

How to be Careful with Covid-19 Counts: A
Practical Guide to Working with Pandemic Panel
Data

Rex W. Douglass

2020-05-02

Contents

1	Executive Summary	5
1.1	Key Takeaways (TLDR)	6
2	Global COVID-19 Count Data	7
2.1	Takeaways	7
2.2	What data are available?	8
2.3	What is their geographic coverage?	9
2.4	What is their temporal coverage?	12
2.5	Where and How do they Disagree?	16
3	Tests	17
3.1	Tested People versus Tested Samples	17
3.2	Interpolate Within Observed	17
3.3	Interplate Prior to Observed	17
3.4	Interpolate Subnationally	17
3.5	Explaining Variation in Testing	17
4	Common Measures of Interest	19
4.1	R0 and R	19
4.2	Case Fatality Rate (CFR)	19
4.3	Percent Positive	19
5	Deaths	21
6	Actual Infections	23
7	Conclusion	25

Chapter 1

Executive Summary

How should we interpret the endless stream of figures and maps of COVID-19 produced by health departments and organizations around the world? For better or worse, we primarily experience large complicated events through counts—How many are dead?; How many are sick?; How many tests did we perform? These are universal questions, immediately accessible to both the producers of information like doctors and scientists and consumers of information like policy makers and citizens. For anyone who regularly works with the answers to those questions, the actual data, you know that every one of those simple numbers in a cell needs a big asterisks pointing to a long footnote explaining all of the problems in developing and using that number. This book is that footnote for COVID-19 counts. It is intended as a practical guide for using COVID-19 data, and all of the regularities and subtle gotchas you can expect to find.

This guide is built around a new resource developed at the Machine Learning for Social Science Lab called the Global Census of Covid-19 Counts (GC3). This is a single normalized and georeferenced aggregation of all of the other public aggregations of COVID-19 counts data available. We are currently aggregating 27 databases, who are in turn scraping and aggregating over ten thousand sources like public statements, news reports, and individual health department websites. Only by mosaicing all of these different resource together (917,149 observations and growing), are we able to finally provide full temporal coverage over the entire COVID-19 pandemic, and full spatial coverage over all countries in the world and in most places states/provinces as well. We are now able to track counts of confirmed cases and deaths in 5,168 locations, and number of tests performed in 1,233 locations.

This book is a deep dive into what problems and opportunities you can expect to find in these data. It is organized in order from simpler issues of data acquisition and aggregation, to more complicated questions of bias and latent true measurement.

1.1 Key Takeaways (TLDR)

You can write citations, too. For example, we are using the **bookdown** package (Xie, 2020) in this sample book, which was built on top of R Markdown and **knitr** (Xie, 2015).

Chapter 2

Global COVID-19 Count Data

2.1 Takeaways

Any data aggregation and cleaning approach will have to deal with the following issues

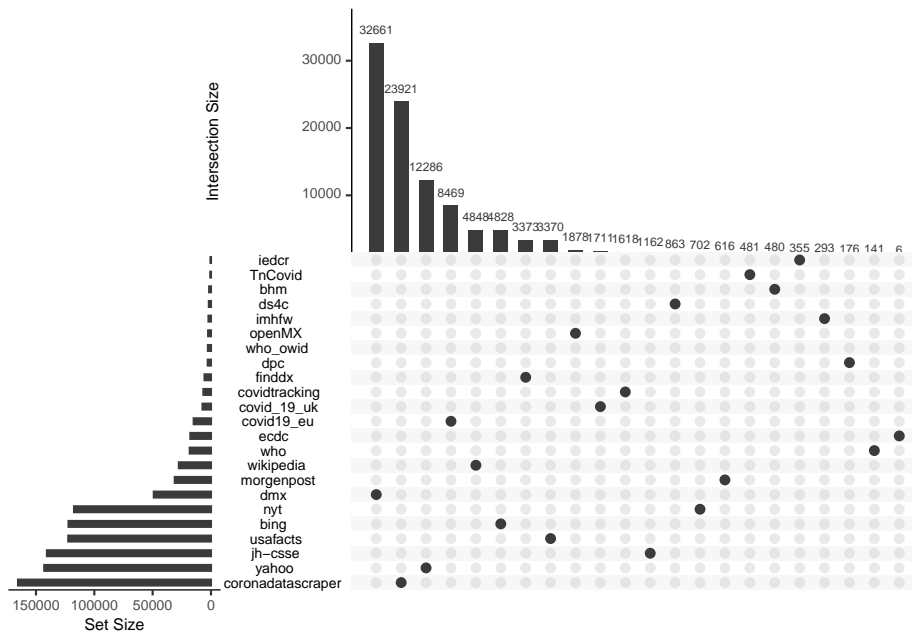
- Missingness
 - Prior to the first reported observation
 - After the last reported observation
 - Within a time series between observations
 - Unbalanced across different sources
- Structural Changes
 - Changes in reporting criteria/definitions
 - Changes in sourcing for underlying data
- Disagreement
 - One or more sources report different numbers
- Errors
 - Outliers
 - Merging errors
- Bias
 - Correlation between missingness and measurement
 - Attenuation bias

2.2 What data are available?

The Global Census of Covid-19 Counts (GC3) currently aggregates 27 databases. The databases vary drastically in size, scope, collection method, and purpose. On the small end are github repositories built around collecting a single country's published statistics, often available in an unstructured form on a government website in a native language. Others are official government statistics reported directly to and compiled by international organizations, like the World Health Organization (WHO) or the European Centre for Disease Prevention and Control. Some are news organizations that collect and compile official government statistics, like the New York Times and Reuters. Nonprofits like the Covidtracking Project compile records on specific issues like testing. Wikipedia provides an interface for a massive army of volunteers to enter in statistics into tabular formats that can later be scraped. The largest and most comprehensive scraping effort is the Corona Data Scraper from the Covid Atlas which only consumes sources directly from government and health organizations (excluding news and wikipedia). These all in turn are then ingested by larger aggregation projects. Johns Hopkins University Center for Systems Science and Engineering (JHU-CSSE) is the most widely used aggregator by downstream projects. Both Microsoft's Bing research Unit and Yahoo! have in turn recently made available their knowledge graph coverage of Covid-19 counts.

Their names, links, and cleaned observation counts appear in the table below.

The unit of observation in our data is the location-day-outcome, e.g. Texas-April 2-Deaths. The Upset plot below shows the number of unique location-day-outcome observations provided by each database along the vertical axis and the number of unique observations found only in that database and no other. Which databases will provide the most unique information is difficult to tell apriori. In general, the databases with the most observations and that rely on direct collection from raw sources rather than aggregation of others, tend to provide the most. For example, Corona Data Scraper provides both the most total and most unique observations. The most unique contributions come from the Corona Data Scraper Project, which might be anticipated by their overall size. The second most unique observations however comes from Wikipedia which is surprising because our treatment of it is currently ad hoc and it should already be ingested by other sources. It goes to show that no single source, or even no small combination of sources, is sufficient to provide full temporal and spatial coverage over even this relatively brief period of the Covid-19 pandemic.



2.3 What is their geographic coverage?

2.3.1 Country Level Data Availability

Despite this major effort by data producers, collectors, and aggregators, there is still major geographic variation in availability across countries. Most notably in availability of counts on number of tests performed, particularly in Central Africa.

2.3.2 State/Province Level Data Availability

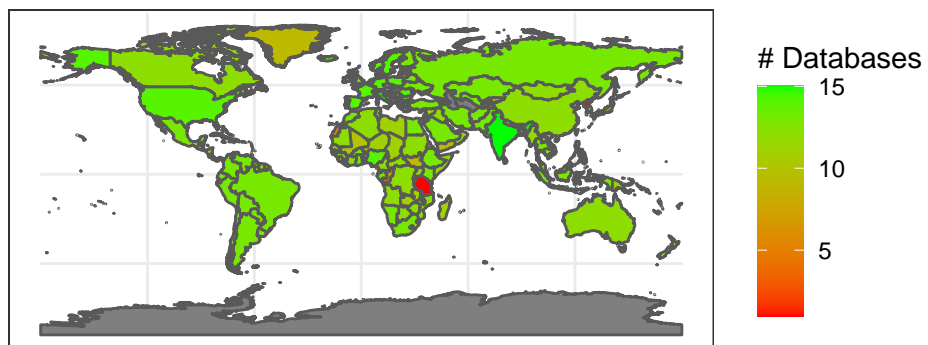
Disparities in coverage across countries is most dramatic at the subnational level.

2.3.3 County District Level Data Availability

This takes a long time to run so we're disabling it until the end

This takes a long time to run so we're disabling it until the end

Number of Databases with Coverage of each Country



Percent of Days with Testing Counts

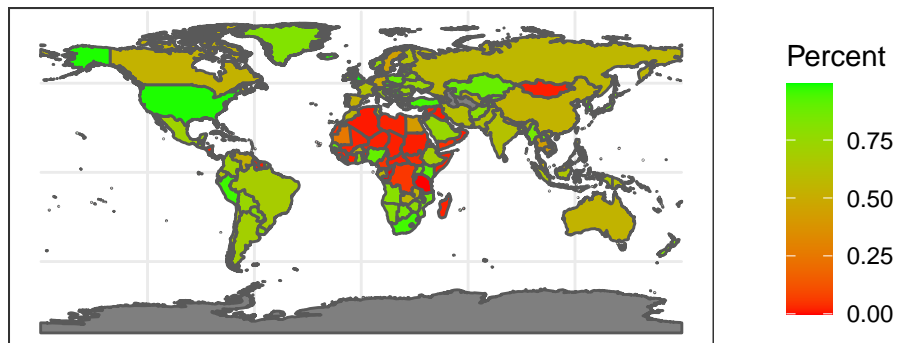


Figure 2.1: Data coverage by country.

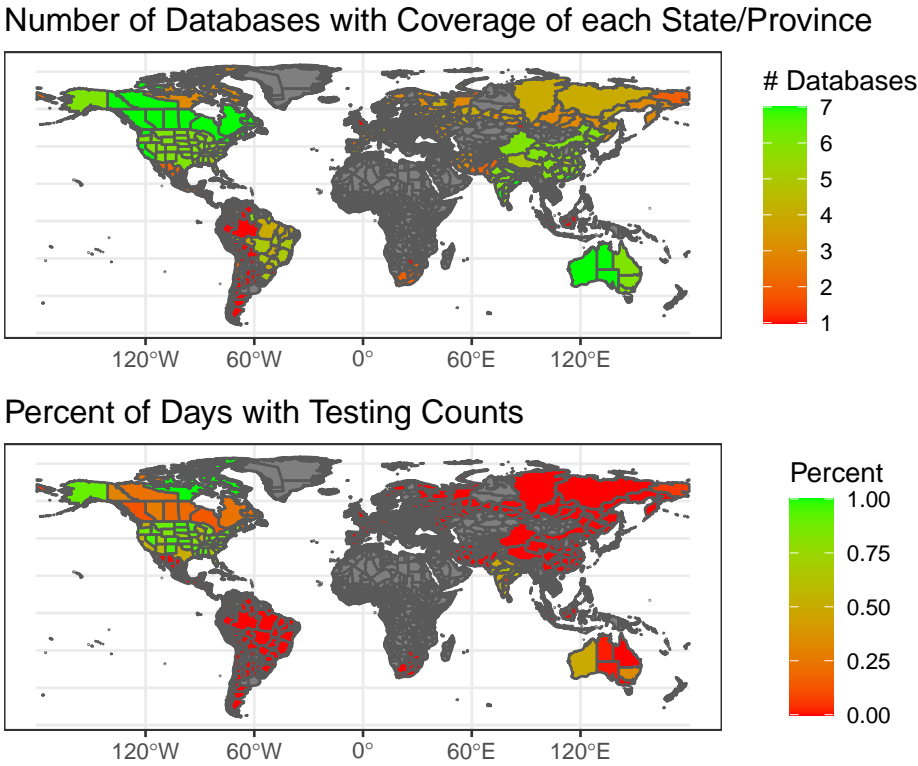


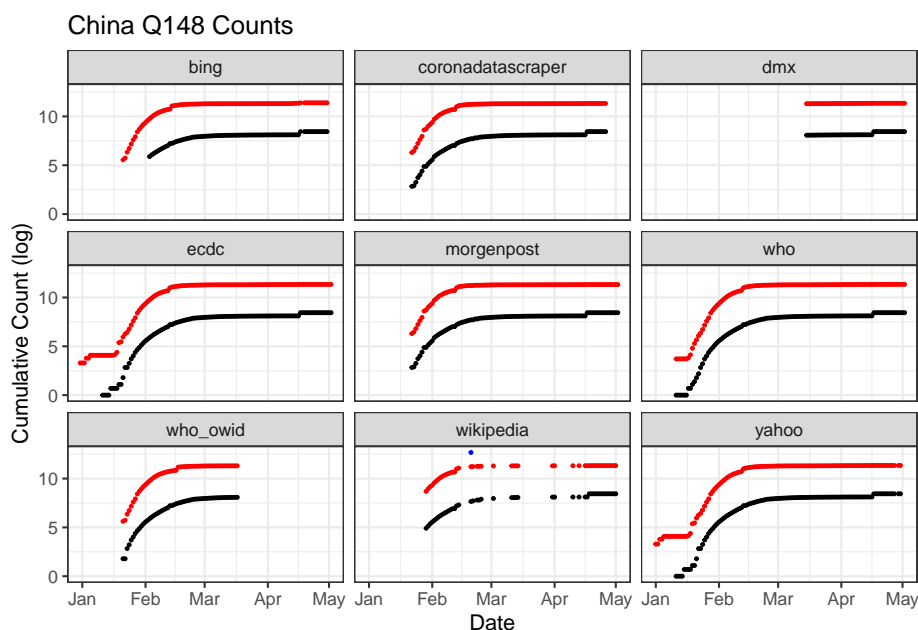
Figure 2.2: Data coverage by State/Province

2.4 What is their temporal coverage?

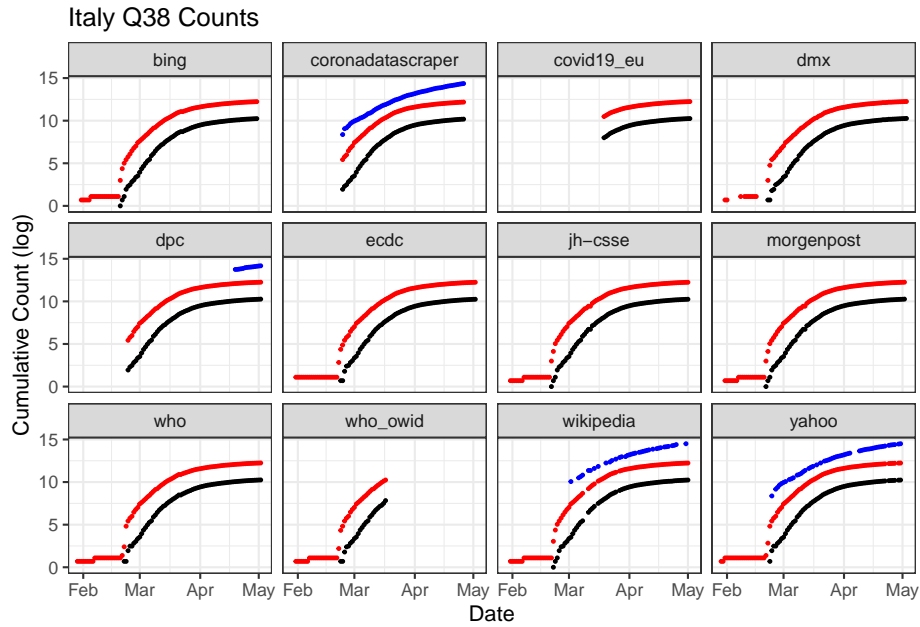
2.4.1 Coverage Across Three Countries

Figures x,y,z illustrate the problem of data coverage for 3 countries, China, Italy, and the U.S.

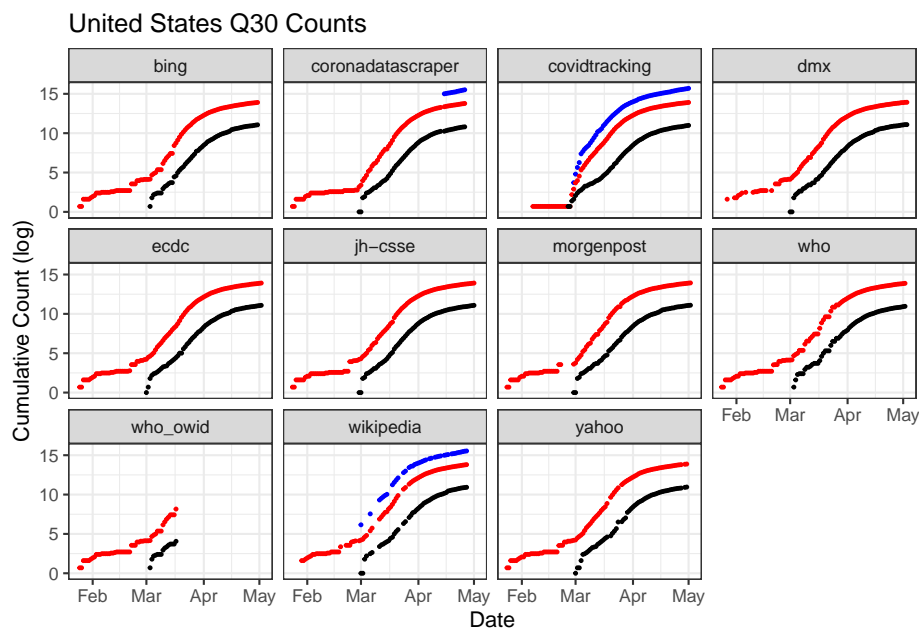
China outright refuses to release daily counts of testing. Only three databases document the beginning of the outbreak, the ECDC, the WHO, and Yahoo. On April 17, China changed its reporting which added 1,290 more deaths for Wuhan city only. The change is not retrospective, it shows up only a sharp discontinuity across multiple datasets.



Italy's coverage across datasets is fairly good and uniform, though there are breaks in coverage of testing for some datasets as well as variation in when each dataset starts tracking testing.

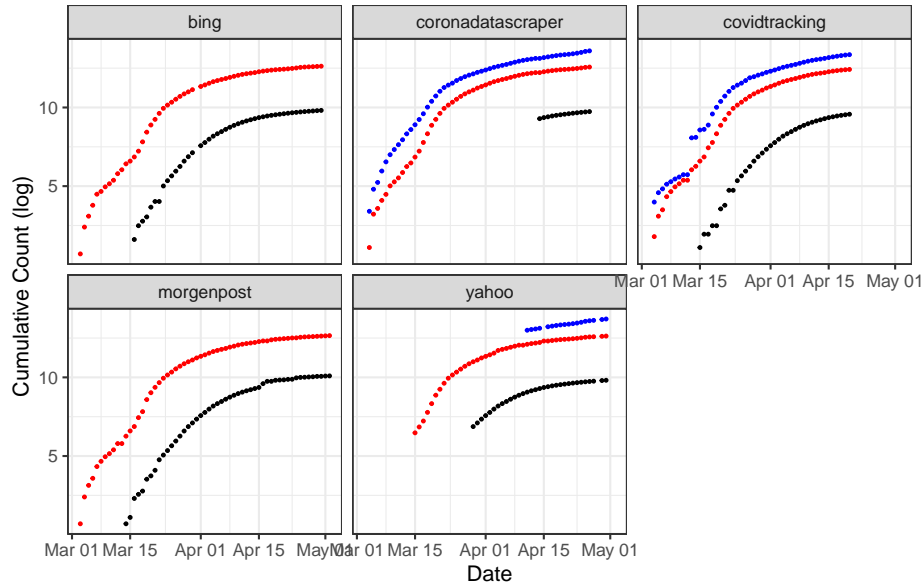


The U.S. has a great deal of coverage, but also a great deal of disagreement in that coverage. There is a stair step pattern in confirmed and deaths for Bing, WHO, and Wikipedia. In others reporting from day to day looks more continuous. There is also a change in reporting in late February that shows us a sharp vertical discontinuity across most datasets, though the size of the jump varies. There is also less temporal coverage of testing than is available from the coronavirus tracking project at the state level. Why those state level estimates aren't totaled and available at the national level is a question.

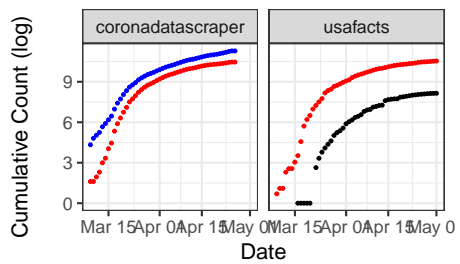


New York has been the most heavily hit by COVID-19 in the U.S. Two sources, CornaDataScraper and the Covid Tracking Project have coverage over nearly the entire period. However, only one shows a sharp discontinuity in testing around March 10th. Digging into that disagreement more, the CTP rates New York’s data release a B quality, coming from snapshots of press conferences and then switching to screenshots of New York’s “Department of Health Covid-19 Tracker” website.

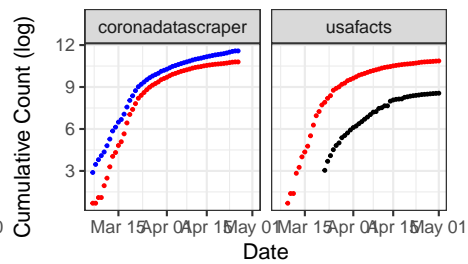
New York State Q1384 Counts



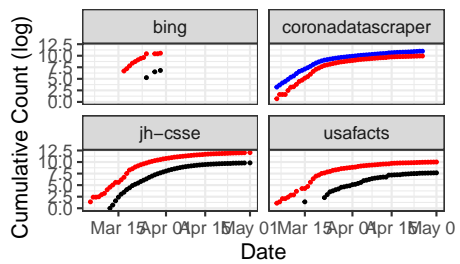
Bronx County (Q855974) Counts



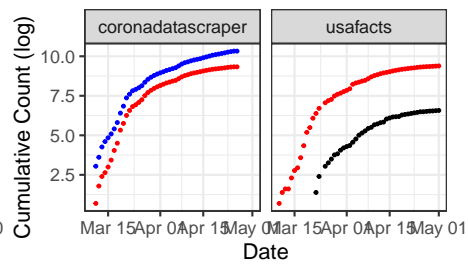
Queens County (Q18419) Counts



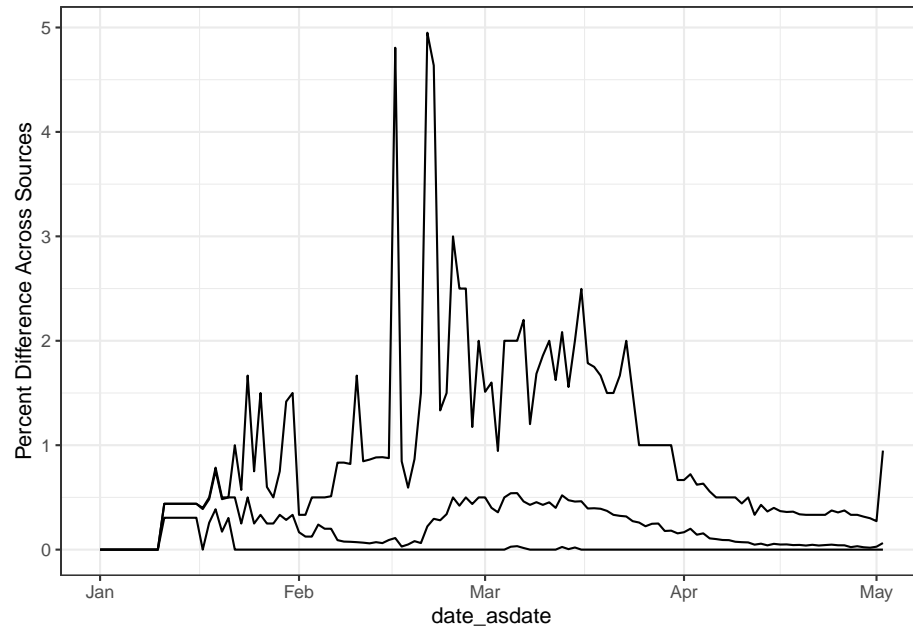
Manhattan (Q500416) Counts



Staten Island (Q11997784) Counts



2.5 Where and How do they Disagree?



Chapter 3

Tests

3.1 Tested People versus Tested Samples

3.2 Interpolate Within Observed

3.3 Interplate Prior to Observed

3.4 Interpolate Subnationally

3.5 Explaining Variation in Testing

South Korea

Vietnam <https://www.reuters.com/article/us-health-coronavirus-vietnam-fight-insi-idUSKBN22B34H>

Chapter 4

Common Measures of Interest

4.1 R_0 and R

4.2 Case Fatality Rate (CFR)

4.3 Percent Positive

Chapter 5

Deaths

Chapter 6

Actual Infections

Chapter 7

Conclusion

Bibliography

Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.

Xie, Y. (2020). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.18.