MILLIMAN WHITE PAPER

Automatic extraction of COVID-19 epidemiological parameters using Natural Language Processing

Remi Bellina
Cedric Mea
Floriane Moy
Alexandre Boumezoued
Amal Elfassihi
Eve Titon

Milliman

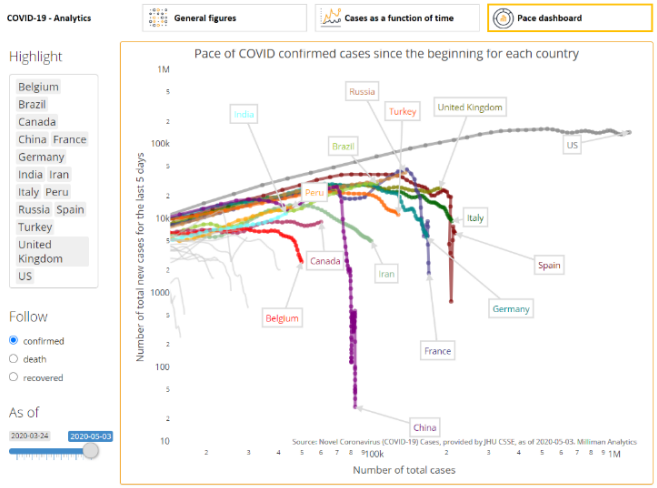
There is still much uncertainty about the current COVID-19 outbreak. Modellers are trying to anticipate the future of this pandemic based on relevant parameters driving its evolution. To this aim, the current scientific literature is a core source of information. However, in a context of exponentially increasing numbers of publications, an exhaustive manual analysis remains out of reach.

While not delving deeply into the full complexities of epidemiological modelling, we illustrate in this paper the potential and the challenges of using Bidirectional Encoder Representations from Transformers (BERT), a Natural Language Processing (NLP) framework, to automate the task of gathering input information and assisting experts for COVID-19 studies. There is a large wealth of public data available to explore plausible future pandemic scenarios. One of the most commonly used data sources is the Johns Hopkins University.[1]

Not only is more detailed information added to this data, but also contributions and modelling from the international scientific sources.[2] Data visualisations, web applications and scientific research papers have flourished on the internet with the aim of understanding, modelling and predicting the virus propagation patterns and their consequences. As an illustration, Figure 1 shows the evolution of the pandemic (by number of cases) in different countries.

While it is particularly interesting to follow the evolution of individual epidemics in this way and to make simple observations, it is nevertheless essential to maintain a certain humility with regard to these representations. There are many pitfalls inherent in the context: discrepancies in different countries' political decisions and medical capabilities, inconsistencies of data, inadequacies of certain modelling assumptions etc.

FIGURE 1: EXAMPLE OF A MILLIMAN COVID-19 TRACKING DASHBOARD



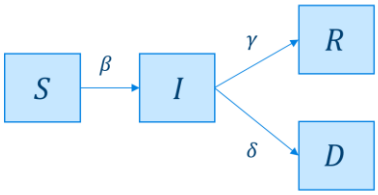
Resorting to dedicated epidemiological models appears to be the right approach to properly understand the dynamics of the pandemic. Multiple models have been implemented and shared recently to provide forecasts of infected, recovered and deaths related to COVID-19 for a variety of countries. An example of a compartmental model of the so-called Susceptibles, Infected, Recovered and Deaths (SIRD) type is illustrated in Figure 2. These compartmental models allow simulations of the evolution of the pandemic over time.[3] A SIRD model, as shown in Figure 2, is of course simplistic, and other factors are vital to producing realistic results (e.g., regional heterogeneity such as hospital capacity limits, modelling of subsequent waves, changes in mortality as the healthcare system becomes strained etc.). This model will thus serve as an illustration.

[1] The 2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository by the Johns Hopkins Center for Systems Science and Engineering (CSSE) is available at https://github.com/CSSEGISandData/COVID-19.

[2] As an example, Kaggle's platform is currently hosting a variety of challenges focused on better understanding COVID-19. See https://www.kaggle.com/covid19.

[3] Such models go back to Kermack and McKendrick (1927). See references in N. Bacaër, A Short History of Mathematical Population Dynamics. Springer-Verlag London Limited 2011.

FIGURE 2: ILLUSTRATION OF A BASIC SIRD MODEL



In order to provide reasonable forecasts and sensitivities, such SIRD models have to be fed by appropriate parameters, which can be summarized as follows in the simplest setting possible:

- A contamination rate β driving the pace at which susceptibles are contaminated by infected over time
- A recovery rate γ specifying the percentage of infected who recover per time unit
- A fatality rate δ driving deaths of infected over time

Those parameters are rarely provided in a homogeneous form in terms of estimate or units; as an example, the fatality rate can be provided in the form of a case fatality rate (CFR), as the number of deaths divided by the number of cases, or in the form of a "true" mortality rate per time unit, say by 10 person-days.

Also, they are not often separately provided, because a major focus is dedicated to the so-called basic reproduction number R_0, as the average number of people a given infected individual infects in a fully susceptible population. The estimation of this parameter R_0 can change over time with the evolution of the underlying parameters (contamination, recovery and fatality rates), especially due to the implementation of lockdown measures. In the modelling example above, the reproduction number can be written as follows:

$$R_0 = \frac{\beta}{\gamma + \delta}$$

Setting the parameter values required is usually based on the existing literature. At this point it appears that different studies will provide different estimates, depending on the methodology retained to capture those parameters, the conditions of the study (individuals sampled, testing process, experience metrics) and, of course, the region/population studied. As such, it appears that there is no "true" value to consider, rather a range of possible values provided to the research community by different teams facing different contexts.

Also, in the first early phase of the pandemic, new reports are quickly published and the high volume is difficult to be efficiently digested by modelling teams to recover appropriate parameters. In this context, we propose to illustrate the benefit of NLP to automate the extraction of parameters of interest. The NLP approach can't be self-sufficient, in the sense that expert clinical and epidemiological knowledge is required not only to assess the quality of each publication but also to review and interpret the values in the studies. We believe that the NLP approach nevertheless provides an efficient way to synthesize large-scale bibliographies as explained and illustrated in this paper.

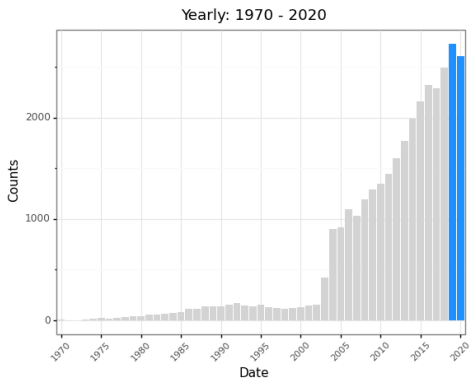
TAKING ADVANTAGE OF AN EXPLODING LITERATURE

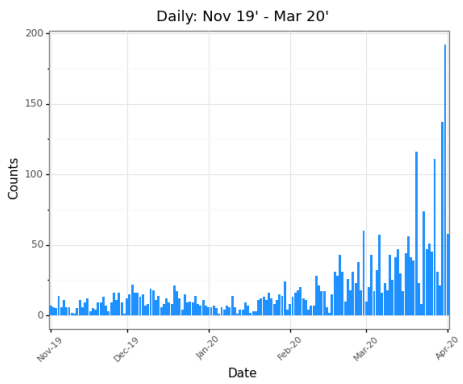
In this paper, we focus on the COVID-19 Open Research Data set (CORD-19).[4] CORD-19 is a free and open-source database containing more than 50,000 research articles (most of them available in a .json format with their full texts) about the SARS-CoV-2 and other similar viruses. The data set is distributed by the Allen Institute for AI and is updated on a weekly basis.

Figure 3 displays the evolution of the number of scientific papers available in CORD-19 as of 10 April 2020: since the onset of SARS-CoV in 2003 and after the 2009 swine flu pandemic, the study of pandemics related to coronaviruses has been constantly growing.

The number of such scientific papers increased at a linear pace between 2003 and 2019. As of the end of March 2020, the number of publications in 2020 almost reached the level for the entire year of 2019. The daily publications show a growth at an exponential rate between mid-January 2020 and today: it is hard to manually keep track of the many releases.

FIGURE 3: PAPERS DISTRIBUTION BY DATE OF PUBLICATION





[4] Available at https://pages.semanticscholar.org/coronavirus-research.

How can NLP and the BERT Q&A algorithm help?

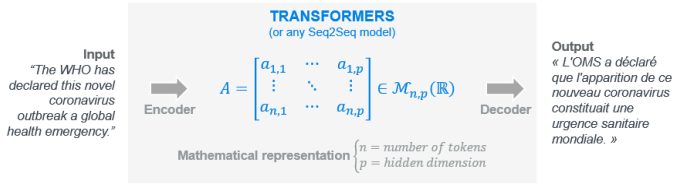
Deep learning has experienced strong growth over the last decade due to increasingly large data sets and progress in computing capacity. Even if underlying methods were already generalized in the field of computer vision, advances in Natural Language Processing (NLP) are more recent. For instance, embeddings (such as FastText or Glove) now allow words to be represented by higher-dimensional numerical vectors and Recurrent Neural Network models go further by adding information from the surrounding word sequences. The publication of the BERT algorithm greatly enhanced the toolbox of data scientists to process non-structured text data.

WHAT IS BERT?

BERT stands for Bidirectional Encoder Representations from Transformers[5] and is a natural language algorithm pretrained by Google AI. It provides mathematical representations (matrix forms) of input text sequences, taking into account grammatical and semantic relationships among different words. As indicated in its name, BERT is based on an algorithmic architecture called Transformers, introduced in 2017.[6] Transformers are sequence-to-sequence (Seq2Seq) algorithms that transform one sequence into another.

One of the most common application examples of this type of Seq2Seq model is the automated text translation. The implementation of Seq2Seq models requires the use of two other architectures: encoders and decoders. Encoders transform an input sequence into mathematical representations, which are then used by the decoder to produce the output sequence. Figure 4 shows an example of a process for an automatic text translation.

FIGURE 4: SEQ2SEQ MODELS OVERVIEW FOR MACHINE TRANSLATION

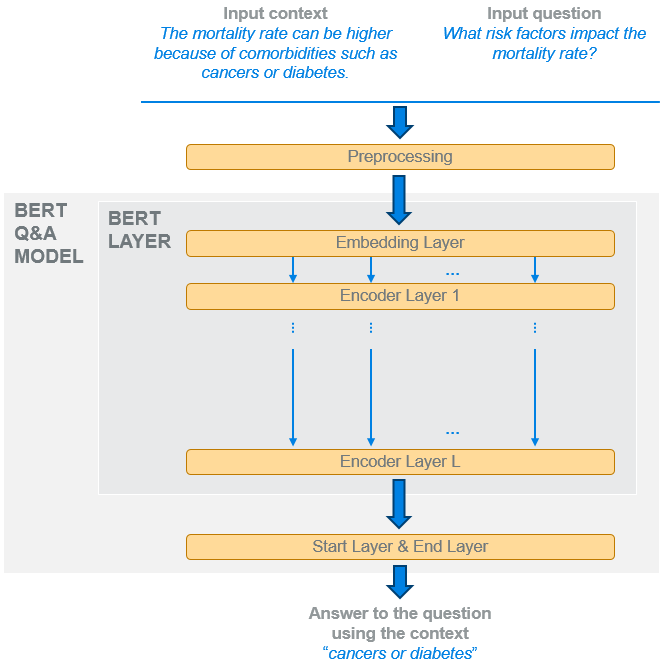


Milliman internal source.

The BERT algorithm relies only on the encoders of the Transformers architecture (decoders are not included). It uses several encoder layers that are stacked together: the output of encoder i becomes the input of encoder $i + 1$. The characteristics of the different encoders vary with the version of the BERT algorithm. The basic version, known as BERT-Base, has the following specifications: $n = 512$ (number of tokens in a sentence, a token being a basic word-like element), $p = 768$ (hidden embedding size of different tokens) and $L = 12$ (number of encoder layers). Encoders use a mechanism called "Attention," allowing the algorithm to better capture the dependencies that may exist between words in the input sequence.

Figure 5 illustrates the overall simplified BERT workflow, from the input sequence to the final output. It processes multiple operations through a huge ensemble of stacks based on neural networks. The output of the BERT layer is a mathematical representation (similar to the output of the encoder in Figure 4): this is one of the strengths of BERT because this matrix can be used to perform many NLP tasks. For instance, BERT is used today for many different applications including text classification or question-answering (Q&A) problem, which is the algorithm of interest here.

FIGURE 5: INFERENCE WORKFLOW FOR BERT Q&A MODEL



Milliman internal source.

BERT APPLICATION FOR Q&A

The question-answering algorithm requires a question and a context as inputs, as illustrated in Figure 5. It searches for the answer to the question inside the context and returns the computed answer as output.

As mentioned above, the BERT layer is a pretrained model. The different weights included in the model have already been optimised. When using the model for a new task, such as for the Q&A algorithm, these parameters need to be updated regarding the specific use case: this is called the fine-tuning process.

[5] Available at https://arxiv.org/pdf/1810.04805.pdf.

[6] In the publication "Attention Is All You Need," available at https://arxiv.org/pdf/1706.03762.pdf.

In order to implement BERT fine-tuning for question-answering, extra layers, called the start and end layers, are added to the process—both are needed to transform the output matrix and obtain the desired answer by score maximisation. The whole process requires a training data set containing many contexts with questions and their corresponding expected answers. The Stanford Question Answering Data set (SQuAD)[7] addresses these requirements. This database, with more than 100,000 questions, has been created specifically for the fine-tuning of the BERT Q&A algorithm.

All of these operations are, however, very time-consuming and require lots of computing resources (powerful GPU or TPU). In our analysis, we implemented a BERT Q&A model for a prediction (inference) purpose using the optimised parameters given by the Transformers library.[8]

Automatic extraction of COVID-19 epidemiological parameters

Processing the scientific literature in a specific domain such as COVID-19 requires a rigorously defined set of dictionaries of key words corresponding to the SIRD parameters of interest, in order to automatically select specific scientific papers and extract the relevant sentences. The mortality dictionary would, for instance, include "mortality rate," "death rate," "time to death" or "illness onset to death," whereas the key words like "infection rate," "contamination rate" or "spread rate" would be part of the infection dictionary.

Because each parameter depends on the context in which it is calculated (estimation method, geographical area, type of individuals etc.), getting some contextual sentences around the estimation of the parameters can be useful.

EXTRACTION OF THE MORTALITY RATE

Using the BERT Q&A algorithm described in the previous section, we could go further than extracting significant sentences around the SIRD parameters and instead focus on the extraction of the parameter values.

There are three ways of estimating the mortality rate and each of them is commonly used in the literature:

- CFR (number of deaths divided by number of cases)
- CFR without bias (recovering a "true" probability of death given the infected status, obtained by removing the individuals infected for a short time)
- A true mortality rate (driving the number of deaths per time unit)

It is particularly known for the mortality rate that such different estimates can lead to different results.[9] Although theoretically a true mortality rate (per time unit) is to be used in the model, this is often not available because a refined measurement of the exposure at risk is required (time from illness onset to death). Several references provide comparisons of such different methods to compute the mortality rate in the case of COVID-19.[10]

Below are some mortality rate values and examples of corresponding contexts we could extract from papers in an automated way using the Q&A BERT algorithm:

- **0.2%:** "According to the China CDC11, the case fatality rate (CFR) was 0.2% at the end of January 2020."[11]
- **2.3%:** "The overall case fatality rate (CFR) was 2.3% in the entire cohort but significantly higher (6%, 7.3% and 10.5% respectively) in patients with hypertension, diabetes and CVD."[12]
- **3.7%:** "A total of 185 [11.5%] patients were admitted to intensive care unit (ICU) while the overall case fatality rate (CFR) was 3.7%."[13]

However, not all articles among the large panel available in CORD-19 clearly detail and explicitly state their estimation approaches. Most of them don't provide any so-called "context" around their calculations.

Therefore, even if these values are informative, they can't be aggregated without more context on the estimation methodology. This example shows the potential of extracting the context along with the relevant parameter values.

EXTRACTION OF R_0

As opposed to the other SIRD parameters, the definition of the reproduction number seems to be a well-known concept and it is more universal, as it is a number without unit. In that sense, we believe it provides a good illustration for a BERT Q&A application. It varies over time and conditions (hygiene measures, social distancing etc.), however, showing again the importance of linking the value extraction with the context of the publications.

The example below, extracted using the BERT algorithm, comes from a paper14 in the Journal of Travel Medicine that studied the Diamond Princess cruise ship experience:

"Based on the modelled initial R_0 of 14.8, we estimated that without any interventions within the time period of 21 January to 19 February, 2920 out of the 3700 (79%) would have been infected. […] Isolation and quarantine therefore prevented 2307 cases, and lowered the R_0 to 1.78."

[7] Available at https://rajpurkar.github.io/SQuAD-explorer/.

[8] Available at https://huggingface.co/transformers/model_doc/bert.html.

[9] Böttcher, L., Xia, M., & Chou, T. Why estimating population-based case fatality rates during epidemics may be misleading. MedRxiv preprint.

[10] Jung, S. et al. Real-time estimation of the risk of death from novel coronavirus (COVID-19) infection: Inference using exported cases. Journal of Clinical Medicine.

[11] Caramelo, F. et al. Role of temperature and humidity in the modulation of the doubling time of COVID-19 cases. MedRxiv preprint.

[12] Bansal, M. Cardiovascular disease and COVID-19. Diabetes & Metabolic Syndrome: Clinical Research & Reviews.

[13] Fang, Z. et al. Clinical characteristics of coronavirus disease 2019 (COVID-19): An updated systematic review. MedRxiv preprint.

[14] Rocklöv, J., Sjödin, H. & Wilder-Smith, A. COVID-19 outbreak on the Diamond Princess cruise ship: Estimating the epidemic potential and effectiveness of public health countermeasures.

The context extracted here illustrates the impact of isolation and quarantine on virus reproduction in the Diamond Princess, which was one of the early outbreaks of COVID-19.

In addition to the political decisions, other interesting contexts can be the type of individuals or the geographical areas that are affected by the virus. The sentence below was automatically extracted from a MedRxiv preprint:[15]

"We estimated the initial basic reproduction number for South Korea, the Guangdong province and mainland China as 2.6 (95% confidence interval (CI): (2.5, 2.7)), 3.0 (95%CI: (2.6, 3.3)) and 3.8 (95%CI: (3.5, 4.2)), respectively […]."

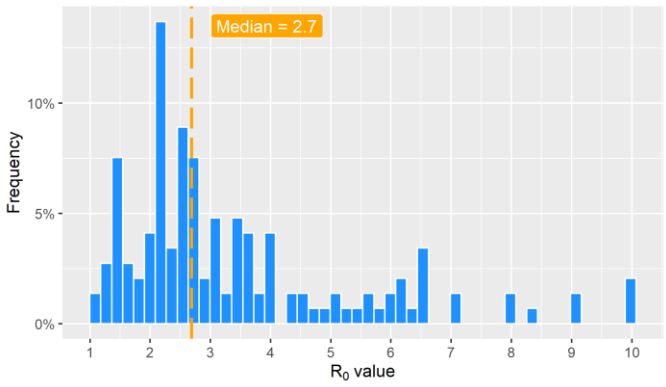
Such extracted sentences are important to understand R_0 values. They can help evaluate the impact of different risk factors on the virus propagation and its consequences.

EXTRACTED DISTRIBUTION OF R_0

We now focus on the extraction of the R_0 parameter values. We implemented the BERT Q&A algorithm on the entire panel of scientific papers. Before applying the algorithm to a specific context, we selected the relevant papers, which focus on the COVID-19 propagation and address the SIRD models topic.

Figure 6 presents the histogram corresponding to the extraction of R_0 values from the selected scientific papers' text bodies, after removing extreme values and those reported below the value of 1, which relate to final pandemic stages. Around 150 values have automatically been extracted and the median of that sample is equal to $R_0 = 2.7$.

FIGURE 6: DISTRIBUTION OF THE R_0 PARAMETER



The distribution of the R_0 values shows that it is roughly concentrated between 1 and 4, although a significant part is spread beyond. It is an important question how this uncertainty and different estimates would impact quantities of interest, such as the number of cases and deaths.

The analysis of this impact is the purpose of the last section. It is worth emphasising that the extracted values are aggregated across different populations and geographical areas, and are therefore not appropriate to use in modelling for a particular population or situation. In practice, the extracted context (see again the examples discussed above) should guide the modeller to sort or rank the publications depending on the relation they have to the particular modelling situation (country, population, time at which the study has been done etc.).

Impacts on SIRD modelling output

As mentioned above, SIRD-type models can be implemented to give insights on the dynamics of the COVID-19 pandemic, by dividing the population into four different groups (susceptibles, infected, recovered and deaths) and simulating the virus propagation over time.

SETTING THE SIRD MODEL

The dynamics for each group of individuals are modelled using the following differential equations:

$$\begin{aligned}
\frac{\mathrm{d}S}{\mathrm{d}t} &= -\frac{\beta}{N} I \cdot S \\
\frac{\mathrm{d}I}{\mathrm{d}t} &= \frac{\beta}{N} I \cdot S - \gamma I - \delta I \\
\frac{\mathrm{d}R}{\mathrm{d}t} &= \gamma I \\
\frac{\mathrm{d}D}{\mathrm{d}t} &= \delta I
\end{aligned}$$

These differential equations are discretised and simulations are run with specific initial values: a susceptible population with size $S(0) = N - 1 = 1\mathrm{M}$, a single infected individual $(I(0) = 1)$ and no initial recovery or death $(R(0) = D(0) = 0)$.

The fatality and recovery parameters are taken as follows:

- Fatality rate: $\delta = 1\%$
- Recovery rate: $\gamma = 5\%$

Assuming these constant values for these two parameters δ and γ is a strong structural choice for our case study and we are aware that they vary in practice. Indeed, measures like social distancing impact the fatality and recovery parameters, because the individuals at risk—older, with higher body mass index (BMI) or comorbidities—are in practice more likely to fulfil the strict distancing guidelines.

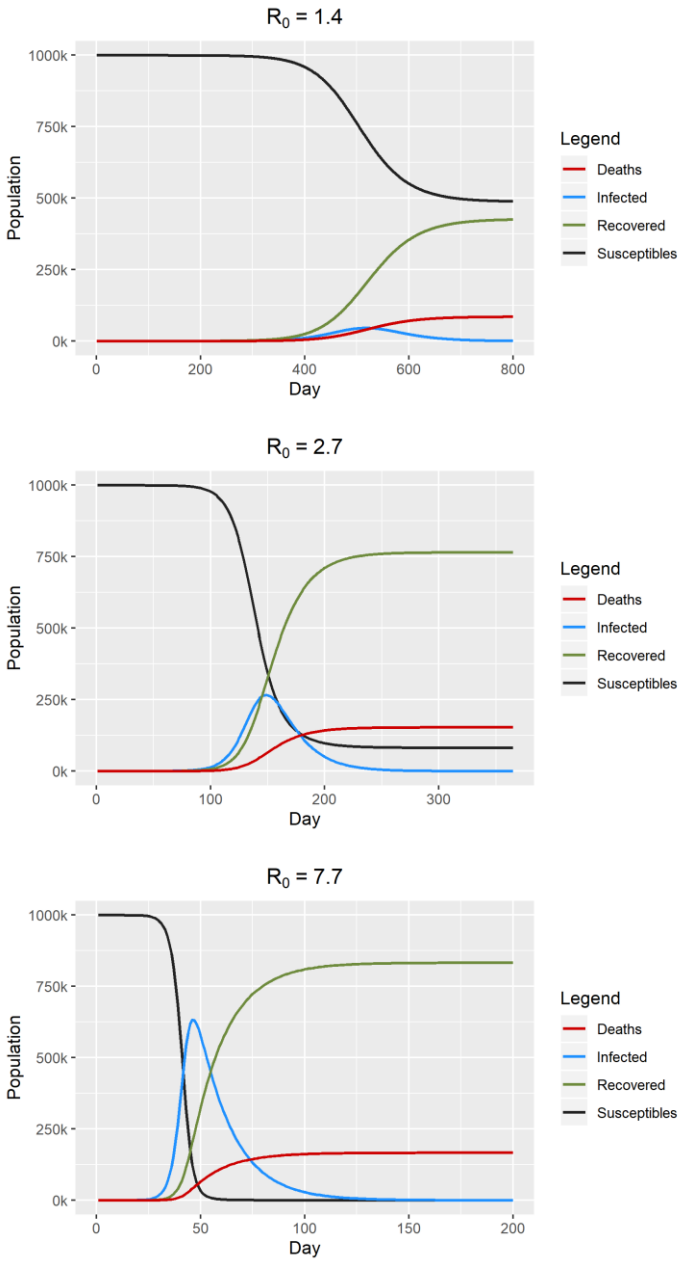
By definition of the basic reproduction number, the contamination parameter is given by $\beta = R_0 \cdot (\delta + \gamma)$. The observed variations in R_0 are related to the containment measures (attempting to manage the spread of the pandemic) and it seems natural to consider the contamination rate to be directly impacted and to vary.

[15] Tang, B. et al. Lessons drawn from China and South Korea for managing COVID-19 epidemic: Insights from a comparative modelling study

FROM THE R_0 DISTRIBUTION TO THE MODEL OUTPUTS

Figure 7 depicts epidemiological trajectories based on three values of R_0: 1.4, 2.7 and 7.7, which are, respectively, the 5th percentile, the median and the 95th percentile of the R_0 distribution.

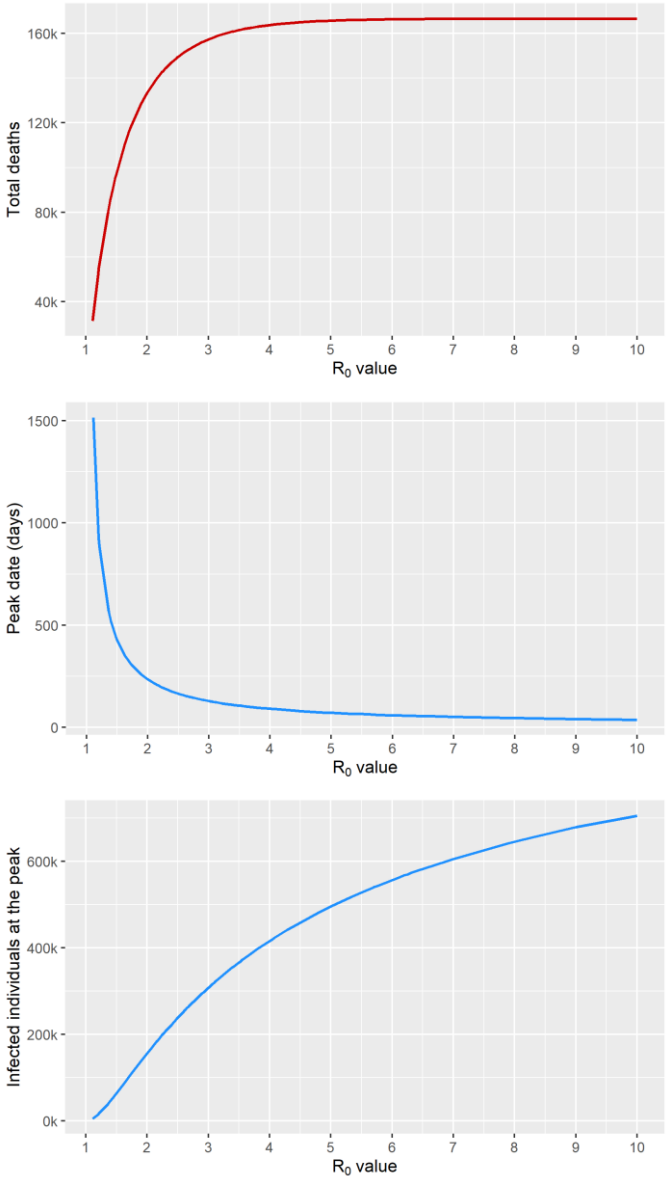
FIGURE 7: EPIDEMIOLOGICAL SCENARIOS



As expected, the value of the reproduction number drives the timing and magnitude of the infected peak (in blue), as well as the resulting number of deaths (in red) and recoveries (in green). The number of susceptibles (in black) indicates the number of individuals who have not yet been infected by the virus and is decreasing at a dramatically faster pace with large values of R_0.

Figure 8 represents key variables (total number of deaths, peak date and number of infected individuals at the peak) as functions of R_0, as forecasted by the SIRD model described above.

FIGURE 8: INFLUENCE OF THE R_0 ON KEY VARIABLES



It is shown that the total number of deaths has a linear growth before slowing down and plateauing. Thus, the total number of deaths is practically unchanged from a certain threshold value of R_0, but can be closely managed by lowering the R_0 below this threshold.

Note that in this illustrative model, and as we discussed before, this toy model does not include a number of factors that are critical to obtaining realistic results. For instance, hospital capacity is assumed to be unlimited, which is not the case in practice. Moreover, the peak date in number of days since the beginning of the epidemic has an exponential decay and the number of infected individuals at the peak is an increasing

function of R_0. Hence, it is relevant to introduce countering measures (e.g., social distancing) in order to lower R_0 and therefore to delay the peak date, and to reduce the number of infected to be treated at the peak of the epidemic.

WHAT HAS BEEN LEARNED?

The implementation of Natural Language Processing for the scientific literature written about the COVID-19 pandemic has many applications. It helps identify relevant papers, not in terms of quality but in terms of content, and extracts both parameter values and informative contexts. By aggregating these parameters estimated by different research teams within the community, we can obtain the range of observed values in a set of papers. If those papers are selected coherently in relation to the context of the specific modelling work, the resulting values can provide a valuable measure of the uncertainty of key model parameters. Measuring this uncertainty is particularly important in early stages of pandemics to refine estimations by efficiently processing the available research papers.

This approach is definitely an interesting and helpful tool to manage the large volume of research papers that are continuously published but it is of course not self-sufficient. Clinical and epidemiological knowledge is required to evaluate the quality of the input studies and understand their limitations, and experts need to take a critical look when interpreting the parameters. Especially in the context of the research around the COVID-19 epidemic, it can't replace the expertise of qualified individuals.

Still, we believe our approach provides an efficient way to be reactive to new publications. Natural Language Processing can assist epidemiologists and modellers in exploring and digesting research papers and in identifying general trends and discrepancies in the scientific community. This widens the toolbox available for experts and leaves the way open to further applications and insights.

References

- Jacob Devlin, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
- Yuwen Zhang & Zhaozhuo Xu. (2019). BERT for Question Answering on SQuAD 2.0.
- Source of the CORD-19 database: is available at https://pages.semanticscholar.org/coronavirus-research.

In a crisis, clarity can be elusive. When the new normal bears little resemblance to the past, and data and information for decision making are in short supply, organisations turn to their trusted advisors to help them find a way forward. For more than 70 years, Milliman has been among the most trusted of those advisors to the life insurance industry.

Our ongoing efforts are helping our clients understand, anticipate and respond to the full range of possible impacts from this public health crisis. Milliman is advising the full spectrum of stakeholders to help them answer important business questions. We are:

- Helping insurers make certain they have **adequate financial reserves and sufficient capital**
- Supporting insurers with the adjustment of their **enterprise risk management and Own Risk and Solvency Assessment (ORSA) frameworks**
- Measuring how this crisis affects **life insurance products' design, pricing, valuation and reinsurance**

The need for professional insight and advice is now more important than ever. For more information about how Milliman can help your organisation find clarity in this time of uncertainty, contact your Milliman consultant or email us at COVID19@milliman.com.

Milliman

Milliman is among the world's largest providers of actuarial and related products and services. The firm has consulting practices in life insurance and financial services, property & casualty insurance, healthcare, and employee benefits. Founded in 1947, Milliman is an independent firm with offices in major cities around the globe.

milliman.com

CONTACT

Remi Bellina
remi.bellina@milliman.com

Alexandre Boumezoued
alexandre.boumezoued@milliman.com

Amal Elfassihi
amal.elfassihi@milliman.com

Cedric Mea
cedric.mea@milliman.com

Floriane Moy
floriane.moy@milliman.com

Eve Titon
eveelisabeth.titon@milliman.com

© 2020 Milliman, Inc. All Rights Reserved. Milliman makes no representations or warranties to the reader with respect to the information contained in this document ("Information") or to any other person or entity, as to the accuracy, completeness or merchantability of the Information. The reader of this document should not construe any of the Information as investment, legal, regulatory, financial, accounting or other advice and persons should consult qualified professionals before taking any specific actions. Milliman shall not be liable to the reader of the Information or any person or entity under any circumstances relating to or arising, in whole or in part, from any circumstance or risk (whether or not this is the result of negligence), or, for any losses, damages or other damages caused in connection with the publication of the Information or its distribution. The holder of this document agrees that it shall not use Milliman's name, trademarks or service marks, or refer to Milliman directly or indirectly in any media release, public announcement or public disclosure, including in any promotional or marketing materials, customer lists, referral lists, websites or business presentations without Milliman's prior written consent for each such use or release, which consent shall be given in Milliman's sole discretion.

This Information contained therein is protected by Milliman's and the authors'/co-authors' copyrights and must not be modified or reproduced without express consent.