Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Method of Estimating Time-to-Recovery for a Disease Caused by a Contagious Pathogen like SARS-CoV-2 using a Time Series of Aggregated Case Reports

Stavros Pitoglou (spitoglou@biomed.ntua.gr)
National Technical University of Athens https://orcid.org/0000-0002-5309-4683
Dimitrios-Dionysios Koutsouris
National Technical University of Athens

Research article

Keywords: sars-cov-2, time-to-recovery, covid-19

Posted Date: March 28th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-19556/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Abstract

Background: During the outbreak of a disease caused by a pathogen with unknown characteristics, the uncertainty of its progression parameters can be reduced by devising methods that, based on rational assumptions, exploit available information in order to provide actionable insights.

Methods: In this study, performed few (~6) weeks into the outbreak of COVID-19 (caused by SARS-CoV-2), data publicly available on the Internet including daily reported cases of confirmed infections, deaths and recoveries are fed into an algorithm that matches confirmed cases with deaths and recoveries, in order to calculate average time-intervals. Unmatched cases are adjusted based on the matched cases calculation.

Results: The mean time-to-recovery calculated from all globally reported cases was found 18.01 days (SD 3.31 days) for the matched cases and 18.29 days (SD 2.73 days), taking under consideration the adjusted unmatched cases as well.

Conclusion:The experimental results indicate that the proposed method, in combination with expert knowledge and informed calculated assumptions, could provide a meaningful calculated average time-to-recovery figure, which can be used as an evidence-based estimation to support the containment and mitigation policy decisions.

Trial registration: Not applicable.

1. Background

There are many “known unknowns” in the progression patterns of an infectious disease caused by a pathogen that has not been widely spread before, and its characteristics have not been extensively studied. One of them is the time-to-recovery (or recovery time), t_{rec}, the time that passes from the onset of the disease before an infected individual moves to the recovered class having fought off the infection.

When a disease has the characteristics of a pandemic, an accurate estimation of this time interval is essential. From a disease modelling perspective, it translates to the probability of an individual moving from the ***I***nfected to ***R***ecovered class, being dependent on how long they have been infected, and is an essential parameter to any mathematical disease model that contains I→R dynamics (S→I→R,S→E→I→R, etc.)(1). Furthermore, it provides a measure of how long any infected individual is likely to transmit the disease, so it could help responding authorities to make informed and evidence-based decisions on containment measures/policies, such as quarantine periods, population mobility restrictions, etc.

One of the main restrictions is that, during the outbreak, and due to the extensively multi-centred data flow (from hundreds of distressed healthcare facilities), data about the confirmation, deaths and recoveries are reported in an aggregated fashion (daily totals of cases per category), rendering effectively impossible the data analysis on a per case basis.

In this paper, an algorithmic approach/method is proposed that overcomes the above restriction, exploring the hypothesis that a reliable and actionable estimate of the time-to-recovery for an infectious disease could be calculated from available reported data, even during the early stages of the outbreak.

In order to come up with a meaningful approach given such limited data and severe information loss due to their aggregated form, a series of assumptions must be made:

- The disease progression follows the simple model, depicted in Fig. 1
- It is taken as granted that each case reported is subject to regular follow-up, ensuring that the time of recovery or death will be recorded and appended to the available aggregated cumulative datasets without exception
- The mean time to recovery is calculated irrespective specific patient strata (e.g. age, gender, etc.). This is consistent with the simplification commonly used by disease modellers that the recovery rate (which is the inverse of the infectious period) is constant (1).
- The mean time to recovery for patients that recover is equal to the mean time to death in the mortal case occasions ($t_{rec} = t_d$), as it is often considered plausible to assume that mortality occurs towards the end of the infectious period (1).
- Confirmed cases that are not yet matched (considered still ill) at the time of analysis will have time-to-recovery equal to the mean calculated from the matched cases.

2. Methods

Data Source

In response to the public health emergency caused by SARS-CoV-2, Johns Hopkins University developed an interactive web-based dashboard hosted by their Center for Systems Science and Engineering (CSSE), in order to visualize and track reported cases in real-time (2). The data sources include the World Health Organization (WHO) (3), the Centers of Disease Control and Prevention (CDC) (4), the European Centre for Disease Prevention and Control (ECDC) (5) and the China's National Health Commission (NHC) (6). All the data collected and displayed are made freely available GitHub repository (7). The raw data sources, in the form of "comma-delimited/separated files"(CSV), that were used for the subsequent experiments, are presented in Table 1.

Table 1
Source Datafiles

Case Category	File [a]
Confirmed	time_series_19-covid-Confirmed.csv
Deaths	time_series_19-covid-Deaths.csv
Reported	time_series_19-covid-Recovered.csv

[a] URL path: https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series/{}

Data Processing

The data preparation part of the algorithm processes the daily cumulative data from the source mentioned above, in order to create a time series with the daily newly reported cases per category. Next, using these time series, it creates three (3) sets that contain one data point for each reported case (confirmed, deaths, and recovered, respectively), using as value the index of the date that it was reported. Specifically, in our experimental case, the available data start on 22nd Jan 2020; thus, all cases reported on this date are appended to the set with the value '1', cases reported on 23rd Jan 2020 with the value '2', etc.

Each confirmed case is matched with the earliest unmatched reported case of death or recovery, whichever comes earlier, with the death cases taking precedence when there are both unmatched death and recovery cases in the same date. The difference between the dates of each pair is calculated, and the cases that take part in this calculation are removed from the respective sets. This procedure is iterated until there are no death and recovery cases left. At this point, the mean of the intervals of the matched case pairs is calculated.

For the rest of the confirmed cases, the difference from the last available date is calculated and, if it exceeds the previously computed average, it is appended to the final set unchanged. If not, it is replaced by the average.

The final set includes estimated time intervals between time of confirmation and time of recovery for all confirmed cases to date. From this complete set, the overall mean time-to-recovery is calculated. As new data reports come in daily, this figure can be updated to provide enhanced estimation accuracy based on a growing body of evidence.

Experiments

For the purpose of checking the main hypothesis experimentally, a computer program was written in Python programming language, able to apply the algorithm described above to the available data (from

22nd Jan to 9th Mar 2020). Moreover, there was the ability to choose any particular country, cluster of countries or global figures to run the calculations on.

The reported results include (i) Global figures (ii) Mainland China as the “cradle” of the infection and the largest available dataset (iii) Italy and Iran as two countries with widespread of the disease (iv) US, UK, Spain, France, Germany, and the Netherlands as countries with highly probable domestic sustained spread of the disease and relevant high quality of healthcare service infrastructure.

3. Results

The results of the experimental runs for the selected countries/clusters/regions are presented in Table 2 (relevant graphs in Fig. 2).

The mean time-to-recovery calculated from all globally reported cases is 18.01 days (SD 3.31 days) for the matched cases. Taking under consideration the adjusted unmatched cases as well, based on the assumptions as mentioned above, the mean recovery time is adjusted to 18.29 days (SD 2.73 days)

As expected, Mainland China that accounts for the majority of the reported cases at this time (80735 out of 113583, 71.08%) is the dominant subset, affecting substantially the global figures.

Table 2
Results

	Matched			**All** [a]		
	#	mean	SD	#	mean	SD
China	61855	17.81	3.31	80735	18.52	3.39
Netherlands	3	8	0	321	8.01	0.10
France	31	11.84	5.51	1209	11.84	0.87
US	30	18.77	5.86	605	18.77	1.28
Spain	60	8.18	1.17	1073	8.18	0.27
UK	22	9.59	3.79	321	9.59	0.97
Iran	2631	4.37	0.82	7161	4.40	0.51
Germany	20	15.60	3.30	1176	15.60	0.42
Italy	1188	7.73	1.16	9172	7.74	0.42
Global	66508	18.01	3.31	113583	18.29	2.73

[a] matched plus adjusted unmatched

4. Discussion

The accuracy of the figures provided by the proposed methodology is directly dependent on the plausibility of the underlying assumptions, as reported in the Introduction section. Besides that, the specific conditions of the experimental leg of this study, the specific disease (COVID-19), its stage and the available data, introduce several limitations regarding the approach, the methodology and the experiments conducted in this study. As far as the data quality is concerned, it should be noted that they are provided to the public strictly for educational and academic research purposes, as they rely upon publicly available data from multiple sources that do not always agree. Regarding the purpose of this article, the accuracy of the reporting is presumed adequate. However, even if the raw reports are aggregated, confirmed and appended to the source data files in an accurate and timely manner, there is no guarantee for the quality or the homogeneity of the methodology used in the originating health facilities that compile and report the case numbers to the surveillance authorities.

This fact, combined with the early stage of the experiments leading to small peripheral datasets, provides a plausible explanation for the variability that is observed between individual countries. Factors like delay in first reports due to low initial awareness could explain lower means in some countries in this early stage. However, the part of the cumulative deviation caused by temporary reasons or measurement noise is expected to get smaller with time and figures will tend to converge. This tendency can be disrupted by systematic affecting factors, like different definitions of recovery (due to medical approach or –sometimes- political reasons)

Finally, COVID-19 is the first epidemic for which official data of recovered cases are collected and publicly available, which makes the proposed calculation approach possible, but also means that there is no practical way to evaluate it by applying it to historical data of analogous epidemics, such as SARS and MERS. Thus, the only way of evaluation lies in future work, involving retrospective case studies with non-aggregated datasets.

5. Conclusions

In this paper, plausible assumptions led to a method that utilizes raw aggregated data, available via surveillance reporting routes even at the early stages of a disease outbreak, in order to calculate, with the minimum possible uncertainty, the average time-to-recovery of an infected individual. The method was experimentally tested, resulting in an estimation of this vital time interval relative to COVID-19, with limited (~ 6 weeks) data. The preliminary experimental calculations support the hypothesis that the proposed methodology could provide meaningful and actionable estimations to facilitate evidence-based disease spread forecasting and containment measure design.

List Of Abbreviations

COVID-19	Coronavirus Disease-2019
SARS-CoV-2	Severe Acute Respiratory Syndrom- CoronaVirus-2

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The data used in this research are publicly available by the Center for Systems Science and Engineering (CSSE) of John Hopkins University (Github repository: https://github.com/CSSEGISandData/COVID-19).

Competing interests

The authors declare that they have no competing interests.

Funding

This research has not received any funding.

Authors' contributions

SP: conceived and designed the study, analysed the data and drafted the manuscript. D-DK: oversaw the methodology and revised the draft manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

References

1. Keeling MJ, Rohani P. Modeling infectious diseases in humans and animals. Modeling Infectious Diseases in Humans and Animals. 2011.
2. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis [Internet]. 2020 [cited 2020 Mar 9];3099(20):19–20. Available from: http://dx.doi.org/10.1016/S1473-3099(20)30120-1

3. WHO. Coronavirus disease 2019 (COVID-19) situation reports [Internet]. [cited 2020 Mar 9]. Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports
4. Coronavirus Disease 2019 (COVID-19) | CDC [Internet]. [cited 2020 Mar 9]. Available from: https://www.cdc.gov/coronavirus/2019-ncov/index.html
5. ECDC. Situation update worldwide, as of 9 March 2020 08:00 [Internet]. [cited 2020 Mar 9]. Available from: https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases
6. NHC. □□□□ [Internet]. [cited 2020 Mar 9]. Available from: http://www.nhc.gov.cn/xcs/yqtb/list_gzbd.shtml
7. CSSEGISandData/COVID-19: Novel Coronavirus (COVID-19) Cases, provided by JHU CSSE [Internet]. [cited 2020 Mar 9]. Available from: https://github.com/CSSEGISandData/COVID-19

Figures

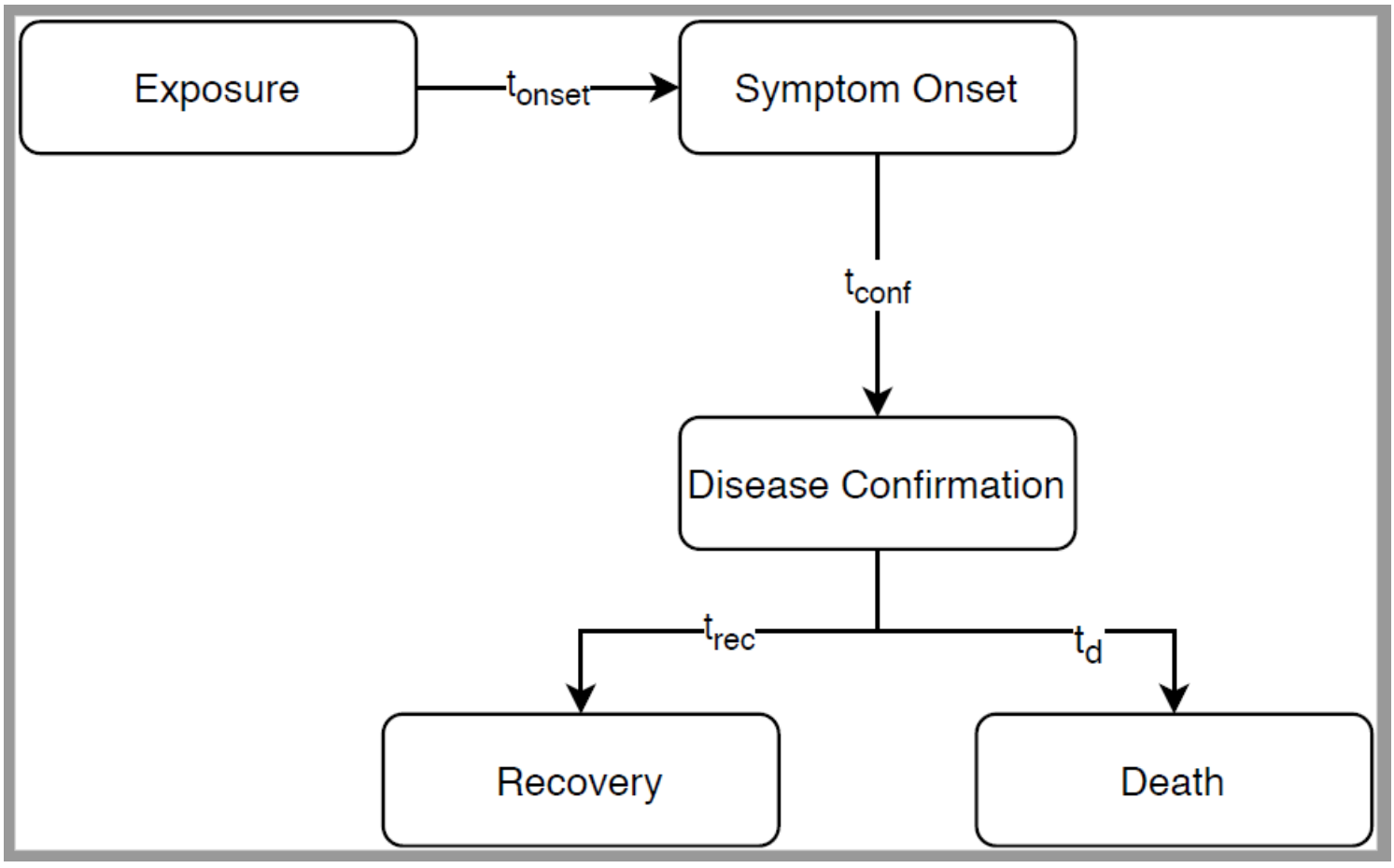

Figure 1

Model of Disease Progression

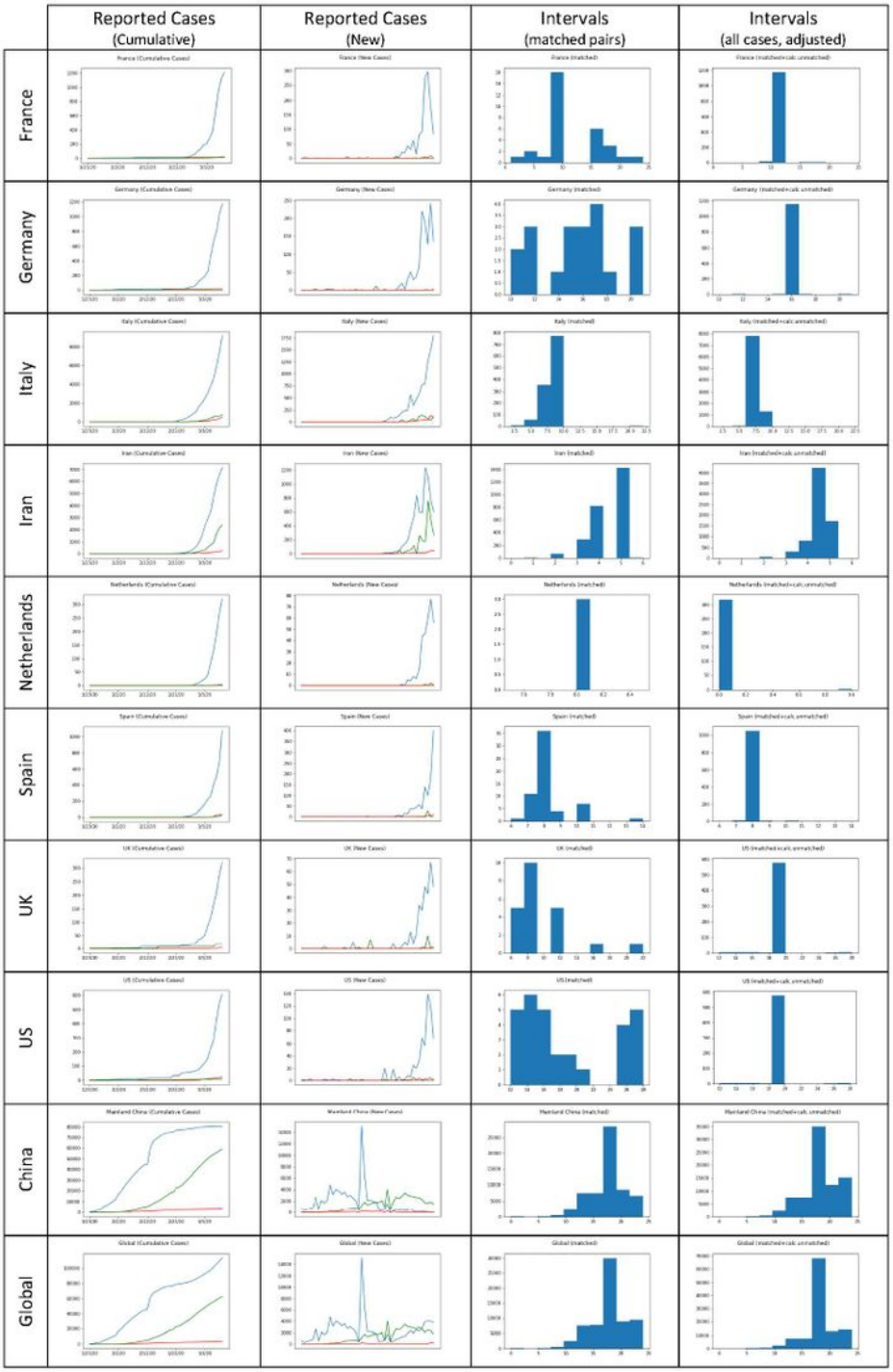

Figure 2

Reported cases diagrams and calculated interval distributions.