Weight of Statistical Evidence

Detection and Correction of Publication Bias

Servan Grüninger

Supervisor: Prof. Dr. Stephan Morgenthaler (EPFL)

External expert: Prof. Dr. Reinhard Furrer (University of Zurich)

Master's Programme in Computational Science and Engineering

Écublens, July 9th

Outline

Publication Bias—The Bane of Scientific Publishing

Many studies land in the file drawer (Image: Geckoboard)

The Woozle Effect

Pooh and Piglet tracking down the elusive Woozle (Image: Ernest H. Shepard)

Influenza Surveillance in the USA

Virology: total & influenza respiratory specimens per week

https://www.cdc.gov/flu/weekly/overview.htm

Influenza Surveillance in the USA

Virology: total & influenza respiratory specimens per week

ILI patients: % of total per week & state

https://www.cdc.gov/flu/weekly/overview.htm

Influenza Surveillance in the USA

Virology: total & influenza respiratory specimens per week

ILI patients: % of total per week & state

Mortality: influenza related deaths

https://www.cdc.gov/flu/weekly/overview.htm

Influenza Surveillance in the USA

Virology: total & influenza respiratory specimens per week

ILI patients: % of total per week & state

Mortality: influenza related deaths

Hospitalisation: influenza related hospitalisations

https://www.cdc.gov/flu/weekly/overview.htm

Influenza Surveillance in the USA

Virology: total & influenza respiratory specimens per week

ILI patients: % of total per week & state

Mortality: influenza related deaths

Hospitalisation: influenza related hospitalisations

Geographic spread of influenza: within each state

https://www.cdc.gov/flu/weekly/overview.htm

Influenza Surveillance in the USA

Strengths:

- reliable
- comparably fast (ca. 1-2 weeks)
- virological "ground truth"
- severe cases covered

Influenza Surveillance in the USA

Strengths:

- reliable
- comparably fast (ca. 1-2 weeks)
- virological "ground truth"
- severe cases covered

Weaknesses:

- only tip of the iceberg
- comparably slow (ca. 1-2 weeks)
- weak cases are missed

Twitter as complementary information source

Virology: total & influenza respiratory specimens per week

-> **ILI patients:** % of total per week & state

Mortality: influenza related deaths

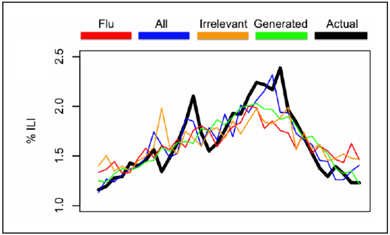
Hospitalisation: influenza related hospitalisations

-> **Geographic spread of influenza:** within each state

https://www.cdc.gov/flu/weekly/overview.htm

Flu surveillance via twitter: two approaches

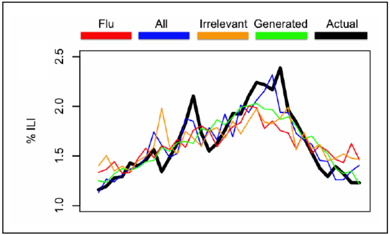
correlate aggregated tweets with aggregated CDC data



Bodnar and Salathé [2013]

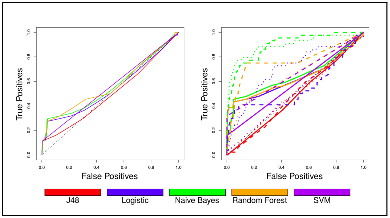
Flu surveillance via twitter: two approaches

correlate aggregated tweets with aggregated CDC data



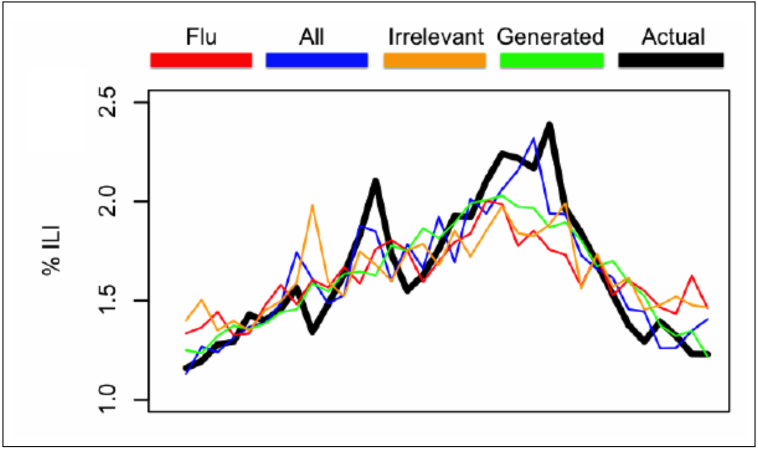
Bodnar and Salathé [2013]

correlate individual tweets with disease state of individuals



Bodnar et al. [2014]

Use aggregate data to detect the flu



Bodnar and Salathé [2013]

Use aggregate data to detect the flu

Strengths:

- ”easy” with enough data

Use aggregate data to detect the flu

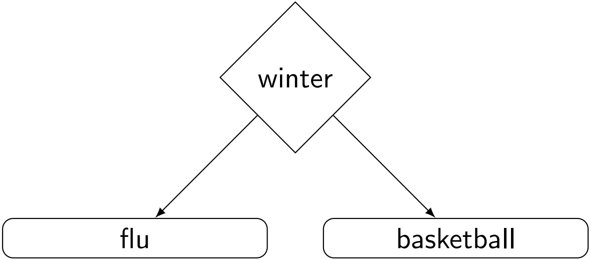
Strengths:

- ”easy” with enough data

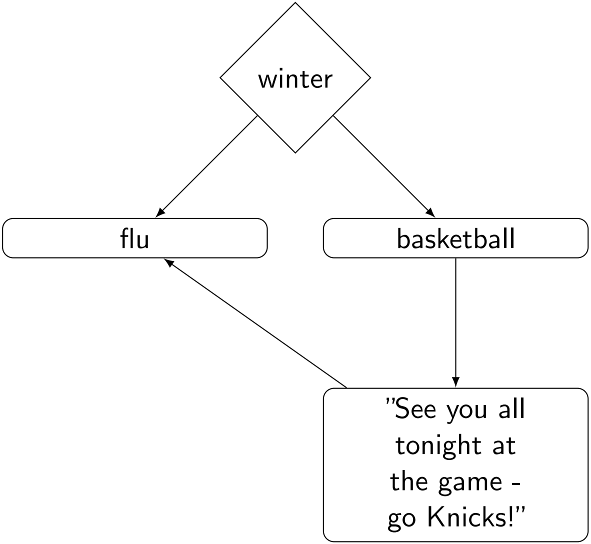
Weaknesses:

- prone to overfitting
- random data performed as good as tweets with flu related tweets
- temporal & spatial division of data influence model

The dangers of confounders

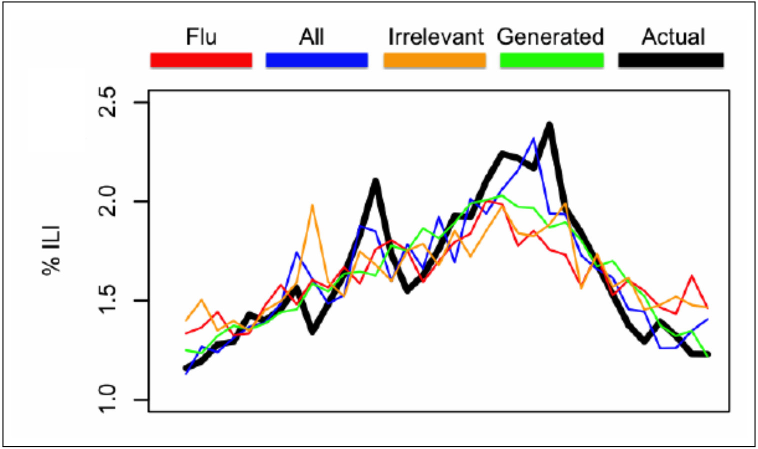


The dangers of confounders



The cautionary tale of Google Flu Trends

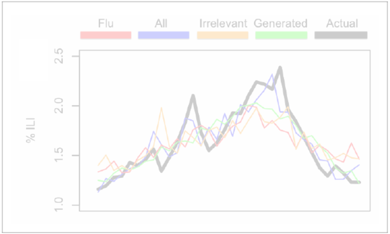
Overfitting: The bane of machine learning



Bodnar and Salathé [2013]

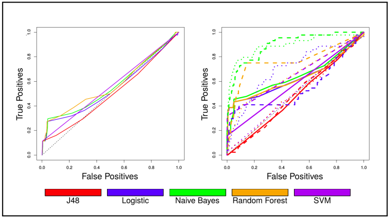
Flu surveillance via twitter: two approaches

correlate aggregated tweets with aggregated CDC data



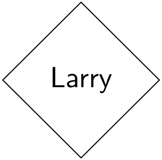
Bodnar and Salathé [2013]

correlate individual tweets with disease state of individuals

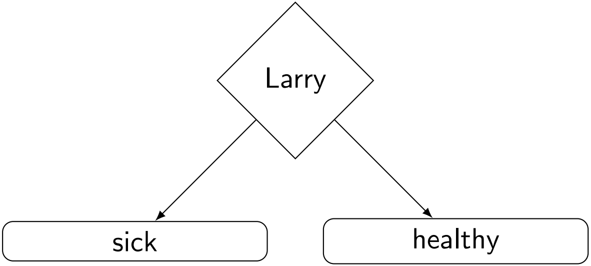


Bodnar et al. [2014]

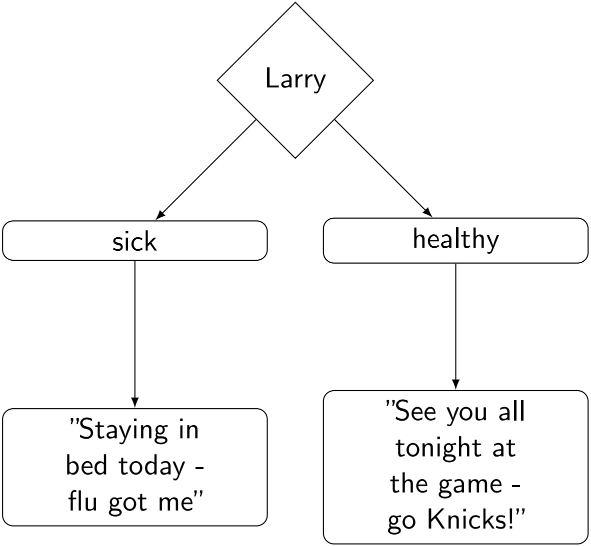
Use individual-level data to detect the flu



Use individual-level data to detect the flu



Use individual-level data to detect the flu



On the ground validation of online diagnosis with Twitter and medical records [Bodnar et al., 2014]

- 37'599 tweets from 104 accounts

On the ground validation of online diagnosis with Twitter and medical records [Bodnar et al., 2014]

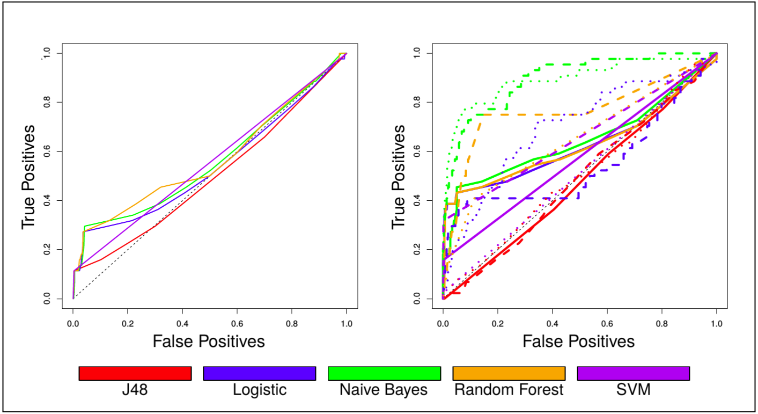
- 37'599 tweets from 104 accounts
- 1609 tweets from 35 users w/ medically diagnosed flu in 2011/2012 season

On the ground validation of online diagnosis with Twitter and medical records [Bodnar et al., 2014]

- 37'599 tweets from 104 accounts
- 1609 tweets from 35 users w/ medically diagnosed flu in 2011/2012 season
- ranked list of 12'393 most common keywords

On the ground validation of online diagnosis with Twitter and medical records [Bodnar et al., 2014]

- 37'599 tweets from 104 accounts
- 1609 tweets from 35 users w/ medically diagnosed flu in 2011/2012 season
- ranked list of 12'393 most common keywords
- rank established by different machine learning methods

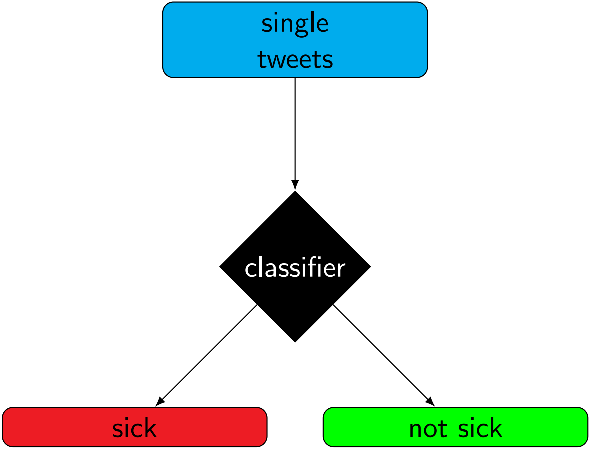


Bodnar et al. [2014]

The classifier model

- naive Bayes classifier: $p(C_k \mid \mathbf{x}) = \frac{p(C_k)\ p(\mathbf{x}|C_k)}{p(\mathbf{x})}$
- 100 most predictive keywords
- accuracy: 89.72%, AUC: 0.8544

Tweet classification: workflow



Strengths and weaknesses of the Twitter classifier

Strengths:

- ”ground truth” available
- high internal validity

Strengths and weaknesses of the Twitter classifier

Strengths:

- "ground truth" available
- high internal validity

Weaknesses:

- small data set (-> low external validity?)
  - 104 individuals w/ influenza, 122 individuals w/o influenza
  - only 1609 tweets from 35 users in "sick" category
- low temporal resolution (one month time window)
- ethical concerns (anonymity might be compromised)

Use individual data to detect the flu

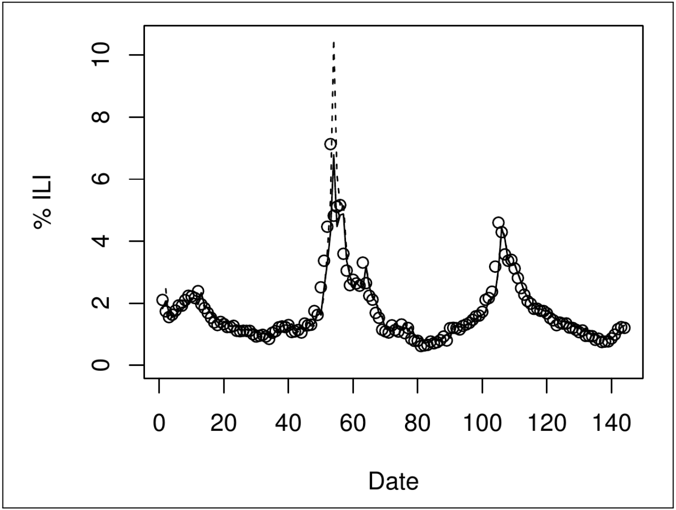
Strengths:

- ”ground truth” available
- high internal validity

Weaknesses:

- small data set (-> low external validity?)
  - 104 individuals w/ influenza, 122 individuals w/o influenza
  - only 1609 tweets from 35 users in “sick” category
- low temporal resolution (one month time window)
- ethical concerns (anonymity might be compromised)

A first proof of principle



Bodnar [2015]

An adventure in reproducibility

http://blog.revolutionanalytics.com/2014/10/introducing-rrt.html

To replicate or to reproduce?

Reproducibility according to Goodman et al. [2016]:

- methods reproducibility
- results reproducibility (replicability)
- inferential reproducibility

The three main goals:

- Assess the validity of the Twitter classifier for large data sets (results and inferential reproducibility)
- Reproduce key findings from Bodnar [2015] (methods reproducibility)
- Ensure the methods reproducibility of this thesis

The Twitter Data Set

The nature of the data beast

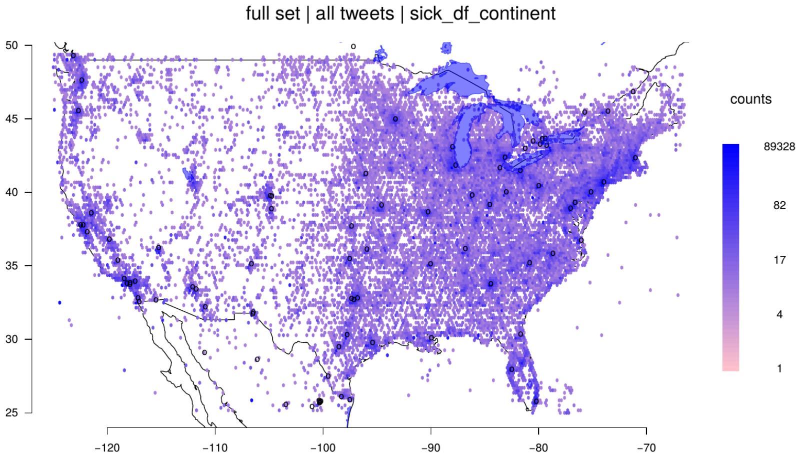
- `all_tweets`: the whole set of rated tweets (2'847'039'672 tweets)
- `one_hundred`: the rated tweets of those users who sent at least 100 tweets (42'611'004 tweets)
- `sick_users`: the rated tweets of all those users who sent at least one sick tweet (4'131'650 tweets)

The nature of the data beast

```
##           userID longitude latitude       time sick state
## [1,] 1000007198 -86.34844 39.63168 1424580963    0    30
## [2,] 1000007198 -86.34844 39.63168 1424580963    0    30
## [3,] 1000009051 -87.63464 24.39631 1409880397    0    56
## [4,] 1000009051 -87.63464 24.39631 1409880397    0    56
## [5,] 1000010509 -90.14008 29.86666 1394405061    0    36
## [6,] 1000010509 -90.13791 29.88957 1411750890    0    36
```

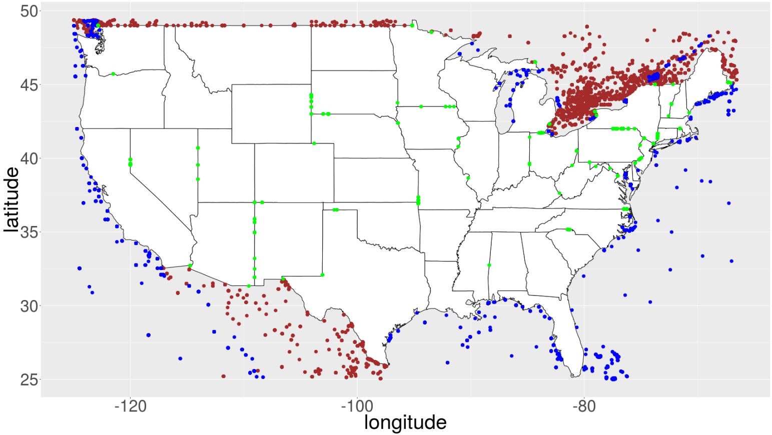
```
         used (Mb) gc trigger  (Mb) max used  (Mb)
Ncells 409223 21.9     847687  45.3   641597  34.3
Vcells 828218  6.4   23885155 182.3 23324386 178.0
```

The sick_users data set - full



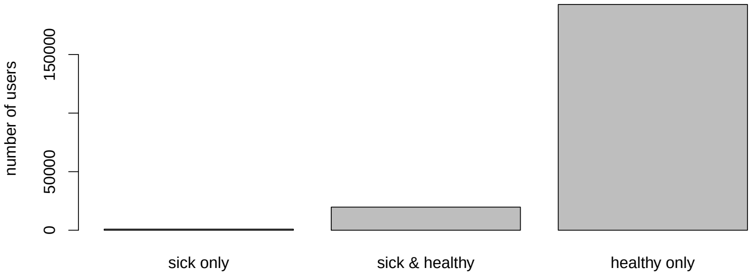
4'131'650 tweets from 222'446 users before pre-processing

sick_users: pruned tweets

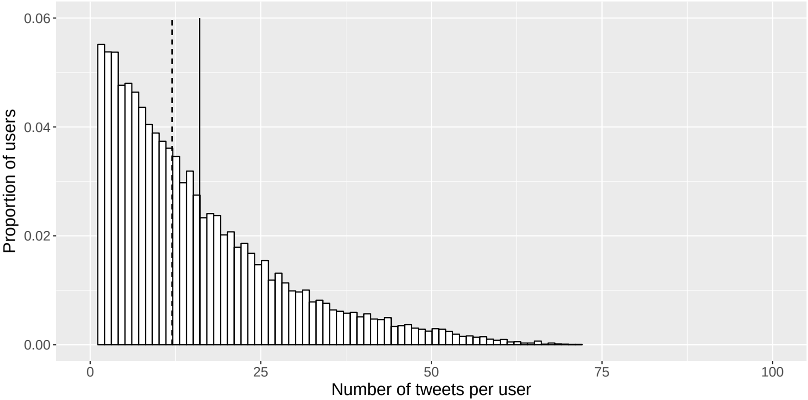


3'696'989 tweets from 213'426 users after pre-processing

`sick_users`: only few sick people in the data set

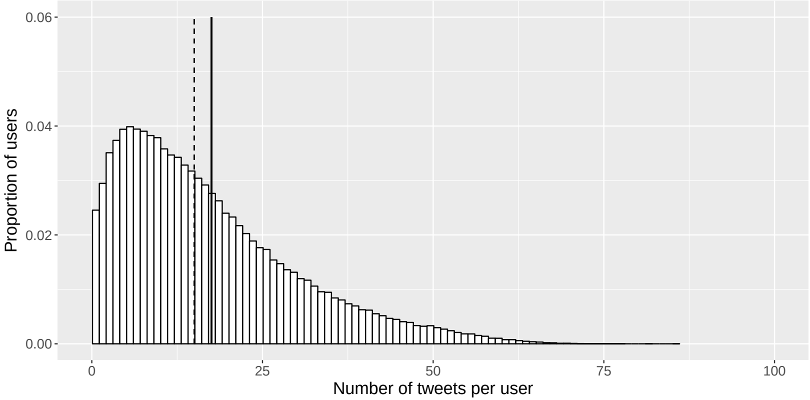


`sick_users`: differences between sick and healthy



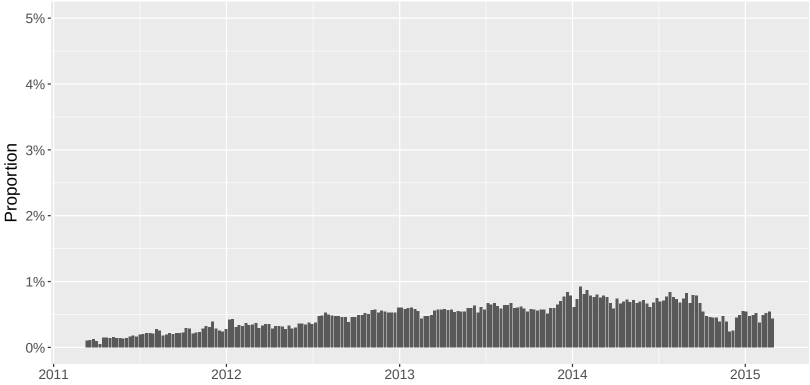
Mean = 16.01 (solid line); median = 12 (dashed line)

sick_users: differences between sick and healthy



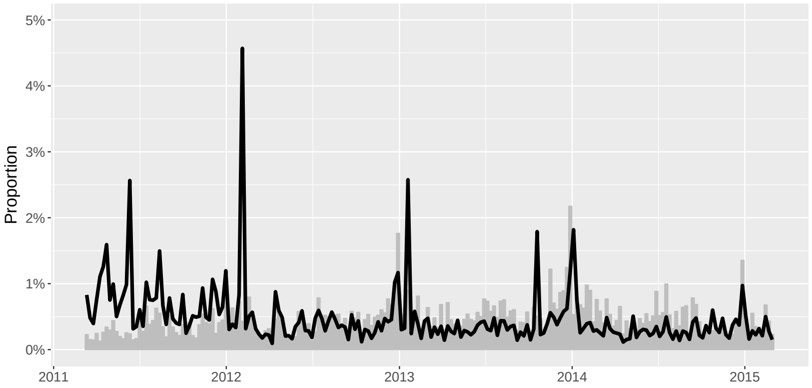
Mean = 17.53 (solid line); median = 15 (dashed line)

The `all_tweets` data set



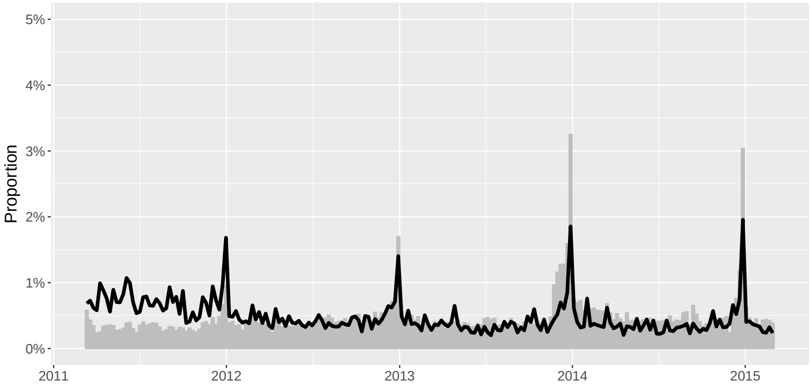
2’847’039’672 tweets sent by 16’015’981 users before pruning;
2’764’210’962 tweets and 15’229’049 users after pruning

all_tweets: sick tweets over time

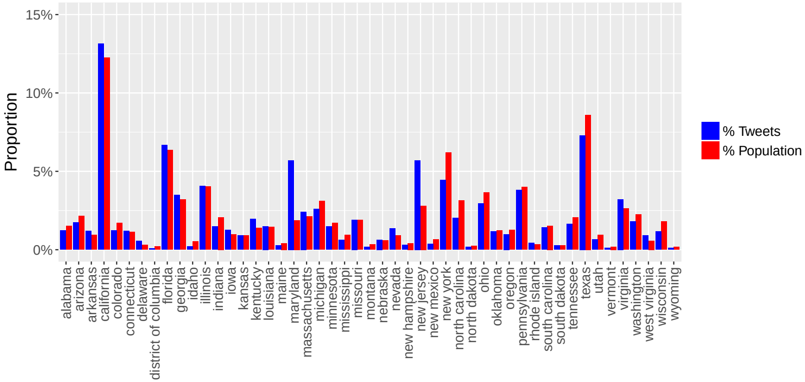


1’189’809 sick tweets from max. 27’052 users

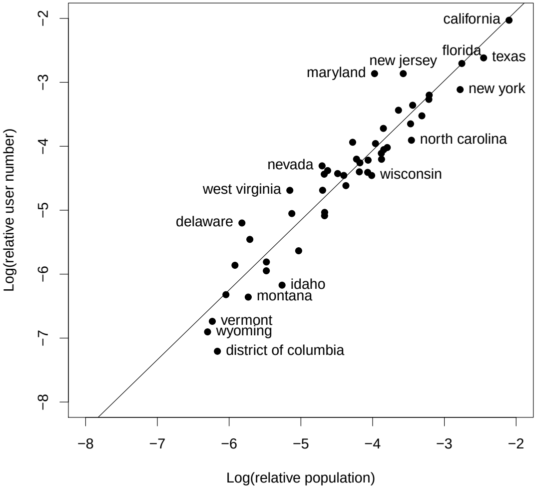
all_tweets: sick users over time



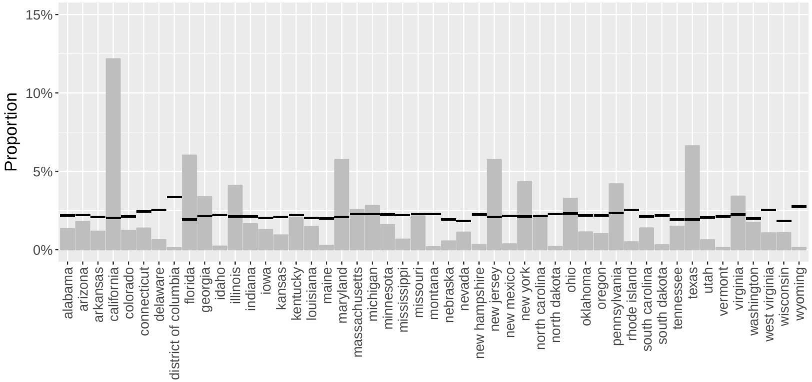
all_tweets: total users over space



all_tweets: total users over space



all_tweets: sick users over space



Competing For Gold

CDC ILI rates: the gold standard

Influenza-like illnesses are defined as:

- fever (body temperature of 37.8°C or greater) AND

CDC ILI rates: the gold standard

Influenza-like illnesses are defined as:

- fever (body temperature of 37.8°C or greater) AND
- cough and/or sore throat AND

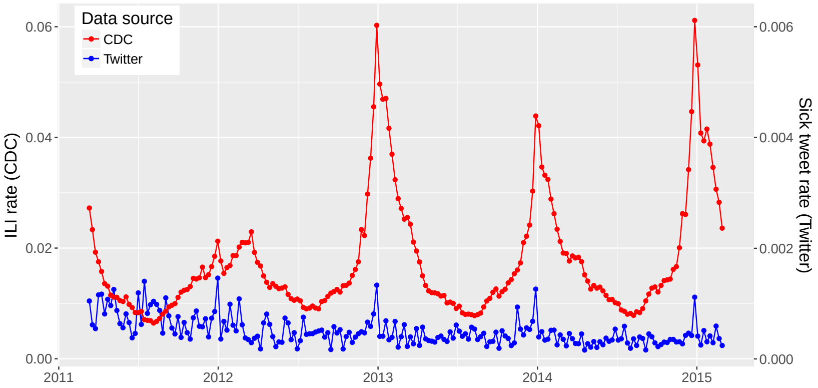
CDC ILI rates: the gold standard

Influenza-like illnesses are defined as:

- fever (body temperature of 37.8°C or greater) AND
- cough and/or sore throat AND
- no other known cause

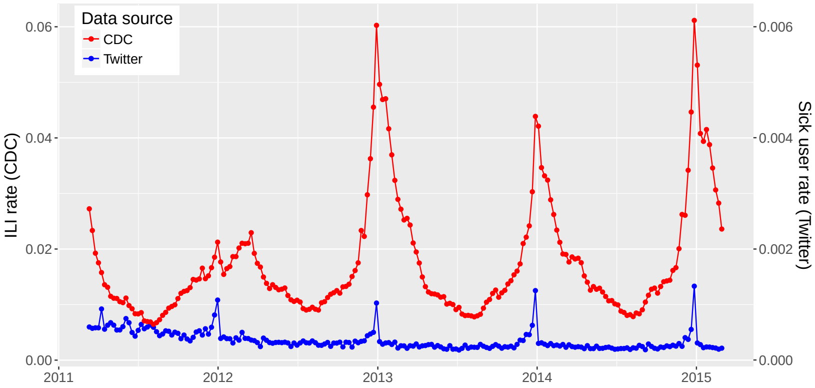
ILI rates are defined as the percentage of patients who visit ILINet sentinal clinics and show ILI symptoms.

CDC vs. Twitter: tweet-based



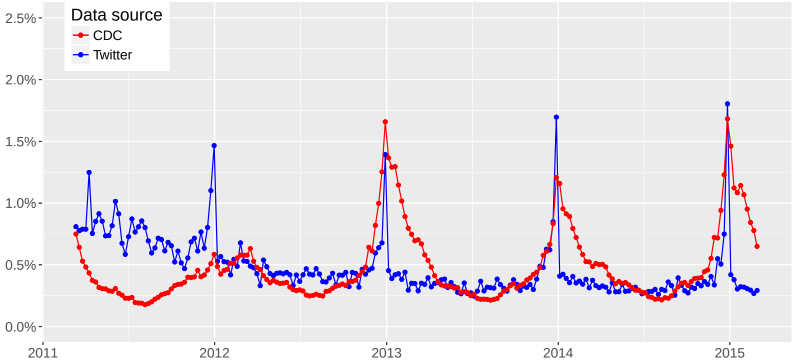
CDC ILI rates compared with Twitter sick tweet rate

CDC vs. Twitter: user-based

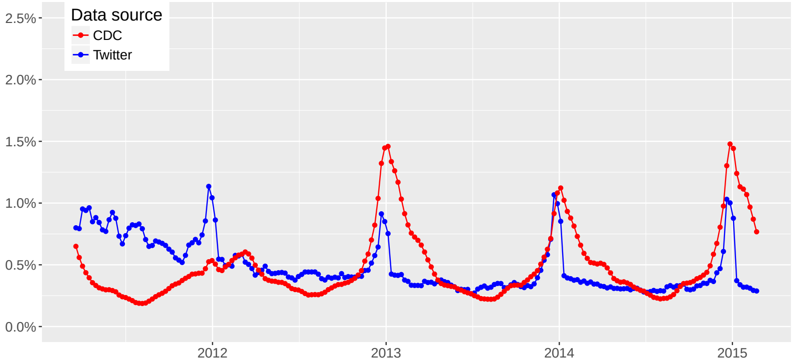


CDC ILI rates compared with Twitter sick user rates

CDC vs. Twitter: normalised and user-based



CDC vs. Twitter: normalised, user-based, and smoothed



CDC flu activity levels

0 insufficient data

1 below baseline

2 less than 1 SD above baseline

3 less than 2 SD above baseline

...

9 less than 8 SD above baseline

10 more than 8 SD above baseline

CDC flu activity levels - Baseline Calculation

CDC calculation

$$\text{baseline} = \frac{1}{n}\sum_{i=1}^{n} x_i + 2s$$

x_i = percentage of ILI patients among in week i;

n = # of non-influenza weeks in previous three season

s = sample standard deviation

CDC flu activity levels - Baseline Calculation

CDC calculation

baseline $= \frac{1}{n} \sum_{i=1}^{n} x_i + 2s$

$x_i =$ percentage of ILI patients among in week i;

$n = \#$ of non-influenza weeks in previous three season

$s =$ sample standard deviation

adopted for Twitter data set

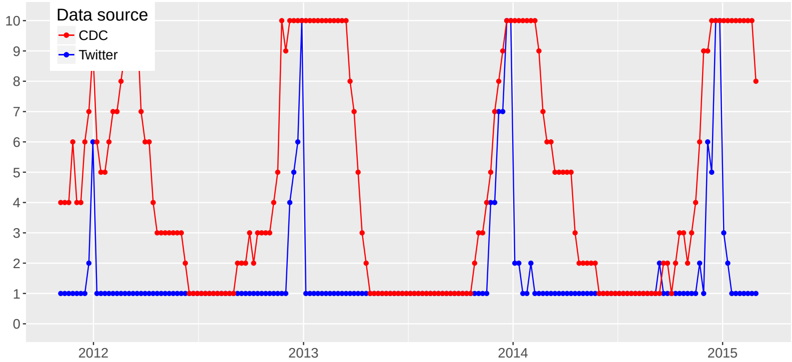
baseline $= \frac{1}{n} \sum_{i=1}^{n} x_i + 2s$

$x_i =$ number of tweets labelled "sick" in week i;

$n = \#$ of non-influenza weeks in previous summer (Jun, Jul, Aug & Sep)

$s =$ sample standard deviation

CDC vs. Twitter: based on activity levels



CDC ILI activity levels: spatiotemporal activity

Twitter flu activity levels: spatiotemporal activity

CDC vs. Twitter: spatiotemporal comparison

Exercises in Reproducibility

Reproducing key figures from Bodnar [2015]

- SIR model parameters β and γ

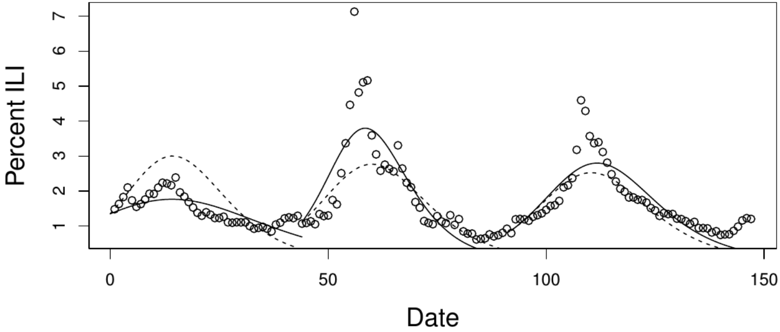
Reproducing key figures from Bodnar [2015]

- SIR model parameters β and γ
- SIR model figure

Reproducing key figures from Bodnar [2015]

- SIR model parameters β and γ
- SIR model figure
- full Twitter model figure

SIR model in Bodnar [2015]



SIR model based on yearly (solid) and combined (dashed) parameters

SIR model theory

- $\frac{dS}{dt} = -SI\beta$

SIR model theory

- $\frac{dS}{dt} = -SI\beta$
- $\frac{dI}{dt} = -SI\beta - I\gamma$

SIR model theory

- $\frac{dS}{dt} = -SI\beta$
- $\frac{dI}{dt} = -SI\beta - I\gamma$
- $\frac{dR}{dt} = I\gamma$

SIR model theory

- $\frac{dS}{dt} = -SI\beta$
- $\frac{dI}{dt} = -SI\beta - I\gamma$
- $\frac{dR}{dt} = I\gamma$

β and γ are estimated by minimising the residual sum of squares:

$$\mathsf{RSS} = \sum_t (I_{\gamma,\beta}(t) - I_{\mathsf{CDC}}(t))$$

SIR model: data basis for calculation

- `cdcoffset`
- `full_base`
- `full_autocor`
- `full_autocor2`
- `full_both`
- `full_both2`

Optimisation done with a simple grid-search algorithm.

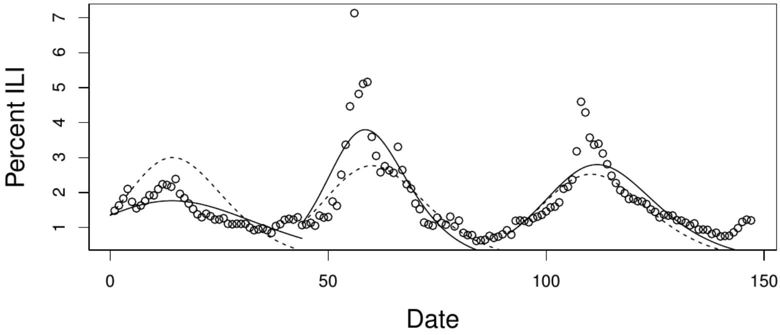
National best-fit parameters from the CDC (white) and Twitter data (grey)

Year	γ	β	RSS
2011-2012 (Bod)	0.4	0.44	0.00041
	0.37	0.41	0.00036

National best-fit parameters from the CDC (white) and Twitter data (grey)

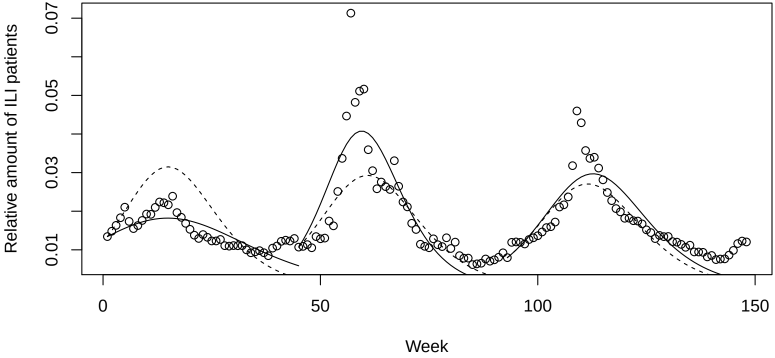
Year	γ	β	RSS
2011-2012 (Bod)	0.4	0.44	0.00041
	0.37	0.41	0.00036
2011-2012 (Gru)	0.17	0.17	0.00010
	0.12	0.12	0.00013

SIR model in Bodnar [2015]

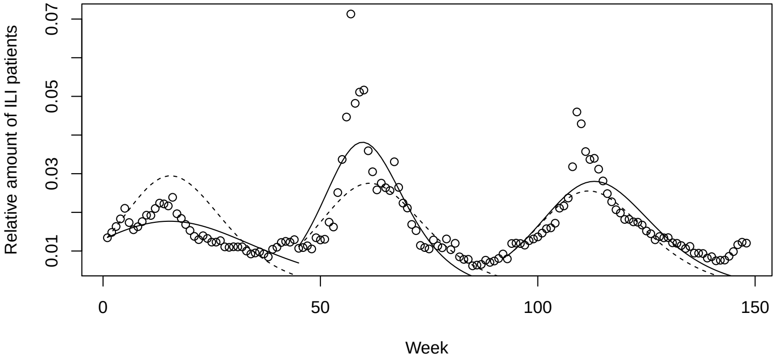


SIR model based on yearly (solid) and combined (dashed) parameters

SIR model: based on `cdcoffset`

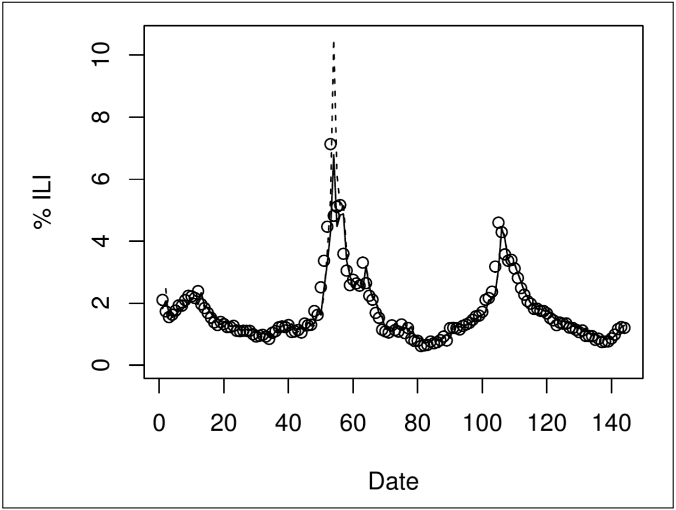


SIR model: based on `full_both2`



Full model consisting of AR(2) model based on CDC data and Twitter base model

Full Twitter model from Bodnar [2015]

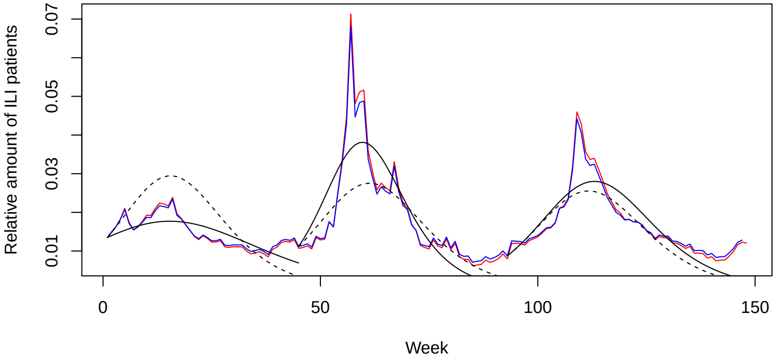


Bodnar [2015]

Full Twitter model from Bodnar [2015]

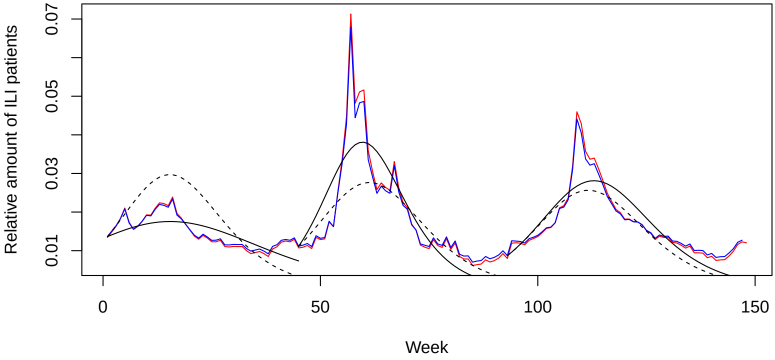
$$I_{\mathsf{full}}(t+1) = a \cdot I_{\mathsf{CDC}}(t-1) + b \cdot I_{\mathsf{CDC}}(t) + c \cdot I_{\mathsf{Twitter}}(t) + d.$$

Full Twitter model from Bodnar [2015]



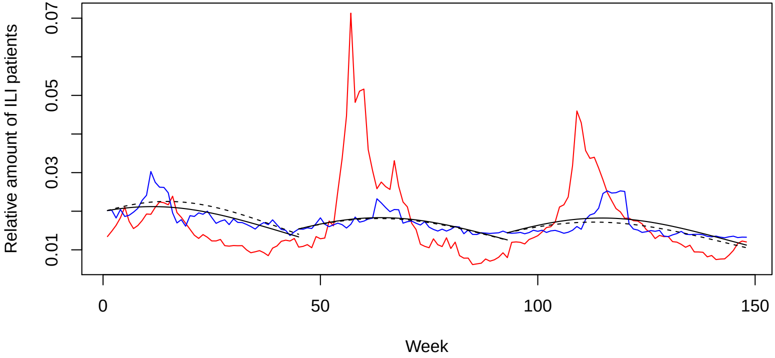
Full model consisting of AR(2) model based on CDC data and Twitter base model

CDC AR(2) model

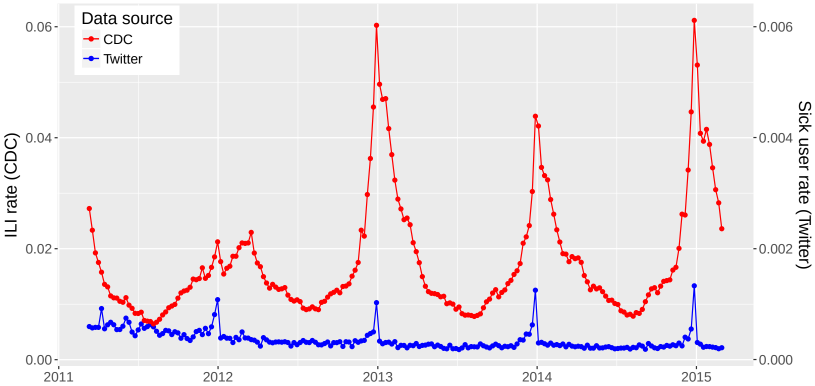


Model consisting of AR(2) model based on CDC data

Twitter base model



CDC vs. Twitter: user-based



CDC ILI rates compared with Twitter sick user rates

Discussion

Discrepancies between Grüninger [2017] and Bodnar [2015]

With regard to:

- methods reproducibility: figures & tables not reproducible

Discrepancies between Grüninger [2017] and Bodnar [2015]

With regard to:

- methods reproducibility: figures & tables not reproducible
- results reproducibility: Classifier results did not fit CDC rates

Discrepancies between Grüninger [2017] and Bodnar [2015]

With regard to:

- methods reproducibility: figures & tables not reproducible
- results reproducibility: Classifier results did not fit CDC rates
- inferential reproducibility: Classifer detects flu peaks, but is only of complementary use

Discrepancies between Grüninger [2017] and Bodnar [2015]

With regard to:

- methods reproducibility: figures & tables not reproducible
- results reproducibility: Classifier results did not fit CDC rates
- inferential reproducibility: Classifer detects flu peaks, but is only of complementary use

Most likely explanation: Data set used by Bodnar [2015] was pruned, filtered, or otherwise transformed.

Support for "Transformation hypothesis"

- stark discrepancies with regard to basic parameters of the two dat sets (e.g. mean tweets sent, total number of users, relative distribution of sick users)

Differences in data sets

	Bodnar	Grüninger	Note
# users	15'560'328	15'229'049	raw

Differences in data sets

	Bodnar	Grüninger	Note
# users	15'560'328	15'229'049	raw
# tweets	2'732'174'105	2'847'039'672	raw

Differences in data sets

	Bodnar	Grüninger	Note
# users	15’560’328	15’229’049	raw
# tweets	2’732’174’105	2’847’039’672	raw
# sick users (2011-2015)	182’801	27’052	processed

Differences in data sets

	Bodnar	Grüninger	Note
# users	15'560'328	15'229'049	raw
# tweets	2'732'174'105	2'847'039'672	raw
# sick users (2011-2015)	182'801	27'052	processed
mean tweet rate	175.59 tw/pp	31.42 tw/pp	per week

Differences in data sets

	Bodnar	Grüninger	Note
# users	15’560’328	15’229’049	raw
# tweets	2’732’174’105	2’847’039’672	raw
# sick users (2011-2015)	182’801	27’052	processed
mean tweet rate	175.59 tw/pp	31.42 tw/pp	per week
median tweet rate	10 tw/pp	32.53 tw/pp	per week

Differences in data sets

	Bodnar	Grüninger	Note
# users	15'560'328	15'229'049	raw
# tweets	2'732'174'105	2'847'039'672	raw
# sick users (2011-2015)	182'801	27'052	processed
mean tweet rate	175.59 tw/pp	31.42 tw/pp	per week
median tweet rate	10 tw/pp	32.53 tw/pp	per week
# total users in 2011	45'086	175'382	processed

Support for "Transformation hypothesis"

- stark discrepancies with regard to basic parameters of the two dat sets (e.g. mean tweets sent, total number of users, relative distribution of sick users)
- sick rates in Bodnar [2015] were 10-times larger than in my set
- `sick_user` subset

Discrepancies between Grüninger [2017] and Bodnar [2015]

With regard to:

- methods reproducibility: figures & tables not reproducible
- results reproducibility: Classifier results did not fit CDC rates
- inferential reproducibility: Classifer detects flu peaks, but is only of complementary use

Another explanation: Data set used by Bodnar [2015] was the same, but there were crucial gaps and/or fundamental errors in reporting.

Why not to trust Grüninger [2017]?

- no possibility to re-run classifier

Why not to trust Grüninger [2017]?

- no possibility to re-run classifier
- large, unwieldy data-set, analysed by novice data scientist

Why not to trust Grüninger [2017]?

- no possibility to re-run classifier
- large, unwieldy data-set, analysed by novice data scientist
- no access to complete code and processed data base used by Bodnar [2015]

Why to trust Grüninger [2017]?

- analysed & aggregated data in various ways (spatial, temporal) and with various methods

Why to trust Grüninger [2017]?

- analysed & aggregated data in various ways (spatial, temporal) and with various methods
- key statistics of aggregated data resemble real-world Twitter statistics more closely

Why to trust Grüninger [2017]?

- analysed & aggregated data in various ways (spatial, temporal) and with various methods
- key statistics of aggregated data resemble real-world Twitter statistics more closely
- Because the literature and Bodnar [2015] say so

Recommendations

- Assessing validity of analysis of Grüninger [2017]

Recommendations

- Assessing validity of analysis of Grüninger [2017]
- re-establishing functionality of Twitter classifier

Recommendations

- Assessing validity of analysis of Grüninger [2017]
- re-establishing functionality of Twitter classifier
- testing classifier with representative subset of Twitter users

Stop.

References I

Todd Bodnar and Marcel Salathé. Validating models for disease detection using twitter. *Proceedings of the 22nd international conference on World Wide Web companion*, (699-702), 2013. bibtex: bodnar_validating_2013.

Todd Bodnar, Victoria C. Barclay, Nilam Ram, Conrad S. Tucker, and Marcel Salathé. On the ground validation of online diagnosis with Twitter and medical records. pages 651–656. ACM Press, 2014. ISBN 978-1-4503-2745-9. doi: 10.1145/2567948.2579272. URL http://dl.acm.org/citation.cfm?doid=2567948.2579272. bibtex: bodnar_ground_2014.

Todd Bodnar. Data science with social media for epidemiology and public health, 2015.

Steven N. Goodman, Daniele Fanelli, and John P. A. Ioannidis. What does research reproducibility mean? *Science Translational Medicine*, 8(341): 341ps12–341ps12, June 2016. ISSN 1946-6234, 1946-6242. doi: 10.1126/scitranslmed.aaf5027. URL http://stm.sciencemag.org/content/8/341/341ps12.