Covid-19 Probabilistic Surveillance of a Nation's Population: a Proposal of a Quick Project

Peter J.F. Lucas
University of Twente, Enschede, the Netherlands
LIACS, Leiden University, the Netherlands
Email: peter.lucas@utwente.nl; plucas@liacs.nl

23rd April, 2020

Abstract

Instead of the current practice of collecting information on Covid-19 epidemiology by counting Covid-19 hospital admissions and deaths, information about the geographical spread of Covid-19 can also be obtained by letting an AI-based smartphone app compute the likelihood whether or not a citizen has mild or severe Covid-19. This will be presented together with a clinical advice to the app's user in an easy to understand way. By sending the resulting posterior probability distribution of Covid-19 with the smartphone's GPS-location, and the person's age to a web-based database server over a secure connection, Covid-19 information at the level of a nation's population can be collected, offering a more precise picture of the actual Covid-19 geographical distribution and hotspots. The described approach imposes minimal infringement of privacy, as no personal data are collected centrally.

Keywords: Covid-19 surveillance, Artificial Intelligence, epidemiology, clinical decision-making, citizen empowerment.

Gaining insight into the geography of Covid-19

The Covid-19 pandemic has confronted national governments with several important challenges. The question of how to monitor outbreaks and spread of the disease in the nation's human population is certainly not the least significant one of these. Clearly, the most effective way to obtain precise insight into the impact of Covid-19 at the level of population health is simply to carry out **laboratory tests on the presence of SARS-CoV-2** (the virus that causes Covid-19) by taking nasal and throat swaps, either by systematic sampling of the entire population or by allowing anyone with beginning symptoms and signs of the disease to be tested in this way. Unfortunately, the availability of testing facilities have been lagging behind. In addition, the outcome of the test only tells that you have the corona virus, not whether you are or will become seriously ill from it.

At the moment, many governments are monitoring the geographical distribution of Covid-19 in their nation based on **counting hospital admissions** of tested severe Covid-19 patients in conjunction with keeping track of the number of seriously ill Covid-19 patients treated in the intensive care units and the number of deaths due to Covid-19[1]. Global information of the Covid-19 pandemic is currently maintained, regularly updated, and disseminated through the internet by the John Hopkins Covid-19 Resource Center [2].

An alternative to laboratory testing and counting SARS-CoV-2 tested people is to let **the citizen diagnose the presence of Covid-19 themselves by using a probabilistic model**

[1]Update for the Netherlands (17th September, 2020): since about 3 months, lab testing has been extended to people with symptoms of Covid-19. However, the lab capacity is still insufficient, indicating that the government has been overly optimistic of its organisational competence. This clearly indicates that the policy described in the present document, which already might have be implemented in April 2020, is much more realistic.

that is made available in an **app on the smartphone** that provides the person with information about how likely it is they have or have not mild or severe Covid-19. When this probabilistic information is combined with data about the GPS-location of the smartphone, together with rough information about the age of the person, the triple

$$(\text{Prob. dist. Covid-19}, \text{GPS-location}, \text{Age-group})$$

can be sent over an encrypted connection and collected by (government-controlled)[2] web-based databases and used to provide information about the distribution of mild and severe Covid-19 on the map of the country by visualising the probability distribution on the location, as shown in Figure 1, where that information was collected [5].

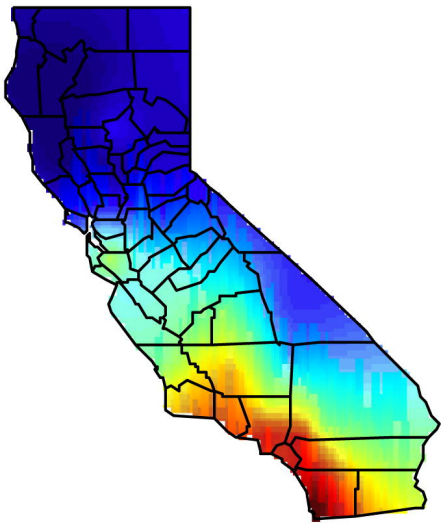

Figure 1: Visualisation of the geographical probability distribution of an infectious disease; similar mappings can be used to visualise the geographical probability distribution of mild and severe Covid-19 on the map of a country.

This proposal is about how to implement the strategy described above — diagnostic feedback at the level of the citizen with Covid-19 surveillance with minimal privacy infringement at the global level — as quickly as possible in the face of all the limitations of the current situation. It is very different from the in many countries (UK, Australia, the Netherlands) adopted **contact-tracing app**, which works by collecting information of nearby smartphones through bluetooth.

Design of a Covid-19 monitoring and reporting framework

At the time of writing there is not much reliable information available on the signs and symptoms confirmed (by lab tests), with the exception of several (non-peer-reviewed) papers in top-clinical journals (e.g. paper [4] and [9]). The advantage of the summary statistics provided in these papers is that they offer detailed information on the relative importance of signs and symptoms of Covid-19 in patients. The disadvantage is that without access to the original raw data, only the probabilistic information reported in the papers can be used. This imposes rather strict constraints on the modelling methods that can be employed. However, as will be demonstrated below, even with these limitation and with some additional assumptions it is feasible to develop a reasonably

[2] As an alternative, the strategy described here can be implemented by a BigTech company, such as Google, Facebook, and Amazon, which has the advantage of obtaining a global worldwide overview of spread of the disease.

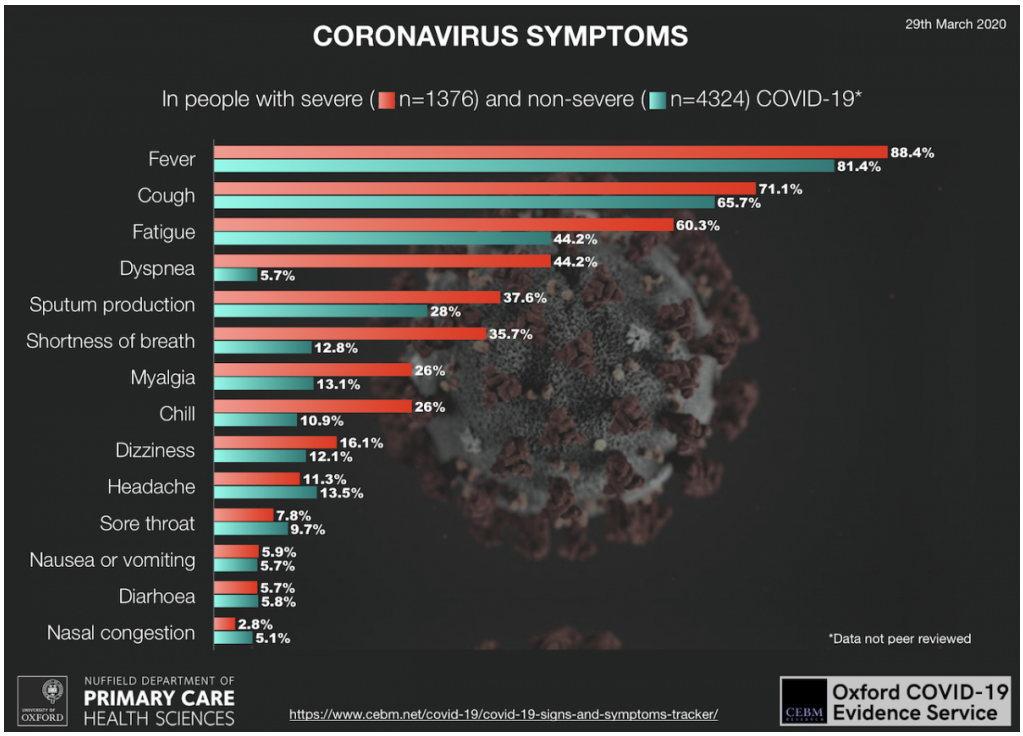

Figure 2: Overview of signs and symptoms of Covid-19, derived from [9].

reliable probabilistic model that can be used to predict the likelihood of the presence of mild or severe Covid-19 in a given patient.

Based on the paper by C. Huang et al. [9], which reviews the **signs and symptoms of confirmed Covid-19 cases** in Wuhan, the Oxford-based Covid-19 Evidence Centre [1] has produced a summary of the most important clinical features of the disease, which is reproduced in Figure 2.

The envisioned use of such a probabilistic model is as a foundation of population surveillance of the geographical outbreak and spread of Covid-19. The proposed infrastructure for personalised Covid-19 status feedback and collecting geographical data is shown in Figure 3, and is inspired by related research of the author's research group on mHealth [8, 10]. As the picture shows, it is assumed that a Bayesian network is embedded in a person's smartphone. The Bayesian-network algorithms compute the likelihood of having no, mild, or severe Covid-19, based on present **signs and symptoms** entered by the user, which is presented in an attractive and easy to understand way to the smartphone user with additional advice whether or not it is wise to contact a GP. In addition, a state is assumed to be added to the Covid-19 variable, indicating that the disease is present, but asymptomatic.

The advantage of a Bayesian network is that if certain evidence is not entered by the user, the model is able to use prior probabilistic information rather than make particular assumptions. Measurements consist of **body temperature** and **oxygen saturation**. However, it is up to the user whether these measurements are actually made. Using the Bayesian network it is in addition possible to predict which feature will be the most informative one in contributing to the diagnosis, and this feature can be used to request additional information from the app's user after some initial input (see also below).

Figure 3: Infrastructure for personalised Covid-19 feedback and collecting geographical Covid-19 data.

Construction of a simple Bayesian network

Based on the information summarised in Figure 2 that comes from the paper in The Lancet [9], a Bayesian network (BN) [3, 7, 11] was designed with the following assumptions:

- Only the most important signs and symptoms S where included in the current version of the BN, with symptoms occurring in at least 25% of patients with either mild or severe Covid-19. In addition, the selected symptom S was expected to have a likelihood ratio $\lambda(S \mid \text{Covid-19}) = P(S = \text{yes} \mid \text{Covid-19} = \text{severe})/P(S = \text{yes} \mid \text{Covid-19} = \text{mild}) > 1.08$ (The likelihood ratio of the selected symptoms and signs vary from 1.085 for fever to 7.75 for dyspnea).

- Some of the signs and symptoms, such as nasal congestion, have a likelihood ratio lower than 1 ($\lambda(\text{nasal congestion} \mid \text{Covid-19}) = 0.549$), and could be added to the model to reverse to distribution of probability mass between severe and mild Covid-19. However, these symptoms are relatively rare.

- It is assumed that all signs and symptoms are conditionally independent given the presence or absence of Covid-19, with the exception of the variable Age and body temperature (BodyTemp). Note, however, that all signs and symptoms are dependent of each other through the Covid-19 variable.

- The conditional probability distributions $P(S \mid \text{Covid-19})$, with $\text{Covid-19} \in \{\text{mild}, \text{severe}\}$, and S each of the features included in the network, were obtained from [9].

- The prevalence of Covid-19 is assumed to be 0.5%, meaning that absence of Covid-19 represents those people who are healthy or with any other respiratory disease, such as the common cold, flu, etcetera.

- The prior probability distribution of Age was roughly based on the CBS numbers for the Netherlands. (Not yet that precise, but it is not hard to refine the numbers.)

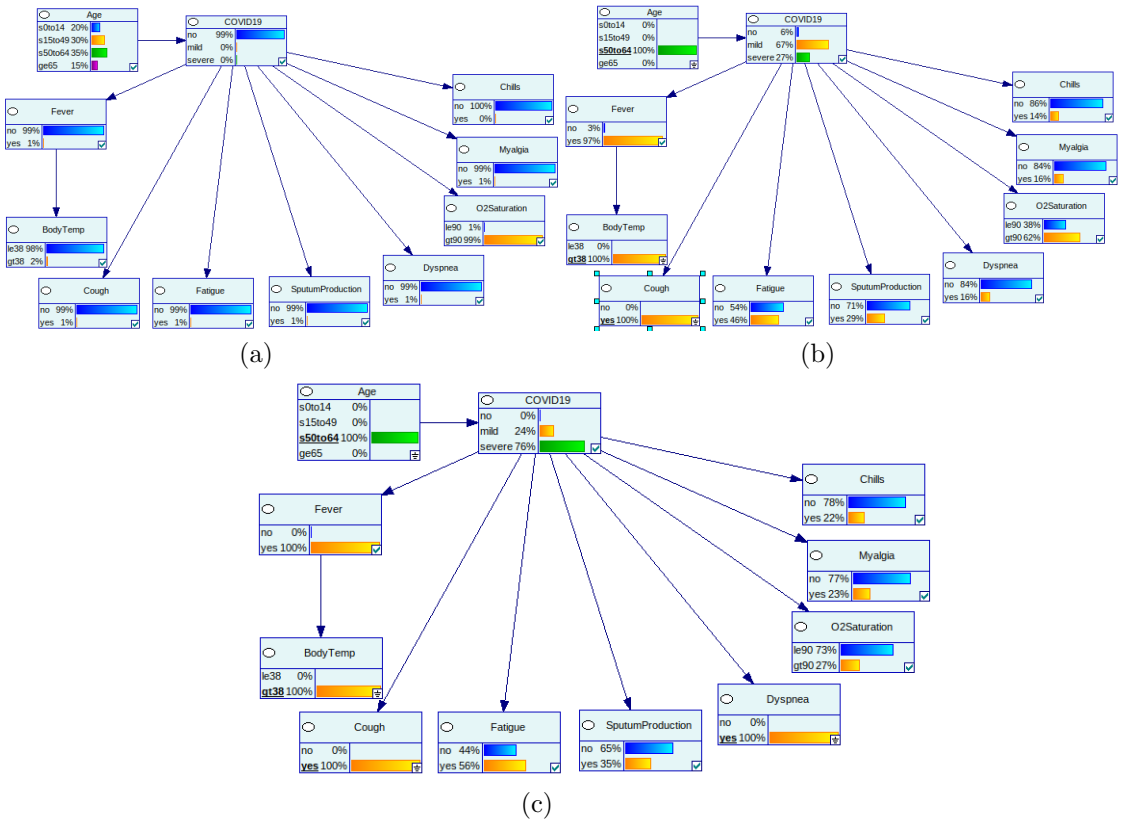

Figure 4: Covid-19 Bayesian network: (a) prior marginal distribution of all the variables; (b) posterior distribution of Covid-19 with evidence Age between 50 and 65, body temperature above 38, and coughing; (c) posterior distribution of Covid-19 with same evidence as (b) and in addition dyspnea.

Figure 4(a) depicts the computed marginal probability distribution of the individual variables, based on the specification of the conditional probability distributions, reviewed above. These probabilities give an overview of the entire population; by for example entering a value of Age, one can zoom in on the conditional probabilities of any of the variables given the value for Age (not shown here). Note that the software that has been used to visualise the network structure and the probabilities rounds off the decimal fractions of the probabilities to the closest number (0 or 1).

Figure 4(b) demonstrates a person's likely Covid-19 diagnosis when entering age (between 50 and 65), body temperature higher that 38 centigrade, and the symptom of coughing. Note that the posterior probability of mild Covid-19 changes from the prior of almost 0% to 67%, and severe Covid-19 from almost 0% to 27%. Finally, Figure 4(c) shows that when entering the additional symptom of dyspnea, the probability of severe Covid-19 increases from 27% to 76%, as one would expect clinically.

This model is still an incomplete attempt of developing a Bayesian network for the prediction of the presence of Covid-19, but the assumptions mentioned above are not unrealistic. It is possible to add other signs and symptoms (for example dizziness seems useful) and also comorbidities and immunodeficiency could be added, as the literature provides the relevant information. A typical example of a comorbidity that will have great impact on the severity of Covid-19 is COPD. One problem is that the more features you add the more difficult it becomes to maintain the conditional

independence assumptions[3]. However, for COPD there is much known, hence it is likely that the model can be further improved, for example to express the probabilistic relationship between COPD, sputum, cough and fever. Also more attention to differentiating between Covid-19 and flu (now part of the 'no' value of the Covid-19 variable) would be a valuable refinement. All of these improvements would be interesting exercises.

Challenges of disease surveillance

In this report it is assumed that a **citizen of a country obtains feedback about the likelihood of presence of mild or severe Covid-19** from a smartphone app, but the main purpose of making an app with the Bayesian network embedded in an app is **to monitor the population for detecting new outbreaks as early as possible**. For this it is only needed that the triple

$$(\text{Prob. dist. Covid-19}, \text{GPS-location}, \text{Age-group})$$

is collected centrally. The age information might be useful to get information about required protection of particular groups. In addition it might be useful to in addition add a unique identifier so that it is possible to follow the progress of Covid-19 in the individual (possibly until hospital admission). However, collecting only the above-mentioned triple of information has the advantage of **minimal infringement of privacy**.

Questions and answers

Compliance with the EU General Data Protection Regulation The design of the app described above conforms to the "data protection by design" principle of the EU General Data Protection Regulation (GDPR), in that the minimum amount of data is collected for the purpose of Covid-19 surveillance. Clearly the posterior probability of Covid-19 is needed to determine whether there is an outbreak in a particular geographical region. GPS-location and age may under certain conditions offer enough information to uniquely identify a person. However, there is no need to collect the GPS-location as precisely as GPS allows; for example, information about the location of the subject may be sufficient in a circle around the subject with a radius of 100 m. Furthermore, age is collected in 4-5 different age groups, hence making it impossible to identify a subject by age only. As a consequence, privacy is ensured, and the triple of data collected is supposted to be designed in such a way that an individual cannot be identified by the triple only, to ensure compliance to the GDPR.

Are the collected data reliable? It is likely that the collected data will not be entirely reliable, because some people will act as fake patients. However, it is likely that this will happen at random, and you will see a similar percentage of fake registration across the entire country. This will allow to subtract the fake percentage from the total data yielding a reliable overview of infected people in the country.

Is collecting all the entered evidence instead of the triple useful? Instead of just centrally collecting the information triple mentioned above, it is in principle also possible to collect all entered information, e.g. to carry out epidemiological research. However, a clear disadvantage is that the information concerns persons of whom the presence of Covid-19 has not been confirmed by laboratory tests.

[3] A more extensive Bayesian network model is described in doi:10.1101/2020.07.15.20154286 and can be downloaded, [6]

Will people use the app? As the app is not yet available, it is hard to forecast its usage. However, the advantage of the app is that people get feedback about the likelihood of future development of mild or severe Covid-19, which is actually the information people want to hear from their GP. In this case, the probabilistic conclusions that are presented, are based on clinical scientific research. Of course, the feedback must be presented in an easy to understand way.

How easy is it to implement the platform? As similar software already exists, for example at the company MonitAir (`www.monitair.com`) or Google (`www.google.com`), and also the author has research software that is suitable, with the addition of web-based database servers, the implementation of the platform can be done quite rapidly. Needed is a revision of the software, careful scrutiny of the developed Bayesian network model based on a review of the clinical literature on Covid-19, and setting up an evaluation with a group of interested users.

Is the provided Bayesian network not too simple? The design of the current Bayesian network was purely based on the information provided in the clinical literature. When raw reliable (i.e. with SARS-CoV-2 lab-tested) patient data would have been available, a model that would better fit the data could have been designed by combining structure learning with parameter learning. Without access to such data, the model offers a starting point that may be good enough for its purpose, but as soon as clinical data of confirmed Covid-19 patients are available, a more reliable model can be constructed. A more elaborate model is discussed in [6].

Is there still a role for the medical doctor? The purpose of the app is *not* to replace a medical doctor, but simply to give citizens easy access to clinically useful knowledge of Covid-19. When it is likely that the person has contracted severe Covid-19, the app may in fact advise to contact the GP or go to the hospital (depending on the healthcare organisation of the country).

References

[1] `https://www.cebm.net/covid-19`.

[2] `https://coronavirus.jhu.edu/map.html`.

[3] R.G. Cowell, A.P. Dawid, S.L. Lauritzen, D.J. Spiegelhalter (1999). *Probabilistic Networks and Expert Systems*. New York: Springer.

[4] W. Guan, Z. Ni, Yu Hu, et al. (2020). Clinical Characteristics of Coronavirus Disease 2019 in China, *The New England Journal of Medicine*, doi:10.1056/NEJMoa2002032.

[5] S.I. Hay, K.E. Battle, D.M. Pigott, et al. (2013). Global mapping of infectious disease. *Phil Trans R Soc B* 368: 20120250, doi:10.1098/rstb.2012.0250.

[6] N. Fenton, S. McLachlan, P.J.F. Lucas, K. Dube, G. Hitman, M. Osman, E. Kyrimi, M. Neil (2020). A privacy-preserving Bayesian network model for personalised COVID19 risk assessment and contact tracing. *medRχiv*, doi:10.1101/2020.07.15.20154286.

[7] D. Koller and N. Friedman (2009). *Probabilistic graphical models : principles and techniques*. Cambridge, Massachusetts: The MIT Press.

[8] M. van der Heijden, P.J.F. Lucas, B. Lijnse, Y.F. Heijdra, T.R.J. Schermer (2013). An autonomous mobile system for the management of COPD. *Journal of Biomedical Informatics* 46: 458–469, doi:10.1016/j.jbi.2013.03.003.

[9] C. Huang, Y. Wang, X. Li, et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*, 395: 497-506, doi:10.1016/S0140-6736(20)30183-5.

[10] M. Velikova, J. Terwisscha van Scheltinga, P.J.F. Lucas, M. Spaanderman (2014). Exploiting causal functional relationships in Bayesian network modelling for personalised healthcare. *International Journal of Approximate Reasoning*, 55(1): 59–73, doi:10.1016/j.ijar.2013.03.016.

[11] J. Pearl (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. San Mateo, CA: Morgan Kaufmann.