, even if this curvature is unlikely to continue the same way going forwards (i.e. the cubic model is likely not good for predicting)? Is it preferable to have a simple message to convey (i.e. use a linear trend model) or is it important to convey the complexities of this trend over the past 10 years?

Before this question is answered, the data are checked for exponential trend, log transformed, and a linear regression is run on them and plotted with the observed data. The adjusted R^2 is 96.0%. The AIC is -287.6 and not comparable to those on the non-transformed data. The plot does not look very different to that of the linear model (Figure 32).

Figure 32. Number of confirmed salmonellosis cases by month and exponential model for the trend, taking month into account as a seasonal indicator term, EU/EEA, 2010–2019



Judging from the R^2 and the plot (Figure 32), the exponential model does not provide any benefit over the polynomial models (including the polynomial model of order 1, the linear model).

Based on the above information and with the study objectives in mind, while the cubic model is interesting, it is preferable to have a simple message to convey a general trend over the 10-year period. Therefore, the model with the linear trend is chosen as the pre-final model, with month as a seasonal indicator.

Step 8: Build the model

The model was built in the previous step. It is a simple linear regression model with time modelled as a linear trend and month as a seasonal indicator.

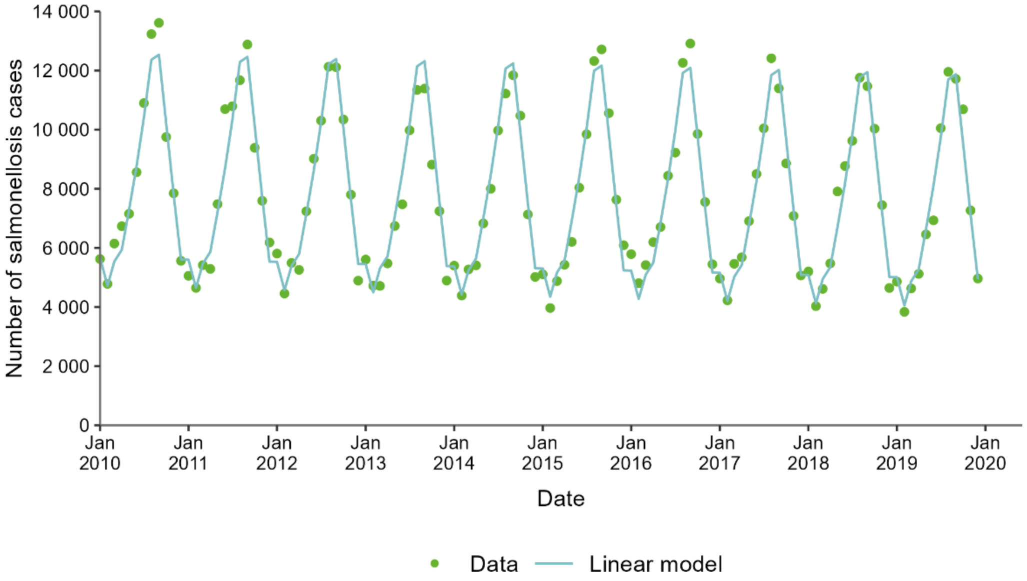
Step 9: Validate the model

Next, regression diagnostics are used to validate the model. Linear regression is being used, so the model must be assessed against assumptions of linearity, homoscedasticity, independence, and normality of error terms.

Linearity

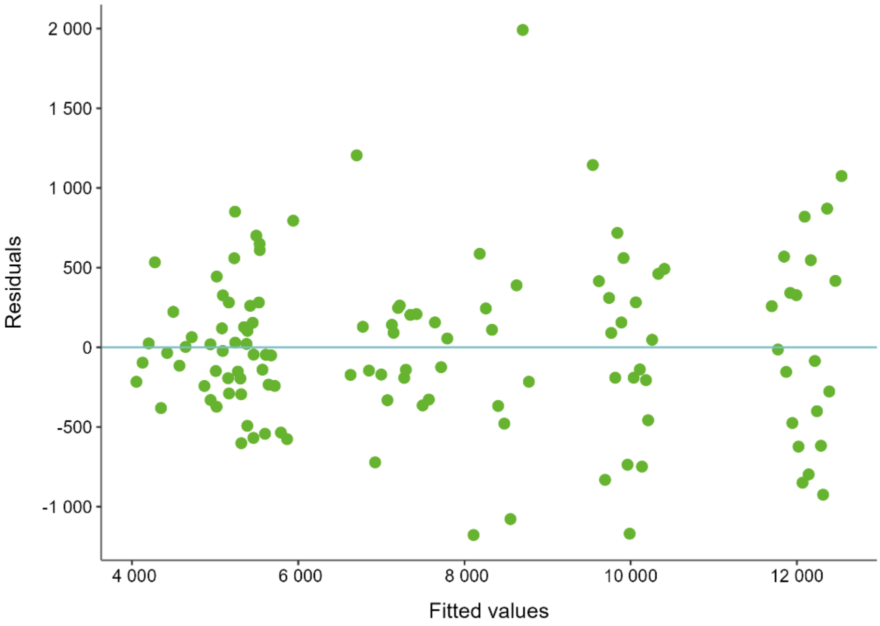
To validate the linearity assumption, it is checked whether the data points are symmetrically distributed around the model. This is more or less the case (see Figure 33). There are some peaks in the data that are not adequately taken into account by the model. These may be years with particularly large outbreaks.

Figure 33. Number of confirmed salmonellosis cases by month and linear model for the trend, taking month into account as a seasonal indicator term, EU/EEA, 2010–2019



Then the residuals are checked against the model (Figure 34).

Figure 34. Plot with fitted values versus residuals to check the linearity assumption

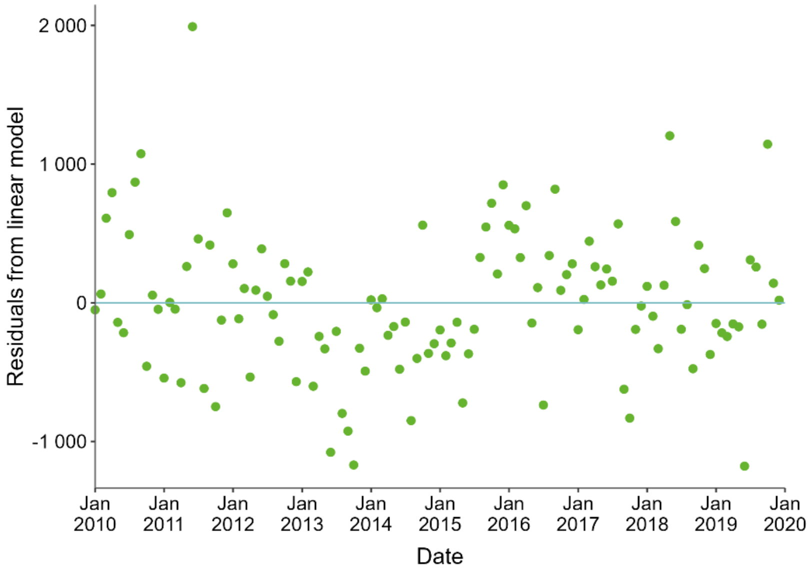


As seen in Figure 34, the data points should be reasonably distributed symmetrically around the horizontal.

Homoscedasticity

To validate the homoscedasticity assumption, the plot of residuals in Figure 34 is inspected against the model over time (Figure 35).

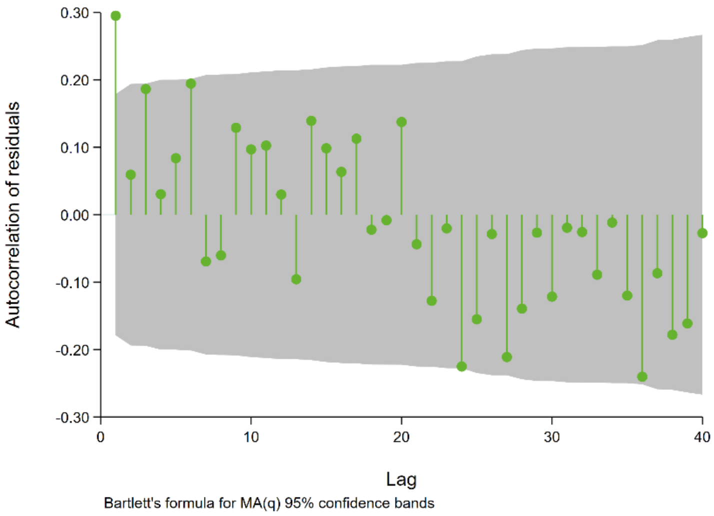
Figure 35. **Residuals from linear model over time to check the homoscedasticity assumption**



As seen in Figure 35, the residuals are not entirely symmetrical over time. There appears to be a time period, around 2013 to 2015, where the residuals are consistently lower than the model.

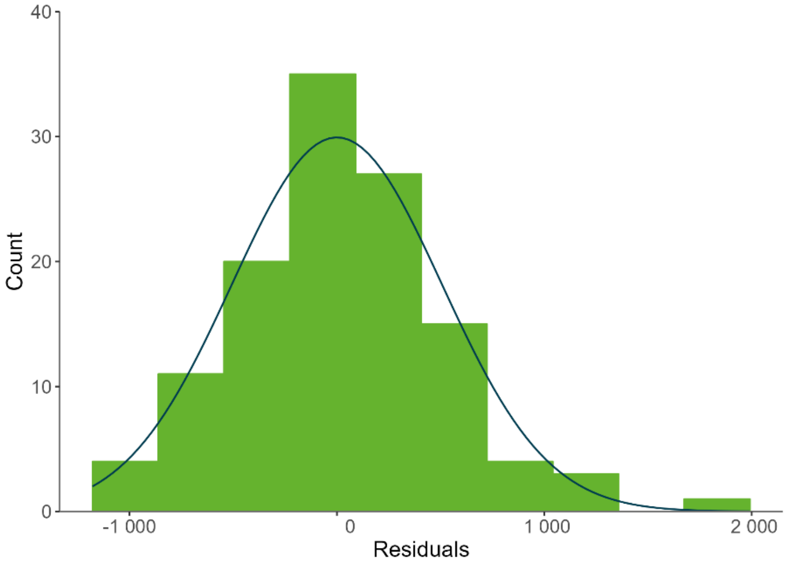
Independence
To validate the independence assumption, the residuals over time are checked (Figure 35) and an autocorrelation graph of the residuals is created (Figure 36). There does not appear to be any strong signal of pattern or autocorrelation between residuals.

Figure 36. **Autocorrelation graph to check the independence of residuals**



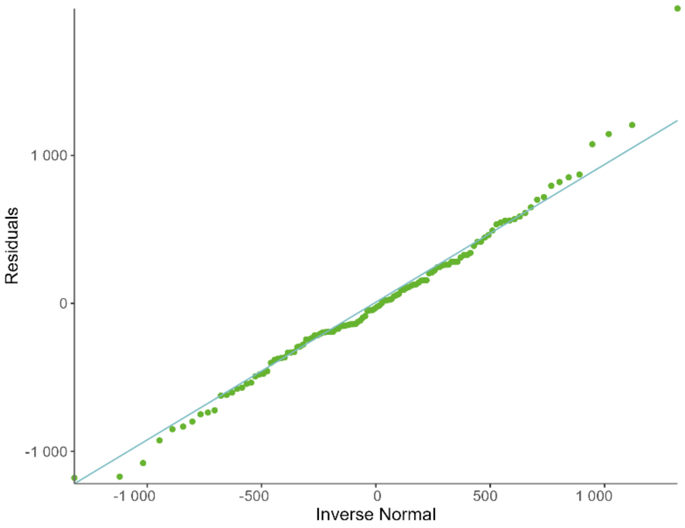
Normality assumption
To validate the normality assumption, a histogram of the residuals is created (Figure 37). Note that a normal density has been added to the plot.

Figure 37. Histogram to check whether the residuals are normally distributed



As seen in Figure 37, the residuals appear to be reasonably normal, except in the positive end of the residuals, where they have a bit of a right-hand tail. Again, this is likely to be due to increased outbreaks in some summer months during the study period. A normal probability plot is also created (Figure 38).

Figure 38. Normal probability plot to check whether the residuals are normally distributed



As seen in Figure 38, the residuals follow the normal distribution very well, except (not unexpectedly) when the residuals are very large. Again, it is possible that this is due to large outbreaks in certain years compared to others.

The verdict

The data meet the linear regression assumptions to a certain extent. In the homoscedasticity plot, it can be seen that there are a few time points where the residuals are not symmetrical over time. However, overall in the graph there does not appear to be an increase or decrease in variability over time.

It is important to be clear, when communicating inferences from the model, that it has some limitations. If the homoscedasticity assumption is violated, then the coefficients are still valid, but there may be biased standard errors (meaning that the p-values may not be reliable).

As the absolute values of the coefficients do not exceed the standard errors and there is only one parameter in the model, it does not appear that the model has been overfitted.

Step 10: Interpret the model

The final model indicates that, for each increase in year between 2010 and 2019, there is an average decrease in salmonellosis cases in the EU/EEA by month-year of around six cases, resulting in a yearly decrease of 74 cases, after taking monthly seasonality into account. This trend is statistically significant.

In terms of seasonality, compared to February, there is an increase of 7 682 and 7 862 cases in August and September, respectively; an increase of 1 233 to 5 718 cases in April, May, June, July, October and November; and an increase of 824 to 952 cases in December, January and March (Table 5).

Table 5. Model summary

	Coefficient	Standard error	p-value	Lower 95% CI	Upper 95% CI
β_1 (Time)	-6.2	1.4	<0.001	-9.0	-3.4
Month 1	949.3	239.1	<0.001	475.4	1 423.3
Month 2	Reference	–	–	–	–
Month 3	824.6	239.1	0.001	350.6	1 298.5
Month 4	1 233.2	239.1	<0.001	759.2	1 707.2
Month 5	2 595.1	239.1	<0.001	2 121.1	3 069.1
Month 6	4 081.4	239.2	<0.001	3 607.3	4 555. 5
Month 7	5 718.4	239.2	<0.001	5 244.3	6 192.6
Month 8	7 682.3	239.2	<0.001	7 208.0	8 156.6
Month 9	7 862.0	239.3	<0.001	7 387.6	8 336.3
Month 10	5 540.7	239.3	<0.001	5 066.2	6 015.2
Month 11	3 129.9	239.4	<0.001	2 655.2	3 604.5
Month 12	952.1	239.5	<0.001	477.3	1 426.9
β_0(Constant)	4 730.9	186.7	<0.001	4 360.7	5 101.1

CI: confidence interval

Part C. Trend analysis across multiple geographical areas (countries, regions, etc.)

1. Considerations for trend analysis across multiple geographical areas

1.1 Why combine data from different countries?

At ECDC, experts most often analyse surveillance data reported from individual countries. A common research question is: What is the trend in the EU/EEA over the past years?

To answer this question, individual country data must be combined. This is also called 'pooling' data. There are different methods for combining or pooling data, which depend on the research questions. These questions are described in Part A, Section 2, 'What type of question is trend analysis appropriate for?'.

In the ECDC context, data are pooled to better understand EU/EEA-wide trends for given health topics. Pooling data also ensures a higher sample size and can lead to estimation of trends with greater precision. However, a greater precision is not guaranteed, if there is underlying heterogeneity.

1.2 When should data from different countries be combined?

While it is technically easy to pool data from different countries, it may not always be advisable. If underlying data are heterogeneous, pooling may not be appropriate.

ECDC has published the guidance document, 'Managing heterogeneity when pooling data from different surveillance systems' [5]. The heterogeneity of the underlying data should be assessed prior to carrying out the trend analysis so that an informed decision can be made whether to pool data from all countries or some countries, or not to pool data at all.

The different heterogeneities within the underlying country data are outlined in the aforementioned ECDC guidance document [5]. Heterogeneities related to data collection and data collection systems that have the biggest impact on a trend analysis of pooled data are:

- heterogeneities in system design (changes in stability in systems over time),
- differences in data sources,
- surveillance systems established in different populations,
- differences in case definitions, and
- missing data in one or more countries (missing data issues are covered in Part D, 'Trend analysis with missing data').

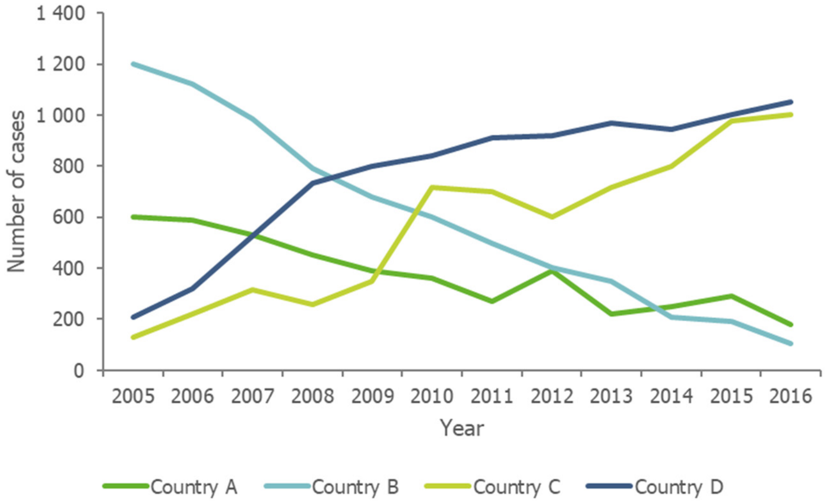
It is important to assess the country data qualitatively to understand if heterogeneities are present, and to evaluate their impact on the trend analysis. Does it make sense to include a country that collects data very differently than other countries? If groups of countries use different case definitions, should the trend analysis be carried out by group, combining those that use the same case definition? Pooling should only be undertaken if the assumption is met that data are comparable.

Heterogeneities (i.e. differences in countries' overall trends) can also be related to disease determinants. The impact of these heterogeneities needs to be considered for the trend analysis and trend analysis methods. Heterogeneities related to disease determinants include:

- **Time period** (e.g. countries with different peak seasons of disease incidence or countries with years with particularly high outbreak cases)
- **Environmental factors** (e.g. certain countries' climates have a higher incidence of West Nile virus)
- **Interventions** (e.g. differences in vaccination programmes)
- **Population characteristics** (e.g. differences in the population structure of a country, such as a very young or a very old population).

The heterogeneities in data collection and data collection systems are referred to as 'nuisance', but the heterogeneities in disease determinants are referred to as 'real'. What is being measured with a pooled-country trend if there are large differences in underlying trends (Figure 39 provides an example of this)? How useful is a pooled-country analysis if there are very different trends across countries (see Part C, Section 2.3.4, 'Stratified analysis')? Depending on the study objectives, it may be more informative and interesting to present the trend analysis by country (Figure 39), rather than the trend analysis of the pooled data.

Figure 39. Number of cases of disease X, countries A–D, 2005–2016



2. Theory

2.1 Introduction

If the research question is 'What is the trend in the EU/EEA over the past years?', there are different methods to answer this and different interpretations for each method. The main question is whether to take individual countries into account and, if so, how to do this.

There are three main methods for dealing with multi-country data in the context of trend analysis, each with different sets of assumptions. In the context of this document, these are:

- crude pooled analysis,
- stratified analysis and
- adjusted pooled analysis.

2.1.1 Crude pooled analysis

In a crude pooled analysis, all data are pooled together and the EU/EEA is considered as one country. Inter-country differences are not taken into account (e.g. case numbers may be declining in one country and rising in another).

2.1.2 Stratified analysis

There is no pooling in a stratified analysis. The trend is calculated for each country individually.

2.1.3 Adjusted pooled analysis

The adjusted pooled analysis finds a middle ground between the crude pooled and the stratified analyses. It does not assume that the EU/EEA is one country, but does assume some similarities between the countries in the region, which can be modelled. There are different statistical methods to estimate an adjusted pooled trend, but this guidance document recommends a random-effects meta-analysis approach.

Each of these types of analyses (crude pooled, stratified and adjusted pooled) have different interpretations, strengths and weaknesses, which are outlined in Part C, Section 2.3, 'Crude pooled analysis of data from different countries'.

2.2 Absolute change in numbers or rates versus percentage change in numbers or rates

As seen in Part B, Sections 5.1, 'Count data' and 5.2, 'Other types of data', it is possible to perform trend analysis using counts or rates.

The percentage change can also be measured in numbers or rates, either by modelling the data with an exponential trend or by using an alternative regression type (e.g. Poisson regression or, if overdispersion is present, negative binomial regression).

When using absolute numbers and linear regression, information is obtained on the burden of disease in the country or region. This information should be properly interpreted. For example, an average increase of 900 cases per year for disease X could be a lot in a country with a small population, whereas, the same number could be little in the context of a country with a larger population. For a country, this provides information that is useful for public health decision-making (e.g. estimating resources that may be required for hospital beds or allocated to interventions or treatments). This indicator, however, is not easy to use for the purposes of comparison without knowing the population size. In this case, a standardised indicator would be preferred (e.g. incidence, where the absolute number of cases is divided by the population). For example, if linear regression and rates are being used and individual country data are taken into account, a decrease of 20 cases per 100 000 population per year for an average-sized country is easy to understand and to compare against other countries. Other standardised indicators include proportions (e.g. proportion of multidrug-resistant tuberculosis) or rates that are standardised even further (e.g. age-standardised rates).

However, even the change in rate per year may not be meaningful if there are countries with vastly differing underlying rates. In such instances, obtaining a metric of annual percentage change may be more useful. A country with a very low rate of disease and another with a very high rate of disease are unlikely to have the same absolute change in rate per year. However, their annual percentage change may be the same.

2.3 Crude pooled analysis of data from different countries

2.3.1 Crude pooled analysis trends based on absolute numbers and linear regression

A crude analysis approach is often used when carrying out a trend analysis of pooled data. In this approach, the data are pooled without taking into account differences in individual country data. In the context of ECDC, country-level data are often pooled together to obtain an EU/EEA total. The EU/EEA is treated as one country and any clustering (which would be caused by differences in individual country trends) is ignored.

Method

When carrying out a crude trend analysis of pooled data, all data are summed together without taking underlying country structures into account and an overall total is calculated. Using the example from Figure 39, there are four countries with individual data, as shown in Table 6. As seen in Figure 39, two countries show a positive trend and two countries show a negative trend.

Table 6. Number of cases of disease X by year, countries A–D, 2005–2016

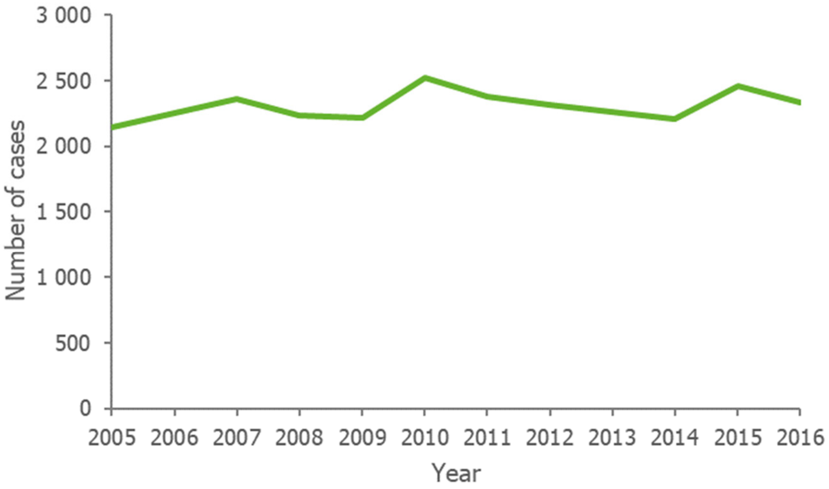
Year	Country A	Country B	Country C	Country D
2005	600	1 200	130	210
2006	590	1 120	220	320
2007	530	987	317	525
2008	450	789	258	735
2009	390	681	347	800
2010	360	600	717	840
2011	270	499	700	910
2012	390	402	600	920
2013	220	350	718	970
2014	250	210	800	945
2015	290	190	979	1 000
2016	180	104	1 000	1 050

If a crude trend analysis is performed, then the total number of cases of disease X, by country, is summed for each year to create a pooled total by year. This results in the Total column of Table 7, and has been visualised in Figure 40 as well.

Table 7. **Number of cases of disease X by year, countries A–D and total, 2005–2016**

Year	Country A	Country B	Country C	Country D	Total
2005	600	1 200	130	210	2 140
2006	590	1 120	220	320	2 250
2007	530	987	317	525	2 359
2008	450	789	258	735	2 232
2009	390	681	347	800	2 218
2010	360	600	717	840	2 517
2011	270	499	700	910	2 379
2012	390	402	600	920	2 312
2013	220	350	718	970	2 258
2014	250	210	800	945	2 205
2015	290	190	979	1 000	2 459
2016	180	104	1 000	1 050	2 334

Figure 40. **Number of cases of disease X by year, total of countries A–D, 2005–2016**

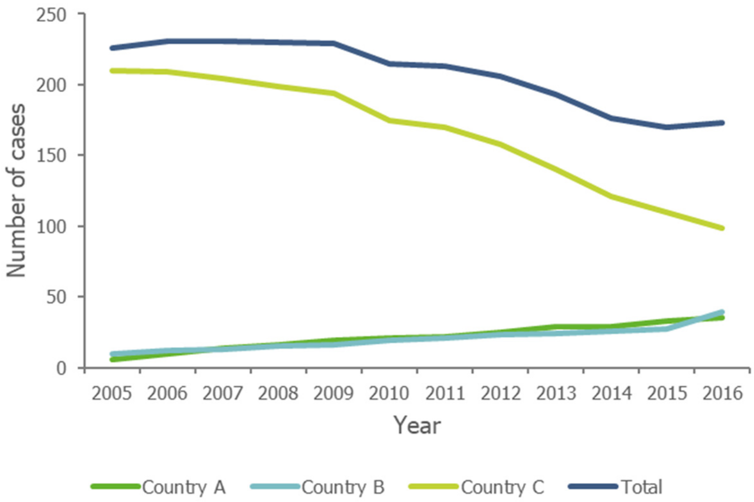


Interpretation

With linear regression, the resulting pooled trend is in fact the sum of the trends in the underlying data. When performing a crude pooled analysis, the data are treated as if they came from one country and any underlying trends are ignored. The pooled trend is reasonably stable over time.

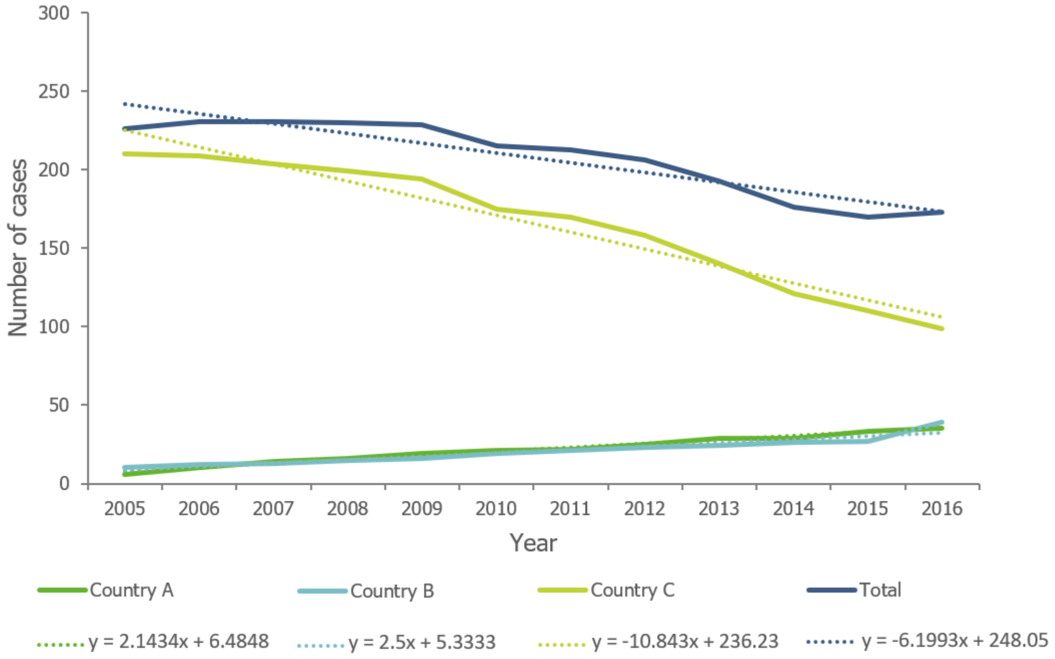
Another example can be seen in Figure 41, where there are two countries with small numbers and upward trends, and a third country (country C) with a strong downward trend. The crude pooled trend is the sum of the trends in each country, which results in a pooled downward trend, even though two countries have upward trends and only one has a downward trend.

Figure 41. Number of cases of disease X, countries A–C and total, 2005–2016



The pooled trend coefficient is the sum of the individual country coefficients. In Figure 42, the trend equations are added to the individual country and pooled trends. The sum of the trend coefficients of countries A to C are $2.5 + 2.1 - 10.8$, which equals the pooled trend coefficient: -6.2. The same is true for the intercept.

Figure 42. Number of cases of disease X and trend equations, countries A–C and total, 2005–2016



In this example, linear regression was used, and a linear trend was modelled as a polynomial trend to the order 1. The interpretation is the same for trends of polynomial regression of higher orders (the coefficients of each order are totalled in the crude pooled trend).

However, this would no longer hold if different countries had different types of trends (e.g. country A follows a linear trend, country B a cubic trend and country C a quadratic trend). In that case, a pooled trend analysis would not be appropriate.

The results obtained from linear regression of the total number of cases by country and by time (in years) are in Table 8.

Table 8. Model summary

	Coefficient	Standard error	p-value	Lower 95% CI	Upper 95% CI
β_1 (Time)	-6.20	0.75	<0.001	-7.87	-4.53

CI: confidence interval

The output can be interpreted as: each year there is a statistically significant $(p < 0.05)$ decrease of 6.2 cases of disease X in the overall population of countries A to C, not taking any country-specific trends into account. This trend is statistically significant (95% CI: -7.87 to -4.53).

Strengths
The strength of this method is that it is very simple to carry out and can easily be incorporated into routine analysis. It is also easy to understand for the reader.

Limitations
This analysis assumes that the data come from one country and does not take individual country trends into account. Depending on the study objective, this may be what is required; in other circumstances, this may be a limitation. In this scenario, absolute numbers are used. This provides information on the burden of disease, but these numbers may not be helpful for understanding how the trend compares to another region (e.g. comparing the EU/EEA to the Americas or a particular country).

If making a comparison with another region is the aim, it may be more useful to use a different indicator than yearly trend in absolute numbers.

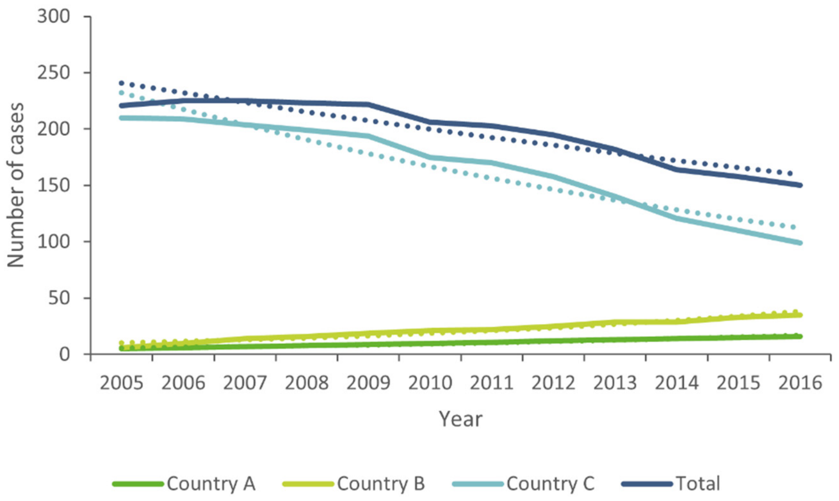
2.3.2 Crude analysis trends based on absolute numbers and other regression types

In addition to linear regression, trend analysis can also be performed with other regression types, such as Poisson or negative binomial regression (if overdispersion is present). See Part B, Section 5.4, 'Poisson regression' on the use of regression types other than linear regression, in terms of assumptions and which research questions they answer.

Method
As with linear regression, when Poisson regression is chosen, all data are totalled without taking underlying country structures into account. An overall total is calculated.

Figure 43 presents the Poisson models of three countries from Figure 41, and the total. Note that, as per Part B, Section 5.4, the Poisson assumptions must be met. If not, negative binomial regression should be considered.

Figure 43. Number of cases of disease X and Poisson model, countries A–C and total, 2005–2016



Interpretation
The Poisson regression of the crude pooled analysis (the overall total of the countries) gives the following output (Table 9).

Table 9. **Model summary**

	Coefficient	Standard error	p-value	Lower 95% CI	Upper 95% CI
β_1 (Time)	-0.030	0.006	<0.001	-0.041	-0.018

CI: confidence interval

In the crude analysis model, the coefficient of the Poisson model is -0.030 (95% CI: -0.041 to -0.018). This means that the expected number of cases of disease X, in the overall population of countries A to C, decreases significantly $(p < 0.05)$ by a factor of $exp(-0.030) = 0.9$ (95% CI: 0.96–0.98).

Strengths

This annual percentage change can be a useful metric, as it is comparable across groups that have very different absolute numbers. It can answer the research question: on average, what is the annual percentage change in the number of cases of disease X?

Limitations

As before, this crude pooled analysis assumes that the data come from one region and does not take individual country trends into account. Depending on the study objective, this may be what is required; in other circumstances, this may be a limitation.

2.3.3 Crude analysis of trends based on rates

As mentioned in Part B, Section 5, 'Choice of statistical model for the outcome', a trend analysis can be performed using rates in two different ways:

- **Modelling rates with linear regression**: If rates are large enough (not too close to zero) and appear to be changing linearly, and if the variance around the trend is stable over time, linear regression can be used. The study question is: what is the trend in absolute rate over time?
- **Modelling rates with Poisson (or negative binomial) regression:** If rates are changing exponentially and the variance around the trend appears to be increasing or decreasing, Poisson (or negative binomial) regression can be used. The study question is: what is the percentage change in trend over time?

Method

When carrying out a crude analysis of trends based on rates, the absolute number of cases of disease X are totalled for each country and each time unit. The same is done for the population variable. This gives the total number of disease notifications for each time unit, and the total population of the countries for each time unit (Table 10).

Table 10. **Number of cases of disease X, by country and pooled, including population data, by year, countries A–C, 2005–2016**

Country	Year	Number of cases	Population
A	2005	5	835 164
B	2005	6	1 083 990
C	2005	210	81 914
A	2006	6	837 516
B	2006	10	1 100 063
C	2006	209	83 975
A	2007	7	840 812
B	2007	14	1 107 588
C	2007	204	86 201
A	2016	16	888 877
B	2016	35	1 154 551
C	2016	99	89 489

Country	Year	Number of cases	Population
All	2005	221	2 001 068
All	2006	225	2 021 554
All	2007	225	2 034 601
All	2008	223	2 045 570
All	2009	222	2 054 662
All	2010	206	2 066 919
All	2011	203	2 085 989
All	2012	195	2 097 938
All	2013	182	2 108 507
All	2014	164	2 119 017
All	2015	158	2 126 017
All	2016	150	2 132 917

A rate can be obtained for each year by dividing the count by the population and multiplying it by a metric (e.g. by 100 000 to give the rate per 100 000 population/year) (Table 11). This rate can be used directly in linear regression.

Table 11. Calculating yearly rates of disease X per 100 000 population, totals of countries A–C, 2005–2016

Country	Year	Number of cases	Population	Rate per 100 000 population
All	2005	221	2 001 068	11.04
All	2006	225	2 021 554	11.13
All	2007	225	2 034 601	11.06
All	2008	223	2 045 570	10.90
All	2009	222	2 054 662	10.80
All	2010	206	2 066 919	9.97
All	2011	203	2 085 989	9.73
All	2012	195	2 097 938	9.29
All	2013	182	2 108 507	8.63
All	2014	164	2 119 017	7.74
All	2015	158	2 126 017	7.43
All	2016	150	2 132 917	7.03

Interpretation

The crude pooled analysis of rates using linear regression gives the following output (Table 12).

Table 12. Model summary

	Coefficient	Standard error	p-value	Lower 95% CI	Upper 95% CI
β_1 (Time)	-0.408	0.035	<0.001	-0.487	-0.329

CI: confidence interval

The output can be interpreted as: each year there is a decrease of 0.41 cases of disease X per 100 000 population in the overall population of countries A to C, not taking differences between individual countries into account. This trend is statistically significant (95% CI: -0.49 to -0.33).

The crude pooled analysis using Poisson regression gives the following output (Table 13).

Table 13. Model summary

	Coefficient	Standard error	p-value	Lower 95% CI	Upper 95% CI
β_1 (Time)	-0.043	0.006	<0.001	-0.054	-0.031

CI: confidence interval

In the crude analysis model, the coefficient of the Poisson model is -0.043, with confidence intervals ranging from -0.054 to -0.031. This means that the expected rate of cases of disease X per 100 000 population, in the overall population of countries A to C, decreases each year by a factor of $exp(-0.043) = 0.96$ (95% CI: 0.95–0.97) or by 4%.

Strengths

The strength of this approach is that the population is taken into account. In the linear regression of rates, it is also possible to extrapolate data for a given country. If there is a decrease of 0.41 cases of disease X per 100 000 population per year, and a country has a population of one million, this means that with the same trend one would expect that the country would have a decrease of 4.1 cases per year, given that it follows a similar trend as the overall trend.

In the Poisson regression analysis, there is the same advantage of having taken the population into account. Rather than an absolute increase or decrease in cases, it can determine a percentage increase or decrease.

Limitations

This crude pooled analysis assumes that the data comes from one region and does not take individual country trends into account. Depending on the study objective, this may be what is required; in other circumstances, this may be a limitation.

2.3.4 Stratified analysis

Method

If the data are from several countries, the first step is to assess the methodological and clinical heterogeneities underlying the data. This can be done qualitatively, as has been outlined in Part C, Section 1.2, as well as in the ECDC pooling guidance document [5]. Certain countries may use different case definitions, or the underlying epidemiology of a disease may be very different across countries (e.g. endemic diseases versus diseases only associated with travel).

In this case, a stratified analysis should be considered. This could be a total stratified analysis, where the results from each country are presented separately, or a partial stratified analysis. In a partial stratified analysis, pooled analysis is possible for certain groups of countries that share similar characteristics (e.g. use the same case definition).

In a second step, heterogeneity can be visually and statistically assessed. The trends can be plotted by country. Are they going in the same direction? Do they have the same rate of change? Do the data look like they have the same functional form (e.g. do they all have a linear trend or a quadratic trend)?

Individual trends in countries

In trend analysis, if you are dealing with a large number of countries (e.g. measuring trends in the EU/EEA), it is very likely that some countries will have differing trends. Whether these differences warrant a stratified analysis over a pooled analysis depends on the study objectives and the advice and opinion of a disease expert who understands the heterogeneity of the underlying data. When carrying out a pooled analysis, it is a good idea to provide the country-specific trends (the stratified analysis) alongside the pooled analysis, with a description of potential heterogeneities. This way the reader can better interpret the pooled analysis themselves.

Finally, statistical measures of heterogeneity can be calculated using meta-analysis techniques [11], such as Cochrane's Q-test and the I^2 index.

Cochrane's Q-test is a test statistic outputted during most meta-analysis procedures [12]. It is a non-parametric statistical test that measures whether the individual effects are identical. It follows a chi-squared distribution and the p-value gives information about the assumption of a deviation from a homogeneity of effects (if there is evidence of heterogeneity among studies).

Another measure of heterogeneity is the I^2 index that aims to quantify heterogeneity rather than reducing it to a binary quantity (presence or absence of heterogeneity) when using Cochrane's Q-test p-value [11]. The I^2 index uses Cochrane's Q statistic and the degrees of freedom (the number of studies (k) minus one; $k-1$) to obtain a percentage estimate:

$$I^2 = \frac{Q-(k-1)}{Q} * 100\% \quad \text{for} \quad Q > (k-1)$$

$$I^2 = 0 \quad \text{for} \quad Q \le (k-1)$$

According to the Higgins and Thompson classification, an I^2 index of around 25% indicates low heterogeneity, 50% indicates medium heterogeneity and 75% indicates high heterogeneity between studies [13].

However, these measures must be interpreted with caution, as they are subject to power issues when the number of studies is small (true heterogeneity may not be detected). If the number of studies is large, the Q-test may be statistically significant (indicating there is heterogeneity), but the amount of heterogeneity may not be epidemiologically relevant.

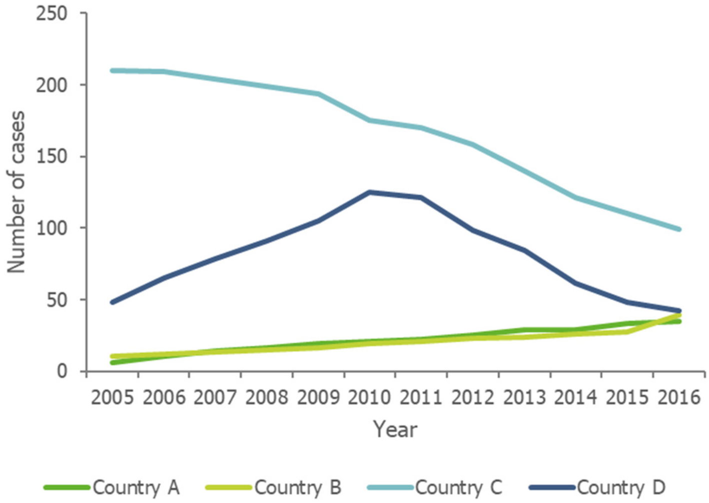
Interpretation

Figure 44 shows the number of cases of disease X for countries A to C, as in the examples in the previous sections. An additional country has been added to this figure, country D, for which the number of cases has a non-linear shape.

A disease expert confirms very different epidemiological situations of disease X in these four countries. There are two countries where cases are rising linearly, one country where cases are decreasing linearly, and one country with an umbrella-shaped curve.

Question: How would you best model the trend in disease X for country D?

Figure 44. Number of cases of disease X in countries A–D, 2005–2016



While there is an argument for pooling countries A and B together, pooling only two countries may not add value. Though heterogeneity of the trends may already be apparent, the statistical measures of heterogeneity should be calculated. For this, one must assume the same type of trend (e.g. a linear trend, even though the trend of country D has more of a quadratic shape).

> When comparing linear trends between countries, it is entirely possible that statistical heterogeneity is not detected, even if the data follow a completely quadratic trend. It may be that the linear trend forced on the country with the quadratic trend is similar to that of the linear trend of the other countries.

The Cochrane's Q statistic p-value is < 0.001 and the I^2 is 99.1%. As already assumed, there is considerable heterogeneity within the estimates.

The individual trends can be calculated by country, as summarised in Table 14. To measure heterogeneity, the same type of trend (linear) was assumed for each country. Now, because the trends are not being pooled, different types of trends can be selected according to the study objectives and the characteristics of the data. A linear trend is displayed for countries A to C, and a quadratic trend for country D.

Table 14. Regression outputs for time trend in countries A–D, 2005–2016

	Coefficient	Standard error	p-value	Lower 95% CI	Upper 95% CI
Country A β1 (Time)	2.50	0.09	<0.001	2.30	2.70
Country B β1 (Time)	2.14	0.22	<0.001	1.66	2.63
Country C β1 (Time)	-10.84	0.73	<0.001	-12.48	-9.20
Country D β1 (Time)	29.64	3.83	<0.001	20.97	38.31
Country D β1 (Time^2)	-2.39	0.29	<0.001	-3.04	-1.74

CI: confidence interval

With each increase in time unit, the number of cases is expected to increase by 2.5 in country A and 2.14 in country B. In country C, with each increase in time unit the number of cases is expected to decrease by 10.84. In country D, cases initially increase with time, then stop increasing and start decreasing with time around $t = 6$, following an umbrella-shaped curve.

Strengths

A stratified analysis is useful for understanding the underlying trends within a dataset comprising many countries, regions or other strata. It can be a useful step to really understand the data even if the main analysis is a pooled analysis.

If the underlying countries have very different types of trends (e.g. a mix of linear, quadratic or exponential trends), then a stratified analysis can display these, while a pooled analysis cannot.

Limitations

If the objective of the analysis is to represent the overall data, then a stratified analysis will not be sufficient. Within the dataset there may be similarities between the countries that should be combined, and a pooled analysis would provide additional information.

2.3.5 Adjusted pooled analysis

Method

Different methods exist for combining or pooling data. When there is individual data by country, the methods include one-stage methods and two-stage methods. In one-stage methods, the data are analysed in a single step using multilevel models that take the country-level effect into account. In two-stage methods, a stratified analysis is first performed, as outlined earlier. Each country is analysed separately to obtain the effect of interest (e.g. the trend coefficient and its standard error). Then, in a second step, these individual country-specific estimates are combined using a meta-analysis (this can be a common-effect or a random-effects meta-analysis).

There are many different nomenclatures for multilevel models. They are also commonly referred to as ‘hierarchical models’ or ‘mixed models’.

Often, but not always, similar results are obtained using a one-stage and a two-stage approach [12]. For the purposes of the trend analysis outlined in this guidance document, use of a random-effects meta-analysis approach is recommended.

Strengths

There is a lot of guidance on the meta-analysis method [13,14]. It is easy to understand and communicate. Forest plots can be used to easily visualise the estimations (trends) from each country and show how the average is calculated across them, while multilevel models adopt a more complex approach. There are no common error terms or common effects of potential confounders, so estimates for each country can be more independent from each other. There are also often fewer convergence problems with meta-analysis than with multilevel models.

Limitations

When there is sparse data among individual studies, the assumptions of the random-effects meta-analysis may not be met. With a meta-analysis, country-level effects are not estimated directly (although this is not of interest for most analyses in this guidance document). However, it is possible to do this with meta-regression.

Method

A meta-analysis gives a pooled estimate, but this involves taking country-level trends into account. In the context of trend analysis, the trend can be calculated for each country, as well as its standard error. One can then obtain a summary estimate as a weighted average of the country-specific trends. This averaging is a weighted mean, and the weight depends on the assumptions of the distribution of the true trends.

Most of the time, meta-analysis is carried out using one of two models: a common-effect model or a random-effects model. A common-effect model assumes that there is one true trend common to all countries, and that the country-specific trends deviate from this common trend only due to sampling error. The weight for each country in the summary mean is the inverse of the standard error of their trend. Country-specific trends with lower standard errors have higher weights, while country-specific trends with higher standard errors have lower weights. In the context of surveillance data, there is so much methodological and clinical heterogeneity between countries that this common-effect assumption seems very unlikely to be true.

In the literature, the common-effect meta-analysis is often called a fixed-effect meta-analysis. Both terms can be used interchangeably, but as per Borenstein [15], the name 'common-effect' may be more suitable, as the term 'fixed effects' is often used in a different context in statistics.

A random-effects model assumes that the true effect varies from country to country. However, there are some assumptions of similarity, otherwise a stratified analysis would be performed. The assumption is that the true country-specific trends are normally distributed around a global mean with some between-country variance. The separation of the observed country-specific trends are the result of the between-country variance plus the within-country variance (i.e. the standard error of the study-specific trend that reflects the sampling random error). There are different methods to calculate the between-country variability (a popular one is the DerSimonian-Laird method [16]). Study weights are then calculated in such a way that within-study and between-study variability are minimised. With this type of weight calculation, studies with high standard errors might be given larger weights than they would have received in the common-effect meta-analysis.

In the context of pooling EU/EEA data, there is a high number of countries in the meta-analysis. However, for a low number of countries, it is difficult to measure the between-study variability adequately. Caution should be taken if there is a low number of countries underlying a random-effects meta-analysis. An option to overcome this limitation is to use a Bayesian approach, ideally with the help of a statistician.

Theoretically, in the context of pooling surveillance data from EU/EEA countries, it's likely that there are a variety of true trends rather than a single trend (meaning that a random-effects approach is the most appropriate method). However, if a random-effects meta-analysis is performed and the common-effect assumption is actually true, the same results as a common-effect meta-analysis will be obtained.

To carry out a meta-analysis (common-effect or random-effects), calculate the trend and the standard error for each country. These estimates can then be used with meta-analysis commands/functions in standard software (e.g. Stata or R). The summary estimate and its standard error and confidence intervals will be calculated, as well as measures of heterogeneity.

Producing a forest plot can also be beneficial. This is a plot in which the measures for each study are displayed along with their weights, as well as the summary measure. A forest plot of the data from the example of the number of cases of disease X in countries A to D in Figure 44 and Table 14 is shown in Figure 45 (for a common-effect analysis) and in Figure 46 (for a random-effects analysis).

Figure 45. Adjusted pooled analysis (common-effect meta-analysis) and forest plot of trend of disease X, countries A–D, 2005–2016

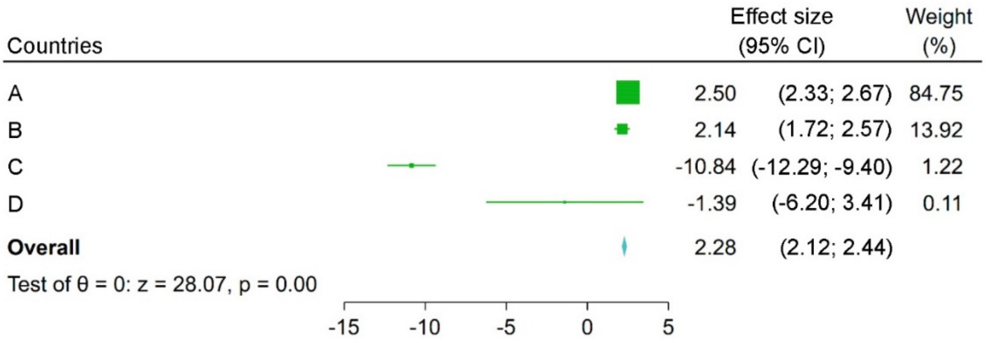
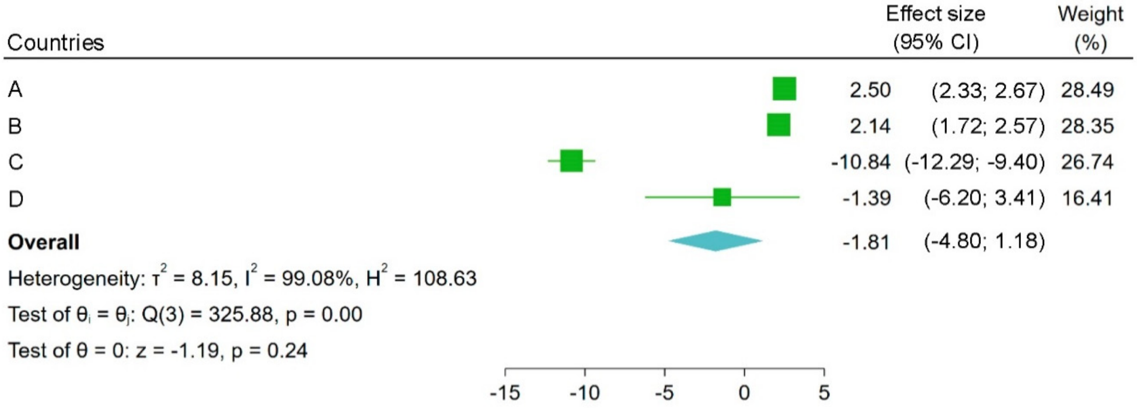


Figure 46. Adjusted pooled analysis (random-effects meta-analysis) and forest plot of trend of disease X, countries A–D, 2005–2016



Interpretation

In Figure 45, the results of a common-effect meta-analysis are displayed for the linear trend of reported cases of disease X in countries A to D. The forest plot visualises the coefficients and their confidence intervals, and the size of the estimate in the graph corresponds to their weight in the overall estimate. The summary estimate of trend in the common-effect meta-analysis is 2.28 (95% CI: 2.12–2.44). As mentioned before, there is considerable heterogeneity, with the I^2 at 99% and Cochrane's Q statistic's $p - value \ < 0.001$. The results of the common-effect meta-analysis are shown for comparison, as—theoretically—it is not the model that would have been chosen at the outset. The random-effects meta-analysis is the preferred method, the results of which are shown in Figure 46.

In the random-effects meta-analysis, using the DerSimonian-Laird approach, each country has the same trend and confidence interval as before, but the weights are very different. The weights are more equal across countries, compared to the common-effect meta-analysis. The summary estimate of the trend is -1.81 cases (95% CI: -4.8 to 1.18) per year. The measures of heterogeneity are the same as before. The summary estimate differs by 4.09 cases (2.28 compared with -1.81) per year, and the two trends are in different directions. In the common-effect meta-analysis, there is a slight upward trend that is statistically significant, as the confidence interval does not include the zero. On the other hand, in the random-effects meta-analysis, there is a slightly downward trend. This downward trend is not statistically significant, as the confidence interval clearly includes zero. Common-effect and random-effects meta-analyses can give different results and the inferences made from them can be different. Therefore, it is important to be confident in the assumptions underlying the meta-analyses.

The random-effects summary estimate of -1.81 means that, on average, there is a decrease in reported cases of disease X of 1.8 cases per year. This decrease is not statistically significant.

There is an important difference when interpreting the adjusted pooled analysis (random-effects meta-analysis approach) summary estimate compared to the crude pooled analysis estimate when using absolute numbers. In the crude pooled analysis estimate, the trend pertains to the total trend by year for all countries together. The random-effects summary estimate for trend pertains to the trend for an average country in that group. Using a standardised measure, such as a rate, would render these types of estimates more comparable.

The pooled approach using different outcome measures (e.g. rates) or different regression types (e.g. Poisson regression) is similar to the example above, and is not covered in this theoretical discussion. Rates are covered in the examples (see Part C, Section 4, 'Examples').

Strengths

The important strength of the adjusted pooled approach with random-effects meta-analysis is that individual country trends are taken into account, which means the differences between countries are accounted for. This could be important in the context of surveillance data across countries, as there is likely to be a variety of different trends. The analysis is also easy to carry out and convey to a reader.

Limitations

As with any model, assumptions are made. The assumptions in a random-effects model are that the estimates obtained from the countries are part of a wider distribution of countries. This means that it allows for similarities between estimates of countries. If heterogeneity is very high, then these similarities may not be there.

In Figure 46, a random-effects meta-analysis is carried out using only four countries. While meta-analysis with few units is commonly seen in publications, a low number of countries underlying the between-study variability may result in erroneous outputs and inferences.

3. Algorithm for carrying out a trend analysis taking country into account

Step 1: Define the study objectives

When deciding whether to take individual country trends into account, the objectives of the study or analysis are key. Is there interest in determining the average change in the number of cases of a given disease in the EU/EEA as a whole in recent years? If so, then it may not be necessary to take country data into account. Is the average rate of a given disease across the EU/EEA—taking differences in rates between countries into account—of interest? If so, then an adjusted pooled analysis with random-effects meta-analysis approach may be the best method.

Step 2: Understand the data

Understanding the data is always a key point in an analysis, but this is particularly important when trying to summarise several data sources. Reading the ECDC pooling guidance document [5] is recommended, as it discusses the summarisation of data in terms of:

- **heterogeneity in data collection and data collection systems:**
  - system design
  - data sources
  - populations underlying the surveillance systems
  - case definitions
- **missing data**
- **heterogeneity related to disease determinants:**
  - time period (e.g. seasonal and inter-seasonal data collection)
  - environment
  - disease interventions
  - population characteristics.

Obtaining this information will greatly increase the understanding of the data and will inform how pooling will help or hinder the study objectives. If possible, this information should be obtained from a disease expert.

Step 3: Plot the data

Plot the raw data by country

Plotting the raw data by country in a line graph can provide insights into the underlying trends by country. If there are a lot of countries within the data, creating several plots of groups of countries may provide a better overview.

Questions to consider: Does it look like the data has the same shape of trend (e.g. linear trend, quadratic trend, exponential trend) across countries? Does it look like the data has the same direction of trend (upward or downward)? Does it look like the data has the same magnitude of trend (steep slopes, slight inclines, or declines)?

Plot the raw data by country and plot the trend models by country on top

Start with a linear trend. In separate plots, display the raw data and other trends that look like they fit some of the data (e.g. quadratic or exponential trends).

Questions to consider: Is there a type of trend model that best fits the data? Are there countries where this trend does not fit? How many countries does this concern?

Step 4: Decide how you will pool the data in the main analysis

Based on the study objectives, the information on heterogeneity of the underlying systems and disease determinants, and the plots, decide whether the data are too heterogeneous to pool. An informed qualitative assessment may be more useful than basing this decision on statistical measures. Statistics should be part of the decision, not the only tool used for it.

If some or all of the data are too heterogeneous and the information on individual country trends may be more valuable for the reader, a stratified analysis or pooled analysis by subgroup can be carried out.

If the study objective is to obtain a value for the EU/EEA as a whole, and the underlying country differences are not important, a crude pooled analysis can be carried out, regardless of the underlying heterogeneities by country.

If the study objective is to obtain a measure of the trend in the EU/EEA that takes individual country data into account and provides a measure that is useful for comparison between countries, a pooled analysis can be considered if the underlying heterogeneity is within acceptable limits.

Pooled analysis despite large underlying heterogeneities

With surveillance data, there are often substantial underlying heterogeneities. However, study objectives may indicate that a pooled analysis is necessary. In this case, a pooled analysis can still be carried out, but it is important to document the suspected heterogeneity and the limitations of the pooling. This way the reader can interpret the pooled results in context.

Step 5: Choose the time unit of analysis, consider inclusion of seasonal terms, check for non-linearity and decide on the preliminary regression type

Choosing the time unit of analysis, considering inclusion of seasonal terms, checking for non-linearity and deciding regarding the preliminary regression type can be carried out in the same way as the algorithm in Part B, Section 6, 'Algorithm for performing the trend analysis'. This can either be overall (if a crude pooled analysis is carried out) or country-by-country (if a stratified analysis is carried out).

If a pooled analysis is chosen, then a consensus type of trend (linear, quadratic, exponential, etc.) and a consensus regression type must be selected, based on the data from all countries. As in Figure 44, there may be one or two countries for which a different trend fits better than the consensus trend (e.g. they have a more quadratic trend, compared to a linear consensus trend). Depending on the study objectives, this better fit can be ignored and a random-effects meta-analysis can still be carried out with the consensus trend. However, if there are many countries with different types of trends or different directions of trends, one might need to reconsider whether the underlying heterogeneities are too large for a pooled analysis.

Step 6: Build and validate the model

Model building and validation can be conducted following the same instructions as the algorithm in Part B, Section 6. This can be either overall (if a crude pooled analysis is being carried out) or country-by-country (if a stratified analysis is being carried out).

Step 7: Interpret the model

Now that the model is final, the regression outputs can be used to interpret it. What type of trend does it have? What is its magnitude? Is it a statistically significant trend at the 5% level? Is the trend easy to communicate?

It is important to provide information on underlying heterogeneities so the reader can better interpret the pooled model. It may make sense to present both stratified and pooled results, so the reader has the most information possible.

4. Examples

4.1 Example: HIV notification rates in the EU/EEA, taking country into account, 2007–2016

This example uses the algorithm in the previous section (Part C, Section 3, 'Algorithm for carrying out the trend analysis taking country into account') to demonstrate the process of obtaining a trend in HIV notification rates in the EU/EEA, taking individual country trends into account. The dataset is freely available on ECDC's 'Surveillance Atlas of Infectious Diseases' [4].

Step 1: Define the study objectives

When deciding whether to take country data into account, the study or analysis objectives are key. In this example, the change in HIV notification rates across the EU/EEA are of interest, taking country trends into account. This means that the clustering of cases by country should be considered, rather than assuming that the data arises from a single system within the EU/EEA as one large entity. Rates are of interest, as they are standardised to population size and may give a more meaningful result.

Step 2: Understand the data

Together with a disease expert and using ECDC's pooling guidelines, the data are summarised in terms of the different potential types of underlying heterogeneities in data collection systems and types of heterogeneity related to disease determinants.

It is concluded that HIV notification systems are longstanding and well-established in all the countries participating in the surveillance. While there are some differences in data collection methods and disease determinants, they are sufficiently homogenous to warrant pooling.

Step 3: Plot the data

The raw data are plotted by country in two graphs (Figures 47 and 48) for a better overview.

Figure 47. HIV notification rates by country in the EU/EEA, 2007–2016

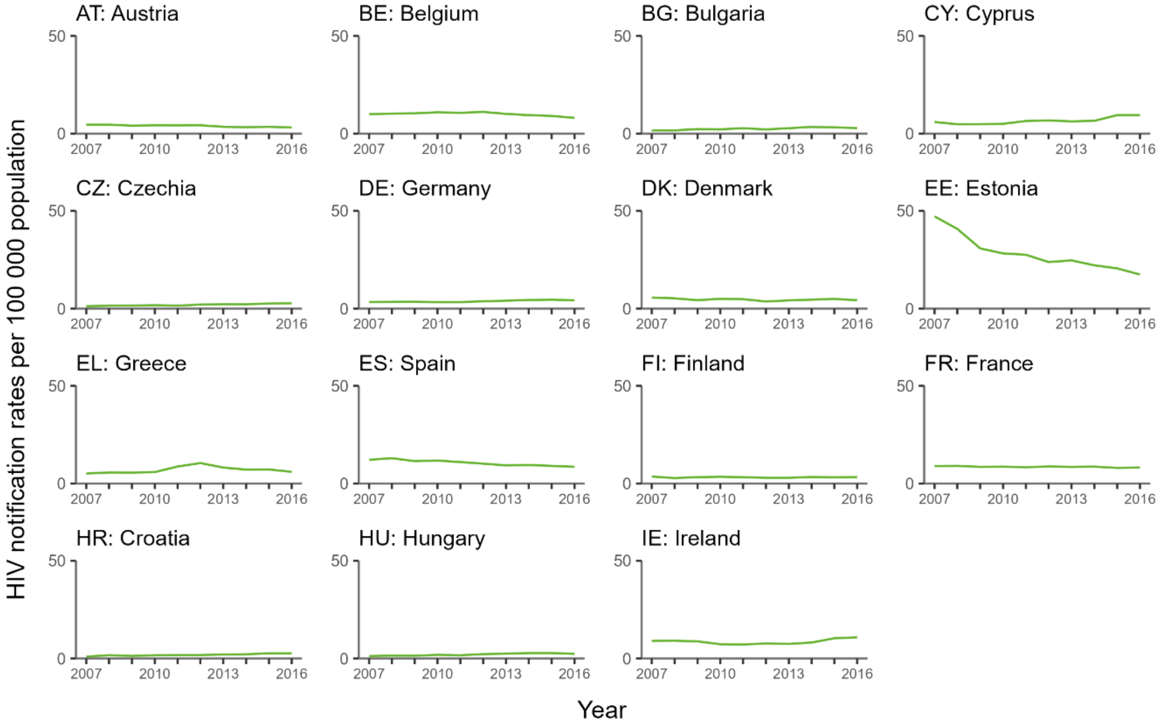
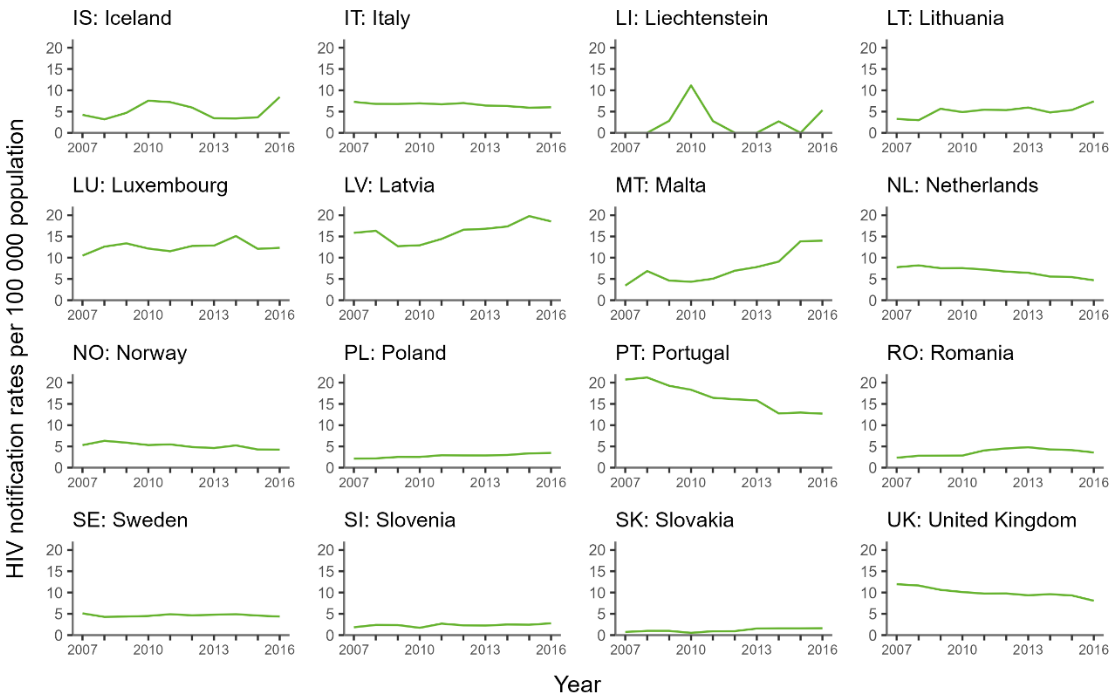


Figure 48. HIV notification rates by country in the EU/EEA, 2007-2016



As seen in Figures 47 and 48, there is a mix of slight upward and downward trends. Overall, it looks like the consensus trend in the data is a linear trend. Exceptions are some countries where the data follow curves that mainly have lower rates: Greece, Iceland, Liechtenstein and Malta. A few countries have a strong upward or downward trend (e.g. Estonia and Portugal).

The raw data and the trend models are then plotted by country (Figures 49 and 50).

Figure 49. HIV notification rates by country and linear trend in the EU/EEA, 2007–2016

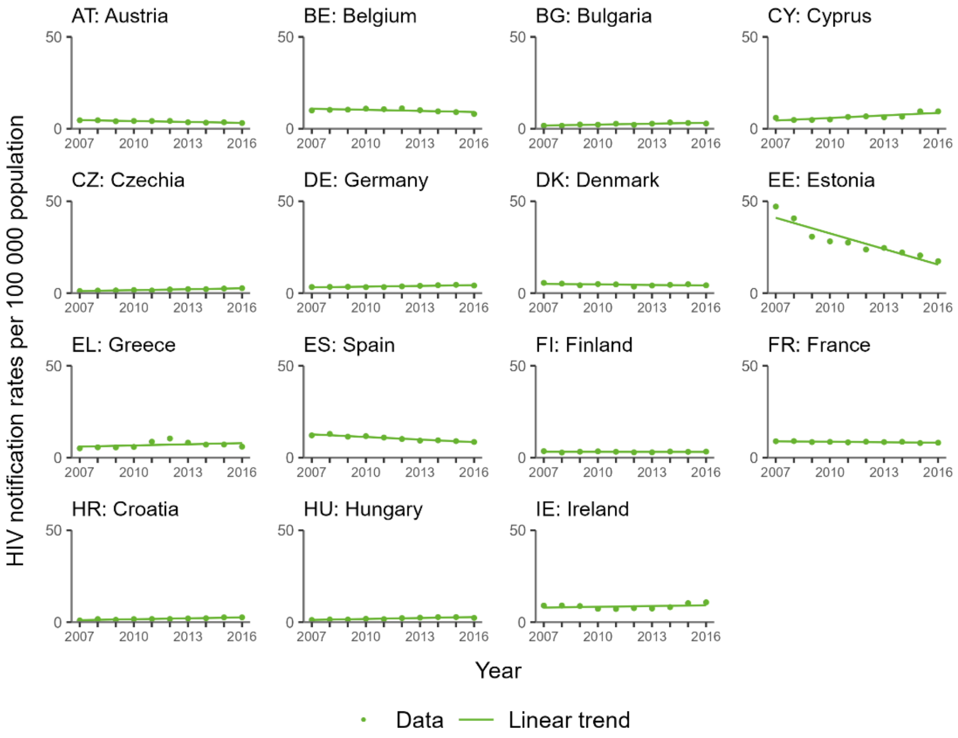
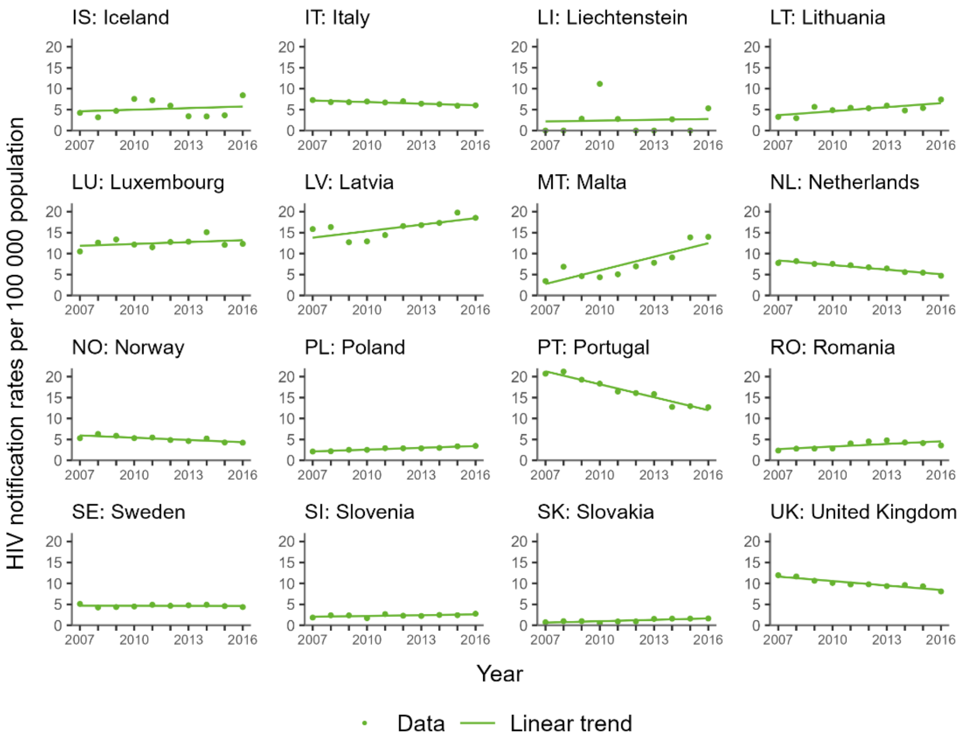


Figure 50. HIV notification rates by country and linear trend in the EU/EEA, 2007–2016



As seen in Figures 49 and 50, it seems that the linear trend fits well for many countries. There are only a few countries where the model does not fit the data well (Iceland, Liechtenstein, Malta), and a few countries where the trend fits most data points but there are some outlying data points (Estonia, Greece).

It may also be beneficial to model the HIV notification rates per 100 000 population with a quadratic trend modelled over the data (Figures 51 and 52).

Figure 51. HIV notification rates by country and quadratic trend in the EU/EEA, 2007–2016

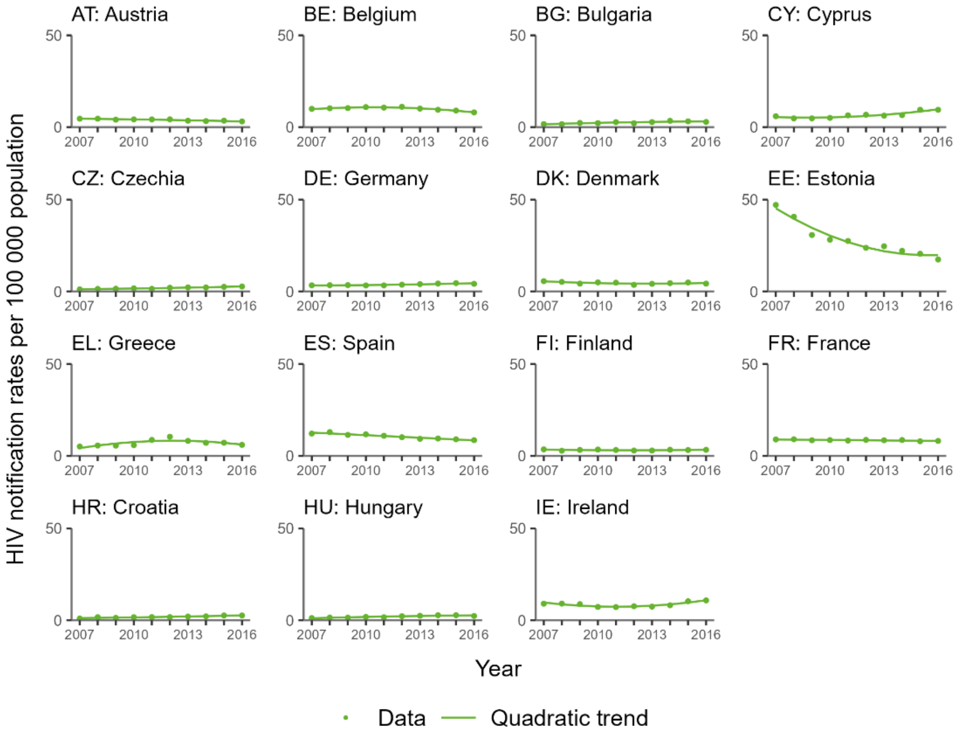
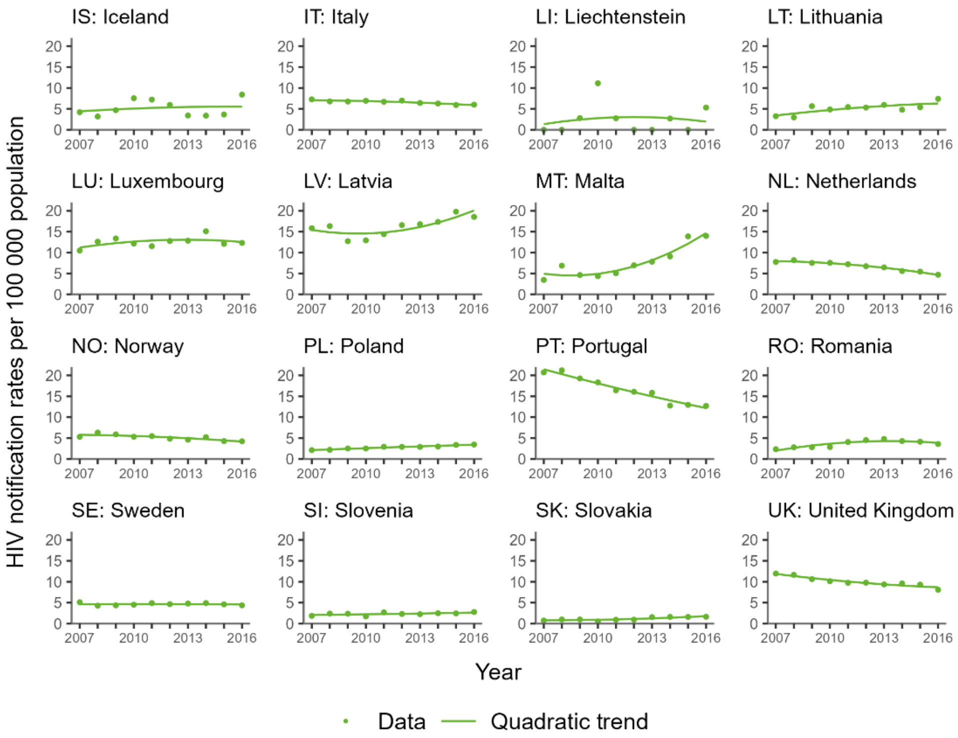


Figure 52. HIV notification rates by country and quadratic trend in the EU/EEA, 2007–2016



As seen in Figures 51 and 52, the quadratic trend fits well for many countries. However, consider that a model with a polynomial of a higher order will always fit the data better, and decide whether this better data fit is worth the extra complexity. The quadratic trend model does not improve the data fit substantially in countries where the linear trend was identified as not fitting well (Iceland, Liechtenstein, Malta).

Step 4: Decide how to pool the data in the main analysis

As mentioned in Step 2, together with the disease expert it was decided that the heterogeneity between systems was acceptable. As the study objective is to obtain a trend estimate that takes individual country trends into account, it may be necessary to proceed even if there is a high level of heterogeneity. It is possible to comment on the heterogeneity when presenting the model.

Therefore, it is decided that the data will be pooled using a random-effects meta-analysis to take individual country trends into account. Given the level of heterogeneity, however, the individual country results will also be presented alongside the pooled estimate.

Step 5: Choose the time unit of analysis, consider inclusion of seasonal terms, check for non-linearity and decide on preliminary regression type

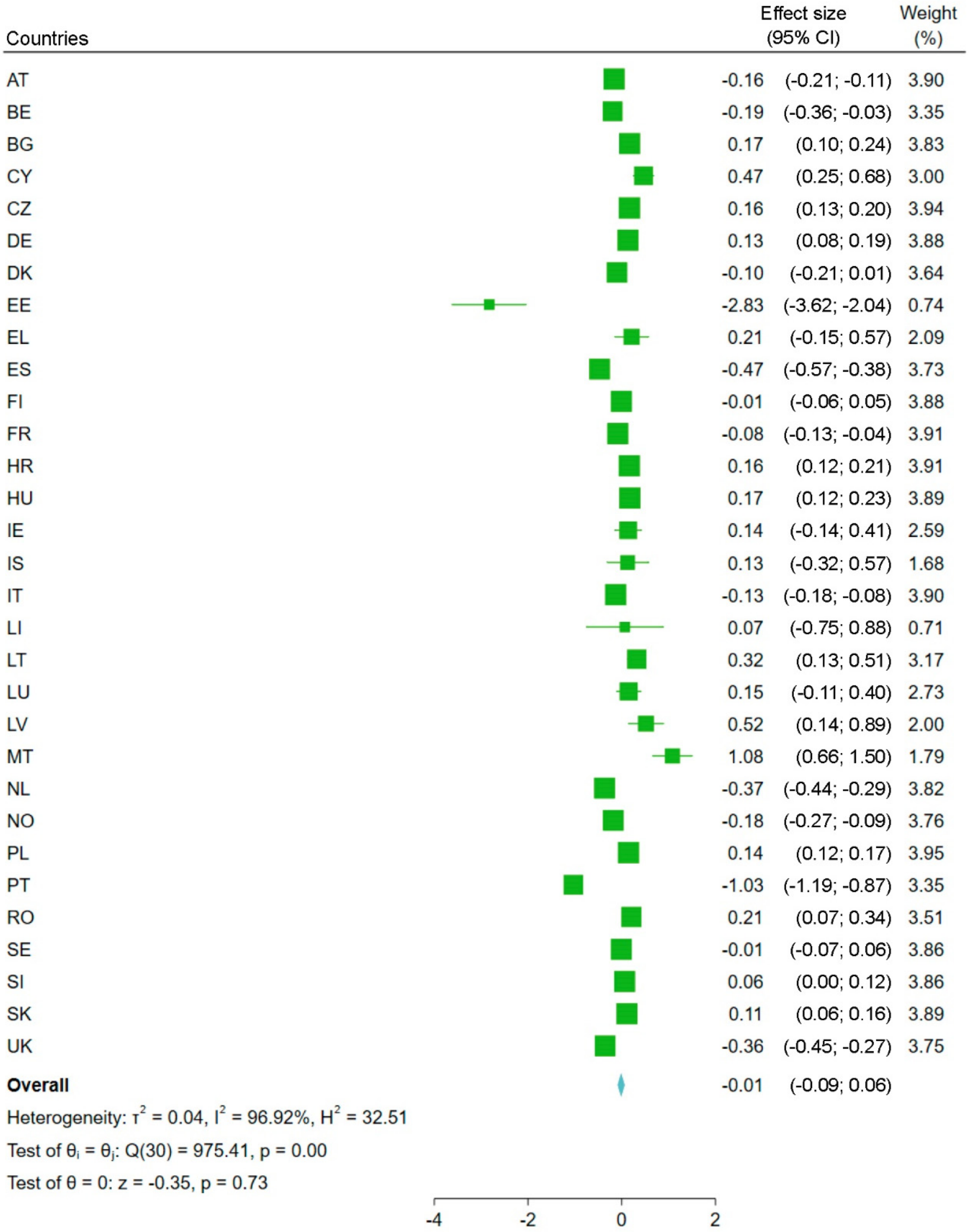
Step 3 already included preliminary thoughts on the time units used for analysis, preliminary regression types, inclusion of seasonal terms and checks for non-linearity when the models were plotted over the data. For this HIV notification data, yearly rates are of interest and therefore seasonal terms will not be included.

As a meta-analysis will be carried out, the same model must be chosen for each country. Rates are of interest and they follow a reasonably linear pattern. They appear to be high enough to use a linear regression, so yearly rates are modelled using linear regression. This confirms what was considered in step 3, when plotting the linear model against the data.

Step 6: Build and validate the model

The trend and standard error for each country are then captured and a meta-analysis is carried out using these. A random-effects meta-analysis is used with the DerSimonian-Laird method. A forest plot is also created (Figure 53). This provides information on weight per country and makes it easy to spot outliers.

Figure 53. Adjusted pooled analysis (random-effects meta-analysis) and forest plot of trend of HIV notification rates in the EU/EEA, 2007–2016



Countries	Effect size (95% CI)		Weight (%)
AT	-0.16	(-0.21; -0.11)	3.90
BE	-0.19	(-0.36; -0.03)	3.35
BG	0.17	(0.10; 0.24)	3.83
CY	0.47	(0.25; 0.68)	3.00
CZ	0.16	(0.13; 0.20)	3.94
DE	0.13	(0.08; 0.19)	3.88
DK	-0.10	(-0.21; 0.01)	3.64
EE	-2.83	(-3.62; -2.04)	0.74
EL	0.21	(-0.15; 0.57)	2.09
ES	-0.47	(-0.57; -0.38)	3.73
FI	-0.01	(-0.06; 0.05)	3.88
FR	-0.08	(-0.13; -0.04)	3.91
HR	0.16	(0.12; 0.21)	3.91
HU	0.17	(0.12; 0.23)	3.89
IE	0.14	(-0.14; 0.41)	2.59
IS	0.13	(-0.32; 0.57)	1.68
IT	-0.13	(-0.18; -0.08)	3.90
LI	0.07	(-0.75; 0.88)	0.71
LT	0.32	(0.13; 0.51)	3.17
LU	0.15	(-0.11; 0.40)	2.73
LV	0.52	(0.14; 0.89)	2.00
MT	1.08	(0.66; 1.50)	1.79
NL	-0.37	(-0.44; -0.29)	3.82
NO	-0.18	(-0.27; -0.09)	3.76
PL	0.14	(0.12; 0.17)	3.95
PT	-1.03	(-1.19; -0.87)	3.35
RO	0.21	(0.07; 0.34)	3.51
SE	-0.01	(-0.07; 0.06)	3.86
SI	0.06	(0.00; 0.12)	3.86
SK	0.11	(0.06; 0.16)	3.89
UK	-0.36	(-0.45; -0.27)	3.75
Overall	-0.01	(-0.09; 0.06)	

Heterogeneity: $\tau^2 = 0.04$, $I^2 = 96.92\%$, $H^2 = 32.51$

Test of $\theta_i = \theta_j$: Q(30) = 975.41, p = 0.00

Test of $\theta = 0$: z = -0.35, p = 0.73

AT: Austria; BE: Belgium; BG: Bulgaria; CI: confidence interval; CY: Cyprus; CZ: Czechia; DE: Germany; DK: Denmark; EE: Estonia; EL: Greece; ES: Spain; FI: Finland; FR: France; HR: Croatia; HU: Hungary; IE: Ireland; IS: Iceland; IT: Italy; LI: Liechtenstein; LT: Lithuania; LU: Luxembourg; LV: Latvia; MT: Malta; NL: the Netherlands; NO: Norway; PL: Poland; PT: Portugal; RO: Romania; SE: Sweden; SI: Slovenia; SK: Slovakia; UK: the United Kingdom.
The UK was a Member State of the European Union (EU) at the time of collating the data for this report. The UK withdrew from the EU on 31 January 2020.

Step 7: Interpret the model

The pooled estimate is -0.01 (95% CI: -0.09 to 0.06). This means that, on average, the rate of HIV notifications in the EU/EEA (taking individual country trends into account) from 2007 to 2016 decreased by 0.01 cases per 100 000 population a year. However, since it is not statistically significant, the hypothesis that the overall trend was stable cannot be excluded.

4.2. Example: Hepatitis B notification rates in the EU/EEA, taking country into account, 2008–2017

This example uses the steps in the algorithm in Section C.3 to demonstrate the process of obtaining a trend in hepatitis B notification rates in the EU/EEA that takes individual country trends into account. The dataset is freely available on the ECDC 'Surveillance Atlas of Infectious Diseases' [4]. Note that, for the purposes of this example, the countries Belgium and Croatia have not been included, as they do not report data every year.

Step 1: Define the study objectives

When deciding whether to take individual country trends into account, the objectives of the study or analysis are key. In this example, the change in hepatitis notification rates across the EU/EEA is of interest, taking country-level data into account. This means that the clustering of cases by country must be considered, rather than assuming that the data arises from a single system within the EU/EEA as one large entity. Rates are of interest, as they are standardised to population size and may give a more meaningful result.

Step 2: Understand the data

Together with a disease expert and using ECDC's pooling guidelines, the data are summarised in terms of the different potential types of underlying heterogeneities in data collection systems and types of heterogeneity related to disease determinants.

The conclusion is that hepatitis notification systems are quite diverse in the countries participating in the surveillance, particularly when hepatitis overall is of interest, rather than by acute or chronic phase. In addition to these underlying data collection heterogeneities, there are also considerable differences in disease determinants by country. It is decided that the data are not sufficiently homogenous to warrant pooling and that more information might be obtained if the trends are presented by country separately.

Step 3: Plot the data

Plot the raw data by country

The data are plotted in two graphs for a better overview (Figures 54 and 55).

Figure 54. Hepatitis B notification rates by country in the EU/EEA, 2008–2017

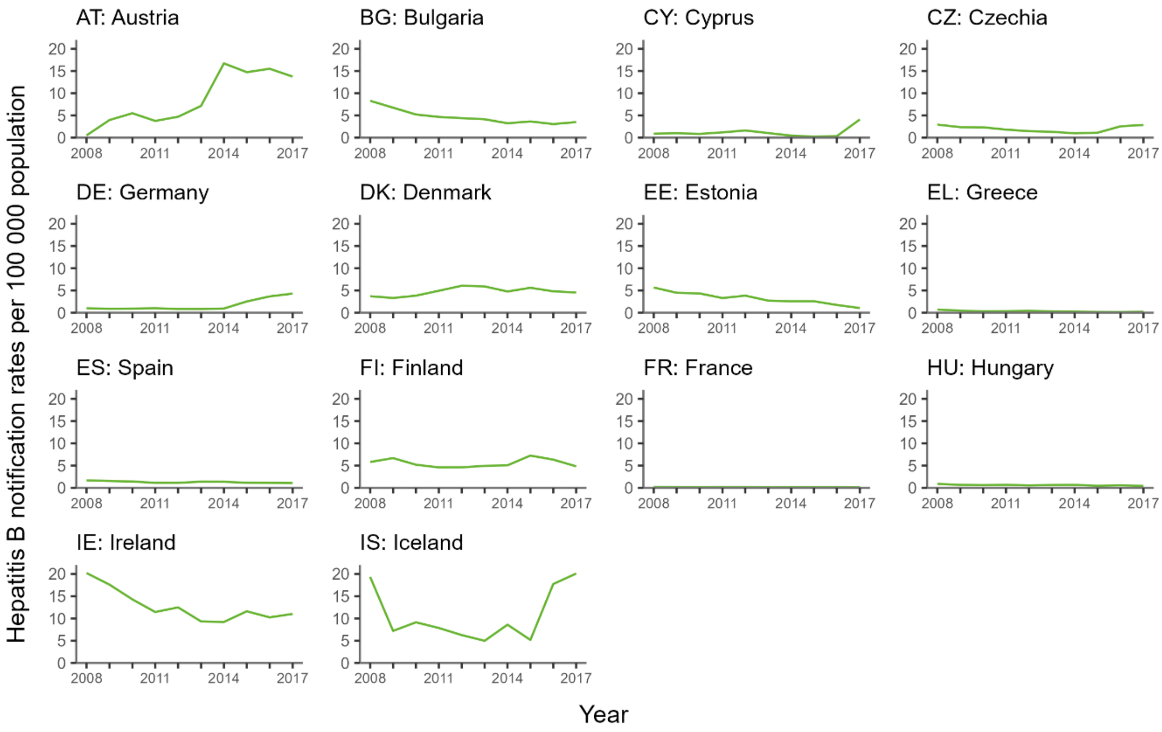
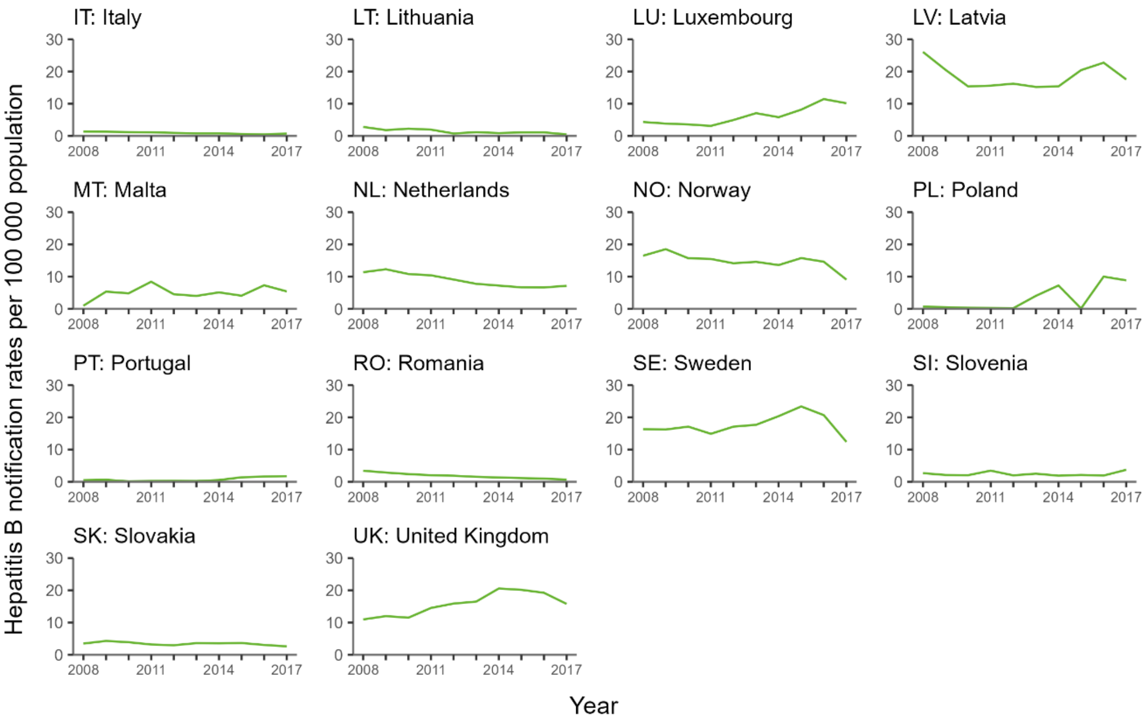


Figure 55. **Hepatitis B notification rates by country in the EU/EEA, 2008–2017**



As seen in Figures 54 and 55, there are some distinctly different trends by country: linear trends, quadratic trends, upward trends, downward trends and no change over time.

Plot the raw data by country and plot the trend models by country
First, a linear trend is plotted over the hepatitis B notification rates per 100 000 population (Figures 56 and 57).

Figure 56. **Hepatitis B notification rates by country and linear trend in the EU/EEA, 2008–2017**

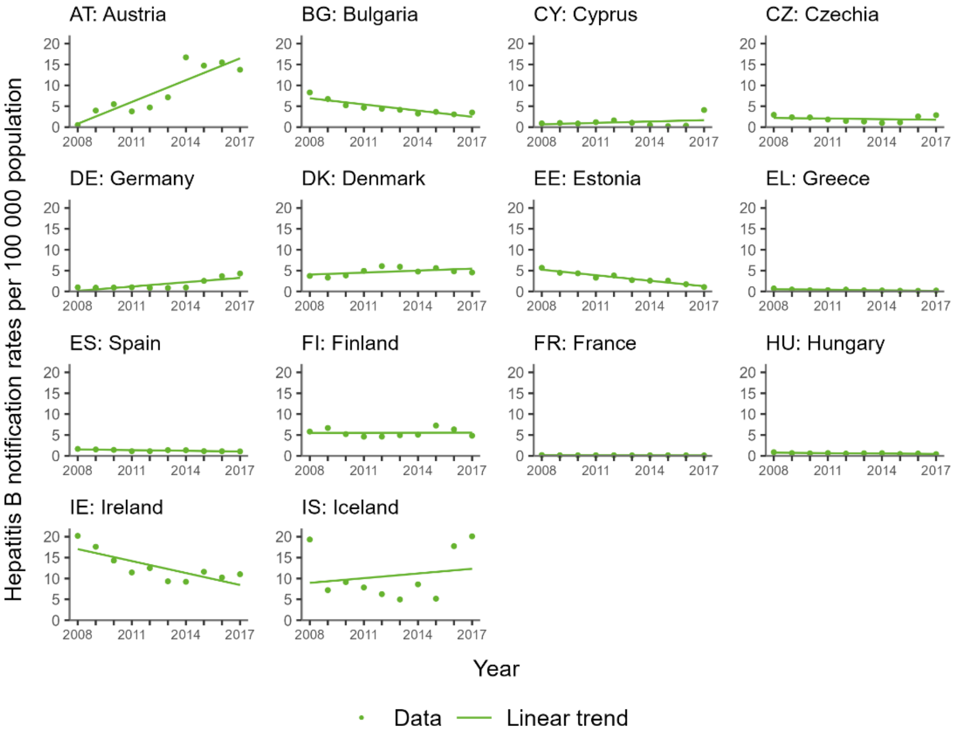
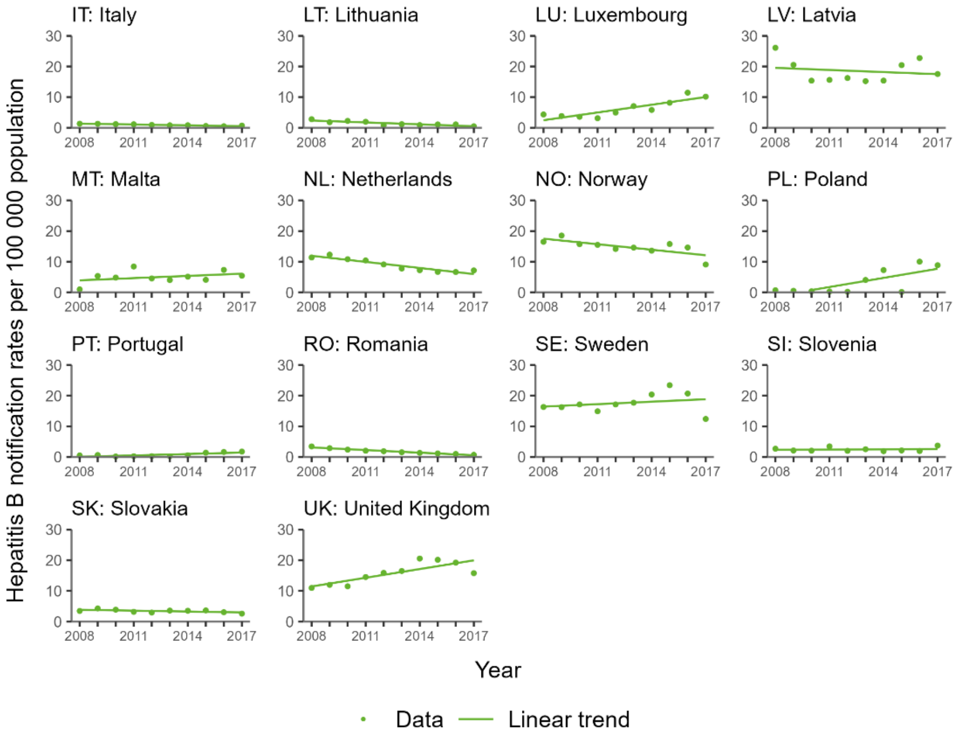


Figure 57. **Hepatitis B notification rates by country and linear trend in the EU/EEA, 2008–2017**



As seen in Figures 56 and 57, the linear trend fits well for many countries, but not all. Notably, it does not fit Austria, Iceland, Ireland, Latvia, Poland or Sweden.

Next, a quadratic trend is plotted over the hepatitis B notification rates per 100 000 population (Figures 58 and 59).

Figure 58. **Hepatitis B notification rates by country and quadratic trend in the EU/EEA, 2008–2017**

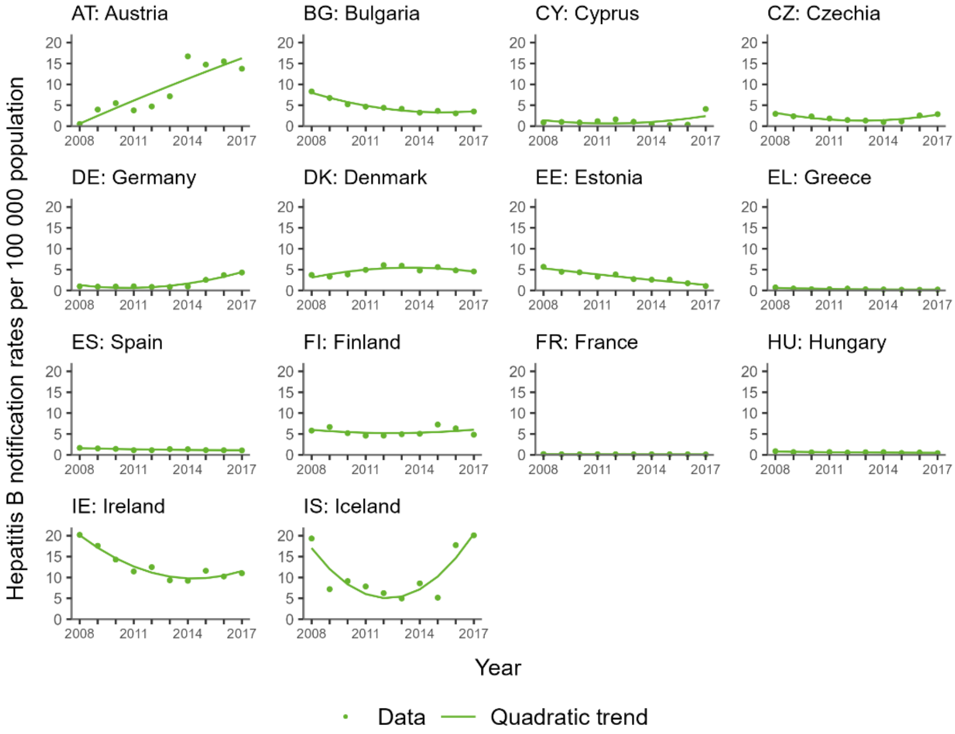
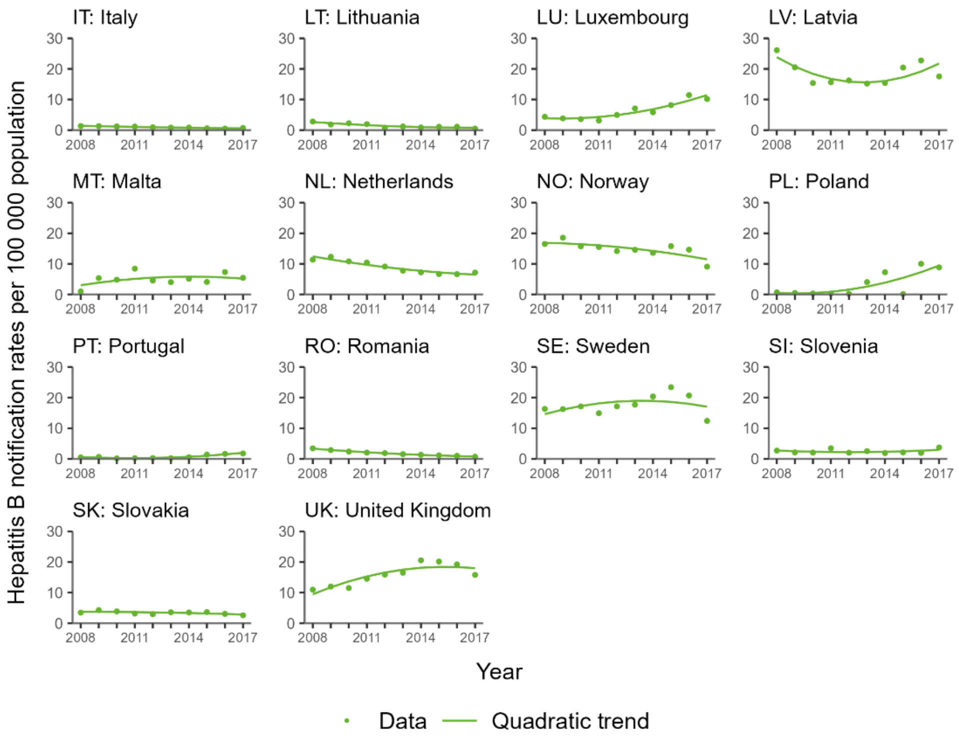


Figure 59. **Hepatitis B notification rates by country and quadratic trend in the EU/EEA, 2008–2017**



As seen in Figures 58 and 59, the quadratic trend fits well for many countries. However, a model with a polynomial of a higher order will always fit the data better. Therefore, a decision needs to be made as to whether this better data fit is worth the extra complexity.

Step 4: Decide how to pool the data in the main analysis

In Step 2, it was decided together with the disease expert that the heterogeneity between systems was quite large and a pooled analysis was not necessarily warranted or useful. A random-effects meta-analysis is used to calculate the Q-statistic, the p-value and the I^2. These indicate a very high heterogeneity by country. For the study's purposes, it may also be useful to describe the trend by country.

In this situation, the decision is made not to pool the data and to instead carry out a stratified analysis.

Step 5: Choose the time unit of analysis, consider inclusion of seasonal terms, check for non-linearity and decide on preliminary regression type

Step 3 already included a discussion on choosing the time unit of analysis, considering the inclusion of seasonal terms, checking for non-linearity when plotting models over the data and deciding on the preliminary regression type. For this hepatitis notification data, yearly rates are of interest and, therefore, seasonal terms will not be included.

A decision was made with the disease expert to graph the data by country in the report, so the readers can see the pattern of trend in the data. After a discussion, it was also decided that instead of modelling the best trend for each country (e.g. a quadratic trend for Ireland, a linear trend for France), a linear trend would be used with linear regression for each country. By doing this, sacrifices are being made in terms of model validity (a linear trend does not fit the Irish data very well, for example), but an overall summary of trend may be useful to the report's readers.

Step 6: Build and validate the model

Step 3 already included a linear regression for linear trend by country. As mentioned in Step 5, it is accepted that there will be issues in model validity.

Step 7: Interpret the model

Along with the graphs of the actual data by country, the rates are reported by year, along with the linear trends and their 95% CI (Table 15).

These results reflect the rates and linear trends for hepatitis B overall, by country, from 2008 to 2017 for EU/EEA countries. The rates have different magnitudes and directions. As a next step, an analysis by acute hepatitis and chronic hepatitis could be carried out, in which case a pooled analysis might be useful and feasible.

Table 15. Hepatitis B notification rates by country and linear trend in the EU/EEA, 2008–2017

Country	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	Linear trend, 2008–2017 (95% CI)
Austria	0.5	4.0	5.5	3.8	4.7	7.1	16.7	14.7	15.5	13.7	1.7 (1.1 to 2.4)
Bulgaria	8.3	6.7	5.2	4.7	4.4	4.1	3.2	3.7	3.1	3.5	-0.5 (-0.7 to -0.3)
Cyprus	0.9	1.0	0.9	1.2	1.6	1.0	0.5	0.2	0.4	4.1	0.1 (-0.1 to 0.4)
Czechia	2.9	2.4	2.3	1.8	1.5	1.3	1.0	1.1	2.6	2.9	0.0 (-0.2 to 0.1)
Denmark	3.7	3.3	3.8	4.9	6.1	5.9	4.8	5.6	4.8	4.6	0.2 (0.0 to 0.3)
Estonia	5.7	4.5	4.4	3.3	3.8	2.7	2.6	2.6	1.7	1.1	-0.4 (-0.5 to -0.4)
Finland	5.8	6.7	5.2	4.6	8.3	6.7	5.2	7.3	6.3	4.8	0.0 (-0.2 to 0.2)
France	0.2	0.1	0.1	0.2	0.9	1.0	0.9	0.1	0.1	0.1	0.0 (0.0 to 0.0)
Germany	1.0	0.9	0.9	1.0	0.9	0.9	0.9	2.5	3.7	4.3	0.3 (0.2 to 0.5)
Greece	0.7	0.5	0.3	0.3	0.5	0.3	0.2	0.2	0.2	0.2	0.0 (-0.1 to 0.0)
Hungary	0.9	0.7	0.6	0.7	2.9	2.4	2.3	0.5	0.6	0.4	0.0 (-0.1 to 0.0)
Iceland	19.3	7.2	9.1	7.9	3.7	3.3	3.8	5.2	17.7	20.1	0.4 (-1.0 to 1.7)
Ireland	20.2	17.6	14.3	11.4	1.0	0.9	0.9	11.6	10.2	11.0	-1.0 (-1.5 to -0.5)
Italy	1.3	1.3	1.2	1.1	0.9	0.8	0.8	0.6	0.5	0.7	-0.1 (-0.1 to -0.1)
Latvia	26.1	20.5	15.4	15.6	16.2	15.2	15.4	20.4	22.8	17.5	-0.2 (-1.1 to 0.6)
Lithuania	2.8	1.8	2.3	2.0	0.8	1.2	0.9	1.1	1.1	0.5	-0.2 (-0.3 to -0.1)
Luxembourg	4.3	3.9	3.6	3.1	5.0	7.1	5.8	8.2	11.5	10.2	0.8 (0.5 to 1.2)
Malta	1.0	5.4	4.8	8.4	4.6	4.0	5.1	4.1	7.3	5.4	0.2 (-0.2 to 0.7)
Netherlands	11.4	12.3	10.8	10.4	9.1	7.8	7.2	6.7	6.6	7.2	-0.7 (-0.8 to -0.5)
Norway	16.5	18.5	15.7	15.5	14.2	14.6	13.6	15.8	14.6	9.1	-0.6 (-1.0 to -0.2)
Poland	0.7	0.5	0.3	0.3	0.2	4.0	7.3	0.1	10.0	8.9	1.0 (0.4 to 1.6)
Portugal	0.5	0.6	0.2	0.3	0.3	0.2	0.5	1.4	1.6	1.8	0.2 (0.1 to 0.2)
Romania	3.4	2.9	2.4	2.0	1.9	1.5	1.3	1.2	1.0	0.7	-0.3 (-0.3 to -0.2)
Slovakia	3.5	4.3	3.9	3.2	2.9	3.6	3.5	3.6	3.0	2.6	-0.1 (-0.2 to 0.0)
Slovenia	2.7	2.1	2.1	3.5	2.0	2.5	1.9	2.1	1.9	3.7	0.0 (-0.1 to 0.2)
Spain	1.7	1.5	1.4	1.1	0.5	4.0	5.5	1.1	1.1	1.1	-0.1 (-0.1 to 0.0)
Sweden	16.3	16.2	17.1	14.9	17.1	17.7	20.4	23.4	20.7	12.4	0.3 (-0.4 to 1.0)
United Kingdom	11.0	12.0	11.5	14.5	15.9	16.5	20.6	20.2	19.2	15.8	0.9 (0.5 to 1.4)

CI: confidence interval.

Part D. Trend analysis with missing data

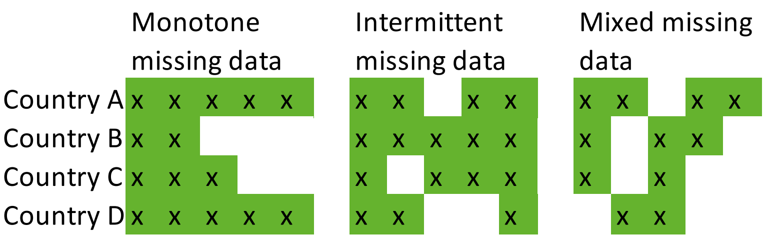
1. Theory

In the context of trend analysis with surveillance data, only two variables are often used: the count variable (number of reported cases by health topic) and the time unit. When working with rates, there is also information on population. In general, there are no missing data on population or on time unit. However, as surveillance data are not always perfect, there can often be missing data for the count variable for a given time unit.

1.1 Patterns of missing data

In the context of surveillance data, there can be three missing data patterns: monotone, intermittent and mixed missing data patterns (Figure 60).

Figure 60. **Different patterns of missing data**



	Monotone missing data					Intermittent missing data					Mixed missing data				
Country A	x	x	x	x	x	x	x		x	x	x	x		x	x
Country B	x	x				x	x	x	x	x	x		x	x	
Country C	x	x	x			x		x	x	x	x		x		
Country D	x	x	x	x	x	x	x			x		x	x		

The green cells with an 'x' indicate observed data and the white cells indicate missing data.

Missing data are **monotone** if all missing values are found after a certain cut-off time, while all values before the first instance of missing data are available. Monotone missing data can occur if an entity providing data drops out. In the context of surveillance data at ECDC, this is a country that no longer reports data. For example:

- The United Kingdom is no longer part of the EU/EEA, and may no longer report data to The European Surveillance System (TESSy).
- A country that has discontinued surveillance of a particular disease will no longer have data to report.

While methods for monotone missing data are available, it may be more important to decide if it makes sense for the trend analysis to include a country that is no longer part of European-level surveillance.

Missing data are **intermittent** if there are observed values for certain time units after missing values occur. In the context of surveillance data at ECDC, a country may not report data for a certain year if there is a very high workload, potentially due to a high incidence of the disease itself, or due to other factors (e.g. the COVID-19 pandemic).

Missing data are **mixed** if there is a mix of monotone and intermittent missing data.

1.2 Mechanisms of missing data

Mechanisms of missing data are different from patterns in missing data—the mechanisms relate to the statistical relationship between observations and the probability of missing data. There are three classifications of missing data mechanisms:

- missing completely at random (MCAR)
- missing at random (MAR)
- missing not at random (MNAR)

If data points are **missing completely at random (MCAR)**, the probability of missingness in a variable is not associated with its true value (larger values in that variable are not more likely to become missing than smaller values or vice versa). The variable is also not associated with any other variable in the dataset, meaning that it is not statistically related in any way to data points that are not missing. For instance, if only a random selection of patients in a study might have additional laboratory tests performed on them (e.g. only a random selection of influenza-positive patients might have their virus genetically characterised, as characterisation is quite expensive). In the context of MCAR data, if the data are analysed with observed values only, this will result in a loss of precision, but not in bias. In the context of surveillance data, counts could be MCAR if there is an unexpected problem with the surveillance system and a country fails to report data for a given time unit (e.g. there is an issue with the fridge holding samples, and therefore specimens for a given time unit could not be tested).

If data points are **missing at random (MAR)**, the probability of missingness in a variable is not associated with its true value (larger values in that variable are not more likely to become missing than smaller values or vice versa). However, the probability of missingness could be associated with other observed variables in the dataset (i.e. whether or not data points are MAR is conditional on observed data). The probability of missingness in a variable is also independent of its own unobserved values. For example, whether an HIV diagnosis test is performed might depend on the characteristics of the person related to the propensity of having a test (age, sex, profession, drug use, etc.). However, as the test has not yet been performed, it cannot depend on the results of the test. If MAR data with only observed values are analysed, this can result in a loss of precision and potentially also bias, depending on how the values are associated with other values. In the context of surveillance data, there may be MAR data if there is an increase in disease incidence in one country, which may also be seen in other countries.

If data points are **missing not at random (MNAR)**, the missing value is related to the variable itself, even after taking other variables into account. For example, individuals with higher incomes may be less likely to report the value of their income. Or, when there is a large outbreak of a given disease, reporting may be interrupted/missed due to capacity reasons. Therefore, the missingness depends on the value itself, and the observed data are not representative of the population under study. Because of this, if MNAR data with only observed values are analysed, this can result in a loss of precision as well as bias.

1.3 Overview of methods to handle missing data

1.3.1 Complete case analysis

Complete case analysis is a technique where data with missing values are dropped. It is sometimes known as 'list-wise deletion'. In the context of surveillance data, this might include restricting to a study period where there are no missing data or dropping a country or countries with missing data. A variation of this might be the use of a higher-level time unit if the information is missing in lower-level time units.

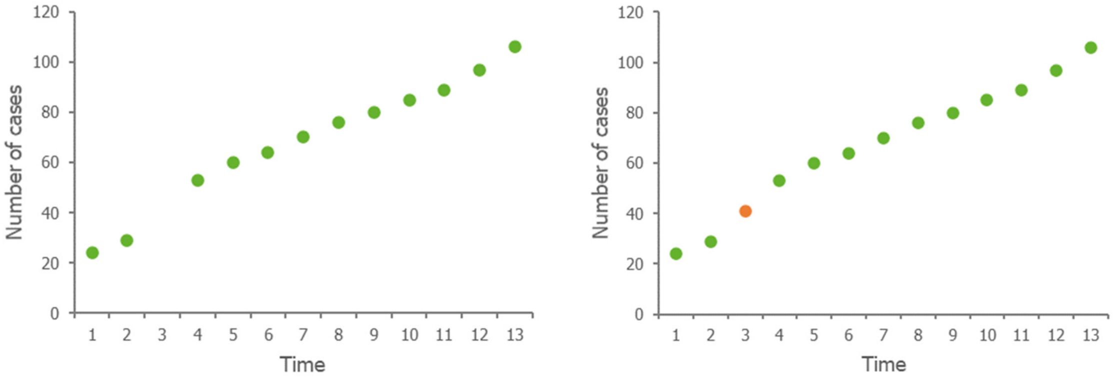
1.3.2 Ignoring missing data points

When carrying out a trend analysis in statistical software, one option is to ignore the missing data points. Trend models estimate a continuous function over a continuous time variable, but only using data reported at certain time points. For example, when yearly cases of a disease are modelled, the values are placed on the actual number of the year (i.e. on 2010, 2011, 2012, etc.). However, the function is calculated over continuous time, which means the number of cases expected at any time of the year can be estimated. The fact that there are no data observed at a specific time point is therefore not a problem. The same might occur if certain years are missing (although too many consecutive missing years can pose a problem in estimating a reliable trend).

1.3.3 Single imputation methods

Linear interpolation can be a popular choice of imputation for time series data. In this method, the missing value or values between two time points with observed data are simply obtained by creating a straight line through these points. The orange data point in Figure 61 below was imputed by drawing a straight line through the non-missing data points on its sides.

Figure 61. **Linear interpolation of a missing data point**



This is an intuitive way of accounting for missing data. However, this method, along with other single imputation methods, is not the most statistically sound. This is because it does not take the uncertainty around these estimates into account and tends to underestimate the standard errors. In addition, the interpolation may be less useful with more consecutive missing data points.

There are many other single imputation methods, although not all are appropriate in the time series analysis setting.

1.3.4 Multiple imputation methods

Multiple imputation methods impute values in a variable Y with a prediction model (often a regression) using other variables (X_1, X_2, etc.). This imputation is repeated several times, creating a different dataset of imputed values each time. In each dataset, the trend of the variable of interest is estimated and the variability between these trends accounts for the uncertainty around the imputed values. The results from the different datasets must be adequately combined to provide a summary estimate. Standard errors are calculated in such a way as to take the variability between the datasets into account. The individual estimates and standard errors are combined using Rubin's rules [14].

The validity of multiple imputation depends on whether it is possible to adequately model the distribution of the variables with missing data using the observed data. There are many assumptions underlying this and, where possible, it is a good idea to carry out multiple imputation with the help of a statistician.

If the data permit a multiple imputation and the method is properly carried out, then it can help avoid bias and potentially improve precision in an analysis.

A popular method of multiple imputation, which is recommended here, is multiple imputation using chained equations. Here multiple variables are imputed iteratively using a sequence of univariable imputation models. All variables except the one to be imputed are included in the prediction equation used for the imputation. Each variable with missing values is imputed by a separate model. Here a different model can be specified for each variable and the use of predictive mean matching is recommended (as indicated in the ECDC guidance document [5] for rates and counts). This approach may be useful if data are not normally distributed. It combines linear regression imputation methods with nearest-neighbour imputation methods (the number of nearest neighbours will need to be set).

1.4 Amount of missing data

There is a great deal of interest in the questions 'How much missing data are required before a complete case analysis is biased?' and 'What is the upper threshold of missingness after which a multiple imputation is no longer meaningful?' While some cut-off points in terms of proportion of missingness have been suggested, there is a lack of evidence to support them [17].

In terms of longitudinal data, such as time series data, the proportion of missing data alone may not be meaningful. It is important to describe the missing data overall, as well as for each geographical unit and time unit (see Part D, Section 1.5, 'Analysis'). For example, as seen in Figure 62, the top scenario and the bottom scenario have the same amount of missing data. But the implications of carrying out a complete case analysis or multiple imputation are quite different (as it is potentially more difficult to perform a multiple imputation if large chunks of data are missing).

Figure 62. Two scenarios of the same proportion of missing data, countries A–D

Country A	x	x					x	x	x	x
Country B	x	x	x	x	x	x	x	x	x	x
Country C	x	x	x	x	x	x	x	x	x	x
Country D	x	x	x	x	x	x	x	x	x	x

Country A	x	x		x	x	x	x	x	x	x
Country B	x	x	x	x	x	x	x	x		x
Country C	x	x	x	x	x	x		x	x	x
Country D	x	x	x		x	x	x	x	x	x

The green cells with an 'x' indicate observed data and the white cells indicate missing data.

Some researchers suggest that multiple imputation can be used to provide unbiased estimates and greater precision than a complete case analysis, regardless of the proportion of missing data (particularly if auxiliary information is available) [18].

Further information on missing data mechanisms might be of interest, including their mathematical definitions. Relevant resources include Little RJA, Rubin DB (2002) [19]; Carpenter JR, Kenward MG (2013) [20]; and Buuren Svan (2018) [21]. A useful overview can also be found in Sterne JAC, White IR, et al. (2009) [22].

1.5 Analysis

The following examples use the ECDC HIV yearly notification dataset for EU/EEA countries from 2007 to 2016, which has no missing data (this dataset is freely available from ECDC's 'Surveillance Atlas of Infectious Diseases' [4]). For illustration purposes, two further datasets have been created from these data:

- **Dataset 1:** The same original data, but with 5% of data points missing (randomly selected).
- **Dataset 2:** The same original data, but with 4% of data points missing in chunks, from two countries only.

These datasets have quite a low proportion of missing data. This is because, in the context of surveillance data, yearly count data most often do not have a large proportion of missing data.

In the original dataset, which had no missing data, the linear trend for a crude pooled analysis (where all counts by country are totalled and the data are analysed as if they were from a single country) is 100.5 cases per year (95% CI: -123.1 to 324.0). The trend for a pooled analysis using random-effects meta-analysis is 0.8 cases per year (95% CI: -1.9 to 3.5). This latter estimate relates to the average trend across the countries in the dataset, and it differs in magnitude from the crude pooled analysis.

1.5.1 Descriptive analysis of missing data

An important step is to describe the missing data in the dataset.

Plotting the missing data by time and country

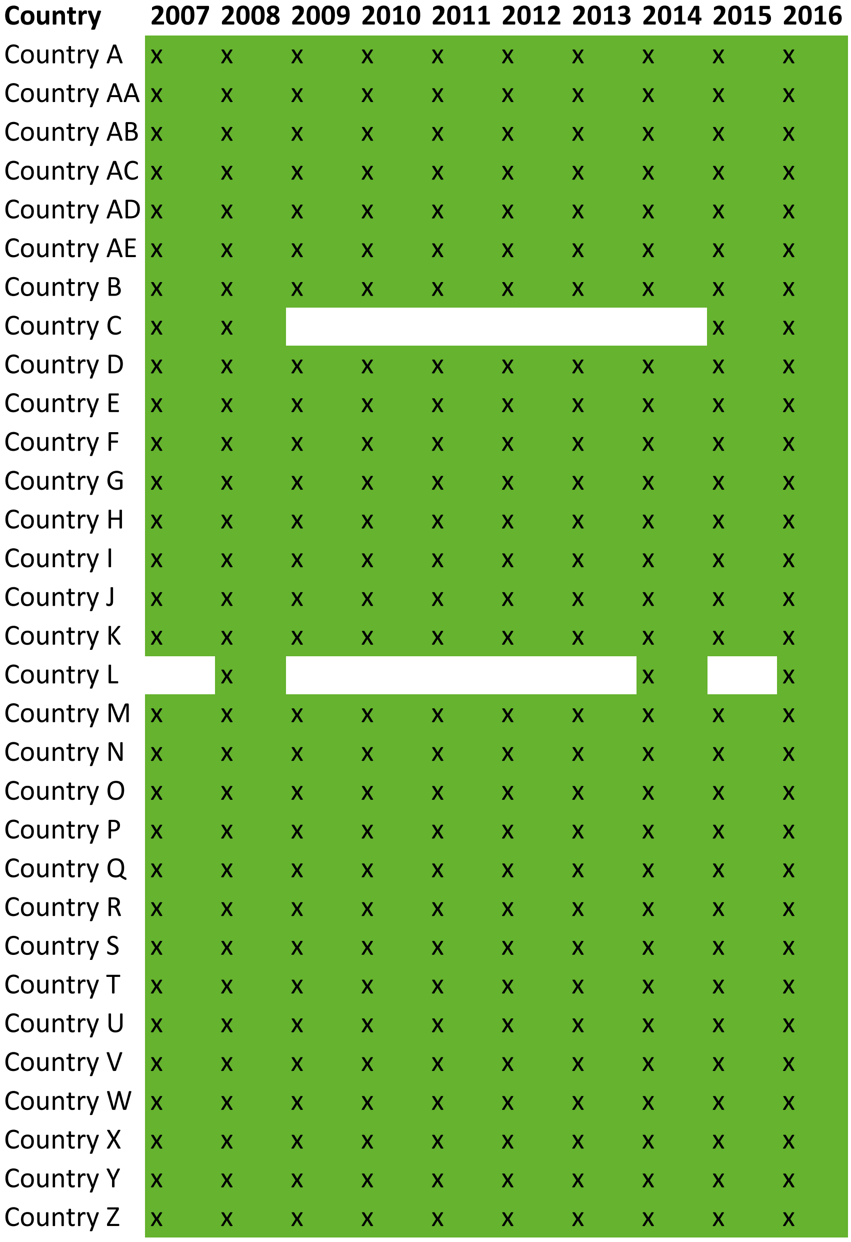
In a longitudinal dataset such as surveillance data with few variables, it is possible to plot the data by time and country in relation to its missingness (Figures 63 and 64). The visual aspect helps to provide an understanding of the distribution of missingness over time and country. It also helps in identifying any clusters of missing data and patterns in missing data.

Figure 63. Observed and missing data of Dataset 1 (5% missing data)

Country	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Country A	x	x	x	x	x	x	x	x	x	x
Country AA	x	x	x	x	x	x	x	x	x	
Country AB	x	x	x	x	x	x	x	x	x	x
Country AC	x	x	x	x		x	x	x	x	
Country AD	x	x	x	x	x	x	x	x	x	x
Country AE	x	x	x	x	x	x	x	x	x	
Country B	x		x	x	x	x	x	x	x	x
Country C	x	x	x	x	x	x	x	x	x	x
Country D	x	x	x		x	x	x	x	x	x
Country E	x	x	x	x	x	x	x	x	x	x
Country F	x	x	x		x	x	x	x	x	x
Country G	x	x	x	x	x	x		x	x	x
Country H	x	x	x	x	x	x	x	x	x	x
Country I	x	x	x		x	x	x	x	x	x
Country J	x	x	x	x	x	x	x	x	x	x
Country K	x	x	x	x	x	x	x	x	x	x
Country L	x	x		x	x	x	x	x	x	x
Country M	x	x	x	x	x	x	x	x	x	x
Country N	x	x	x	x	x	x	x	x	x	x
Country O	x	x	x	x	x	x	x	x	x	x
Country P	x	x	x	x	x	x		x	x	x
Country Q	x	x	x	x	x	x		x	x	x
Country R	x	x	x	x	x	x	x	x	x	x
Country S	x	x	x	x	x	x	x	x	x	x
Country T	x	x	x	x	x	x	x	x	x	x
Country U	x	x	x	x	x	x	x	x	x	x
Country V	x	x	x	x	x	x	x	x	x	x
Country W	x	x	x	x	x	x	x	x	x	x
Country X		x	x	x	x	x	x	x	x	x
Country Y	x	x	x	x		x	x	x	x	x
Country Z	x	x	x	x	x		x	x	x	x

The green cells with an 'x' indicate observed data and the white cells indicate missing data. For the country names, 'AA' to 'AE' were added to the standard English alphabet so the 31 EU/EEA countries could be represented anonymously using letters.

Figure 64. Observed and missing data of Dataset 2 (4% missing data)

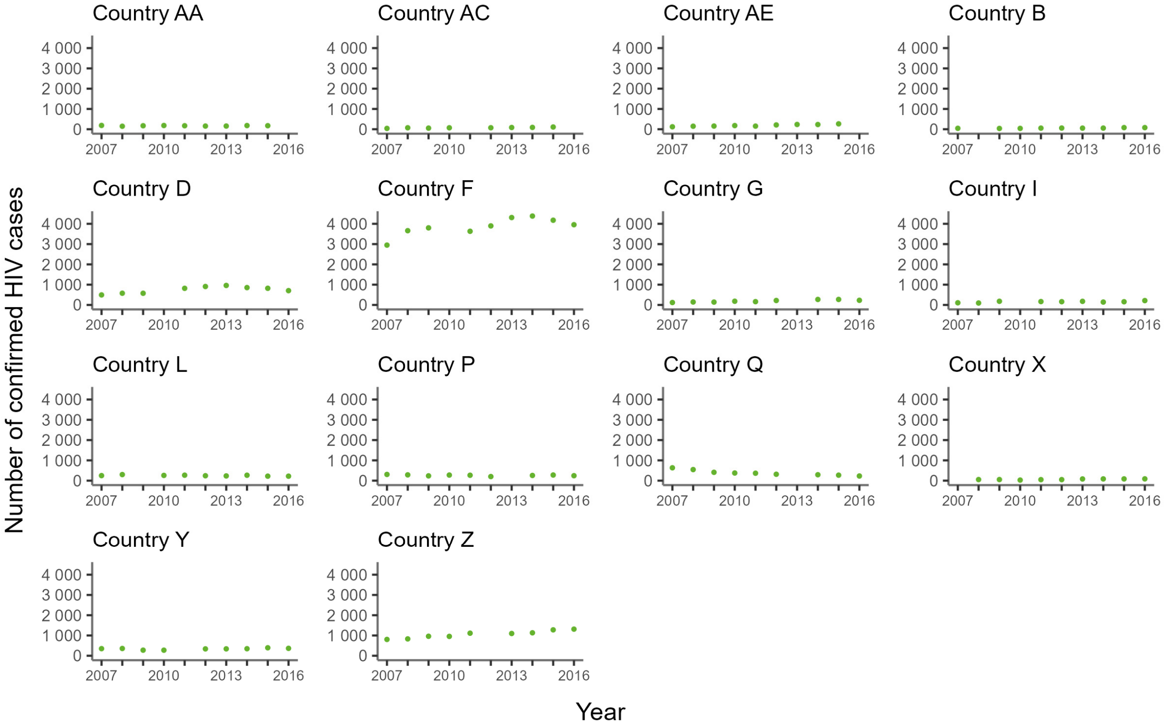


Country	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Country A	x	x	x	x	x	x	x	x	x	x
Country AA	x	x	x	x	x	x	x	x	x	x
Country AB	x	x	x	x	x	x	x	x	x	x
Country AC	x	x	x	x	x	x	x	x	x	x
Country AD	x	x	x	x	x	x	x	x	x	x
Country AE	x	x	x	x	x	x	x	x	x	x
Country B	x	x	x	x	x	x	x	x	x	x
Country C	x	x							x	x
Country D	x	x	x	x	x	x	x	x	x	x
Country E	x	x	x	x	x	x	x	x	x	x
Country F	x	x	x	x	x	x	x	x	x	x
Country G	x	x	x	x	x	x	x	x	x	x
Country H	x	x	x	x	x	x	x	x	x	x
Country I	x	x	x	x	x	x	x	x	x	x
Country J	x	x	x	x	x	x	x	x	x	x
Country K	x	x	x	x	x	x	x	x	x	x
Country L		x						x		x
Country M	x	x	x	x	x	x	x	x	x	x
Country N	x	x	x	x	x	x	x	x	x	x
Country O	x	x	x	x	x	x	x	x	x	x
Country P	x	x	x	x	x	x	x	x	x	x
Country Q	x	x	x	x	x	x	x	x	x	x
Country R	x	x	x	x	x	x	x	x	x	x
Country S	x	x	x	x	x	x	x	x	x	x
Country T	x	x	x	x	x	x	x	x	x	x
Country U	x	x	x	x	x	x	x	x	x	x
Country V	x	x	x	x	x	x	x	x	x	x
Country W	x	x	x	x	x	x	x	x	x	x
Country X	x	x	x	x	x	x	x	x	x	x
Country Y	x	x	x	x	x	x	x	x	x	x
Country Z	x	x	x	x	x	x	x	x	x	x

The green cells with an 'x' indicate observed data and the white cells indicate missing data. For the country names, 'AA' to 'AE' were added to the standard English alphabet so the 31 EU/EEA countries could be represented anonymously using letters.

Plots of the observed data by country, with missing data, can also provide useful information (see Figure 65). Is the missing data in what looks like it might be a peak in the number of cases? Or a trough?

Figure 65. Number of confirmed HIV cases reported among countries with missing data, Dataset 1, 2007–2016



For the country names, 'AA' to 'AE' were added to the standard English alphabet so the 31 EU/EEA countries could be represented anonymously using letters.

1.6 Quantifying completeness

In this type of surveillance data, the numbers and proportion of missing data can be quantified overall by country and time unit. Often, instead of documenting what is missing, 'completeness' is documented. If an analysis is being carried out with seasonal terms, the time unit can also be broken down and the numbers and proportion of complete data described by month/week (whichever seasonal term is being used) and year.

Tables 16 and 17 outline the number and proportion of complete data for Datasets 1 and 2 by year and country.

Table 16. Completeness of Datasets 1 and 2, by year

	N/total (%) Dataset 1	N/total (%) Dataset 2
Overall	295/310 (95)	297/310 (96)
Years with complete data	2/10 (20)	2/10 (20)
2007	30/31 (97)	30/31 (97)
2008	30/31 (97)	31/31 (100)
2009	30/31 (97)	29/31 (94)
2010	28/31 (90)	29/31 (94)
2011	29/31 (94)	29/31 (94)
2012	30/31 (97)	29/31 (94)
2013	28/31 (90)	29/31 (94)
2014	31/31 (100)	30/31 (97)
2015	31/31 (100)	30/31 (97)
2016	28/31 (90)	31/31 (100)

Table 17. Completeness of Datasets 1 and 2, by country

	N/total (%) Dataset 1	N/total (%) Dataset 2
Overall	295/310 (95)	297/310 (96)
Countries with any missing data	17/31 (55)	29/31 (94%)
Country A	10/10 (100)	10/10 (100)
Country AA	9/10 (90)	10/10 (100)
Country AB	10/10 (100)	10/10 (100)
Country AC	8/10 (80)	10/10 (100)
Country AD	10/10 (100)	10/10 (100)
Country AE	9/10 (90)	10/10 (100)
Country B	9/10 (90)	10/10 (100)
Country C	10/10 (100)	4/10 (40)
Country D	9/10 (90)	10/10 (100)
Country E	10/10 (100)	10/10 (100)
Country F	9/10 (90)	10/10 (100)
Country G	9/10 (90)	10/10 (100)
Country H	10/10 (100)	10/10 (100)
Country I	9/10 (90)	10/10 (100)
Country J	10/10 (100)	10/10 (100)
Country K	10/10 (100)	10/10 (100)
Country L	9/10 (90)	3/10 (30)
Country M	10/10 (100)	10/10 (100)
Country N	10/10 (100)	10/10 (100)
Country O	10/10 (100)	10/10 (100)
Country P	9/10 (90)	10/10 (100)
Country Q	9/10 (90)	10/10 (100)
Country R	10/10 (100)	10/10 (100)
Country S	10/10 (100)	10/10 (100)
Country T	10/10 (100)	10/10 (100)
Country U	10/10 (100)	10/10 (100)
Country V	10/10 (100)	10/10 (100)
Country W	10/10 (100)	10/10 (100)
Country X	9/10 (90)	10/10 (100)
Country Y	9/10 (90)	10/10 (100)
Country Z	9/10 (90)	10/10 (100)

For the country names, 'AA' to 'AE' were added to the standard English alphabet so the 31 EU/EEA countries could be represented anonymously using letters.

The completeness/missing data are quite different in Datasets 1 and 2. It is important to understand the reasons for missing data and to describe them. Documenting reasons for missing data can form part of the descriptive analysis of missing data.

The ideal situation, of course, is not to have any missing data. If it is possible to obtain any of the missing data before the analysis, then every effort should be made to do so.

Though most data contain missing values, there is often still interest in carrying out an analysis. Part D, Section 1.3 describes two pragmatic approaches to analysing missing data for the purposes of the type of surveillance data that is considered in this guidance document.

1.7 Complete case analysis

If a complete case analysis is carried out for Dataset 1, all the countries and years where there are missing counts of HIV notifications are dropped. As seen in Table 18, only 17 of 31 countries can be included. If a complete case analysis is carried out for Dataset 2, two countries are dropped because only these have missing data.

Table 18. Pooled trend for the complete dataset and the complete case analysis for Datasets 1 and 2

Dataset	Countries included (N)	Crude pooled analysis Coefficient (95% CI)	Adjusted pooled analysis Coefficient (95% CI)
Complete dataset	31	100.5 (-123.1 to 324.0)	0.8 (-1.9 to 3.5)
Dataset 1	17	-111.8 (-270.9 to 47.4)	-5.3 (-9.1 to -1.5)
Dataset 2	29	2.5 (-224.2 to 229.2)	0.8 (-1.9 to 3.5)

CI: confidence interval

1.7.1 Results and interpretation

When excluding 14 countries in the complete case analysis for Dataset 1, the number of HIV notifications decreases by 112 per year in the EU/EEA in the crude pooled analysis. This is a downward trend that is not statistically significant (Table 18). This is substantially lower than in the crude pooled analysis of the complete dataset. It is possible that the randomly generated dataset had missing values in countries with upward trends in particular.

In the adjusted pooled analysis (the meta-analysis approach), the summary estimate is -5.3 HIV notifications per year, suggesting that there is an annual decrease of 5.3 cases per country. The estimate is statistically significant, compared to the complete dataset where there was an increase of 0.8 cases. The inferences around these results are different.

As illustrated in Part C, Section 2.3, ‘Crude pooled analysis of data composed from different countries’, a crude pooled analysis treats the data as if it were reported from one region. Therefore, when using counts, the trend from the crude pooled analysis relates to the trend (the number of cases by which it decreases or increases) of the entire entity. When carrying out an adjusted pooled analysis (the meta-analysis approach), the summary estimate for the trend is calculated based on the average trend for a given country. These measures for counts are not directly comparable.

In Dataset 2, the crude pooled analysis of the number of HIV notifications increased by 2.5 per year, which is not statistically significant. This is lower than the crude pooled analysis of the complete dataset. The adjusted pooled analysis had the same value of coefficient and confidence intervals as the complete dataset.

In this example, when carrying out a complete case analysis, the results are quite dissimilar for both Dataset 1 and Dataset 2 in the crude pooled analysis. But results in the adjusted pooled analysis (the meta-analysis approach) were nearly identical for Dataset 2. It might be worth exploring whether using a standardised measure, such as rates, would provide a more similar output in the crude pooled analysis.

If a large number of countries are dropped from a trend analysis, such as in Dataset 1, it is likely that the trend analysis is no longer representative of all EU/EEA countries and that the findings can only be extrapolated to the group of 17 countries included. Consequently, it is not likely to provide a suitable answer to the study question: ‘What is the trend of HIV notifications in the EU/EEA?’ However, the analysis will continue for the purpose of this guidance document.

1.7.2 Strengths

The advantage of a complete case analysis is that it is easy to carry out and convey to readers. There may be good reasons for excluding a country with missing data from a trend analysis. Perhaps this country has discontinued surveillance and, therefore, a trend including this country would not be meaningful.

1.7.3 Limitations

Excluding a country or countries from a trend analysis due to missing data may result in a biased trend analysis. This is especially the case if there is a different trend for the excluded country/countries or if there is a different magnitude of cases (if counts are being used, rather than rates).

If a large number of countries is excluded, it may not be possible to extrapolate the findings to the EU/EEA (or other region of interest). It is then only possible to make inferences for the smaller number of included countries.

Ignoring missing data points

Dropping countries with large amounts of missing data and carrying out a complete case analysis seems very intuitive, as shown for Dataset 2. However, if there are many countries with small amounts of missing data, like in Dataset 1, a lot of information is lost by dropping these countries. In the complete case analysis for Dataset 1, 14 countries are dropped. In this situation, it is worth considering ignoring missing data points (see Part D, Section 1.3.2, 'Ignoring missing data points'). This approach can work when carrying out an adjusted pooled analysis (the meta-analysis approach), as trends are calculated at the country level and then pooled. In the context of Dataset 1, this provides a summary estimate of 0.6 cases per year (95% CI: -2.1 to 3.3), which is very close to the corresponding estimate of the complete dataset (0.8 cases per year).

However, it is not recommended to ignore missing data points and carry out a crude pooled analysis. Results will be biased, as those years where countries have not reported data will be artificially lower because of the missing data. Therefore, a trend estimate will be meaningless.

1.8 Multiple imputation of counts using chained equations

It is important that the missing data and the reasons for them are adequately described, so that there is confidence that the chosen imputation is appropriate. There are various methods for multiple imputation. For the purposes of this guidance document, multiple imputation using chained equations with predictive mean matching for the count and rate outcomes is recommended.

There are very often limited predictive variables in surveillance data used to perform an imputation. It is important to discuss the limitations of this method (e.g. whether an imputation on a dataset with available information on age and sex is more valid). Additionally, as mentioned in Part B, Section 5.1, 'Count data', using rates rather than counts may be more useful, as it is a standardised measure. This also applies to an imputation. Rate data can always be transformed into count data by multiplying by the population. This section first illustrates multiple imputation using count data and then using rate data.

There are two potential approaches to imputing missing values in surveillance data with counts or rates by country and time. The missing data points can be imputed based on data points of different time points for the same country, or it can be imputed based on data points from other countries for the given time point.

If the former is chosen, then the hypothesis is that the non-missing data points over time for the same country best informs the missing values. If the latter is chosen, then the hypothesis is that other countries can inform the values for the country with missing data.

This way of using information from neighbouring data points is typical for many spatial models. The chosen method is related to the underlying data (the nature of the disease, the type and amount of missing data). The decision about the method can also be made with a disease expert and, ideally, a statistician. In some statistical software, these methods may not run if the number of parameters exceeds the number of observations (e.g. if the time series is short and there are many countries).

The method illustrated here imputes values based on data from other countries. However, the methods for imputation based on data points from other time periods is similar.

1.8.1 Method

Multiple imputation packages and commands are generally provided with standard statistical software (e.g. the multivariate imputation by chained equations (MICE) package in R and the 'mi impute' command suite in Stata).

Most often, the data must be set for multiple imputation, ensuring that it is in the desired format for longitudinal data. The format is most often wide, with countries as rows and years as columns. It might then be necessary to register the variables for imputation. After that, the multiple imputation may be carried out with a recommended default of five multiple imputations using five nearest neighbours. These parameters can be modified if needed, with the help of a statistician. Note that this will create five imputed values for each missing data point. These multiple values must be considered when calculating an estimate or managing the data, using specific multiple imputation commands/functions.

For example, in Dataset 1, country AC and country Y had missing data in 2011. Country F had missing data in 2010. After imputation, there are five values for the variable 'count' (see Table 19), each with an associated imputation number (column 'Multiple imputation number').

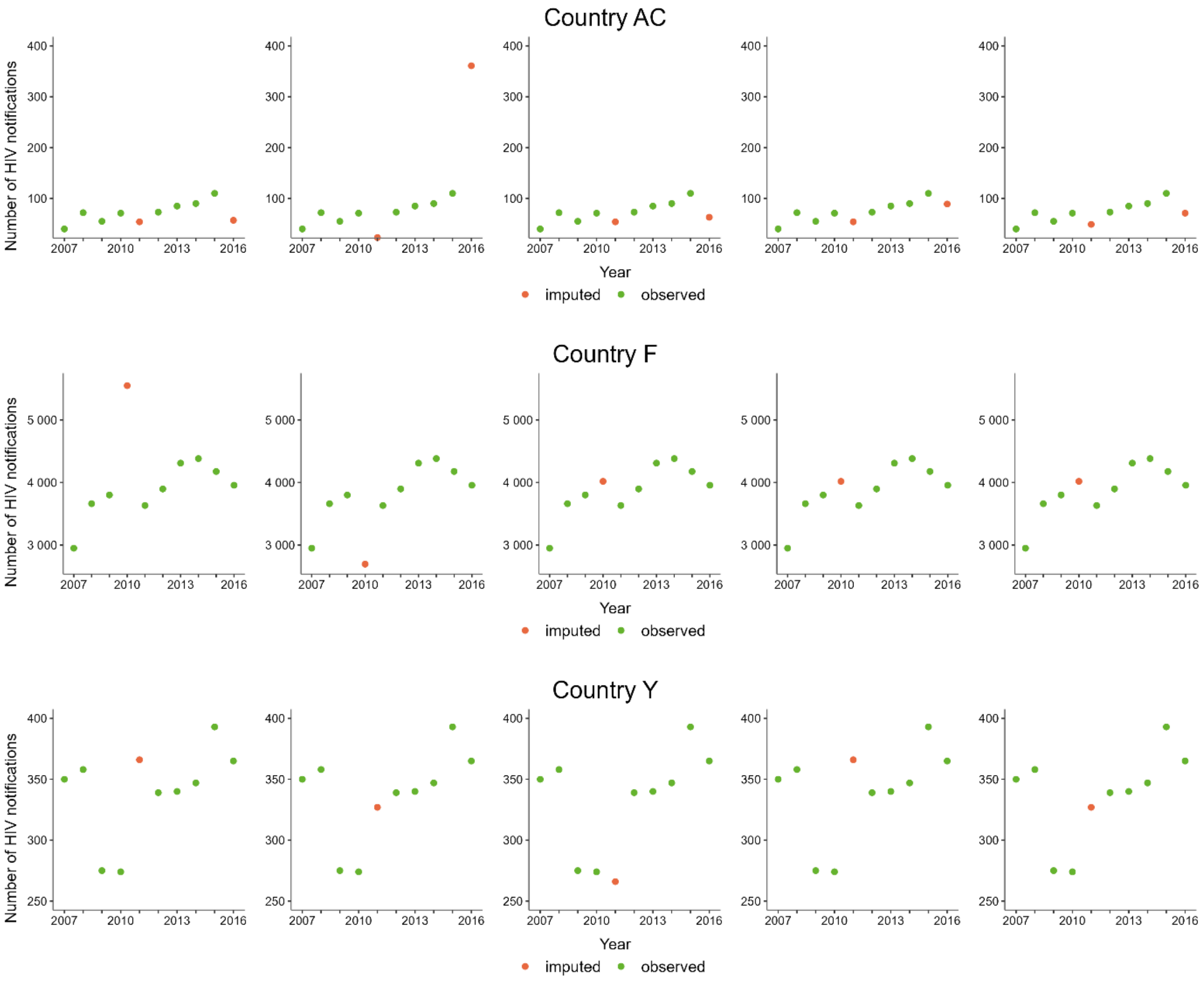
Table 19. Selected output of multiple imputation for Dataset 1 (missing data in 14 countries)

Country	Year	Count	Multiple imputation number
Country AC	2011	54	1
Country AC	2011	23	2
Country AC	2011	54	3
Country AC	2011	54	4
Country AC	2011	49	5
Country F	2010	5 548	1
Country F	2010	2 695	2
Country F	2010	4 018	3
Country F	2010	4 018	4
Country F	2010	4 018	5
Country Y	2011	366	1
Country Y	2011	327	2
Country Y	2011	266	3
Country Y	2011	366	4
Country Y	2011	327	5

After the imputation, it is a good idea to carry out diagnostics. Are there five imputed results for each of the missing values? Plot the observed and imputed data for each country with imputed values and for each imputed value. This means that, if there is one country with missing data, there will be five different graphs (one for each imputation).

In Figure 66, the observed values (green) and imputed values (orange) are plotted for each imputation for a selection of countries in Dataset 1. Not all data points appear to be imputed for all countries, although there is a great deal of variation in observed values in these data (which the imputation reflects).

Figure 66. Selected observed and imputed values by country for each imputation dataset, Dataset 1 (missing data in 14 countries)



When carrying out the trend analysis, specific multiple imputation commands/functions can be used that consider the different structures of the data. In this way, the uncertainty around the imputed values is considered in the estimates.

1.8.2 Results and interpretation

When carrying out a crude pooled analysis on the imputed Dataset 1, the trend analysis shows an annual increase of 154 cases. This is higher than the results in the complete dataset (an annual increase of 101 cases), but the inferences around the results are the same: there is an upward trend that is not statistically significant. The multiple imputation crude pooled analysis appears to provide results that are closer than the complete case analysis, where 14 countries were dropped due to missing data. The adjusted pooled results (meta-analysis approach) with the imputed dataset are also closer to the complete dataset results, compared to the complete case analysis, even if the direction of the trend is different (-0.4 compared to 0.8) (Table 20). When carrying out an imputed analysis, it is always important to present both the imputed results and the complete case analysis results alongside each other [22].

Table 20. Pooled analysis by dataset

	Countries included (N)	Crude pooled analysis Coefficient (95% CI)	Pooled meta-analysis Coefficient (95% CI)
Complete dataset	31	100.5 (-123.1 to 324.0)	0.8 (-1.9 to 3.5)
Complete case analysis	17	-111.8 (-271.9 to 47.4)	-5.3 (-9.1 to -1.5)
Imputed analysis	31	154.4 (-276 to 585.0)	-0.4 (-3.5 to 2.6)

CI: confidence interval

1.8.3 Strengths

If the dataset contains missing data, using multiple imputation is a sophisticated way to attempt to account for bias in estimates. It may lead to less biased estimates than a complete case analysis, where records with missing values are dropped.

1.8.4 Limitations

Multiple imputation is an advanced statistical technique. It requires a good understanding of the underlying reasons for missingness, the appropriateness of imputing and which parameters to use. After imputation, it is crucial to use dedicated imputation functions/commands for estimation and data management that take the imputed structure into account.

Multiple imputation of rates and crude pooled analysis

If rates are the outcome measure, it is straightforward to carry out an adjusted pooled analysis (random-effects meta-analysis) after imputation, as both imputing and trends are calculated at the country level. Carrying out a crude pooled analysis after imputing rates at the country level may take a bit more thought. Follow these steps:

- Carry out the imputation on rates by country, as usual.
- Use multiple imputation functions/commands supplied with standard software to ensure correctness.
- Calculate absolute counts in each imputed dataset in each country
- Sum the counts of all the EU/EEA countries together, so there are five datasets at the end (or the number set in the imputation) per year for the whole EU/EEA.
- Using the EU/EEA population, calculate the rate per year for the whole EU/EEA.
- Using multiple imputation functions/commands, calculate the trend in rates for EU/EEA countries.

2. Algorithm for carrying out a trend analysis on data with missing time points

Step 1: Define the study objectives

Clarity about the study objectives is important when there are missing values in the data. Is there interest in the trend of a given disease in the EU/EEA over a specific set of years?

Step 2: Understand the data

It is important to understand any underlying heterogeneities in the data before carrying out a trend analysis. Obtaining this information will be of great help in understanding how far it makes sense to go with pooling. If possible, obtain this information from a disease expert or experts.

Step 3: Describe the missing data

Create a plot by time and country, indicating where values are missing and not missing. Are there patterns to the missing data? Are the missing data monotone? Or were they missing at the start of the study period? Are the data intermittently missing? Are they mixed?

Quantify the amount of data missing overall, by individual countries and year. Are there certain countries or years where there are a lot of missing values?

Plot the data of the countries with missing values. Where are the missing values? Where is a peak expected? Or a trough? Or is there no discernible pattern?

With the help of the missing data description and the understandings gathered in step 2, try to understand the reasons for missing data. This can be done together with a disease expert. Are countries not reporting data due to an increase in cases? Were they not reporting data because they were not a member of the EU/EEA at the time? Have they stopped reporting data because they have left the EU/EEA?

Additional questions to consider:

- Are there qualitative reasons for excluding a country that fit with the purposes of the analysis?
- Is it possible to obtain the missing data from the countries?
- Are the missing data at lower-level time units? Is it possible to aggregate to higher-level time units?

Step 4: Choose the approach for handling missing data

Depending on the study objectives (step 1) and the nature of missing data (step 3), choose the approach for handling missing data.

There is no simple way to deal with missing data. To decide how to handle the missing data, one must have a very good understanding of the data. Ideally decisions around handling missing data will include consultations with disease experts and statisticians.

Are there large amounts of missing data or missing not at random (MNAR) data?

If there are large amounts of missing data from all countries and/or the reasons for the missingness are related to the missing values themselves (MNAR), it may be determined that the analysis cannot be carried out. In this case, try to obtain more of the missing data before attempting to carry out the analysis.

Are there monotone missing data?

If the data are monotone and missing from a country or the country has stopped reporting entirely (e.g. the country has left the EU/EEA or has discontinued surveillance completely), consider carrying out the analysis excluding this country, if that still meets the objectives of the analysis.

Are there missing data for a specific year or period of years?

Depending on the study objectives, if there is a specific year or period of years where there is a lot of data missing, consider restricting the analysis period and carry out the analysis with a full dataset (a complete case analysis).

Are there missing at random (MAR) data?

Do you consider that the values are MAR? Depending on the objectives of the analysis, consider restricting the analysis to countries without missing values (a complete case analysis), noting that the results pertain to the included countries only. It may also be beneficial to carry out multiple imputation to impute missing values and then carry out the analysis. Results will be inferred on the total number of countries providing data. If so, it is important to compare the results with those of the complete case analysis.

Step 5: Carry out a complete case analysis

If it is determined in step 4 that a complete case analysis is needed, remove countries with missing data. Carry out the analysis among the countries that do not have missing data, determining the trend as defined in the algorithms in Part B, Section 6, 'Algorithm for performing the trend analysis' and Part C, Section 3, 'Algorithm for carrying out the trend analysis taking country into account'. Describe the missing data and the number and proportion of records dropped due to missing values in the report. Explain the inferences that can be made from the data and the limitations.

Step 6: Carry out the multiple imputation

Decide on the use of rates or counts for the multiple imputation. Rates may be a more stable measure of imputation compared to counts, as counts may be subject to greater variation. However, be aware that rounding errors can occur if rates are used for a crude pooled analysis after imputation. If possible, consult a statistician for this procedure.

Step 7: Carry out diagnostics after the imputation

Were the correct number of imputations obtained? Are the values available? Plot the observed and imputed data for those countries with missing values. Does the imputation look adequate?

Step 8: Estimate the trend

Ensure that multiple imputation commands/functions are used when carrying out any analysis on the imputed data. This ensures that the structure of the imputed data is taken into account. Determine the trend as defined in the algorithms in Part B, Section 6 and Part C, Section 3.

Step 9: Report the results

When reporting results from imputed data, ensure that the methods of the imputation are written out. Discuss the limitations of the imputation and provide a description of the missing data and the results of the complete case analysis.

3. Examples

The examples in this section use the steps in the algorithm in Part D, Section 2 to demonstrate the process of obtaining a trend in salmonellosis notifications in the EU/EEA. While the salmonellosis notification dataset is freely available on ECDC's 'Surveillance Atlas of Infectious Diseases' [4], the two datasets used in these examples have been modified to include fictitious missing values. In Dataset 1, for example, there are a lot of missing data for only two countries. In Dataset 2, the missing data are randomly distributed over countries and time. The names of the 29 EU/EEA countries that submitted data to these datasets have been anonymised as 'Country A', 'Country B', etc. (note that 'AA' to 'AC' were added to the standard English alphabet so that the 29 countries could be represented anonymously using letters).

3.1 Example: Salmonellosis notifications across the EU/EEA, taking country trends into account (Missing data: complete case analysis)

Step 1: Define the study objectives

The objectives of the study are to identify any changes in salmonellosis notifications across the EU/EEA, taking country trends into account, using a random-effects meta-analysis.

Step 2: Understand the data

Salmonellosis surveillance systems are reasonably homogeneous across the EU/EEA, with two exceptions. Country W experienced several large salmonellosis outbreaks in 2008 and did not report salmonellosis data at the European level that year or for several years thereafter. Country W restarted surveillance in 2015. A similar pattern was seen for Country V (Country W's neighbour), where—due to several very large salmonellosis outbreaks—reporting at the European level was disrupted for many years.

Step 3: Describe the missing data

The notifications are plotted by time and country, indicating where values are missing and not missing. Figure 67 illustrates that the missing data are restricted to only two countries only (Country V and Country W), where large chunks of data are missing in the middle of the study period.

Figure 67. Observed and missing data of salmonellosis notifications by country and year, EU/EEA, 2007–2016

Country	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Country A	x	x	x	x	x	x	x	x	x	x
Country AA	x	x	x	x	x	x	x	x	x	x
Country AB	x	x	x	x	x	x	x	x	x	x
Country AC	x	x	x	x	x	x	x	x	x	x
Country B	x	x	x	x	x	x	x	x	x	x
Country C	x	x	x	x	x	x	x	x	x	x
Country D	x	x	x	x	x	x	x	x	x	x
Country E	x	x	x	x	x	x	x	x	x	x
Country F	x	x	x	x	x	x	x	x	x	x
Country G	x	x	x	x	x	x	x	x	x	x
Country H	x	x	x	x	x	x	x	x	x	x
Country I	x	x	x	x	x	x	x	x	x	x
Country J	x	x	x	x	x	x	x	x	x	x
Country K	x	x	x	x	x	x	x	x	x	x
Country L	x	x	x	x	x	x	x	x	x	x
Country M	x	x	x	x	x	x	x	x	x	x
Country N	x	x	x	x	x	x	x	x	x	x
Country O	x	x	x	x	x	x	x	x	x	x
Country P	x	x	x	x	x	x	x	x	x	x
Country Q	x	x	x	x	x	x	x	x	x	x
Country R	x	x	x	x	x	x	x	x	x	x
Country S	x	x	x	x	x	x	x	x	x	x
Country T	x	x	x	x	x	x	x	x	x	x
Country U	x	x	x	x	x	x	x	x	x	x
Country V	x	x								x
Country W	x								x	x
Country X	x	x	x	x	x	x	x	x	x	x
Country Y	x	x	x	x	x	x	x	x	x	x
Country Z	x	x	x	x	x	x	x	x	x	x

The green cells with an 'x' indicate observed data and the white cells indicate missing data. For the country names, 'AA' to 'AC' were added to the standard English alphabet so the 29 EU/EEA countries that submitted data could be represented anonymously using letters.

The amount of data missing overall, and by individual countries and years, are then quantified. Overall, 5% of the data are missing and there are only two years with no missing data (Table 21). Data are missing from only two countries (Table 22).

Table 21. Completeness of salmonellosis notifications overall and by year, EU/EEA, 2007–2016

	N/total (%)
Overall	276/290 (95)
Years with complete data	2/10 (20)
2007	29/29 (100)
2008	28/29 (97)
2009	27/29 (93)
2010	27/29 (93)
2011	27/29 (93)
2012	27/29 (93)
2013	27/29 (93)
2014	27/29 (93)
2015	28/29 (97)
2016	29/29 (100)

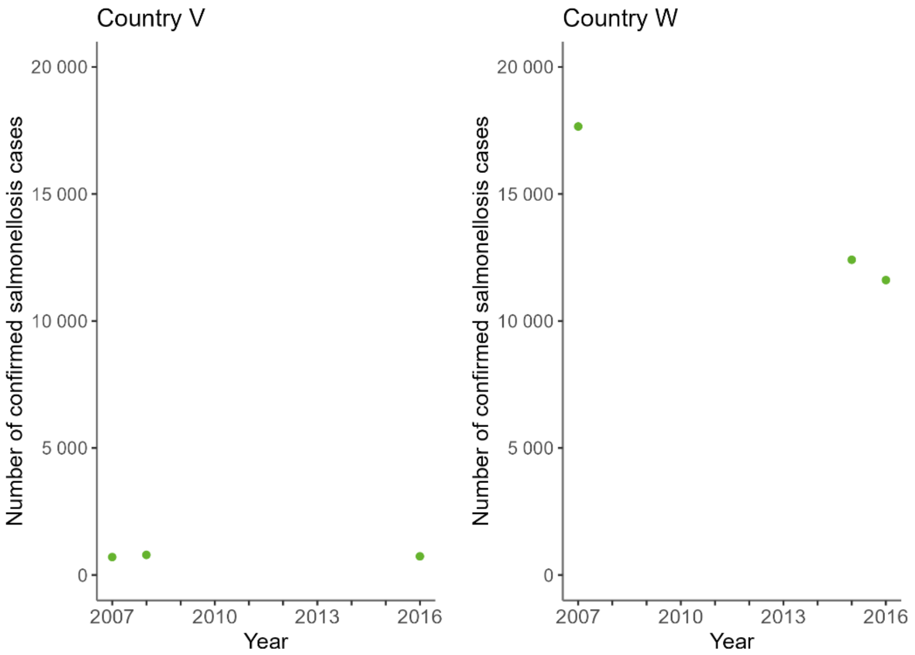
Table 22. Completeness of salmonellosis notifications by country, EU/EEA, 2007–2016

Country	N/total (%)
Countries with complete data	27/29 (93)
Country A	10/10 (100)
Country AA	10/10 (100)
Country AB	10/10 (100)
Country AC	10/10 (100)
Country B	10/10 (100)
Country C	10/10 (100)
Country D	10/10 (100)
Country E	10/10 (100)
Country F	10/10 (100)
Country G	10/10 (100)
Country H	10/10 (100)
Country I	10/10 (100)
Country J	10/10 (100)
Country K	10/10 (100)
Country L	10/10 (100)
Country M	10/10 (100)
Country N	10/10 (100)
Country O	10/10 (100)
Country P	10/10 (100)
Country Q	10/10 (100)
Country R	10/10 (100)
Country S	10/10 (100)
Country T	10/10 (100)
Country U	10/10 (100)
Country V	3/10 (30)
Country W	3/10 (30)
Country X	10/10 (100)
Country Y	10/10 (100)
Country Z	10/10 (100)

For the country names, 'AA' to 'AC' were added to the standard English alphabet so the 29 EU/EEA countries that submitted data could be represented anonymously using letters.

The data of countries with missing values are plotted. Figure 68 illustrates the extent of missing data for countries V and W.

Figure 68. **Number of confirmed salmonellosis cases among countries with missing data, 2007–2016**



Step 4: Choose the approach for handling the missing data

The approach for handling the missing data is chosen based on the study objectives (step 1) and the nature of the missing data (step 3). Given that there are very large amounts of missing data and only for two countries, and as it is apparent that at least some of the missingness may be related to the missing value itself (the reporting ceased when there were many in-country outbreaks), it is decided together with a disease expert that the EU/EEA trend will be estimated excluding these two countries. The text of the report will indicate which countries are included and excluded, and the reasons for exclusion will be stated very clearly.

Step 5: Carry out a complete case analysis

The countries with missing values are excluded and a random-effects meta-analysis is conducted using the remaining countries.

As imputation is not needed in this instance, steps 6 and 7 will be skipped.

Step 8: Estimate the trend

In an earlier stage of this study, using the information in Part B of this document, it would have been decided to model a linear trend using linear regression. The trend estimate obtained is a statistically significant decrease of 57.9 salmonellosis notifications (95% CI: -72.0 to -43.8).

3.2 Example: Salmonellosis notifications across the EU/EEA, taking country into account (Missing data: multiple imputation)

Step 1: Define the study objectives

The study objective is to identify the change in salmonellosis notifications across the EU/EEA, taking individual country trends into account and using a random-effects meta-analysis.

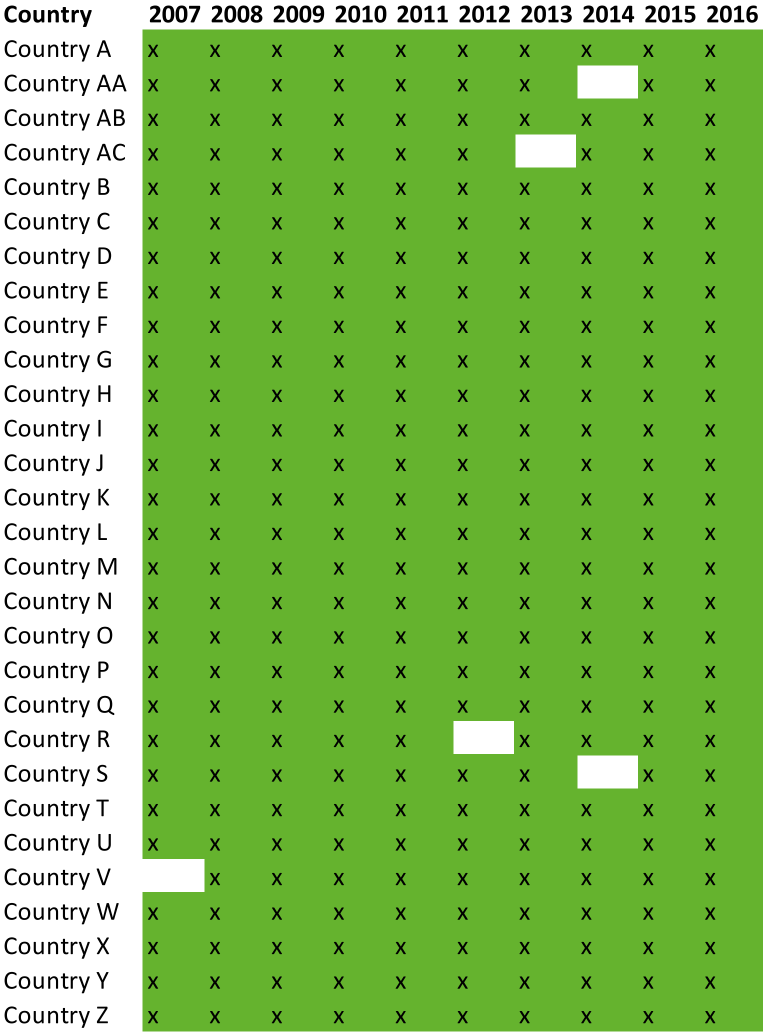
Step 2: Understand the data

A disease expert explains that the salmonellosis surveillance systems are reasonably homogeneous across the EU/EEA, although there are occasional gaps in reporting for countries. This is related to, for example, changeover of staff or public health emergencies not related to salmonellosis outbreaks/cases.

Step 3: Describe the missing data

The data are plotted by time and country, indicating where values are missing and not missing. Figure 69 illustrates that the missing values are few, with only one missing value per country. There does not appear to be a year that is particularly associated with missing values.

Figure 69. Observed and missing data of salmonellosis notifications by country and year, EU/EEA, 2007–2016



Country	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Country A	x	x	x	x	x	x	x	x	x	x
Country AA	x	x	x	x	x	x	x		x	x
Country AB	x	x	x	x	x	x	x	x	x	x
Country AC	x	x	x	x	x	x		x	x	x
Country B	x	x	x	x	x	x	x	x	x	x
Country C	x	x	x	x	x	x	x	x	x	x
Country D	x	x	x	x	x	x	x	x	x	x
Country E	x	x	x	x	x	x	x	x	x	x
Country F	x	x	x	x	x	x	x	x	x	x
Country G	x	x	x	x	x	x	x	x	x	x
Country H	x	x	x	x	x	x	x	x	x	x
Country I	x	x	x	x	x	x	x	x	x	x
Country J	x	x	x	x	x	x	x	x	x	x
Country K	x	x	x	x	x	x	x	x	x	x
Country L	x	x	x	x	x	x	x	x	x	x
Country M	x	x	x	x	x	x	x	x	x	x
Country N	x	x	x	x	x	x	x	x	x	x
Country O	x	x	x	x	x	x	x	x	x	x
Country P	x	x	x	x	x	x	x	x	x	x
Country Q	x	x	x	x	x	x	x	x	x	x
Country R	x	x	x	x	x		x	x	x	x
Country S	x	x	x	x	x	x	x		x	x
Country T	x	x	x	x	x	x	x	x	x	x
Country U	x	x	x	x	x	x	x	x	x	x
Country V		x	x	x	x	x	x	x	x	x
Country W	x	x	x	x	x	x	x	x	x	x
Country X	x	x	x	x	x	x	x	x	x	x
Country Y	x	x	x	x	x	x	x	x	x	x
Country Z	x	x	x	x	x	x	x	x	x	x

The green cells with an 'x' indicate observed data and the white cells indicate missing data. For the country names, 'AA' to 'AC' were added to the standard English alphabet so the 29 EU/EEA countries that submitted data could be represented anonymously using letters.

The amount of missing data, overall and by individual countries and years, is quantified. Overall, 2% of the data are missing and only four years have missing data (Table 23). Data are missing from five countries and each country has only one value missing over the study period (Table 24).

Table 23. Completeness of salmonellosis notifications overall and by year, EU/EEA, 2007–2016

Year	N/total (%)
Overall	**285/290 (98)**
Years with complete data	6/10 (60)
2007	28/29 (97)
2008	29/29 (100)
2009	29/29 (100)
2010	29/29 (100)
2011	29/29 (100)
2012	28/29 (97)
2013	28/29 (97)
2014	27/29 (93)
2015	29/29 (100)
2016	29/29 (100)

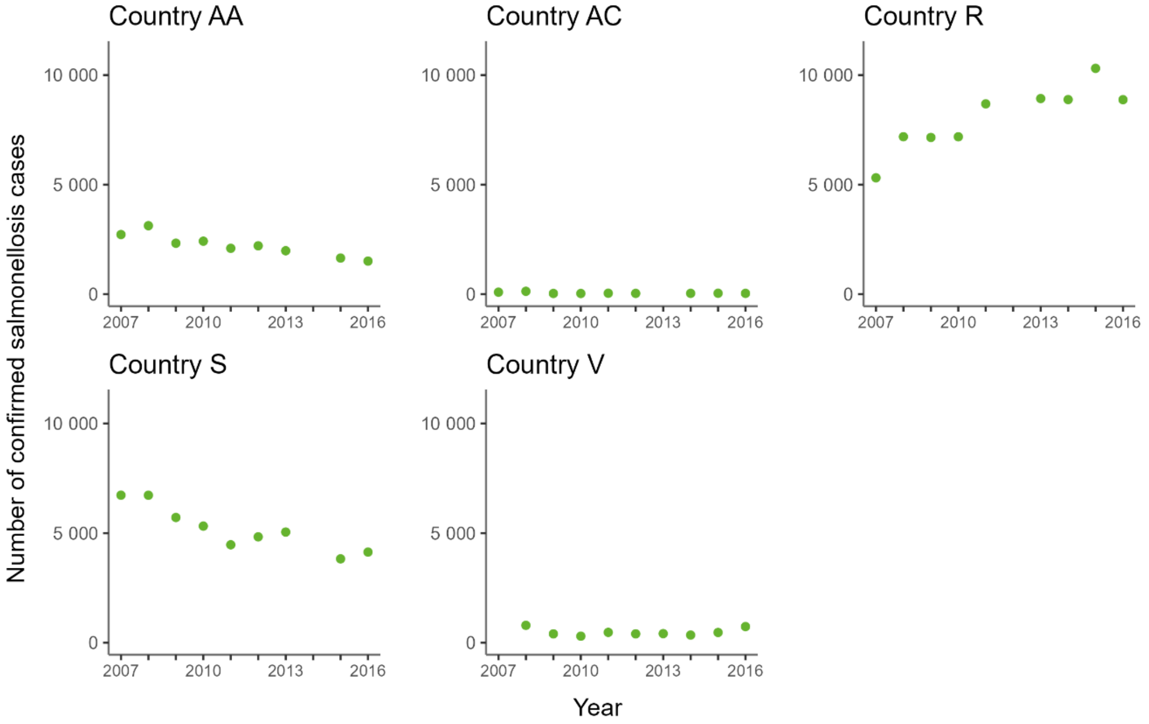
Table 24. Completeness of salmonellosis notifications by country, EU/EEA, 2007–2016

Country	N/total (%)
Countries with complete data	27/29 (93)
Country A	10/10 (100)
Country AA	9/10 (90)
Country AB	10/10 (100)
Country AC	9/10 (90)
Country B	10/10 (100)
Country C	10/10 (100)
Country D	10/10 (100)
Country E	10/10 (100)
Country F	10/10 (100)
Country G	10/10 (100)
Country H	10/10 (100)
Country I	10/10 (100)
Country J	10/10 (100)
Country K	10/10 (100)
Country L	10/10 (100)
Country M	10/10 (100)
Country N	10/10 (100)
Country O	10/10 (100)
Country P	10/10 (100)
Country Q	10/10 (100)
Country R	9/10 (90)
Country S	9/10 (90)
Country T	10/10 (100)
Country U	10/10 (100)
Country V	9/10 (90)
Country W	10/10 (100)
Country X	10/10 (100)
Country Y	10/10 (100)
Country Z	10/10 (100)

For the country names, 'AA' to 'AC' were added to the standard English alphabet so the 29 EU/EEA countries that submitted data could be represented anonymously using letters.

The data of countries with missing values are plotted. The plot does not suggest that the missing data occur at particular peaks or troughs, although this is hard to discern with yearly data (Figure 70).

Figure 70. Number of confirmed salmonellosis notifications among countries with missing data, 2007–2016



For the country names, 'AA' to 'AC' were added to the standard English alphabet so the 29 EU/EEA countries that submitted data could be represented anonymously using letters.

Step 4: Choose the approach for handling the missing data

Depending on the study objectives (step 1) and the nature of the missing data (step 3), an approach for handling the missing data is chosen. The missing data do not seem to be associated with the missing value itself (number of confirmed cases). The amount of missing data within countries is low (only one missing data point each for five countries), but a complete case analysis excluding the five countries with missing data may mean that the trend is no longer representative of the EU/EEA overall, which is part of the study objective. It is decided that a multiple imputation will be carried out, noting that the results of the complete case analysis will be reported alongside it.

Another option that could be considered is to ignore missing data and calculate the trend in those countries with missing data (see Part D, Section 1.3, 'Overview of methods to handle missing data'). This approach works for an adjusted pooled analysis (the meta-analysis approach), but not for a crude pooled analysis.

Step 5: Complete case analysis

Although the main analysis is a multiple imputation, a complete case analysis is carried out so that the results can be compared. The countries with missing values are excluded and a random-effects meta-analysis is carried out with the remaining countries. In an earlier stage of this study, using the information in Part A of this document, it was decided that a linear trend would be modelled using linear regression. The trend estimate obtained is a statistically significant decrease of 55.2 salmonellosis notifications (95% CI: -70.0 to -40.4).

Step 6: Carry out the multiple imputation

First, it must be decided whether rates or counts will be used for the multiple imputation. While rates are a more stable measure for imputation, it's decided that counts will be used, as it is indicated that the populations have been reasonably stable over time. In a secondary analysis (not shown here), the results of the imputation are compared using counts and rates.

The data are set for multiple imputation (this is often part of standard statistical software/packages) and are reshaped into wide format so that countries are in rows and years in columns. This means that the imputation will use years with observed values to impute the missing values. The imputation is carried out with the default values of five imputation databases and five nearest neighbours. The data are then reshaped back to long format (one record per country per year).

Step 7: Carry out diagnostics after the imputation

It is checked that there are five imputed values for the five countries that had missing values, which is the case (Table 25).

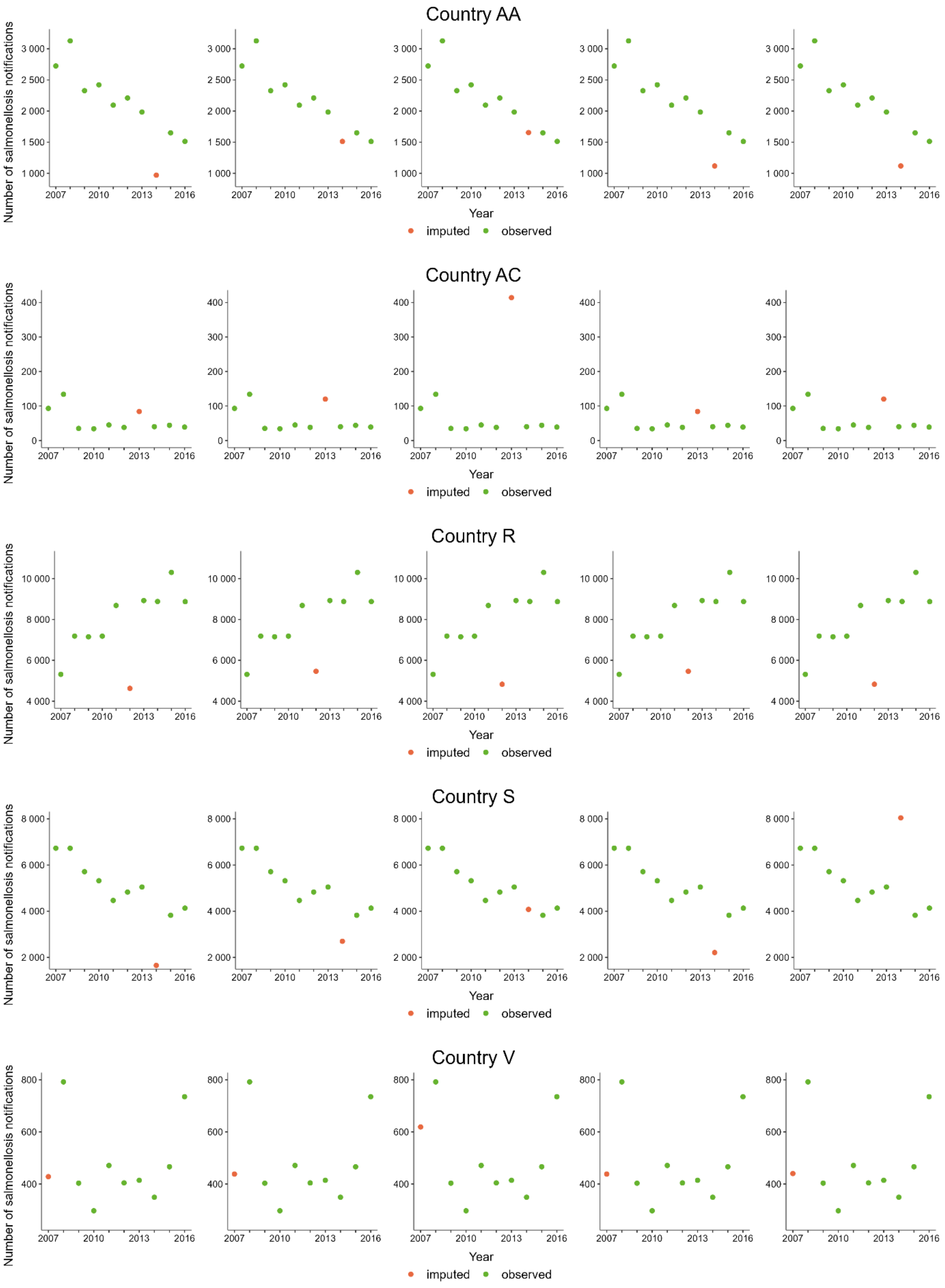
Table 25. The five imputed values for each country with missing data for a given year

Country	Year	Count	Multiple imputation number
Country AA	2014	970	1
Country AA	2014	1 512	2
Country AA	2014	1 654	3
Country AA	2014	1 118	4
Country AA	2014	1 118	5
Country AC	2013	84	1
Country AC	2013	120	2
Country AC	2013	414	3
Country AC	2013	84	4
Country AC	2013	120	5
Country R	2012	4 627	1
Country R	2012	5 462	2
Country R	2012	4 829	3
Country R	2012	5 462	4
Country R	2012	4 829	5
Country S	2014	1 654	1
Country S	2014	2 698	2
Country S	2014	4 078	3
Country S	2014	2 211	4
Country S	2014	8 042	5
Country V	2007	428	1
Country V	2007	438	2
Country V	2007	619	3
Country V	2007	438	4
Country V	2007	440	5

For the country names, 'AA' to 'AC' were added to the standard English alphabet so the 29 EU/EEA countries that submitted data could be represented anonymously using letters.

The observed and imputed data for the countries with missing values are then plotted (Figure 71). The majority of imputed data points look reasonable.

Figure 71. Observed and imputed values by country for each imputation dataset



For the country names, 'AA' to 'AC' were added to the standard English alphabet so the 29 EU/EEA countries that submitted data could be represented anonymously using letters.

Step 8: Estimate the trend

In an earlier stage of this study, using the information in Part A of this document, it was decided to model a linear trend using linear regression. As defined in step 1, a random-effects meta-analysis is carried out, ensuring that multiple imputation commands/functions are used when carrying out the analysis on the imputed data. This means that, when analysing the trend for each individual country, specific multiple imputation commands/functions are used to obtain the coefficient and standard error from the trend analysis. Then the coefficients and standard errors obtained using standard meta-analysis practices are combined. The trend estimate obtained is a statistically significant decrease of 52.6 salmonellosis notifications (95% CI: -66.5 to -38.8).

Step 9: Report the results

In the report, the results of the multiple imputation are presented as the main analysis and the assumptions and limitations of the imputation are discussed. The results of the complete case analysis, which are very similar here, are also reported.

Discussion and key points

This document provides guidance on carrying out trend analysis on surveillance data. Very often, the overarching objective of pooling data and carrying out trend analysis at ECDC is to understand the trend of a given disease over the past 10 years in the EU/EEA.

While simple in nature and frequently used in public health reports, trend analysis requires careful consideration and execution in order to make the most correct inferences around the data. Compared to data from a research study, surveillance data may be less meticulously collected and more subject to heterogeneities over time and between countries/regions. Surveillance datasets may also contain missing data. The objectives, strengths and limitations of the methodological approaches to trend analysis are outlined in this document, providing the reader with an understanding of how each choice can affect the analysis.

Key points for users of this guide to consider include:

- A trend analysis must have clearly defined objectives.
- The first step of any trend analysis is to make sure that you understand the data well, which includes undertaking discussions with data providers and disease experts.
- The next most important step of a trend analysis is to visualise the data.

References

1. European Centre for Disease Prevention and Control (ECDC). Long-term surveillance strategy 2014-2020 (revised). Stockholm: ECDC; 2018. Available at: http://dx.doi.org/10.2900/88780
2. Chatfield C. The analysis of time series: an introduction. Boca Raton: Chapman & Hall/CRC; 2004.
3. European Centre for Disease Prevention and Control (ECDC). Annual Epidemiological Reports (AERs). Stockholm: ECDC; 2023. Available at: https://www.ecdc.europa.eu/en/publications-data/monitoring/all-annual-epidemiological-reports
4. European Centre for Disease Prevention and Control (ECDC). Surveillance Atlas of Infectious Diseases. Stockholm: ECDC; 2023. Available at: https://atlas.ecdc.europa.eu/public/index.aspx
5. European Centre for Disease Prevention and Control (ECDC). Managing heterogeneity when pooling data from different surveillance systems. Stockholm: ECDC; 2019. Available at: https://doi.org/10.2900/83039
6. Kontopantelis E, Doran T, Springate DA, Buchan I, Reeves D. Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis. BMJ. 2015; 350:h2750. Available at: http://www.bmj.com/content/350/bmj.h2750.abstract
7. Bernal JL, Cummins S, Gasparrini A. Interrupted time series regression for the evaluation of public health interventions: a tutorial. Int J Epidemiol. 2016; 46(1):348-55. Available at: https://doi.org/10.1093/ije/dyw098
8. European Centre for Disease Prevention and Control (ECDC). Leptospirosis. 2022. In: Annual Epidemiological Report for 2017. Stockholm: ECDC; 2022. Available at: https://www.ecdc.europa.eu/en/publications-data/leptospirosis-annual-epidemiological-report-2017
9. Hyndman RJ, Athanasopoulos G. Forecasting: principles and practice. Melbourne: OTexts; 2021. Available at: https://otexts.com/fpp3/
10. Burnham KP, Anderson DR. Model selection and multimodel inference: a practical information-theoretic approach. 2nd ed. New York: Springer; 2010.
11. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane Handbook for Systematic Reviews of Interventions. London: Wiley; 2019. Available at: https://doi.org/10.1002/9781119536604
12. Cochran WG. The comparison of percentages in matched samples. Biometrika. 1950; 37(3-4):256-66. Available at: https://doi.org/10.1093/biomet/37.3-4.256
13. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002; 21(11):1539-58. Available at: https://doi.org/10.1002/sim.1186
14. Burke DL, Ensor J, Riley RD. Meta-analysis using individual participant data: one-stage and two-stage approaches, and why they may differ. Stat Med. 2017; 36(5):855-75. Available at: https://doi.org/10.1002/sim.7141
15. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to meta-analysis. Chichester: Wiley; 2009. Available at: https://doi.org/10.1002/9780470743386
16. George BJ, Aban IB. An application of meta-analysis based on DerSimonian and Laird method. J Nucl Cardiol. 2016 2016;23(4):690-2. Available at: https://doi.org/10.1007/s12350-015-0249-6
17. Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York: Wiley; 1987. Available at: https://doi.org/10.1002/9780470316696
18. Madley-Dowd P, Hughes R, Tilling K, Heron J. The proportion of missing data should not be used to guide decisions on multiple imputation. J Clin Epidemiol. 2019; 110:63-73. Available at: https://doi.org/10.1016/j.jclinepi.2019.02.016
19. Little RJA, Rubin DB. Statistical analysis with missing data. Hoboken: Wiley; 2002. Available at: https://doi.org/10.1002/9781119013563
20. Carpenter JR, Kenward MG. Multiple imputation and its application. Chichester: Wiley; 2013. Available at: https://doi.org/10.1002/9781119942283
21. Buuren Sv. Flexible imputation of missing data. 2nd ed. Boca Raton: Chapman & Hall/CRC; 2018.
22. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009; 338:b2393. Available at: https://doi.org/10.1136/bmj.b2393

European Centre for Disease Prevention and Control (ECDC)

Gustav III:s Boulevard 40, 16973 Solna, Sweden

Tel. +46 858601000
Fax +46 858601001
www.ecdc.europa.eu

An agency of the European Union
www.europa.eu

Subscribe to our publications
www.ecdc.europa.eu/en/publications

Contact us
publications@ecdc.europa.eu

Follow us on Twitter
@ECDC_EU

Like our Facebook page
www.facebook.com/ECDC.EU

Publications Office
of the European Union