R Epidemics Consortium and Using Its Packages to Analyze Influenza Data

Jun Cai

Ph.D. Candidate in Ecology
Department of Earth System Science
Tsinghua University

May 21, 2017

清華大學
Tsinghua University

地球系统科学系
Department of Earth System Science

Acknowledgement: Materials on the R Epidemics Consortium (RECON) were partly provided by Dr. Thibaut Jombart, the founder of RECON.

Precursor of RECON — Hackout 3

Hackout 3 (hackout3.ropensci.org)

An R hackthon of analysis and modelling tools for emergency outbreak response

Hackout 3

20-24 June 2016 • Berkeley, CA • USA

MRC Centre for Outbreak Analysis and Modelling

Imperial College London

BERKELEY Institute for Data Science

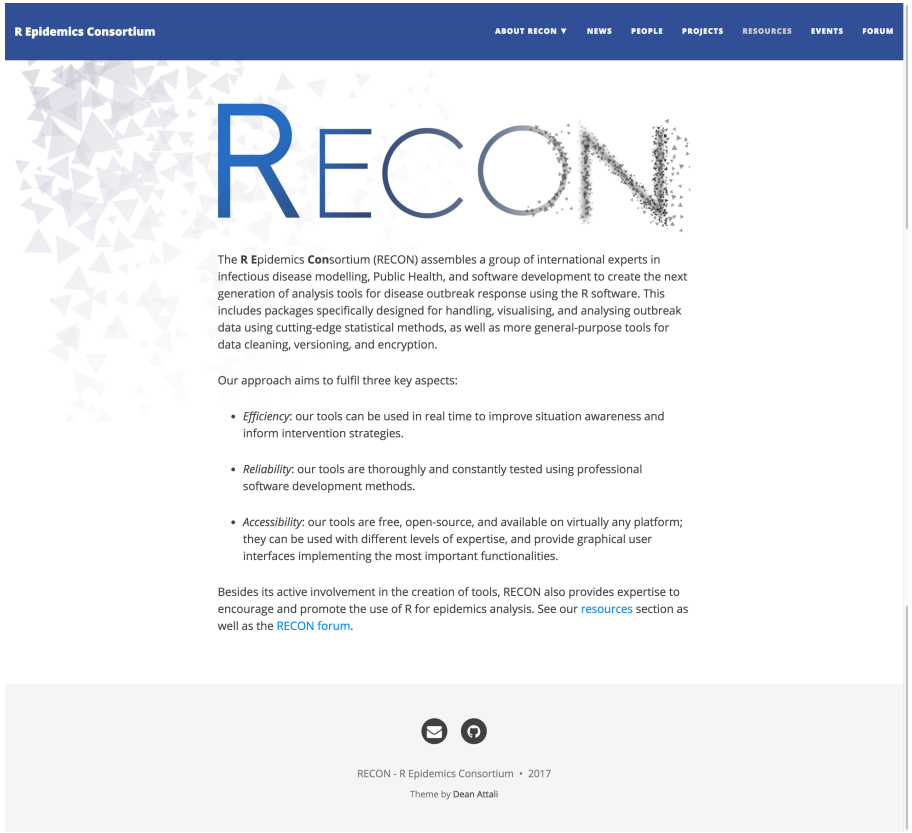
Thibaut Jombart
[organizer]
Imperial College London
// website

Karthik Ram
[organizer]
Berkeley Institute for Data Science (USA), rOpenSci
// website

Marc Baguelin
[co-organizer]
Public Health England & London School of Hygiene and Tropical Medicine (UK)

The **R** **E**pidemics **CON**sortium (RECON)

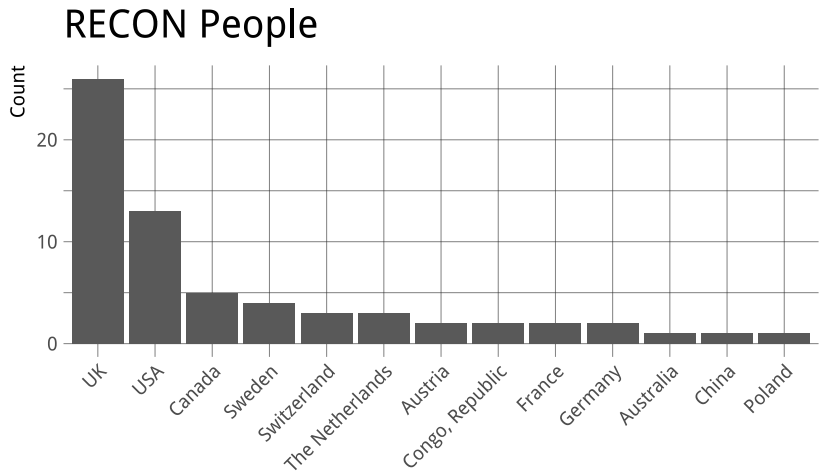
a group of international experts in infectious disease modelling, public health, and software development to create the next generation of analysis tools for disease outbreak response using 



www.repidemicsconsortium.org

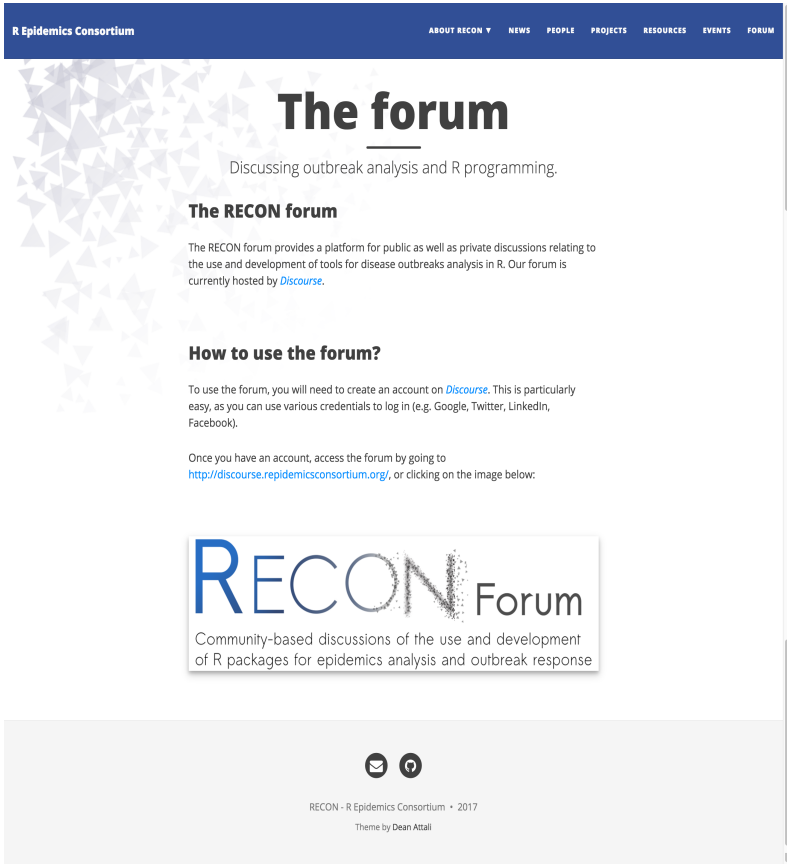
RECON

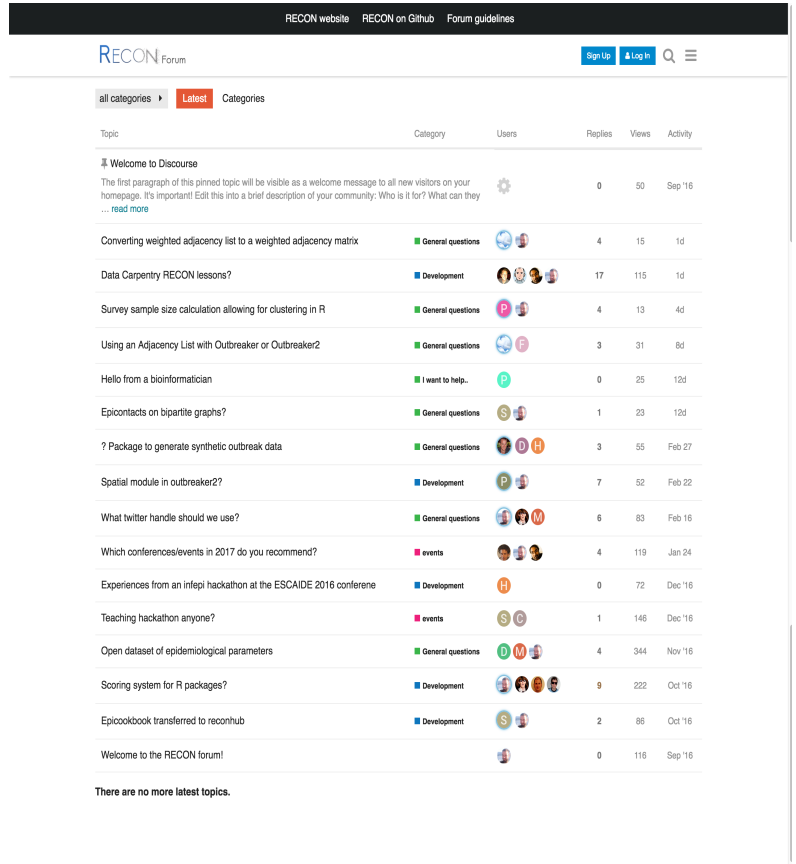
- started 6th September, 2016
- 65 people (57 members, 6 advisory board, 2 administrative staffs)
- 13 countries, $>$ 35 institutions
- 2 new packages (*outbreaks* and *incidence*) released, ~10 - 15 packages in development
- involvement in training programmes starting in 2017 (the **F**ield **E**pidemiology **T**raining **P**rogram (FETP), the **E**uropean **P**rogramme for **I**ntervention **E**pidemiology Training (EPIET), ...)
- public forum, blog, online resources



The RECON forum (discourse.repidemicsconsortium.org)

A platform for public as well as private discussions relating to the use and development of tools for disease outbreaks analysis in R





Join us!

Aims of RECON packages

- **Efficiency**: used in real time to improve situation awareness and inform intervention strategies. cutting-edge, computer-efficient statistical methods and epidemic models
- **Reliability**: thoroughly and constantly tested using professional software development platform (e.g.,GitHub) and methods (e.g., testthat and covr). extensive unit testing, code review
- **Accessibility**: free, open-source, and available on virtually any platform, easy learning curve; extensive documentation, tutorials, GUI implementation

- **Released packages**

outbreaks: collection of outbreak data.

incidence: computation, handling, visualisation and simple modelling of incidence.

- **Up-and-coming packages**

cleanr: rationalised and reproducible data cleaning.

dibbler: investigation of food-borne disease outbreaks.

distcrete: discretized probability distributions.

epicontacts: handling, visualisation and analysis of epidemiological contacts.

epimatch: finding matching patient records across tabular data sets.

outbreaker2: inferring transmission chains by integrating epidemiological and genetic data.

- **Up-and-coming packages (continued)**

vimes: Visualisation and Monitoring of Epidemics, including some outbreak detection algorithms.

- **Related packages**

EpiEstim: quantifying transmissibility throughout an epidemic from incidence time series.

OutbreakTools: basic analysis and visualisation of complex line-list data (to be replaced by *incidence* and *epicontacts*).

EpiJSON: implementation of a generic json format for case outbreak data.

repijson: R package implementing *EpiJSON* format.

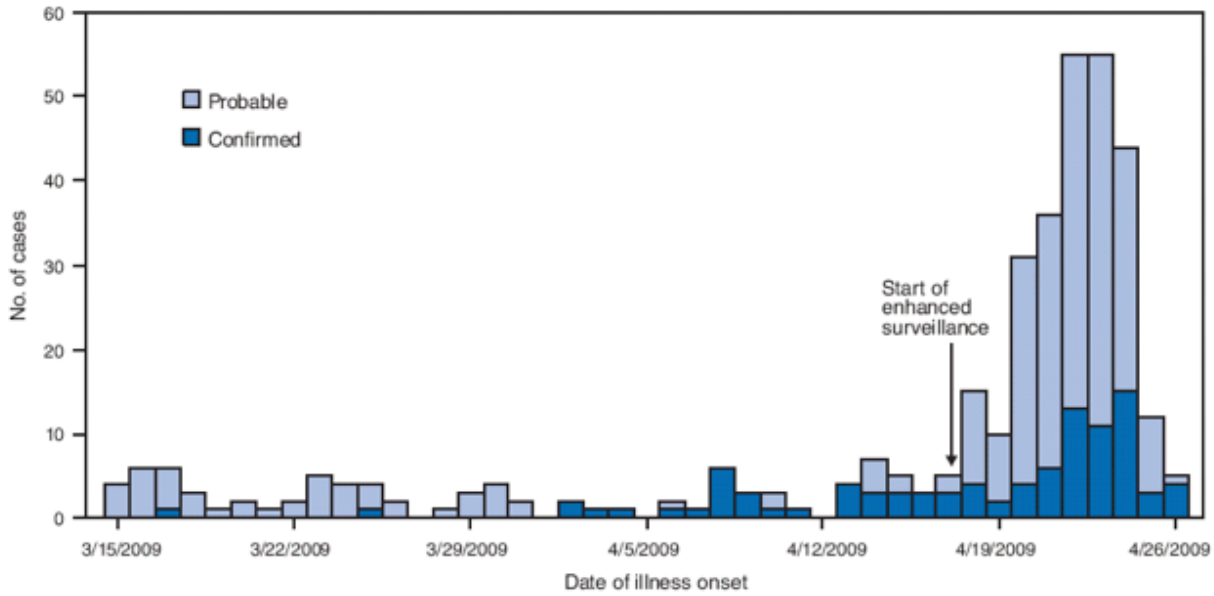
outbreaker: inferring transmission chains using temporal and genetic information.

DSAIDE: Dynamical Systems Approach to Infectious Disease Epidemiology - a Shiny/R based teaching.

Epidemic Curve (Epi Curve)

shows the frequency of new cases over time based on the date of onset of disease.

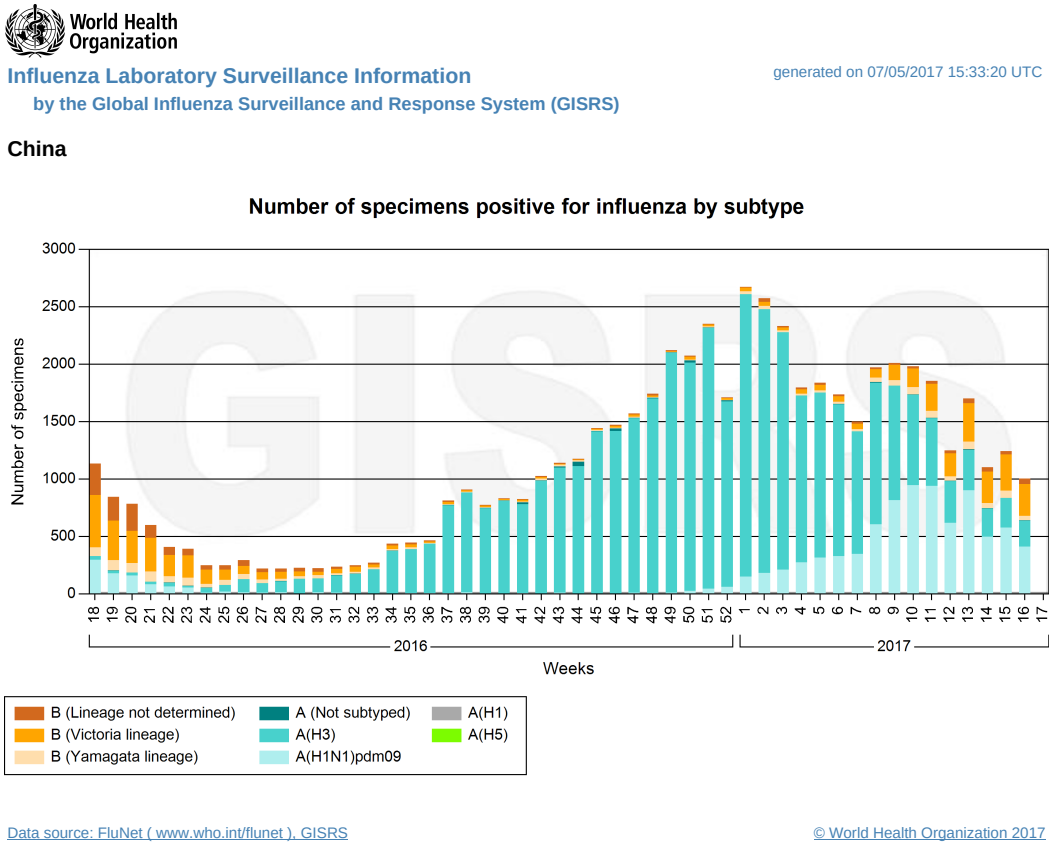
Number of confirmed (N = 97) and probable (N = 260)[1] cases of swine-origin influenza A (H1N1) virus (S-OIV) infection, by date of illness onset — Mexico, March 15 – April 26, 2009 (adapted from Centers for Disease Control and Prevention (CDC) 2009)



- Stacked bar plot
- Daily: date formated tick labels on x-axis

[1] Probable cases for which dates of illness onset are known.

Ubiquitous plot in epidemiological reports



- Stacked bar plot
- Weekly: x-axis tick labels – ISO 8601 week number

Epidemic Curve of Influenza A H7N9 in China, 2013

- By gender
- Daily and weekly

```
library(outbreaks)
head(fluH7N9_china_2013, 6)
```

```
##   case_id date_of_onset date_of_hospitalisation date_of_outcome outcome
## 1       1    2013-02-19                    <NA>      2013-03-04   Death
## 2       2    2013-02-27              2013-03-03      2013-03-10   Death
## 3       3    2013-03-09              2013-03-19      2013-04-09   Death
## 4       4    2013-03-19              2013-03-27            <NA>    <NA>
## 5       5    2013-03-19              2013-03-30      2013-05-15 Recover
## 6       6    2013-03-21              2013-03-28      2013-04-26   Death
##   gender age province
## 1      m  87 Shanghai
## 2      m  27 Shanghai
## 3      f  35    Anhui
## 4      f  45  Jiangsu
## 5      f  48  Jiangsu
## 6      f  32  Jiangsu
```

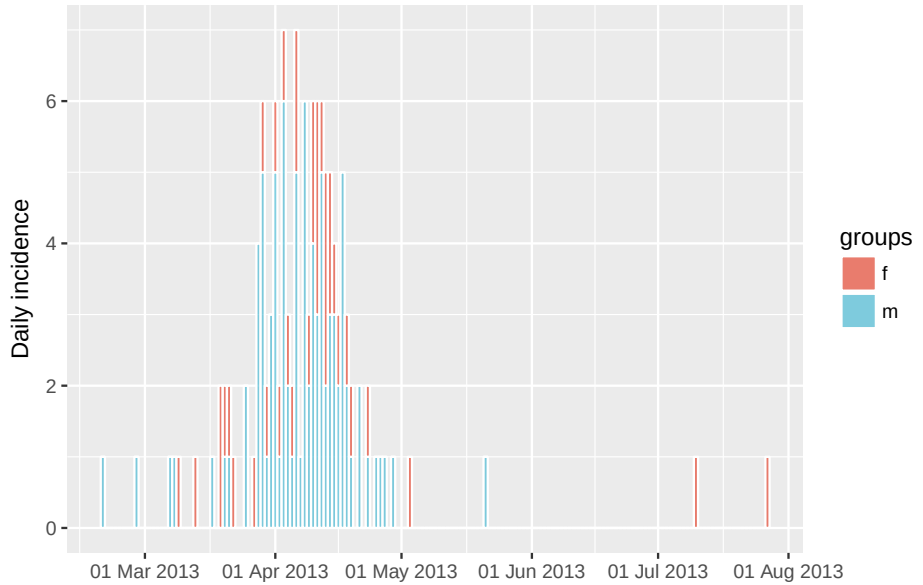
```
range(fluH7N9_china_2013$date_of_onset, na.rm = TRUE)
```

```
## [1] "2013-02-19" "2013-07-27"
```

Plot daily epidemic curve from scratch

```
date.range <- range(df$date_of_onset)
full.date <- seq(date.range[1], date.range[2],
                 by = "day")

# count dated events according to breaks
count.date <- function(dated.events, breaks) {
  df <- data.frame(date = dated.events)
  n <- sapply(breaks, function(b) {
    subdf <- subset(df, date == b)
    nrow(subdf)
  })
  return(data.frame(date = breaks, n))
}

date.list <- tapply(df$date_of_onset, df$gender,
                    count.date, full.date)
date.df <- reduce(date.list, left_join, by = "date")
names(date.df)[-1] <- names(date.list)

# reshape data.frame from wide into long format
library(reshape2)
date.long <- melt(date.df, id.vars = c("date"),
                  variable.name = 'groups',
                  value.name = "counts")
```

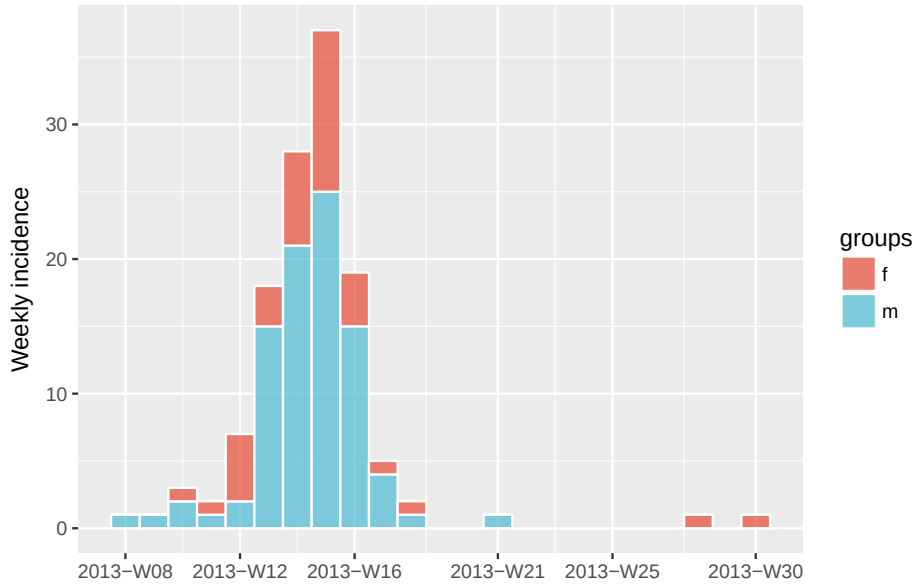
```
library(ggplot2)
library(ggsci)
library(scales)
ggplot(date.long, aes(date, counts, fill = groups)) +
  geom_bar(stat = "identity", position = "stack",
           width = 1, color = "white", alpha = .7) +
  labs(x = "", y = "Daily incidence") +
  scale_fill_npg() +
  scale_x_date(labels = date_format("%d %b %Y"))
```



Plot ISOweek-based weekly epidemic curve from scratch

```
library(ISOweek)
full.week <- unique(ISOweek(full.date))

df1 <- df %>%
  mutate(isoweek = ISOweek(date_of_onset)) %>%
  glimpse()

count.isoweek <- function(isoweek.events, breaks) {
  df <- data.frame(isoweek = isoweek.events)
  n <- sapply(breaks, function(b) {
    subdf <- subset(df, isoweek == b)
    nrow(subdf)
  })
  return(data.frame(isoweek = breaks, n))
}

week.list <- tapply(df1$isoweek, df$gender,
                    count.isoweek, full.week)
week.df <- reduce(week.list, left_join,
                  by = "isoweek")
names(week.df)[-1] <- names(week.list)
# add isoweek starting day
weekdate <- paste0(week.df$isoweek, "-1")
week.df$weekday <- ISOweek2date(weekdate)

# reshape data.frame from wide into long format
week.long <- melt(week.df, id.vars = c("isoweek",
                                       "weekday"),
                  variable.name = 'groups',
                  value.name = "counts")
```

```
# x-axis tick labels
ind <- trunc(seq(1, nrow(week.df), length = 6))
tickind <- week.df$weekday[ind]
ticklab <- week.df$isoweek[ind]
ggplot(week.long, aes(weekday, counts, fill = groups)) +
  geom_bar(stat = "identity", position = "stack",
           width = 7, color = "white", alpha = .7) +
  labs(x = "", y = "Weekly incidence") +
  scale_fill_npg() +
  scale_x_date(breaks = tickind, labels = ticklab)
```



incidence package

incidence: compute, handle, visualise and model incidence of dated events

Thibaut Jombart

Founder of RECON. Statistician and R programmer specialized in outbreak analysis. Imperial College London, UK.

author, creator

Rich Fitzjohn

R developper specialized in data analysis infrastructures. Imperial College London, UK.

author

Jun Cai

PhD candidate interested in influenza transmission dynamics and R programming. Center for Earth System Science, Tsinghua University, China.

contributor

Vignettes

- Overview of the incidence package
- Details of the incidence class
- Customize plots of incidence

Daily plot using *incidence*

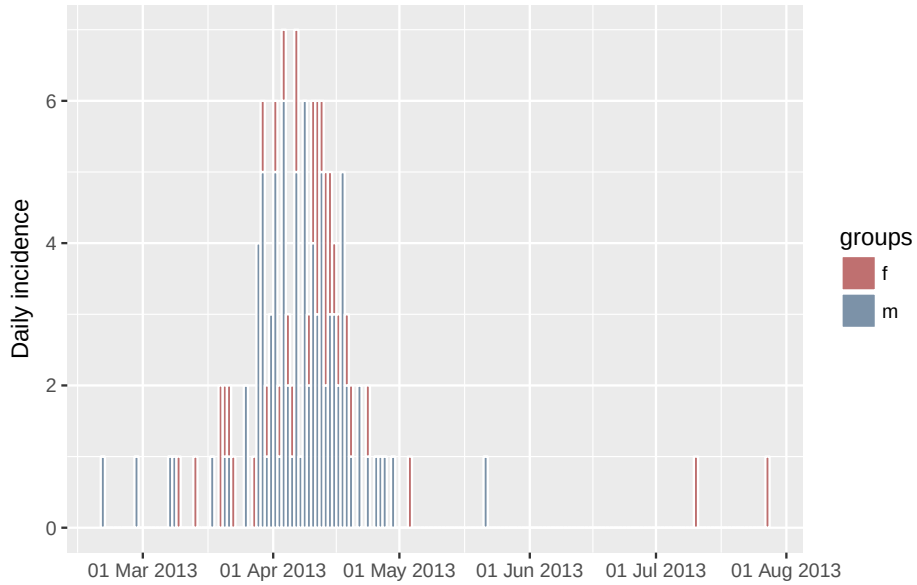
```
suppressPackageStartupMessages(library(tidyverse))
df <- fluH7N9_china_2013 %>%
  filter(!is.na(date_of_onset))

library(incidence)
(inc.gender.daily <- incidence(df$date_of_onset,
                               groups = df$gender,
                               interval = 1))
```

```
## <incidence object>
## [126 cases from days 2013-02-19 to 2013-07-27]
## [2 groups: f, m]
##
## $counts: matrix with 159 rows and 2 columns
## $n: 126 cases in total
## $dates: 159 dates marking the left-side of bins
## $interval: 1 day
## $timespan: 159 days
```

```
inc.gender.daily %>%
  as.data.frame() %>%
  head(3)
```

```
##        dates f m
## 1 2013-02-19 0 1
## 2 2013-02-20 0 0
## 3 2013-02-21 0 0
```

```
plot(inc.gender.daily, border = "white") +
  scale_x_date(labels = date_format("%d %b %Y"))
```



Weekly plot using *incidence*

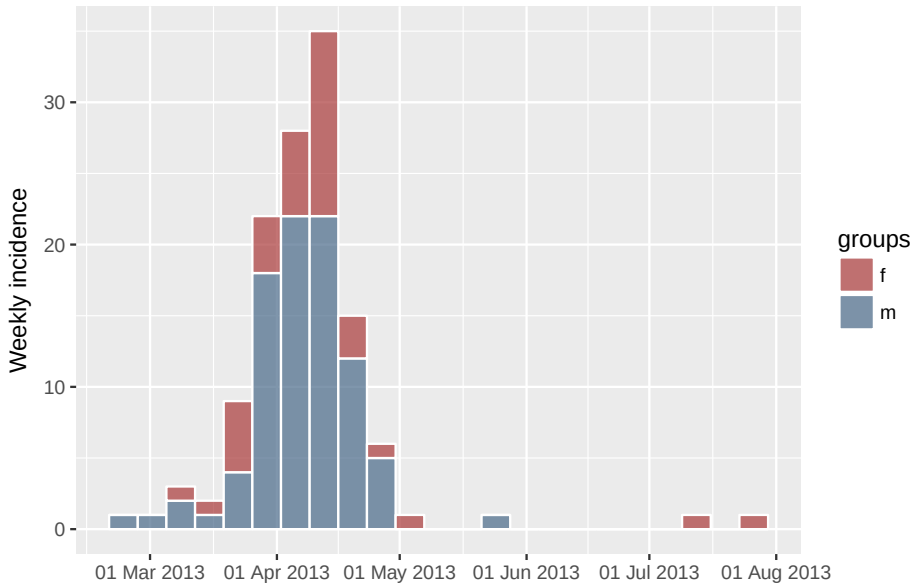
```
# weekly plot, but not calendar week-based
inc.gender.week <- incidence(df$date_of_onset,
                             groups = df$gender,
                             interval = 7,
                             iso_week = FALSE)
inc.gender.week
```

```
## <incidence object>
## [126 cases from days 2013-02-19 to 2013-07-23]
## [2 groups: f, m]
##
## $counts: matrix with 23 rows and 2 columns
## $n: 126 cases in total
## $dates: 23 dates marking the left-side of bins
## $interval: 7 days
## $timespan: 155 days
```

```
inc.gender.week %>%
  as.data.frame() %>%
  head(3)
```

```
##        dates f m
## 1 2013-02-19 0 1
## 2 2013-02-26 0 1
## 3 2013-03-05 1 2
```

```
plot(inc.gender.week, border = "white") +
  scale_x_date(labels = date_format("%d %b %Y"))
```



Pay attention to the red bar starting around 01 May 2013

ISOweek-based weekly plot using *incidence*

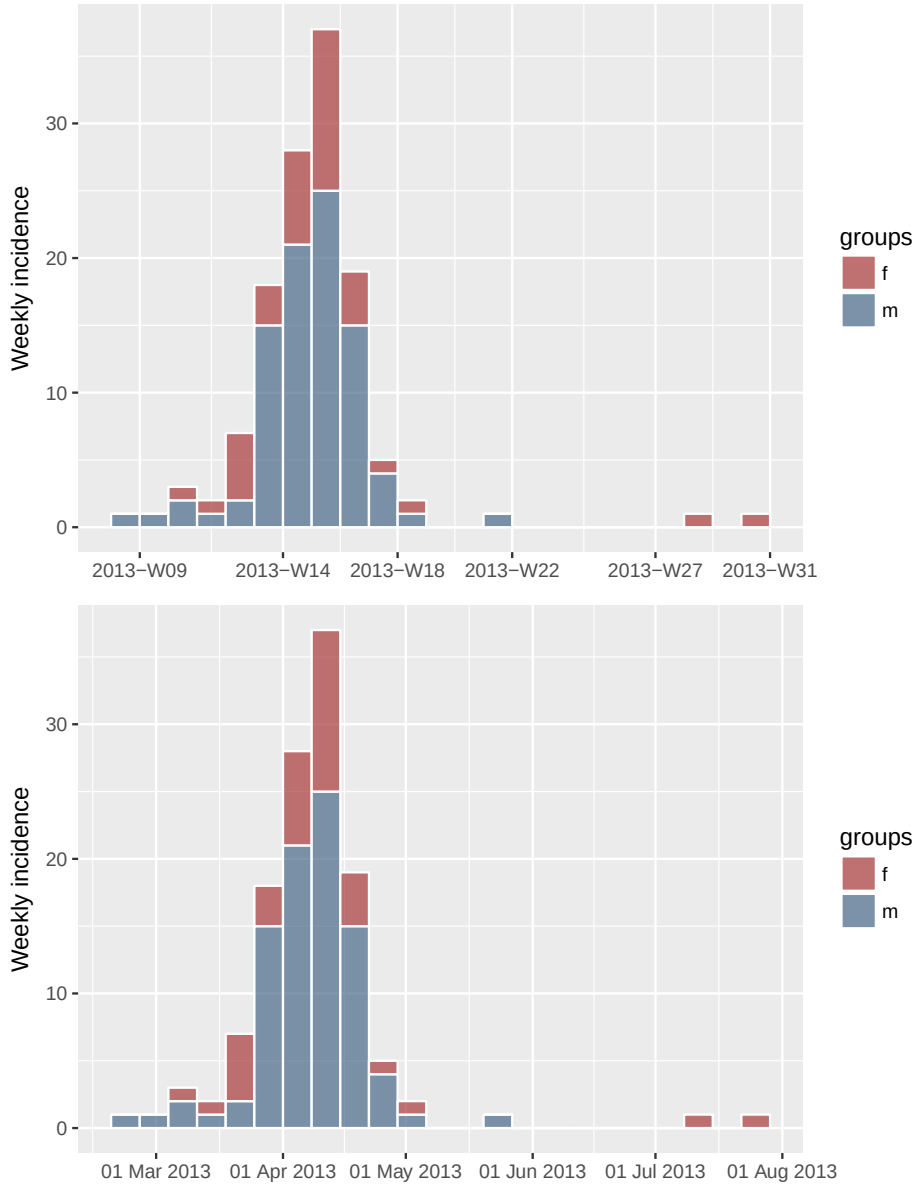
```
# isoweek-based weekly plot
inc.gender.isoweek <- incidence(df$date_of_onset,
                                groups = df$gender,
                                interval = 7,
                                iso_week = TRUE)
inc.gender.isoweek
```

```
## <incidence object>
## [126 cases from days 2013-02-18 to 2013-07-22]
## [126 cases from ISO weeks 2013-W08 to 2013-W30]
## [2 groups: f, m]
##
## $counts: matrix with 23 rows and 2 columns
## $n: 126 cases in total
## $dates: 23 dates marking the left-side of bins
## $interval: 7 days
## $timespan: 155 days
```

```
inc.gender.isoweek %>%
  as.data.frame() %>%
  head(3)
```

```
##         dates isoweeks f m
## 1 2013-02-18 2013-W08 0 1
## 2 2013-02-25 2013-W09 0 1
## 3 2013-03-04 2013-W10 1 2
```

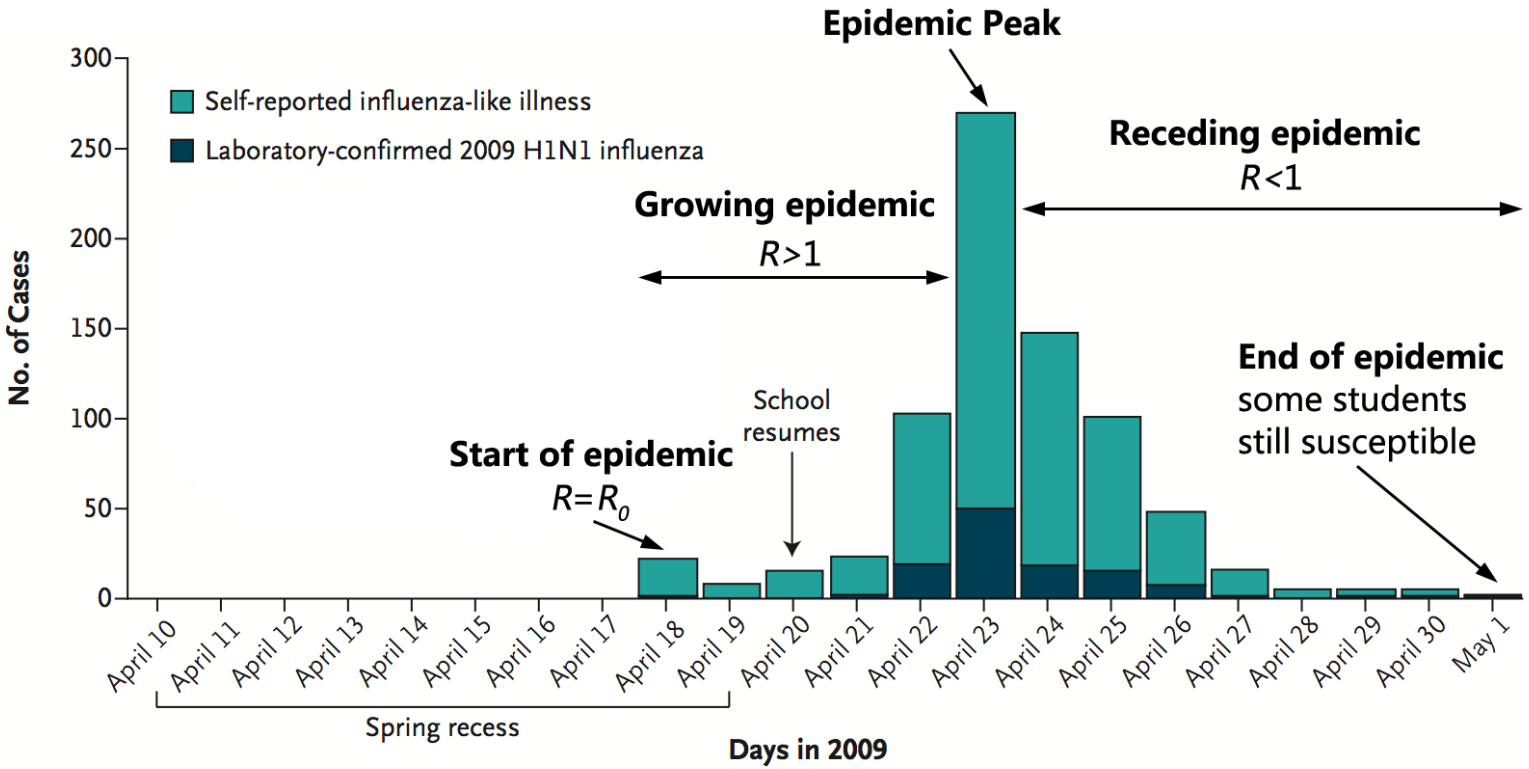
```
# by default label x-axis ticks with isoweek number
plot(inc.gender.isoweek, border = "white")
# label x-axis ticks with isoweek starting date
plot(inc.gender.isoweek, border = "white",
     labels_iso_week = FALSE) +
  scale_x_date(labels = date_format("%d %b %Y"))
```



The Reproduction Number R

the average number of secondary cases caused by a single, typical infected individual in a population with some level of susceptibility. (Nelson and Williams 2013, ch. 6, pp. 133)

- the primary metric used to quantify the transmission of a disease in infectious disease dynamics
- time-varying reproduction number R_t since R changes over time; basic reproduction number R_0 refers to a theoretical time 0 when the entire population is susceptible (Anderson and May 1992, ch. 2, pp. 17).
- R_t provides feedback on the effectiveness of interventions and on the the need to intensify control efforts; reducign R below the threshold value of 1 and as close to 0 as possible (Cori et al. 2013).



Epidemic Curve of Outbreak of 2009 Pandemic Influenza A (H1N1) at a New York City School (modified from Lessler, Reich, and Cummings 2009)

How to easily estimate R_t in real time?

EpiEstim package (in development)

A tool to estimate time varying reproduction numbers from epidemic curves

Anne Cori

Statistician specialized in disease modelling and outbreak response. Imperial College London, UK.

American Journal of Epidemiology
© The Author 2013. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.
Vol. 178, No. 9
DOI: 10.1093/aje/kwt133
Advance Access publication: September 15, 2013

Practice of Epidemiology

A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics

Anne Cori*, Neil M. Ferguson, Christophe Fraser, and Simon Cauchemez

* Correspondence to Dr. Anne Cori, Department of Infectious Disease Epidemiology, MRC Centre for Outbreak Analysis and Modelling, Imperial College London, St Mary's Campus, Norfolk Place, London W2 1PG, United Kingdom (e-mail: a.cori@imperial.ac.uk).

Initially submitted November 26, 2012; accepted for publication May 23, 2013.

The quantification of transmissibility during epidemics is essential to designing and adjusting public health responses. Transmissibility can be measured by the reproduction number *R*, the average number of secondary cases caused by an infected individual. Several methods have been proposed to estimate *R* over the course of an epidemic; however, they are usually difficult to implement for people without a strong background in statistical modeling. Here, we present a ready-to-use tool for estimating *R* from incidence time series, which is implemented in popular software including Microsoft Excel (Microsoft Corporation, Redmond, Washington). This tool produces novel, statistically robust analytical estimates of *R* and incorporates uncertainty in the distribution of the serial interval (the time between the onset of symptoms in a primary case and the onset of symptoms in secondary cases). We applied the method to 5 historical outbreaks; the resulting estimates of *R* are consistent with those presented in the literature. This tool should help epidemiologists quantify temporal changes in the transmission intensity of future epidemics by using surveillance data.

incidence; influenza; measles; reproduction number; SARS; smallpox; software

Abbreviations: CI, credible interval; SARS, severe acute respiratory syndrome.

Estimating R_t using *EpiEstim*

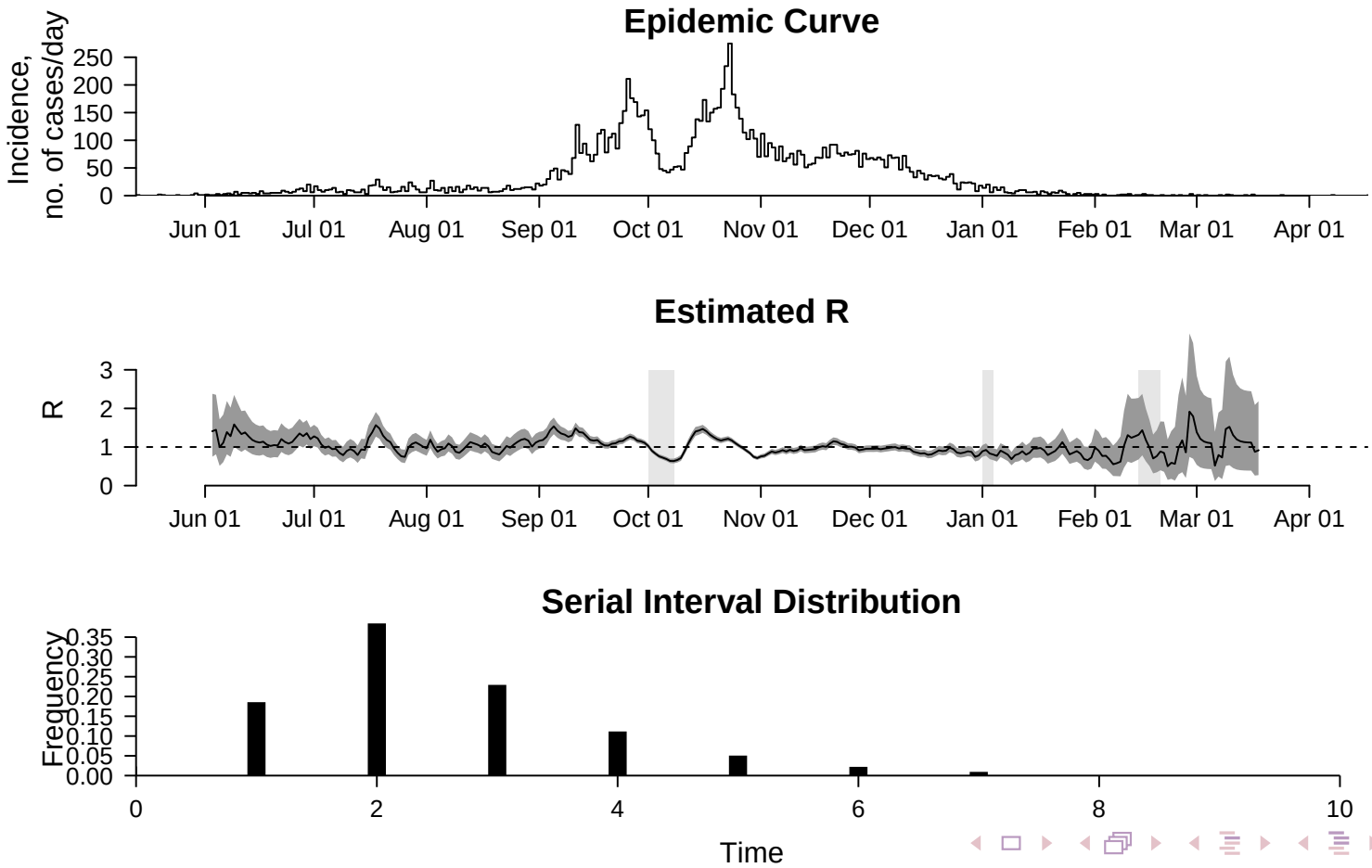
```
load("data/pdmH1N1_beijing_2009.rda")
glimpse(pdmH1N1_beijing_2009)
```

```
## Observations: 11,089
## Variables: 4
## $ card     <dbl> 35333862, 35451598, 35490152, 35491961, 35533905, 356...
## $ onset    <date> 2009-05-13, 2009-05-19, 2009-05-20, 2009-05-19, 2009...
## $ diagnose <date> 2009-05-16, 2009-05-20, 2009-05-22, 2009-05-22, 2009...
## $ type     <chr> "laboratory diagnosed", "laboratory diagnosed", "labo...
```

```
# daily epidemic curve
dec <- as.data.frame(incidence(pdmH1N1_beijing_2009$onset, interval = 1))

# install the latest development version from GitHub
# if (!requireNamespace("devtools")) install.packages("devtools")
# devtools::install_github("annecori/EpiEstim")
library(EpiEstim)
# the instantaneous reproduction number Ri can be estimated from the 16th day
# (May 28, 2009), and only up to the 310th day (March 18, 2010)
# The generation time of influenza A(H1N1)pdm09 had a mean of 2.6 days and
# an SD of 1.3 days, adapted from Yu.etal-Emerg.Infect.Dis.-2012
Ri <- EstimateR(dec$counts, T.Start = 16:304, T.End = 22:310,
                method = "ParametricSI", Mean.SI = 2.6, Std.SI = 1.3)
```

Estimating R_t using *EpiEstim* – Plot

```
# plot epidemic curve, estimated instantaneous reproduction number,
# and serial interval distribution with overlaying holidays
# plot.ec.R.SI() is my own function whose details are not shown here
plot.ec.R.SI(dec, Ri, plot.holidays = TRUE)
```



References

Roy M. Anderson and Robert M. May. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, 1992.

Centers for Disease Control and Prevention (CDC). "Outbreak of swine-origin influenza A (H1N1) virus infection-Mexico, March-April 2009." In: *MMWR* 58.17 (2009), p. 467. ISSN: 1545-861X. URL: https://www.cdc.gov/mmwr/preview/mmwrhtml/mm5817a5.htm.

Anne Cori et al. "A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics". In: *Am. J. Epidemiol.* 178.9 (Nov. 2013), pp. 1505–1512. URL: http://aje.oxfordjournals.org/content/178/9/1505.abstract.

Justin Lessler, Nicholas G. Reich, and Derek A.T. Cummings. "Outbreak of 2009 Pandemic Influenza A (H1N1) at a New York City School". In: *N Engl J Med* 361.27 (Dec. 2009), pp. 2628–2636. ISSN: 0028-4793. DOI: 10.1056/NEJMoa0906089.

Kenrad E. Nelson and Carolyn Williams. *Infectious disease epidemiology*. Jones & Bartlett Publishers, 2013.

Questions?

Contact Information

cai-j12@mails.tsinghua.edu.cn caijunthu@gmail.com

github.com/caijun

blog.tonytsai.name

To download this slides, please scan following QR code: