SOFTWARE TOOL ARTICLE

Epidemic curves made easy using the R package *incidence* [version 1; peer review: 1 approved, 2 approved with reservations]

Zhian N. Kamvar [1], Jun Cai [2], Juliet R.C. Pulliam [3], Jakob Schumacher[4], Thibaut Jombart [1,5]

[1]MRC Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology , School of Public Health, Imperial College London, London, UK
[2]Ministry of Education Key Laboratory for Earth System Modelling, Department of Earth System Science, Tsinghua University, Beijing, 100084, China
[3]South African DST-NRF Centre of Excellence in Epidemiological Modelling and Analysis (SACEMA),, Stellenbosch University, Stellenbosch, South Africa
[4]Gesundheitsamt Reinickendorf, Berlin, Germany
[5]Department of Infectious Disease Epidemiology, London School of Hygiene & Tropical Medicine, London, UK

v1 **First published:** 31 Jan 2019, **8**:139 (https://doi.org/10.12688/f1000research.18002.1)
Latest published: 31 Jan 2019, **8**:139 (https://doi.org/10.12688/f1000research.18002.1)

Abstract
The epidemiological curve (epicurve) is one of the simplest yet most useful tools used by field epidemiologists, modellers, and decision makers for assessing the dynamics of infectious disease epidemics. Here, we present the free, open-source package incidence for the R programming language, which allows users to easily compute, handle, and visualise epicurves from unaggregated linelist data. This package was built in accordance with the development guidelines of the R Epidemics Consortium (RECON), which aim to ensure robustness and reliability through extensive automated testing, documentation, and good coding practices. As such, it fills an important gap in the toolbox for outbreak analytics using the R software, and provides a solid building block for further developments in infectious disease modelling. *incidence* is available from https://www.repidemicsconsortium.org/incidence.

Keywords
epicurve, incidence, epidemics, outbreaks, R

This article is included in the RPackage gateway.

This article is included in the R Epidemics Consortium (RECON) collection.

Open Peer Review

Referee Status: ? ✓ ?

	Invited Referees 1	2	3
version 1 published 31 Jan 2019	? report	✓ report	? report

1 **Benjamin M. Bolker**, McMaster University, Canada

2 **Quirine ten Bosch**, Wageningen University and Research Centre, The Netherlands

3 **Bertrand Sudre**, European Centre for Disease Prevention and Control (ECDC), Sweden

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding authors: Zhian N. Kamvar (zkamvar@gmail.com), Thibaut Jombart (thibautjombart@gmail.com)

Author roles: Kamvar ZN: Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Cai J**: Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Pulliam JRC**: Software, Validation, Writing – Review & Editing; **Schumacher J**: Software, Writing – Review & Editing; **Jombart T**: Conceptualization, Funding Acquisition, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The authors acknowledge financial support from the Global Challenges Research Fund (GCRF) for the project 'RECAP – research capacity building and knowledge generation to support preparedness and response to humanitarian crises and epidemics' managed through RCUK and ESRC (ES/P010873/1), from the UK Public Health Rapid Support Team, which is funded by the United Kingdom Department of Health and Social Care, and from the National Institute for Health Research - Health Protection Research Unit for Modelling Methodology.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Kamvar ZN, Cai J, Pulliam JRC *et al.* **Epidemic curves made easy using the R package *incidence* [version 1; peer review: 1 approved, 2 approved with reservations]** F1000Research 2019, **8**:139 (https://doi.org/10.12688/f1000research.18002.1)

First published: 31 Jan 2019, **8**:139 (https://doi.org/10.12688/f1000research.18002.1)

Introduction

Responses to infectious disease epidemics use a growing body of data sources to inform decision making (Cori *et al.*, 2017; Fraser *et al.*, 2009; WHO Ebola Response Team *et al.*, 2014; WHO Ebola Response Team *et al.*, 2015). While new data—such as whole genome pathogen sequences—are increasingly useful complements to epidemiological data (Gire *et al.*, 2014), epidemic curves—which describe the number of new cases through time (incidence)—remain the most important source of information, particularly early in an outbreak. Specifically epidemic curves(often referred to as 'epicurves') represent the number of new cases per time unit based on the date or time of symptom onset.

While conceptually simple, epicurves are useful in many respects. They provide a simple, visual outline of epidemic dynamics, which can be used for assessing the growth or decline of an outbreak (Barrett *et al.*, 2016; Fitzgerald *et al.*, 2014; Jernberg *et al.*, 2015; Lanini *et al.*, 2014; Nhan *et al.*, 2018) and therefore informing intervention measures (Meltzer *et al.*, 2014; WHO Ebola Response Team *et al.*, 2014; WHO Ebola Response Team *et al.*, 2015). In addition, epicurves also form the raw material used by a range of modelling techniques for short-term forecasting (Cori *et al.*, 2013; Funk *et al.*, 2018; Nouvellet *et al.*, 2018; Viboud *et al.*, 2018) as well as in outbreak detection algorithms from syndromic surveillance data (Farrington & Andrews, 2003; Unkel *et al.*, 2012).

Because of the increasing need to analyse various types of epidemiological data in a single environment using free, transparent and reproducible procedures, the R software (R Core Team, 2017) has been proposed as a platform of choice for epidemic analysis (Jombart *et al.*, 2014). But despite the existence of packages dedicated to time series analysis (Shumway & Stoffer, 2010) as well as surveillance data (Höhle, 2007), a lightweight and *well-tested* package solely dedicated to building, handling and plotting epidemic curves directly from linelist data (e.g. a spreadsheet where each row represents an individual case) is still lacking.

Here, we introduce *incidence*, an R package developed as part of the toolbox for epidemics analysis of the R Epidemics Consortium (RECON) which aims to fill this gap. In this paper, we outline the package's design and illustrate its functionalities using a reproducible worked example.

Methods

Package overview

The philosophy underpinning the development of *incidence* is to 'do the basics well'. The objective of this package is to provide simple, user-friendly and robust tools for computing, manipulating, and plotting epidemic curves, with some additional facilities for basic models of incidence over time.

The general workflow (Figure 1) revolves around a single type of object, formalised as the S3 class **incidence**. **incidence** objects are lists storing separately a matrix of case counts (with dates in rows and groups in columns), dates used as breaks, the time interval used, and an indication of whether incidence is cumulative or not (Figure 1). The **incidence** object is obtained by running the function `incidence()` specifying two inputs: a vector of dates (representing onset of individual cases) and an interval specification. The dates can be any type of input representing dates including **Date** and **POSIXct** objects, as well as numeric and integer values. The dates are aggregated into counts based on the user-defined interval representing the number of days for each bin. The interval can also be defined as a text string of either "week", "month", "quarter", or "year" to represent intervals that can not be defined by a fixed number of days. For these higher-level intervals, an extra parameter—`standard`—is available to specify if the interval should start at the standard beginning of the interval (e.g. weeks start on Monday and months start at the first of the month). `incidence()` also accepts a `groups` argument which can be used to obtain stratified incidence. The basic elements of the **incidence** object can be obtained by the accessors `get_counts()`, `get_dates()`, and `get_interval()`.

This package facilitates the manipulation of **incidence** objects by providing a set of handler functions for the most common tasks. The function `subset()` can be used for isolating case data from a specific time window and/or groups, while the [operator can be used for a finer control to subset dates and groups using integer, logical or character vectors. This is accomplished by using the same syntax as for matrix and data.frame objects, i.e. `x[i, j]` where `x` is the **incidence** object, and `i` and `j` are subsets of dates and groups, respectively.

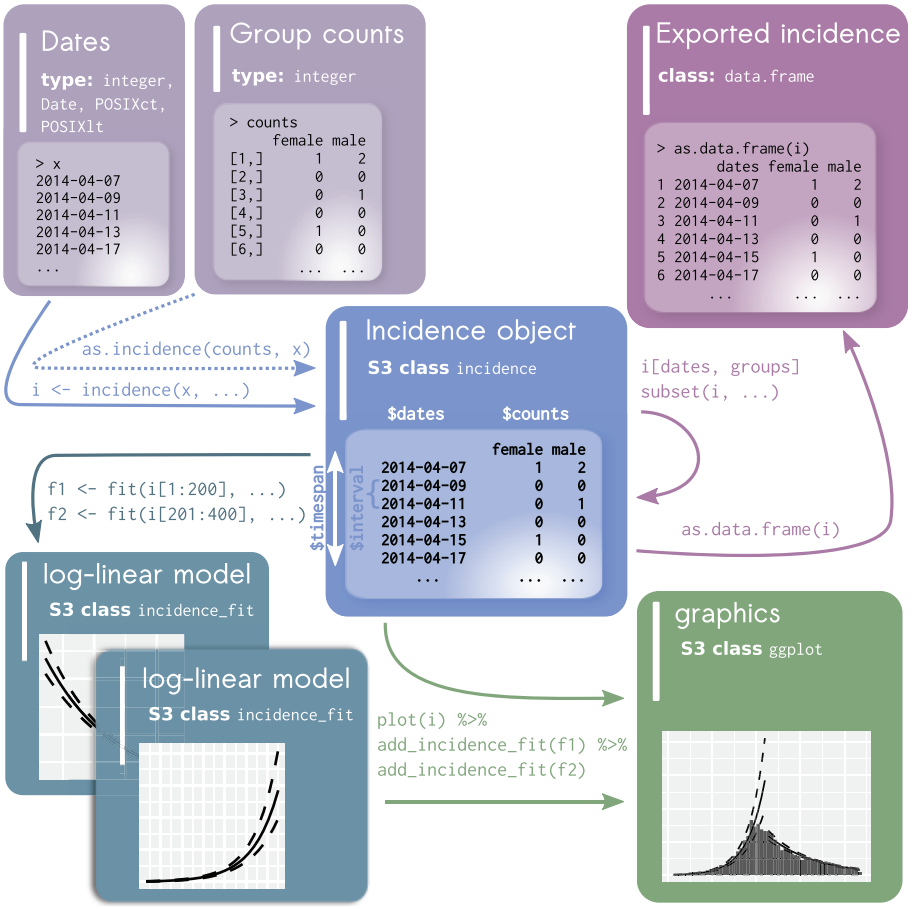

Figure 1. Generalized workflow from incidence object construction to modeling and visualization. The raw data is depicted in the top left as either a vector of dates for each individual case (typical usage) or a combination of both dates and a matrix of group counts. The incidence object is created from these where it checks and validates the timespan and interval between dates. Data subsetting and export is depicted in the upper right. Data visualization is depicted in the lower right. Addition of log-linear models is depicted in the lower left.

The function `pool()` can be used to merge several groups into one, and the function `cumulate()` will turn incidence data into cumulative incidence. To maximize interoperability, **incidence** objects can also be exported to either a matrix using `get_counts()` or a data.frame using `as.data.frame()`, including an option for a 'long' format which is readily compatible with *ggplot2* (Wickham, 2016) for further customization of graphics.

In line with RECON's development guidelines, the *incidence* package is thoroughly tested via automatic tests implemented using *testthat* (Wickham, 2011), with an overall coverage nearing 100% at all times. We use the continuous integration services travis.ci and appveyor to ensure that new versions of the code maintain all existing functionalities and give expected results on known datasets, including matching reference graphics tested using the visual regression testing implemented in *vdiffr* (Henry *et al.*, 2018). Overall, these practices aim to maximise the reliability of the package, and its sustainable development and maintenance over time.

Modeling utilities

Many different approaches can be used to model, and possibly derive predictions from incidence data (e.g. Cori *et al.*, 2013; Nouvellet *et al.*, 2018; Wallinga & Teunis, 2004), and are best implemented in separate packages (e.g. Cori *et al.*, 2013). Here, we highlight three simple functionalities in *incidence* for estimating parameters via modeling or bootstrap and the two specialized data classes that are used to store the models and parameter estimates.

As a basic model, we implement the simple log-linear regression approach in the function `fit()`, which can be used to fit exponential increase or decrease of incidence over time by log-transforming case counts and applying a linear regression on these transformed data. The log-linear regression model is of the form $log(y) = r \times t + b$ where

y is the incidence, *r* is the growth rate, *t* is the number of days since the start of the outbreak, and *b* is the intercept. This approach estimates a growth rate *r* (the slope of the regression), which can in turn be used for estimating the doubling or halving time of the epidemic, and with some knowledge of the serial interval, for approximating the reproduction number, R_0 (Wallinga & Lipsitch, 2007).

In the presence of both growing and decreasing phases of an epidemic, the date representing the peak of the epidemic can be estimated. In *incidence*, this can be done in two ways. The function `estimate_peak()` uses multinomial bootstrapping to estimate the peak, assuming that a) reporting is constant over time, b) the total number of cases is known, and c) the bootstrap never samples zero-incidence days. This function returns the estimated peak with a confidence interval along with the boostrap estimates. Alternatively, the function `fit_optim_split()` can be used to detect the optimal turning point of the epidemic and fit two separate models on either side of the peak. This is done by maximizing the combined mean adjusted R^2 value from the two models (Figure 1, Figure 5).

The `fit()` function returns an **incidence_fit** object and the `fit_optim_split()` function returns an **incidence_fit_list** object, which is a specialized object designed to contain an unlimited number of (potentially nested) **incidence_fit** objects. While the *incidence* package returns **incidence_fit** objects containing log-linear models by default, they can be constructed from any model from which it's possible to extract the growth rate (*r*) and predict incidence along the model. Both object classes can be plotted separately or added to an existing epicurve using the function `add_incidence_fit()` (Figure 5).

Operation

The minimal system requirements for successful operation of this package is R version 3.1.

Use cases

Two worked examples are used to demonstrate the functionality and flexibility of the *incidence* package. The first example illustrates how to compute and manipulate stratified weekly incidence directly from a line-list, while the second example shows how to import pre-computed daily incidence and fit a log-linear model to estimate growth rate (*r*) and doubling time for the growing phase[1].

Example 1: computing and manipulating stratified weekly incidence

In this first example, we use the dataset `ebola_sim_clean` in the *outbreaks* package, which provides a linelist for a fictitious outbreak of Ebola Virus Disease (EVD) that matches some key epidemiological properties (e.g. serial intervals, reproduction numbers) of the West African Ebola outbreak of 2014–2015 (WHO Ebola Response Team *et al.*, 2014).

1) Importing data

First, we load the dataset `ebola_sim_clean` from the *outbreaks* package. The dataset contains 5,829 cases of 9 variables, among which the date of symptom onset (`$date_of_onset`) and the name of the hospital (`$hospital`) are used for computing the weekly epicurves stratified by hospitals.

```
library('outbreaks')

dat1 <- ebola_sim_clean$linelist
str(dat1, strict.width = "cut", width = 76)

## 'data.frame':      5829 obs.  of  9 variables:
##  $ case_id                  : chr  "d1fafd" "53371b" "f5c3d8" "6c286a" ...
##  $ generation               : int  0 1 1 2 2 0 3 3 2 3 ...
##  $ date_of_infection        : Date, format: NA "2014-04-09" ...
##  $ date_of_onset            : Date, format: "2014-04-07" "2014-04-15" ...
##  $ date_of_hospitalisation  : Date, format: "2014-04-17" "2014-04-20" ...
##  $ date_of_outcome          : Date, format: "2014-04-19" NA ...
##  $ outcome                  : Factor w/ 2 levels "Death","Recover": NA NA 2 ..
##  $ gender                   : Factor w/ 2 levels "f","m": 1 2 1 1 1 1 1 1 2 ..
##  $ hospital                 : Factor w/ 5 levels "Connaught Hospital",..: 2 ..
```

[1]Negative values of *r* in incidence are reported as halving times instead of doubling times and decreasing phase instead of growing phase

2) Building the incidence object

The weekly incidence stratified by hospitals is computed by running the function `incidence()` on the Date variable `dat1$date_of_onset` with the arguments `interval = 7` and `groups = dat1$hospital`. The **incidence** object `i.7.group` is a list with class of **incidence** for which several generic methods are implemented, including `print.incidence()` and `plot.incidence()`. Typing **incidence** object `i.7.group` implicitly calls the specific function `print.incidence()` and prints out the summary of the data and its list components. The 5,829 cases (the total number of cases stored in the `$n` component) with dates of symptom onset ranging from 2014-04-07 to 2015-04-27 (spanning from 2014-W15 to 2015-W18 in terms of the ISO 8601 standard for representing weeks) are used for building the **incidence** object `i.7.group`. The `$counts` component contains the actual incidence for defined bins, which is a matrix with one column per group. Here `$count` is a matrix with 56 rows and 6 columns as groups by hospital with 6 factor levels are specified. The bin size in number of days is stored in the `$interval` component. In this example, 7 days suggests that weekly incidence is computed, while by default, daily incidence is computed with the argument `interval = 1`. The `$dates` component contains all the dates marking the left side of the bins, in the format of the input data (e.g. Date, integer, etc.). The `$timespan` component stores the length of time (in days) for which incidence is computed. The `$cumulative` component is a logical indication whether incidence is cumulative or not.

The generic `plot()` method for **incidence** objects calls the specific function `plot.incidence()`, which makes an incidence barplot using the *ggplot2* package. Hence, customization of *incidence* plot can benefit from the powerful graphical language from *ggplot2*.

```
library('incidence')
library('ggplot2')

# compute weekly stratified incidence
i.7.group <- incidence(dat1$date_of_onset, interval = 7, groups = dat1$hospital)
# print incidence object
i.7.group

## <incidence object>
## [5829 cases from days 2014-04-07 to 2015-04-27]
## [5829 cases from ISO weeks 2014-W15 to 2015-W18]
## [6 groups: Connaught Hospital, Military Hospital, other,
##  Princess Christian Maternity Hospital (PCMH), Rokupa Hospital, NA]
##
## $counts: matrix with 56 rows and 6 columns
## $n: 5829 cases in total
## $dates: 56 dates marking the left-side of bins
## $interval: 7 days
## $timespan: 386 days
## $cumulative: FALSE

# plot incidence object
my_theme <- theme_bw(base_size = 12) +
  theme(panel.grid.minor = element_blank()) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5, color = "black"))

plot(i.7.group, border = "white") +
  my_theme +
  theme(legend.position = c(0.8, 0.75))
```

Note that when weekly incidence is computed from dates, like in this example, the ISO 8601 standard weeks are used by default with the argument `standard = TRUE` in the `incidence()` function. Under this situation, an extra component of `$isoweek` is added to the **incidence** object `i.7.group` to store those weeks in the ISO 8601 standard week format "yyyy-Www", and the `$dates` component stores the corresponding first days of those ISO weeks. Meanwhile the x-axis tick labels of the weekly *incidence* plot are in the ISO week format "yyyy-Www" (see Figure 2) rather than in the date format "yyyy-mm-dd" as the argument `labels_iso_week` in the `plot()` function is by default `TRUE` when plotting the ISO week-based **incidence** objects.

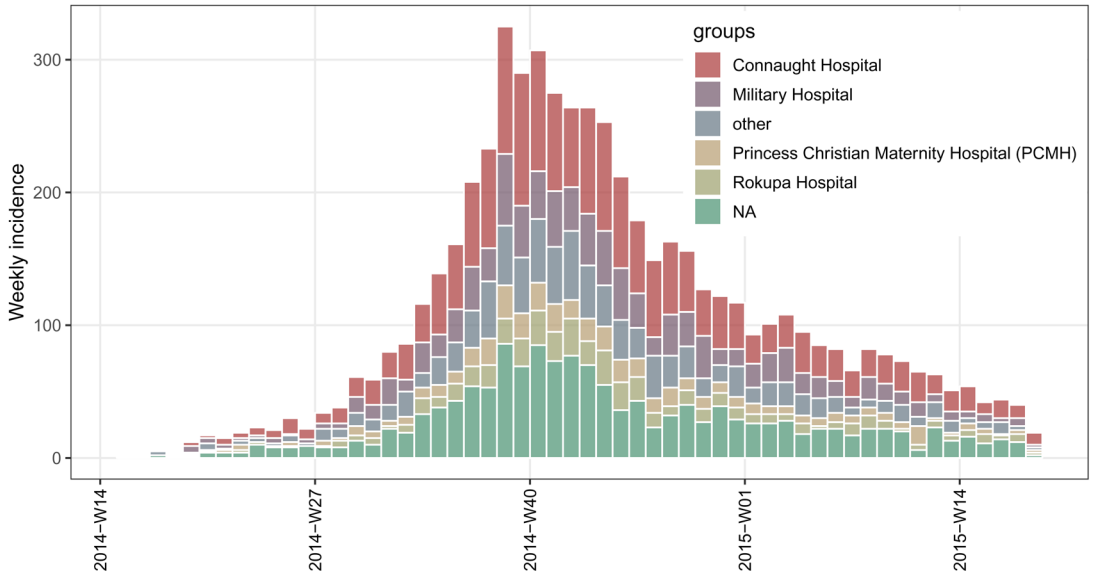

Figure 2. Weekly epicurves stratified by hospitals for the simulated outbreak of EVD.

3) Manipulate the incidence object

In the above visualisation, it can be difficult to see what the dynamics were in the early stages of the epidemic. If we want to see the first 18 weeks of the outbreak in the four major hospitals, we can use the [operator to subset the rows and columns, which represent weeks and hospitals, respectively, in this particular **incidence** object.

```
# plot the first 18 weeks, defined hospitals, and use different colors
i.7.sub <- i.7.group[1:18, grep("Hospital", group_names(1.7.group))]
hosp_colors <- c("#899DA4", "#C93312", "#FAEFD1", "#DC863B")
plot(i.7.sub, show_cases = TRUE, border = "black", color = hosp_colors) +
  my_theme +
  theme(legend.position = c(0.35, 0.8))
```

Here, because of the few numbers of cases in the first few weeks, we have also highlighted each case using `show_cases = TRUE` (Figure 3). We've also used a different color palette to differentiate between the subsetted data and the full data set.

As shown in Figure 2, the missing hospital name (NA) is treated as a separate group, resulting from the default of the argument `na_as_group = TRUE` in the `incidence()` function. This argument can be set to `FALSE` to not include data with missing groups in the object.

Example 2: importing pre-computed daily incidence and fitting log-linear model

The datasets `zika_girardot_2015` and `zika_sanandres_2015` used in the second example are also from the *outbreaks* package. These datasets describe the daily incidence of Zika virus disease (ZVD) in, respectively, Girardot and San Andres Island, Colombia from September 2015 to January 2016. For details on these datasets, please refer to Rojas *et al.* (2016).

1) Import pre-computed daily incidence

`zika_girardot_2015` and `zika_sanandres_2015` are data frames with the same variables `date` and `cases`. In order to obtain a more complete picture of the epidemic dynamics of ZVD in Colombia, we merge these two data.frames into a single one, `dat2`, by variable `date`. As `dat2` is already pre-computed daily incidence rather than a vector of dates such as those in example 1, we can directly convert it into an **incidence** object grouped by geographical locations, `i.group`, by using the `as.incidence()` function. This shows the flexibility of the *incidence* package in making **incidence** objects. Using the `pool()` function, the daily incidence stratified by locations, `i.group`, can be collapsed into an incidence object without groups, `i.pooled`. The stratified and pooled

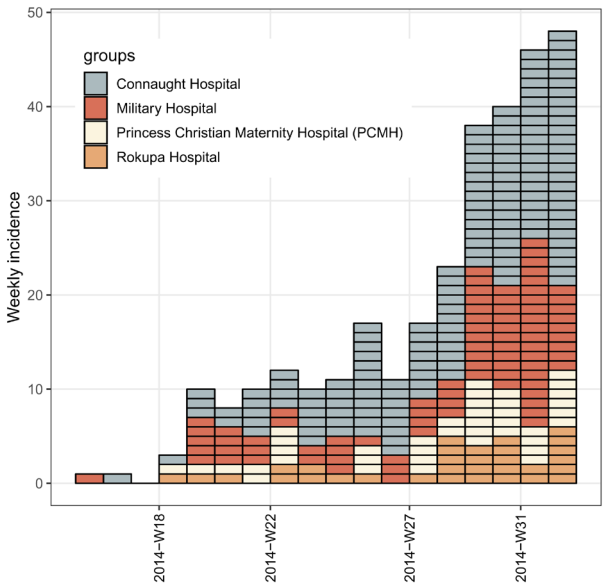

Figure 3. Weekly epicurves stratified by hospitals representing the first eight weeks of simulated outbreak of EVD.

daily incidence plots of ZVD in Colombia are shown in Figure 4, from which we can see that the epidemic of ZVD occurred earlier in San Andres Island than in Girardot.

```
# preview datasets
head(zika_girardot_2015, 3)
```

```
##         date cases
## 1 2015-10-19     1
## 2 2015-10-22     2
## 3 2015-10-23     1
```

```
head(zika_sanandres_2015, 3)
```

```
##         date cases
## 1 2015-09-06     1
## 2 2015-09-07     1
## 3 2015-09-08     1
```

```
# combine two datasets into one
dat2 <- merge(zika_girardot_2015, zika_sanandres_2015, by = "date", all = TRUE)

# rename variables
names(dat2)[2:3] <- c("Girardot", "San Andres")

# replace NA with 0
dat2[is.na(dat2)] <- 0

# convert pre-computed incidence in data.frame into incidence object
# grouped by locations
i.group <- as.incidence(x = dat2[, 2:3], dates = dat2$date)

# pool incidence across two locations
i.pooled <- pool(i.group)
plot(i.group, border = "white")   + my_theme + theme(legend.position = c(0.9, 0.7))
plot(i.pooled, border = "white") + my_theme
```

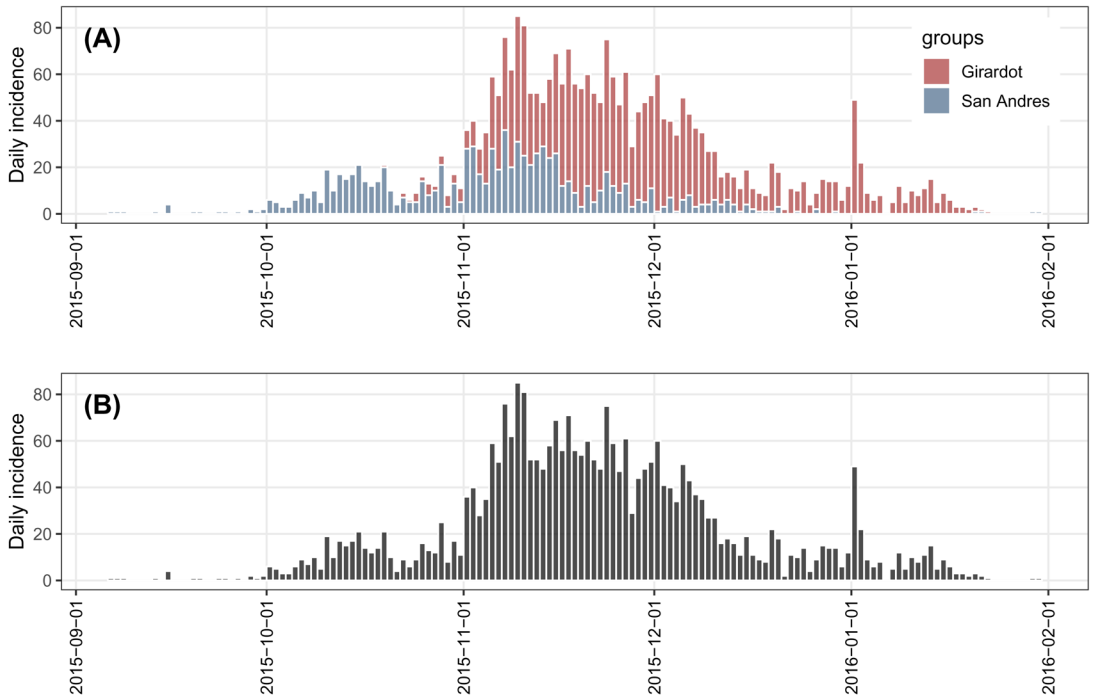

Figure 4. (**A**) stratified and (**B**) pooled daily incidence plots of ZVD in Colombia, September 2015 to January 2016.

As shown in Figure 4B, the pooled daily incidence in Colombia shows approximately exponential phases before and after the epidemic peak. Therefore, we fit two log-linear regression models around the peak to characterize the epidemic dynamics of ZVD in Colombia. Such models can be separately fitted to the two phases of the epicurve of `i.pooled` using the `fit()` function, which, however, requires us to know what date should be used to split the epicurve in two phases (see the argument `split` in the `fit()` function). Without any knowledge on the splitting date, we can turn to the `fit_optim_split()` function to look for the optimal splitting date (i.e. the one maximizing the average fit of both models) and then fit two log-linear regression models before and after the optimal splitting date.

```
library('magrittr')

fos <- fit_optim_split(i.pooled)
fos$split

## [1] "2015-11-15"

fos$fit

## <list of incidence_fit objects>
##
## attr(x, 'locations'): list of vectors with the locations of each incidence_fit object
##
## 'before'
## 'after'
##
## $model: regression of log-incidence over time
##
## $info: list containing the following items:
##   $r (daily growth rate):
##      before       after
##  0.06659200 -0.04813045
```

```
##
##  $r.conf (confidence interval):
##              2.5 %      97.5 %
## before 0.05869968 0.07448432
## after -0.05440018 -0.04186071
##
##   $doubling (doubling time in days):
##   before
## 10.40887
##
##   $doubling.conf (confidence interval):
##           2.5 %   97.5 %
## before 9.305948 11.80836
##
##   $halving (halving time in days):
##    after
## 14.40143
##
##   $halving.conf (confidence interval):
##          2.5 %   97.5 %
## after 12.74163 16.55842
##
##   $pred: data.frame of incidence predictions (129 rows, 6 columns)
```

```
plot(i.pooled, border = "white") %>%
  add_incidence_fit(fos$fit) +
  my_theme
```

The returned object `fos` is a list with 4 components. The `$split` component suggests that the optimal splitting date is 2015-11-15. The `$fit` component is an **incidence_fit_list** containing two **incidence_fit** objects named 'before' and 'after'. These each contain the information extracted from the fitted log-linear regression models. Printing the `$fit` component shows a daily growth rate *r* of 0.067 and its 95% confidence interval (CI) ([0.059, 0.074]), and a doubling time of 10.4 days (95% CI, [9.31, 11.8]) during the first phase, and a daily decreasing rate *r* of -0.048 (95% CI, [-0.054, -0.042]), and a halving time of 14.4 days (95% CI, [12.7, 16.6]) during the second.

The predictions and their 95% CIs from the two **incidence_fit** objects, 'before' and 'after', can be added to the existing incidence plot of `i.pooled` using the piping-friendly function `add_incidence_fit()`. As shown in Figure 5, based on visual comparison of models and data, these two log-linear regression models provide a decent approximation for the actual dynamics of the epidemic (adjusted R^2 = 0.83 and 0.77 for the increasing and decreasing phases, respectively).

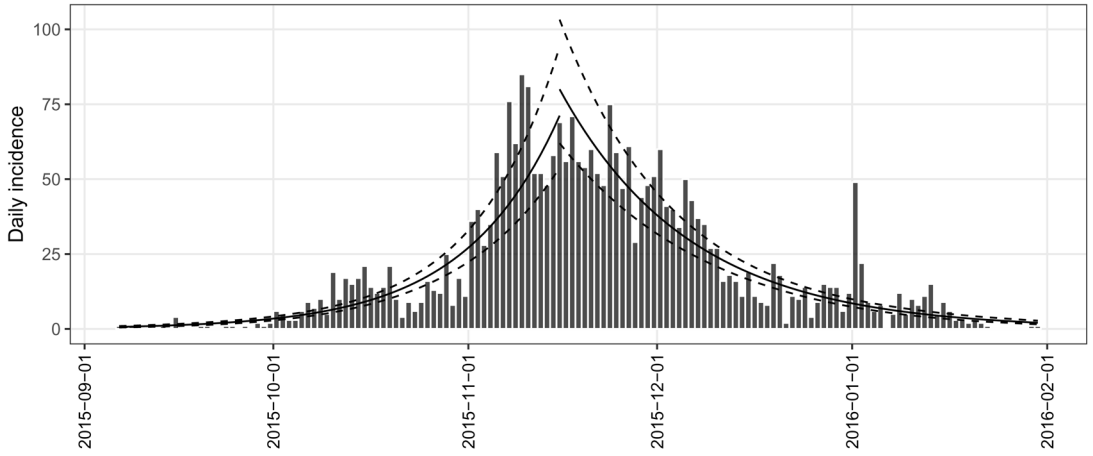

Figure 5. Fit two log-linear regression models, before and after the optimal splitting date.

Conclusion

This article has described the package *incidence* and its features—which include three lightweight data classes and utilities for data manipulation, plotting, and modeling. We have shown that an **incidence** object can flexibly be defined at different datetime intervals with any number of stratifications and be subset by groups or dates. The most important aspects of this package are use-ability and interoperability. For both field epidemiologists and academic modellers, the data received are often in the form of line-lists where each row represents a single case. We have shown that these data can easily be converted to an **incidence** object and then plotted with sensible defaults in two lines of code.

We have additionally shown that because the data are aggregated into a matrix of counts, it becomes simple to perform operations related to peak-finding, model-fitting, and exportation (e.g. using `as.data.frame()`) into different formats. Thus, because it has built-in tools for aggregation, visualisation, and model fitting, the *incidence* package is ideal for rapid generation of reports and estimates in outbreak response situations where time is a critical factor.

Software availability

incidence available from: https://www.repidemicsconsortium.org/incidence Code to reproduce all figures can be found by running demo ("incidence-demo", package = "incidence") from the R console with the incidence package installed.

Source code available from: https://github.com/reconhub/incidence

Archived source code as at time of publication: https://doi.org/10.5281/zenodo.2540217 (Jombart *et al.*, 2019)

Software license: MIT

Data availability

Underlying data

Datasets used in the worked examples are from the *outbreaks* package:

ebola_sim_clean: https://github.com/reconhub/outbreaks/blob/master/data/ebola_ sim_clean.RData

zika_girardot_2015: https://github.com/reconhub/outbreaks/blob/master/data/zika_ girardot_2015.RData

zika_sanandres_2015: https://github.com/reconhub/outbreaks/blob/master/data/zika_sanandres_2015.RData

Grant information

The authors acknowledge financial support from the Global Challenges Research Fund (GCRF) for the project 'RECAP – research capacity building and knowledge generation to support preparedness and response to humanitarian crises and epidemics' managed through RCUK and ESRC (ES/P010873/1), from the UK Public Health Rapid Support Team, which is funded by the United Kingdom Department of Health and Social Care, and from the National Institute for Health Research - Health Protection Research Unit for Modelling Methodology.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

We would like to thank Michael Höhle for discussion about the caveats for `estimate_peak()`, the R developer community for constantly improving our working environment, github for hosting our project, travis, appveyor and codecov for providing free continuous integration resources, and the RECON community.

References

Barrett P, Chaintarli K, Ryan F, *et al.*: **An ongoing measles outbreak linked to a suspected imported case, Ireland, April to June 2016.** *Euro Surveill.* 2016; **21**(27).
PubMed Abstract | Publisher Full Text

Cori A, Donnelly CA, Dorigatti I, *et al.*: **Key data for outbreak evaluation: building on the Ebola experience.** *Philos Trans R Soc Lond B Biol Sci.* 2017; **372**(1721): pii: 20160371.
PubMed Abstract | Publisher Full Text | Free Full Text

Cori A, Ferguson NM, Fraser C, *et al.*: **A new framework and software to estimate time-varying reproduction numbers during epidemics.** *Am J Epidemiol.* 2013; **178**(9): 1505–1512.
PubMed Abstract | Publisher Full Text | Free Full Text

Farrington P, Andrews N: ***Outbreak detection*: Application to infectious disease surveillance**. In *Monitoring the Health of Populations.* Oxford University Press, New York. 2003.
Publisher Full Text

Fitzgerald M, Thornton L, O'Gorman J, *et al.*: **Outbreak of hepatitis A infection associated with the consumption of frozen berries, Ireland, 2013--linked to an international outbreak.** *Euro Surveill.* 2014; **19**(43): pii: 20942.
PubMed Abstract | Publisher Full Text

Fraser C, Donnelly CA, Cauchemez S, *et al.*: **Pandemic potential of a strain of influenza A (H1N1): early findings.** *Science.* 2009; **324**(5934): 1557–1561.
PubMed Abstract | Publisher Full Text | Free Full Text

Funk S, Camacho A, Kucharski AJ, *et al.*: **Real-time forecasting of infectious disease dynamics with a stochastic semi-mechanistic model.** *Epidemics.* 2018; **22**: 56–61.
PubMed Abstract | Publisher Full Text | Free Full Text

Gire SK, Goba A, Andersen KG, *et al.*: **Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak.** *Science.* 2014; **345**(6202): 1369–1372.
PubMed Abstract | Publisher Full Text | Free Full Text

Henry L, Sutherland C, Hong D: **vdiffr: Visual regression testing and graphical diffing**. 2018.
Reference Source

Höhle M: **surveillance: An R package for the monitoring of infectious diseases.** *Comput Stat.* 2007; **22**(4): 571–582.
Publisher Full Text

Jernberg C, Hjertqvist M, Sundborger C, *et al.*: **Outbreak of Salmonella Enteritidis phage type 13a infection in Sweden linked to imported dried-vegetable spice mixes, December 2014 to July 2015.** *Euro Surveill.* 2015; **20**(30): pii: 21194.
PubMed Abstract | Publisher Full Text

Jombart T, Aanensen DM, Baguelin M, *et al.*: **OutbreakTools: a new platform for disease outbreak analysis using the R software.** *Epidemics.* 2014; **7**(0): 28–34.
PubMed Abstract | Publisher Full Text | Free Full Text

Jombart T, Kamvar ZN, Cai J, *et al.*: **reconhub/incidence 1.5 (Version 1.5).** *Zenodo.* 2019.
http://www.doi.org/10.5281/zenodo.2540217

Lanini S, Capobianchi MR, Puro V, *et al.*: **Measles outbreak on a cruise ship in the western Mediterranean, February 2014, preliminary report.** *Euro Surveill.* 2014; **19**(10): pii: 20735.
PubMed Abstract | Publisher Full Text

Meltzer MI, Atkins CY, Santibanez S, *et al.*: **Estimating the future number of cases in the Ebola epidemic--Liberia and Sierra Leone, 2014-2015.** *MMWR Suppl.* 2014; **63**(3): 1–14.
PubMed Abstract

Nhan LNT, Hong NTT, Nhu LNT, *et al.*: **Severe enterovirus A71 associated hand, foot and mouth disease, Vietnam, 2018: preliminary report of an impending outbreak.** *Euro Surveill.* 2018; **23**(46): 1800590.
PubMed Abstract | Publisher Full Text | Free Full Text

Nouvellet P, Cori A, Garske T, *et al.*: **A simple approach to measure transmissibility and forecast incidence.** *Epidemics.* 2018; **22**: 29–35.
PubMed Abstract | Publisher Full Text | Free Full Text

R Core Team: **R: A language and environment for statistical computing**. 2017.
Reference Source

Rojas DP, Dean NE, Yang Y, *et al.*: **The epidemiology and transmissibility of zika virus in girardot and san andres island, colombia, september 2015 to january 2016.** *Euro Surveill.* 2016; **21**(28).
PubMed Abstract | Publisher Full Text | Free Full Text

Shumway RH, Stoffer DS: **Time Series Analysis and Its Applications: With R Examples**. Springer Science & Business Media, 2010.
Reference Source

Unkel S, Farrington CP, Garthwaite PH, *et al.*: **Statistical methods for the prospective detection of infectious disease outbreaks: a review.** *J R Stat Soc Ser A Stat Soc.* 2012; **175**(1): 49–82.
Publisher Full Text

Viboud C, Sun K, Gaffey R, *et al.*: **The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt.** *Epidemics.* 2018; **22**: 13–21.
PubMed Abstract | Publisher Full Text | Free Full Text

Wallinga J, Lipsitch M: **How generation intervals shape the relationship between growth rates and reproductive numbers.** *Proc Biol Sci.* 2007; **274**(1609): 599–604.
PubMed Abstract | Publisher Full Text | Free Full Text

Wallinga J, Teunis P: **Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures.** *Am J Epidemiol.* 2004; **160**(6): 509–516.
PubMed Abstract | Publisher Full Text

WHO Ebola Response Team, Agua-Agum J, Ariyarajah A, *et al.*: **West African Ebola epidemic after one year--slowing but not yet under control.** *N Engl J Med.* 2015; **372**(6): 584–587.
PubMed Abstract | Publisher Full Text | Free Full Text

WHO Ebola Response Team, Aylward B, Barboza P, *et al.*: **Ebola virus disease in West Africa--the first 9 months of the epidemic and forward projections.** *N Engl J Med.* 2014; **371**(16): 1481–1495.
PubMed Abstract | Publisher Full Text | Free Full Text

Wickham H: **ggplot2: elegant graphics for data analysis**. UseR! Springer. 2016.
Reference Source

Wickham H: **testthat: Get started with testing.** *R J.* 2011; **3**(1): 5–10.
Reference Source

Open Peer Review

Current Referee Status: ? ✔ ?

Version 1

Referee Report 07 May 2019

https://doi.org/10.5256/f1000research.19688.r48153

? **Bertrand Sudre**
European Centre for Disease Prevention and Control (ECDC), Solna, Sweden

Major comments

The proposed package is addressing some of the topmost descriptive elements of any epidemiological data set, namely a systematic time-place-person description. With regards to epidemiological curves, a limited number of dedicated packages addressing these aspects were available at the time of this package release (mainly: epitools, https://cran.r-project.org/web/packages/epitools/epitools.pdf [last update: October 26, 2017] and EpiCurve, https://cran.rstudio.com/web/packages/EpiCurve/EpiCurve.pdf [last update: April 24, 2018]). The alternative was a tailor-made, and time consuming, customization based on existing dedicated packages (for instance using customized bar charts geoms from ggplot2). While the data storage and representation is well addressed by the authors, the proposed package offers: i) some basic utilities for outbreak description across time, ii) basic tool for outbreak modelling and iii) a standard for data storage to enhance interoperability between released projects and packages from the R epidemic consortium. In the introduction a short overview of the alternative tools mentioned above should be provided to the reader, together with the new added-values of the current package. It would be an asset in order to ensure a benchmarking analysis with pre-existing resources.

“Here, we introduce *incidence*, an R package developed as part of the toolbox for epidemics analysis of the R Epidemics Consortium (RECON) which aims to fill this gap. In this paper, we outline the package’s design and illustrate its functionalities using a reproducible worked example.”
According to RECON website: package incidence corresponds to “Computation, handling, visualisation and simple modelling of incidence”. It is honourable that the name of a package “incidence” is the choice of the creators of course. It is true that incident cases are all individuals who change to non-disease status from disease, so in this way “incidence” could refer to the occurrence of new cases. In recent modelling papers, the term incidence has been associated to count time series under the same approach, so-called “incidence time series”[1].
Nevertheless, the current name can be misleading for a certain number of epidemiologists. The reason is that in epidemiology the term incidence is traditionally associated to a measure of morbidity, so-called 'Incidence proportion' (or attack rate or risk; for more information see: https://www.cdc.gov/ophss/csels/dsepd/ss1978/lesson3/section2.html). The latter comprised the numerator (= count of case used for a raw epi-curve) but as well a denominator representing the population at risk during the selected time interval. At the first glance, the target audience reading the package name might believe that the package is dedicated to incidence calculation rather than epidemiological curve graphic representation and some basic modelling utilities. Indeed, the package is presented as being able to compute, handle and visualize time-related count data through epi-curve and

additional derived features which are not related to measure of incidence (proportion) strictly speaking. You may wish to consider adding such features and include new capabilities to this package which can cover calculation/representation of incidence in epidemiology. For instance, adding a slot for population data to populate the denominator of an incidence proportion calculation. This can be completed by dedicated graphic outputs with points for each incidence values and line through all data points (geom line). Factor-specific incidence rate (and corresponding CI) can be considered as further extensions (factor: sex-, age-, or any other factor). The advantage is the possibility to overlay and compare several incidence line charts coming for different locations (e.g. attack rate for different health districts or for several population group). It is a suggestion and it is acceptable that the authors would keep the package focus on basic utilities and epicurve representation.

With regards to the structure of the article, it would be easier to start with an example based on simple line listing (see comment on figure 1) with simple epicurve without stratification (with different timing hour, day. week), then move to more complex representation (stratification with various colour in the legend; +/- facetting), and then an example of tailor-made polished figure (see code below). In doing so the figure 1 which would start for a line-list format would be easier to understand.

Minor comments

Section: Author's affiliation

Comment 1:

- "Department of Infectious Disease Epidemiology , School of Public Health, Imperial" - Space to remove after epidemiology
- "Epidemiological Modelling and Analysis (SACEMA),, Stellenbosch University, Stellenbosch" - Double coma to remove

Section: Introduction

Comment 2:

- "Responses to infectious disease epidemics use a growing body of data sources to inform decision making (Cori *et al*., 2017; Fraser *et al*., 2009; WHO Ebola Response Team *et al*., 2014; WHO Ebola Response Team *et al*., 2015). While new data—such as whole genome pathogen sequences—are increasingly useful complements to epidemiological data (Gire *et al*., 2014), epidemic curves—which describe the number of new cases through time (incidence)—remain the most important source of information, particularly early in an outbreak." - Rephrase avoiding four "—" in the same sentence.

Partial genome sequencing is useful to complement epidata/infect. disease epi. If you wish to highlight the WGS/extensive sequencing, consider several aspects:

- Confirm change in infectiousness or epidemiology (CFR, change in clinical presentation, shift in risk group.)
- Relationships between cases (mapping Lassa, Ebola transmission chains ...)
- Emergence and patterns of transmission (vector, host, reservoir)

Suggested examples from the recent literature:

- Kraemer MUG, Cummings DAT, Funk S, Reiner RC, Faria NR, Pybus OG, Cauchemez S. Reconstruction and prediction of viral disease epidemics. Epidemiol Infect. 2018 Nov 5:1-7. doi: 10.1017/S0950268818002881. [Epub ahead of print] PubMed PMID: 30394230; PubMed Central PMCID: PMC6398585.[2]
- Siddle KJ *et al*. Genomic Analysis of Lassa Virus during an Increase in Cases in Nigeria in 2018. N Engl J Med. 2018 Nov.[3]

- Kafetzopoulou LE *et al*, Metagenomic sequencing at the epicenter of the Nigeria 2018 Lassa fever outbreak. Science. 2019 Jan 4;363(6422):74-77.[4]

Comment 3:
- “which describe the number of new cases through time (incidence)—remain” -Proposed change of 'through' to 'over'

Comment 4:
- “important source of information, particularly early in an outbreak.” - Not only. This can be helpful to look at the magnitude and pattern (recurrent environmental sources, detect outliers ...).

Comment 5:
- “Specifically epidemic curves(often” - Missing space.

Comment 6:
- “[…] often referred to as ‘epicurves’) represent the number of new cases […]” - New cases = incident case in epidemiology. Might be more accurate to say, "incident cases of a disease” and add within bracket “(corresponding to the vertical y-axis)”.
- “[…] per time unit based […]” - Consider 'time interval' instead of 'unit' as ‘unit’ can be misunderstood (e.g. unit = hour, min, sec). The interval is at the user's discretion (3 hours = 'time interval’).
- “[…]: on the date or time of symptom onset” - Consider replacing symptoms by “time of illness onset among cases” (and add within bracket “(corresponding to the vertical x-axis)”.

Comment 7:
- “provide a simple, visual outline of epi-demic dynamics, which can be used for assessing the growth or decline of an” - That is the main purpose in practice. You may consider being more assertive, by replacing "can" with "is". Very well-known added-values. No need for six references to support this statement.

Comment 8:
- “[…] by a range of modelling techniques for short-term forecasting (Cori *et al*., 2013; Funk *et al*., 2018; Nouvellet *et al*., 2018; Viboud *et al*., 2018)” - Could you please consider developing with few sentences pointing to ad hoc examples in relation with the features of the cited references.

Comment 9:
- “[…] as well as in outbreak detection algorithms from syndromic surveillance data” - Consider more references that are recent. Note such analytical frameworks are not restricted to syndromic surveillance but to any regularly collected count data from epidemiological time-series in health surveillance system (syndromic or not).

Comment 10:
- “[…] But despite the existence of packages dedicated to time series analysis (Shumway & Stoffer, 2010) as well as surveillance data (Höhle, 2007)” - Time series analysis is a generic term, consider “time series of epidemiological data” to be more accurate. Numerous time-series packages are available for count data that can be "re-used" for human epi.

Comment 11:
- “a lightweight and *well-tested* package solely dedicated to building, handling and plotting epidemic curves directly from linelist data (e.g. a spreadsheet where each row represents an individual case) is still lacking.” - Perhaps “lightweight” is something relative in package development and might be changed. More importantly, as mentioned above, please, consider mentioning other previous packages supporting specifically epi-curve creation (for instance: epitools, EpiCurve). These packages can manage aggregated/no aggregated data with and without factor. Current presented package has an added-value is its interoperability, the presence of simple modelling tools and further graph customization. Stricto sensu, epicurve were able to be done in R form user customization of ggplot2 and cited package.

Section: Methods

Comment 12:
- “[...] some additional facilities for basic models of incidence over time.” - Consider “additional features” instead.

Comment 13:
- “[...] an indication of whether incidence is cumulative or not” - Consider “whether the case count is cumulative”.

Comment 14:
- “[…] representing dates including Date and POSIXct objects”. - The EpiCurve package supports hourly data. This intends to cope with some peculiar situation in which an hourly epi-curve can be of interest (acute food intoxication or environmental source). Could give an example with hourly data (it should be fairly easy as POSIXct is date format in the ggplot2 framework). See an example here: https://www.cambridge.org/core/journals/epidemiology-and-infection/article/unusually-high-illness-s

Comment 15:
- “The dates are aggre-gated into counts based on the user-defined interval representing the number of days for each bin.”- Consider 'user-defined time interval' instead of user-defined interval.
- “number of days for each bin” - “days” is too restrictive. It is just the chosen time interval, it can be the number of weeks, etc.

Comment 16:
- “also accepts a groups argument which can be used to obtain stratified incidence“ -Consider changing stratified incidence by “epidemiological curve".

Comment 17:
- “The basic elements of the incidence object can be obtained by the accessors get_counts(), get_dates(), and get_interval()." - Please, number the number of basic elements for clarity purpose.

Comment 18:
- “The function subset() can be used for isolating case data from a specific time window and/or groups, while the [operator can be used for a finer control to subset dates and groups using integer, logical or character vectors.” - If several functions are to be presented, it is easier to use bullet points to structure the reading.
- Consider the removal 'for isolating case data from a specific' and changing with 'to define'.
- “[“ change to 'indexing operator, to follow the classical denomination https://cran.r-project.org/doc/manuals/r-release/R-lang.pdf.

Comment 19:
- “Figure 1. Generalized workflow from incidence object construction to modeling and visualization.” - The first example in the paper is using a line-list format as data inputs but showed a stratified graphic. To be consistent and easier to follow for the reader, this figure illustrates the flow of such of data type, knowing that line-list is the primary source of epidemiological surveillance. A solution present in the figure is the two types of data inputs (non-aggregated/aggregated count). It would be easier, from a reader point of view, to capture the data flow from the line-lit to the final products proposed by this package.

Comment 20:
- “The function pool() can be used to merge several groups into one,” - Consider ending the sentence and explaining what it does.
- “and the function cumulate() will turn incidence data into cumulative incidence” - Consider changing “cumulative incidence” to “cumulative count of cases”.

Comment 21:

- […]: an option for a 'long' format which is readily compatible with *ggplot2* (Wickham, 2016) for further customization of graphics." - Would it be possible to mention how date format is exported? This might good to elaborate a bit more about date and user's customization with ggplot2. This can be addressed later in the manuscript (see proposal for a theme).

Comment 22:

- "In line with RECON's development guidelines, the *incidence* package is thoroughly" - Add a hyperlink/ref. to the website: https://www.repidemicsconsortium.org/resources/guidelines/

Section: Modelling utilities

Comment 23:

- "Here, we highlight three simple functionalities in *incidence* for estimating parameters via modelling or bootstrap and the two specialized data classes that are used to store the models and parameter estimates." - Consider structuring the following section according to the five elements mentioned (=three functions [fit() , estimate_peak() , fit_optim_split()] and two specialized data classes) using for instance bullet points/subtitles. For each function, the goal, data input, statistical methods and output object(s) can be grouped in single section.

Comment 24:

- "we implement the simple log-linear regression approach in the function fit(), which can […]" - Please add more information about the structure of the 'incidence_fit objects containing log-linear models'

Comment 25:

- "[…] fit exponential increase or decrease of incidence over time by log-transforming case counts … " - Can be simplified to "fit exponential increase or decrease using a linear regression over time on log-transformed case counts…".

Comment 26:

- "where *y* is the incidence, *r* is the growth rate, *t*" - Replace "incidence" with number of new case/incident case.

Comment 27:

- "serial interval" - Consider adding "serial interval of the infectious agent"

Comment 28:

- "uses multinomial bootstrapping to estimate the peak, assuming" - Some explanation about the method and references would be desirable.

Comment 29:

- "Both object classes can be plotted separately or added to an existing epicurve using the function add_incidence_fit() (Figure 5)." - The customization of the epicurve is well described. However, it is not mentioned how to change the layout of the model outcome and confidence interval. Indeed, some users might wish to use an alternative ggplot2 geometric object such as geom_range with a shaded grey semi-transparent band instead of two dotted lines. It would an added-value to provide some capacities or explanation and an example of customization of the layout of the "incidence_fit objects".

Section: Use cases

Comment 30:

- "Two worked examples are used to demonstrate the functionality and flexibility of the *incidence* package. The first example illustrates how to compute and manipulate stratified weekly incidence directly from a line-list". - Consider "The first example illustrates how to create directly from a line-list incidence object in order draw an epicurve of the weekly number of cases with or without stratification on patient characteristics".

Comment 31:

- "while the second example shows how to import pre-computed daily incidence and fit a log-linear model to estimate growth rate (*r*) and doubling time for the growing phase." - Footnote to be included in the section about the function for more clarity.

Example 1: computing and manipulating stratified weekly incidence

Comment 32:

- "The weekly incidence stratified by hospitals is computed by running the function incidence() on the Date vari-able dat1$date_of_onset with the arguments interval = 7 and groups = dat1$hospital." - Consider rephrasing. For instance: "the epicurve with the weekly number of cases by hospital can be computed from the line listing dataframe object (dat1) using the function incidence() on i) the date vari-able (dat1$date_of_onset), ii) by specifying the argument interval of seven days in order to aggregate the number case per week (interval = 7) and iii) including a the line-listing variable for stratification in the argument groups, in this case the hospital name (groups = dat1$hospital).

Comment 33:

- "Here $count is a matrix with 56 rows and 6 columns as groups by hospital" - Missing s at the end of $counts.

Comment 34:

- "The generic plot() method for incidence objects calls the specific function plot.incidence(), which makes an incidence barplot using the *ggplot2* package. Hence, customization of *incidence* plot can benefit from the powerful graphical language from *ggplot2*." - A short explanation and command line to explain how to access the code of the method would be welcome (notably the plot). This would help users to understand which ggplot2 geometric object(s) is used for the bar plot and incidence_fit lines. This would be an asset to understand how to proceed with further customization (within the aesthetic, theme or faceting specifications). This can be proposed at the end through several examples.

Comment 35:

- # plot incidence object
  my_theme <-theme_bw(base_size = 12) +
  theme(panel.grid.minor = element_blank()) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5, color = "black")) plot(i.7.group, border = "white") + my_theme + theme(legend.position = c(0.8, 0.75))

The current example assumes that all users are familiar with ggplot2, notably how to customize the non-data components through the theme. We might suggest to introduce what is a theme in ggplot2 and what it does, for instance "Themes allows modification (content and layouts) of non-data components such as titles, axis labels, legends (position and aspect …), graphics grid lines and backgrounds (Modify components of a theme; ref: https://ggplot2.tidyverse.org/reference/theme.html)".
In addition, mention that theme specifications can "overwrite" some layout specification in the other part of the ggplot2 function. The use in this example of theme_bw appears to clarify what the default built-in theme is (complete themes: https://ggplot2.tidyverse.org/reference/ggtheme.html , other themes https://cran.r-project.org/web/packages/ggthemes/ggthemes.pdf).
To help the reader, you might consider the arguments as displayed in "Modify components of a theme" order as much as possible:

- my_theme <-theme_bw(base_size = 12) +
  theme (
  panel.grid.minor = element_blank,

axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5, color = "black"),
legend.position = c(0.8, 0.75)
)

For this first example, consider setting the legend outside of the main graphic frame, especially for the first figure (instance legend.position="bottom", legend.box = "horizontal").
For the axis labels, the Y is family the “Num. of case of EVD. per week” in the Figure 2.

Comment 36:

- i.7.sub <- i.7.group[1:18, grep("Hospital", group_names(1.7.group))]
  hosp_colors <-c("#899DA4", "#C93312", "#FAEFD1", "#DC863B")
  - Add an intermediate line to illustrate that in this period selection (e.g. print i.7), cases are coming from four hospitals, to hosp_colors would a four colours vector. Keep the legend out of the main graphic frame (top or bottom).

Comment 37:

- “Here, because of the few numbers of cases in the first few weeks, we have also highlighted each case using show_ cases = TRUE (Figure 3)”. - The figure is displaying the individual case which is the standard representation of an epicurve according to number of training programs (US CDC, EPIET …). This is of primary importance in outbreak investigation to able to delineate case number in an easy-to-read visual and often stratified representation by case characteristics (sex, exposure, location …). This type of representation is adopted by default in package EpiCurve (see https://cran.r-project.org/web/packages/EpiCurve/vignettes/EpiCurve.pdf) as well as US CDC training example of epicurve (https://www.cdc.gov/training/quicklearns/createepi/). Representation of case as square can be achieved with the current package as illustrated in at the end of the package vignette (see: vignette ("customize_plot", package="incidence") in the section Applying the style of European Programme for Intervention Epidemiology Training (EPIET). It’s understandable that square form can be changed in case of a large dataset making the size of the intervals for the x- and y- axes different.
  It would be an asset for the target audience and the current manuscript to contain a full example of an epicure with case made under square representation (somehow close to the example cited above in the vignette) combined with a full theme following simple and standard representation.

Comment 38

- Consider change in the layout of Figure 2: 'Number of EVD cases stratified by hospital ... between week XX and week YY'.
  Y axis title: 'Number of cases'
  X axis title: 'Week of onset of illness'
  X axis ticks should be made visible for all weeks even the label displayed for any other week.

Comment 39:

- Would it be possible to give some more information to which ggplot2 geometry are used when show_ cases = TRUE is specified? This is of importance to provide the reader with a clue about how both borderline and colour content of each square can be further customized.

Comment 40:

- As standard the label of the X axis time interval is displayed on the left side of the bin. Formally, the ideal position would be right below the bin (as illustrated in the CDC example above and numerous of published epicurve). In the EpiCurve package, the standard representation is not ideal as the X axis tick is right in the middle of the bin. The position of the label can be customized by the user, but would it be possible to look at the option to place the time label under each bin instead of to the left bin tick mark (see an example of ad hoc customization below). Of note, such label position should support figure export resizing.

Comment 41:

Figure 4. (A) stratified and (B) pooled daily incidence plots of ZVD in Colombia, September 2015 to January 2016.
Please consider comments made above for the other figure on figure title and axis labels:

- “Epicurve of the daily number of Zika virus disease between Sep and Jan 2016”. Panel (A) stratified by location and (B) pooled across locations. “
- Y axis title: 'Number of cases'
- X axis title: ' Week of onset of illness'

Some readers might wonder if the faceting capabilities of ggplot2 would be supported or not. An evident alternative is to prepare separated epicurve for the both locations and further combined them using specific package (ggarrange for instance). Would it be possible to add information on this point, and if supported, provide an alternative presentation of the both epicurves using two vertically aligned panels?

Comment 42:

- “Without any knowledge on the splitting date, we can turn to the fit_optim_split() function to look for the optimal splitting date”. - Could you please reconsider the phrasing as “Without any knowledge on the splitting date” which seems not logical when looking retrospectively to an outbreak epicurve and visually identifying the peak of the epidemic wave.

Comment 43:

- library('magrittr') - Provide a short explanation about the dependency(ies) with this package.

Comment 44:

- “The predictions and their 95% CIs from the two incidence_ft objects, ‘before’ and ‘after’, can be added to the existing incidence plot of i.pooled using the piping-friendly function add_incidence_fit().” - Provide more explanation on “the piping-friendly function” aspect. Provide an example of layout customization of the two log-linear regression models (linetype, colour, size).

Comment 45:

- Please find below a proposal for an advanced and custom-made layout epicurve representation using the Zika dataset. In this example, some of the important layout feature of an epicurve are available: i) no space around the limits of both x and Y axes, ii) representation of each case with a square, iii) x label under each bin and make the graphic as light as possible based on Tufte’s advice (less grid as possible, visible labels …).
- In the graphic below, an incidence_fit object can be added for the pooled location with a specific customization to give a complete overview to the package capabilities.

Working example based on Zika diease dataset:

```
head(zika_girardot_2015, 20)
head(zika_sanandres_2015, 20)
dat2 <- merge(zika_girardot_2015, zika_sanandres_2015, by = "date", all = TRUE) # # combine two datasets into one
names(dat2)[2:3] <- c("Girardot", "San Andres") # rename variables
dat2[is.na(dat2)] <- 0 # replace NA with 0
i.group <- as.incidence(x = dat2[, 2:3], dates = dat2$date) # convert pre-computed incidence in data.frame into
str(i.group)

# Graphic
```

```
i.group_zoom_in <- i.group[10:60]

plot(i.group_zoom_in, n_breaks = nrow(i.group_zoom_in),  border = "grey90", show_cases=TRUE) +
 theme_bw() +
 scale_y_continuous(expand = c(0, 0), limits=c(0,max(i.group_zoom_in$counts +1))) +
 scale_x_date(date_breaks = "1 day", date_labels = "%b %d",expand = c(0, 0), limits=
c(as.Date(min(i.group_zoom_in$dates)) , as.Date(max(i.group_zoom_in$dates+1)))) +
 #  scale_fill_discrete(name = "Location:") +
 #scale_x_continuous(expand = c(0, 0),limits= c(as.Date(min(zomm_in$dates)) ,
as.Date(max(zomm_in$dates)))) +
 labs(title = "Number of Zika disease cases",
     subtitle = "Girardot and San Andres municipalities, Colombia. Period: 6 Sep to 11 Oct 2015." ,
     x="Week of onset of illness",
     y="Number of cases",
     fill="Municipalities:") +
 theme(panel.border = element_rect(colour = "white"),
      panel.grid.major.y = element_line(colour = "grey70", linetype="dotted", size =0.5),
      axis.line = element_line(colour = "black", size = 0.7),
      axis.text.x = element_text(angle = 45, hjust = 0.8, size = 6),
      axis.ticks.x =  element_line(size = rel(2)),
      axis.ticks.y =  element_line(size = rel(2)),
      axis.title.y = element_text(margin=margin(0,10,0,0), size=10),
      axis.title.x = element_text(margin=margin(10,0,0,0), size=10),
      plot.title =  element_text(size = 12,face = "bold"),
      plot.subtitle =  element_text(size = 9),
      legend.position="bottom",
      legend.box = "horizontal") +
 coord_equal()
```

Link to plot available here.

Is the rationale for developing the new software tool clearly explained?
Yes. The rational is sounds and the new package “Incidence” is fulfilling the objectives cited. It allows streamlining the manipulation and representation for epidemiological data for epidemiological representation. It expected that this new package would reach a wide audience of epidemiologists and epidemiological data analysts working with R.

Is the description of the software tool technically sound?
Yes. However, some clarifications proposed in the detailed review would enhance the description of the software tool features.

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Overall, yes. Minor adjustments are proposed to allow reader to better access code (e.g. provide example to access full code of incidence (S3) class incidence and its subsequent methods (plot, …)) and offer additional clarification to improve graphic customization (notably around the incidence_fit object and further integration of ggplot2 faceting functionality).

Is sufficient information provided to allow interpretation of the expected output datasets and

any results generated using the tool?
Yes. Several practical and realistic examples are provided by the authors. All examples were tested and further tests with epidemiological additional datasets were conducted and allow to reproduce accurately tool behaviours. In addition to the paper review, the R package documentation (description file, recent reference manual and vignettes) were thoroughly reviewed to assess any discrepancies between the peer-review publication and package documentation on CRAN.

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes.

References
1. Cori A, Ferguson NM, Fraser C, Cauchemez S: A new framework and software to estimate time-varying reproduction numbers during epidemics.*Am J Epidemiol*. 2013; **178** (9): 1505-12 PubMed Abstract | Publisher Full Text
2. Kraemer MUG, Cummings DAT, Funk S, Reiner RC, Faria NR, Pybus OG, Cauchemez S: Reconstruction and prediction of viral disease epidemics.*Epidemiol Infect*. 2018. 1-7 PubMed Abstract | Publisher Full Text
3. Siddle K, Eromon P, Barnes K, Mehta S, Oguzie J, Odia I, Schaffner S, Winnicki S, Shah R, Qu J, Wohl S, Brehio P, Iruolagbe C, Aiyepada J, Uyigue E, Akhilomen P, Okonofua G, Ye S, Kayode T, Ajogbasile F, Uwanibe J, Gaye A, Momoh M, Chak B, Kotliar D, Carter A, Gladden-Young A, Freije C, Omoregie O, Osiemi B, Muoebonam E, Airende M, Enigbe R, Ebo B, Nosamiefan I, Oluniyi P, Nekoui M, Ogbaini-Emovon E, Garry R, Andersen K, Park D, Yozwiak N, Akpede G, Ihekweazu C, Tomori O, Okogbenin S, Folarin O, Okokhere P, MacInnis B, Sabeti P, Happi C: Genomic Analysis of Lassa Virus during an Increase in Cases in Nigeria in 2018. *New England Journal of Medicine*. 2018; **379** (18): 1745-1753 Publisher Full Text
4. Kafetzopoulou L, Pullan S, Lemey P, Suchard M, Ehichioya D, Pahlmann M, Thielebein A, Hinzmann J, Oestereich L, Wozniak D, Efthymiadis K, Schachten D, Koenig F, Matjeschk J, Lorenzen S, Lumley S, Ighodalo Y, Adomeh D, Olokor T, Omomoh E, Omiunu R, Agbukor J, Ebo B, Aiyepada J, Ebhodaghe P, Osiemi B, Ehikhametalor S, Akhilomen P, Airende M, Esumeh R, Muoebonam E, Giwa R, Ekanem A, Igenegbale G, Odigie G, Okonofua G, Enigbe R, Oyakhilome J, Yerumoh E, Odia I, Aire C, Okonofua M, Atafo R, Tobin E, Asogun D, Akpede N, Okokhere P, Rafiu M, Iraoyah K, Iruolagbe C, Akhideno P, Erameh C, Akpede G, Isibor E, Naidoo D, Hewson R, Hiscox J, Vipond R, Carroll M, Ihekweazu C, Formenty P, Okogbenin S, Ogbaini-Emovon E, Günther S, Duraffour S: Metagenomic sequencing at the epicenter of the Nigeria 2018 Lassa fever outbreak. *Science*. 2019; **363** (6422): 74-77 Publisher Full Text

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Referee Report 25 April 2019

https://doi.org/10.5256/f1000research.19688.r45525

Quirine ten Bosch
Quantitative Veterinary Epidemiology, Department of Animal Sciences, Wageningen University and Research Centre, Wageningen, The Netherlands

The authors introduce a package that aids in processing linelist data (*i.e.,* in which each row represents a case) for building, handling, and plotting epidemic curves. This tool fills a gap within the larger epidemics toolbox of the R Epidemics Consortium in 'getting the basics right', or otherwise put, getting data common to outbreak situations in the right format for further analysis.

I agree with the authors that this is a helpful tool. This is particularly true during outbreaks, when linelist data need to be processed quickly and disseminated to relevant actors. The procedures are straightforward and well presented. The examples given in this work are well chosen and provide a good starting point for working with this package. I only have a few comments pertaining to i) the loglinear model implementation and presentation and ii) the integration with other packages.

1. While the package is meant to focus on processing and visualization of data, the authors have added a modeling-capability that estimates epidemic growth rates. The respective function contains an option for estimating the peak of the outbreak and fit the exponential increase or decrease during the epidemic. Basing the time window for exponential growth on the timing of the peak will underestimate the growth rate as growth will no longer be exponential just before the peak of the epidemic. A more careful description in the MS is needed to highlight this and other shortcomings of this approach. Other methods to estimate the best time window, such as those used in the R0-package[1], could prove helpful and be implemented in the package. Further, the choice of fitting the growth rates to the aggregated data rather than the two distinct regions (Girardot and San Andres) is a bit uncomfortable. Particularly so because, as the authors acknowledge, there seem to be two quite distinct outbreaks.

2. The authors highlight the gap within the epidemics toolbox is filled. It would be nice to see a little bit more context on what packages can easily work with their data structures and what analyses can hence readily be done.

Minor comments:

- Page 7: group_names(1.7.group) should be i.7.group

References

1. Obadia T, Haneef R, Boëlle PY: The R0 package: a toolbox to estimate reproduction numbers for epidemic outbreaks.*BMC Med Inform Decis Mak*. 2012; **12**: 147 PubMed Abstract | Publisher Full Text

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: infectious disease modeling, vector-borne disease, multi-host systems

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 03 April 2019

https://doi.org/10.5256/f1000research.19688.r45526

? **Benjamin M. Bolker**
Departments of Mathematics and Statistics and Biology, McMaster University, Hamilton, ON, Canada

The authors have written an R package that does some useful tasks in aggregating and manipulating incidences and plotting them. For the most part the paper is clearly written and technically correct, and the package seems well-constructed. The abstract says that the package is "solely dedicated to building, handling and plotting epidemic curves", although it also fits simple epidemic models to incidence data.

Overall I think this package is sensible and well-designed. My only major concern is with the log-linear fitting procedures; on the one hand, the authors say clearly that there are many fitting procedures which cannot (and should not) all be squeezed into this package, and I suspect that they don't intend these fits to be taken as very precise measures of the growth rate. On the other hand, I'm quite concerned by the possibility that users will take growth rates and confidence intervals estimated by these simplistic fits too seriously. There are a number of delicate issues around the estimation of epidemic growth rates:

- Assume log-Normal incidence (as done by a log-linear fit), or make other distributional assumptions (e.g. Poisson or negative binomial)?

- If using exponential or log-linear fits, what time window should one use to capture enough of the beginning of the epidemic but not bias the growth rate downward by capturing the saturation phase of the epidemic (Ma *et al.*, 2014[1])?

- Should one allow for the influence of both process and measurement error (King *et al.*, 2015[2])?

- Is growth exponential or sub-exponential (Viboud *et al.*, 2016[3])?

I think I would have preferred that, if they were not going to go deeply into this area, that the authors instead provide some sort of non-parametric smooth fit (perhaps constraining the changes to be monotonically increasing before the peak and decreasing after the peak) rather than oversimplifying in this way.

The minimal change that is needed to the document is a stronger set of caveats/warnings to users that the log-linear model may have major shortcomings in some circumstances, and should not be accepted unquestioningly.

The package suggests that the fit() function is extendable to allow other fitting methods ("While the incidence package returns incidence_fit objects containing log-linear models by default, they can be constructed from any model from which it's possible to extract the growth rate (r) and predict incidence along the model"); I wasn't able to figure out how to achieve that goal.

The main use of the package is for converting from line lists to aggregated incidence data. It would be useful (I can't tell if it's possible) to easily be able to aggregate data that are already in date/incidence form to coarser scales, or to approximately disaggregate incidence data.

I had trouble fitting epidemic curves to data sets that only included incidence up to the peak or shortly beyond (it seems that the model was automatically trying to estimate a decline rate as well as an increase rate). This seems like an odd choice for a tool that people may be using to work with data from emerging epidemics that have not yet peaked.

Minor comments:

- p. 3, paragraph 1: missing space before (

- p. 3, paragraph 3: maybe R "language" or "environment" rather than "software" (R defines itself as "a language and environment for statistical computing and graphics")?

- Last paragraph of intro: comma before "which aims to fill this gap"?

- Methods, paragraph 2: Probably don't need to say "numeric and integer" values, "numeric" would suffice (is.numeric(1L) is TRUE in R).

- Methods, paragraph 2: "can not" -> "cannot".

- How do subsets and indexing work with dates?
- Figure 1 caption: delete "both"?
- "created from these where it checks and validates" -> "created from these components after checking and validating" ?
- "[Fitting] of log-linear models is depicted in the lower left"? (Technically, the addition of log-lin models to the plot is depicted along the bottom edge, or in the lower right).
- p. 5: "boo[t]strap".

References

1. Ma J, Dushoff J, Bolker BM, Earn DJ: Estimating initial epidemic growth rates.*Bull Math Biol*. 2014; **76** (1): 245-60 PubMed Abstract | Publisher Full Text
2. King AA, Domenech de Cellès M, Magpantay FM, Rohani P: Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola.*Proc Biol Sci*. 2015; **282** (1806): 20150347 PubMed Abstract | Publisher Full Text
3. Viboud C, Simonsen L, Chowell G: A generalized-growth model to characterize the early ascending phase of infectious disease outbreaks.*Epidemics*. **15**: 27-37 PubMed Abstract | Publisher Full Text

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: epidemic modeling, biostatistics, evolution of virulence

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.