

## Proceedings of Statistics Canada Symposium 2021 Adopting Data Science in Official Statistics to Meet Society's Emerging Needs

### Applying the data science approach to COVID-19 epidemiological modelling to inform PPE demand and supply in Canada

by Jihoon Choi, Deirdre Hennessy, Joel Barnes, Dustin Baker,  
Christina Tucker, Kayle Hatt, Gillian Dawson and James Van Loon

Release date: October 22, 2021



## Applying the data science approach to COVID-19 epidemiological modelling to inform PPE demand and supply in Canada

Jihoon Choi<sup>1</sup>, Deirdre Hennessy<sup>1</sup>, Joel Barnes<sup>1</sup>, Dustin Baker<sup>1</sup>, Christina Tucker<sup>2</sup>, Kayle Hatt<sup>2</sup>, Gillian Dawson<sup>2</sup> and James Van Loon<sup>2</sup>

### Abstract

The outbreak of the COVID-19 pandemic required the Government of Canada to provide relevant and timely information to support decision-making around a host of issues, including personal protective equipment (PPE) procurement and deployment. Our team built a compartmental epidemiological model from an existing code base to project PPE demand under a range of epidemiological scenarios. This model was further enhanced using data science techniques, which allowed for the rapid development and dissemination of model results to inform policy decisions.

Key Words: COVID-19; SARS-CoV-2; Epidemiological model; Data science; Personal Protective Equipment (PPE); SEIR

### 1. Introduction

The global pandemic caused by a novel coronavirus, subsequently named SARS-CoV-2, has initiated an unprecedented surge in demand for personal protective equipment (PPE) in Canada (Eggertson, 2020). PPE, in this context, are items worn to protect the user against exposure to infectious disease, such as surgical masks, gloves, and gowns. During the Coronavirus Disease 2019 (COVID-19) pandemic, PPE has become an essential commodity in various sectors across the country, including but not limited to medical clinics, hospitals, retail stores, public transit systems, and restaurants. Given the importance of PPE in combating the transmission of infection, the Government of Canada faced an urgent need to provide timely, accurate and relevant information on PPE procurement and deployment to the provinces and territories.

In response to this need, Statistics Canada developed the Pan-Canadian Demand and Supply Model in collaboration with Health Canada's COVID-19 Taskforce, Public Services Procurement Canada, the Public Health Agency of Canada, and Innovation, Science and Economic Development Canada, as well as private industry partners, to allow policy makers to estimate PPE demand across the provinces and territories under various pandemic scenarios. The Pan-Canadian Demand and Supply Model is composed of three main parts: (1) the epidemiological model that projects different pandemic progression scenarios by province; (2) the demand model that projects unconstrained demand for PPE based on those epidemiological scenarios and PPE protocols from different sectors of the economy including health and non-health sectors; and (3) the inventory or supply model that integrates data from a few different sources, including provinces and territories, to give near real time on-hand and inbound supply information. With these pictures of both demand and supply, the results are netted to identify PPE supply shortages and generate procurement requirements over the next twelve plus months. In contrast to other epidemiological models used by the Government of Canada, the purpose of the epidemiological model feeding the PPE demand projections was not to predict COVID-

---

<sup>1</sup>Jihoon Choi, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, ON, Canada, K1A 0T6 ([jihoon.choi@statcan.gc.ca](mailto:jihoon.choi@statcan.gc.ca)); Deirdre Hennessy, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, ON, Canada, K1A 0T6 ([deirdre.hennessy@statcan.gc.ca](mailto:deirdre.hennessy@statcan.gc.ca)); Joel Barnes, 150 Tunney's Pasture Driveway, Ottawa, ON, Canada, K1A 0T6 ([joel.barnes@statcan.gc.ca](mailto:joel.barnes@statcan.gc.ca)); Dustin Baker, 150 Tunney's Pasture Driveway, Ottawa, ON, Canada, K1A 0T6 ([dustin.baker@statcan.gc.ca](mailto:dustin.baker@statcan.gc.ca)); <sup>2</sup>Christina Tucker, 70 Colombine Driveway, Ottawa, ON, Canada, K1A 0K9 ([christina.tucker@hc-sc.gc.ca](mailto:christina.tucker@hc-sc.gc.ca)); Kayle Hatt, 70 Colombine Driveway, Ottawa, ON, Canada, K1A 0K9 ([kayle.hatt@hc-sc.gc.ca](mailto:kayle.hatt@hc-sc.gc.ca)); Gillian Dawson, 70 Colombine Driveway, Ottawa, ON, Canada, K1A 0K9 ([gillian.dawson@hc-sc.gc.ca](mailto:gillian.dawson@hc-sc.gc.ca)); James Van Loon, 70 Colombine Driveway, Ottawa, ON, Canada, K1A 0K9 ([james.vanloon@hc-sc.gc.ca](mailto:james.vanloon@hc-sc.gc.ca))

19 cases or to trial non-pharmaceutical interventions, but rather to stress test the PPE supply, under various scenarios of demand.

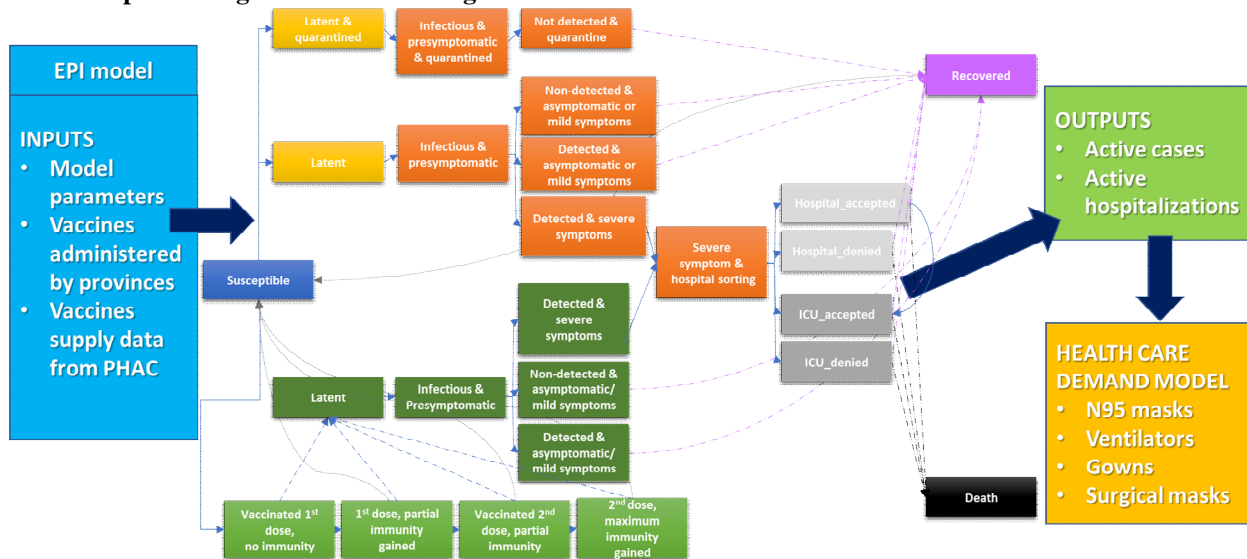
In this article, we will focus on how we developed the epidemiological model while utilizing some elements of the data science approach to quickly iterate epidemiological scenarios and respond to emerging trends in the pandemic. Specifically, we will discuss how we developed the model from an existing [open-source code base](#), optimized calculation processes by utilizing the power of multi-core processors in a cloud computing environment and created visualization tools to systematize reporting of the model outputs for validation and communication.

## 2. Epidemiological model

The Susceptible-Exposed-Infected-Recovered-Deceased-Vaccinated (SEIRDV) model is a classical compartmental model that has been widely used in modelling the spread of infectious diseases, including the COVID-19 pandemic (Ghanam et al., 2020; Korolev, 2021; Piccolomini & Zama, 2020). The base model was initially developed in collaboration with the Public Health Agency of Canada for the purpose of investigating non-pharmaceutical interventions (NPIs) (Ludwig et al., 2020). This model was modified and adapted to become age-structured and province specific, as well as to incorporate vaccination and variants of concern (i.e., Alpha and Delta variants).

In this model, the population flows through the compartments over the course of 1,072 days (January 25, 2020 to January 1, 2023), using a rate parameter defined with ordinary differential equations (ODEs). Initial values of the model parameters are taken from current scientific reports and are further refined through the weekly calibration cycle against real-life COVID-19 statistics. Output from this epidemiological model, such as active cases and hospitalization counts for each province, are forwarded to downstream models, where they are used in estimating PPE demand and supply, see Figure 2.1-1.

**Figure 2.1-1**  
**SEIRDV epidemiological model flow diagram**



## 3. Data science approaches used

### 3.1 Cloud-based delivery

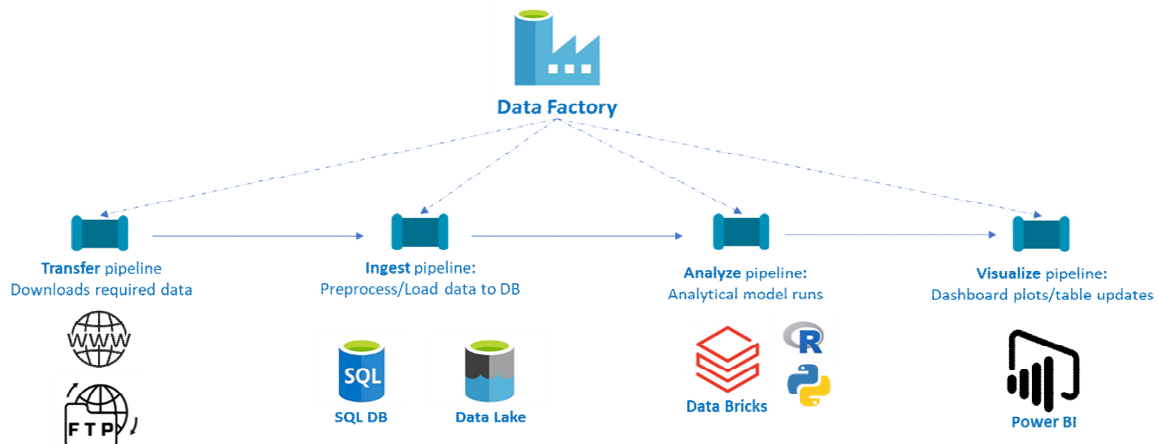
One of the requirements for this model was the need for high level of accessibility for project partners. This was of a critical importance as our deliverable, an interactive dashboard application, was intended to be utilized by numerous internal and external partners with different infrastructural capacities. Moreover, in order to promptly respond to the rapidly changing nature of the pandemic, we had to find a way to deliver frequent updates and patches without interrupting use of the dashboard. As a solution, we decided to host our application on Microsoft Azure (Copeland et al., 2015), a well-established cloud-based computing platform, and delivered the solution to our end-users through the internet, which is similar to the approach seen in the Software as a Service (SaaS) business model.

SaaS is a rapidly growing software delivery model where the software provider hosts the application on a cloud platform and distributes access to users over the internet (Turner et al., 2003). This model is gaining popularity in the Information Technology (IT) field owing to many of its advantages, including superior accessibility and compatibility for users and easier distribution of updates and patches for developers. For instance, since the entire model is hosted on the cloud instead of operating on the user’s device, users do not need to be concerned about infrastructure or storage requirements to access the application. Similarly, because the application is situated in the cloud, applying updates to the software can be done centrally without adversely affecting the user experience. Adapting cloud-based delivery methods have allowed our model and its applications to be highly versatile and scalable, ultimately making it easier for us to react swiftly to the fast-changing conditions of the pandemic.

### 3.2 Production workflow

To maximize scalability, our model has been modularized into multiple pipelines. Figure 3.2-1 shows a simplified diagram of this workflow. These pipelines are overseen by Azure Data Factory, a cloud-based platform that allows developers to monitor and manage the execution of pipelines. When the full model is initiated, a transfer pipeline that utilizes secure file transfer protocol (SFTP) is triggered, which will download relevant data that the model needs from well-established open-source databases, as well as from our internal and external collaborators. These data are then ingested into the Azure Data Lake, a storage repository that holds a vast amount of raw data in its native format. Through an extraction, transform, and load (ETL) cycle, relevant information is extracted from the raw data, transformed into a format that’s suitable for downstream pipelines, and loaded into the SQL database. Consequently, an analytical pipeline is triggered that executes a series of models, including the SEIRDV epidemiological model, and all the outputs are ingested by a visualization pipeline to be displayed in interactive Power BI and R Shiny dashboards for end-users and developers to view.

**Figure 3.2-1**  
**Generalized pipeline flow**



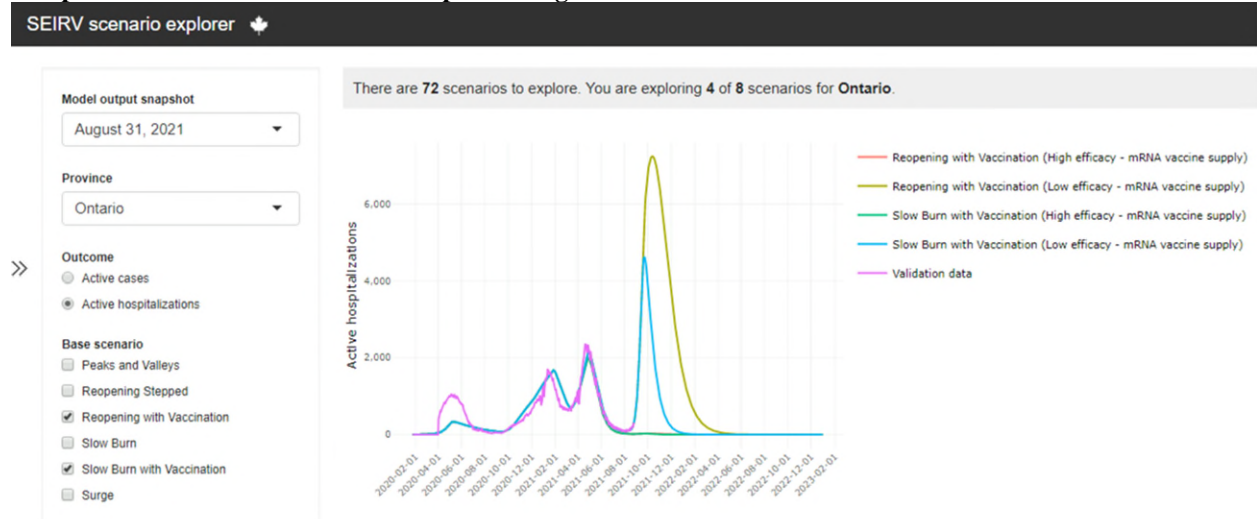
### 3.3 Visualization and validation

Visualization of the model’s output is initially utilized by the quality assurance (QA) team during each of the production cycles. Members of the QA team with subject-matter expertise manually inspect the output plots to check

for any significant outliers or errors. If no issues are found during this inspection cycle, all outputs are pushed to the dashboards.

All tables and plots visualized on the dashboard take an interactive form, where the end-user can explore the output using a built-in control panel. An example of this is the SEIRDV Scenario Explorer powered by R Shiny, which is used for QA and communication (see Figure 3.3-1).

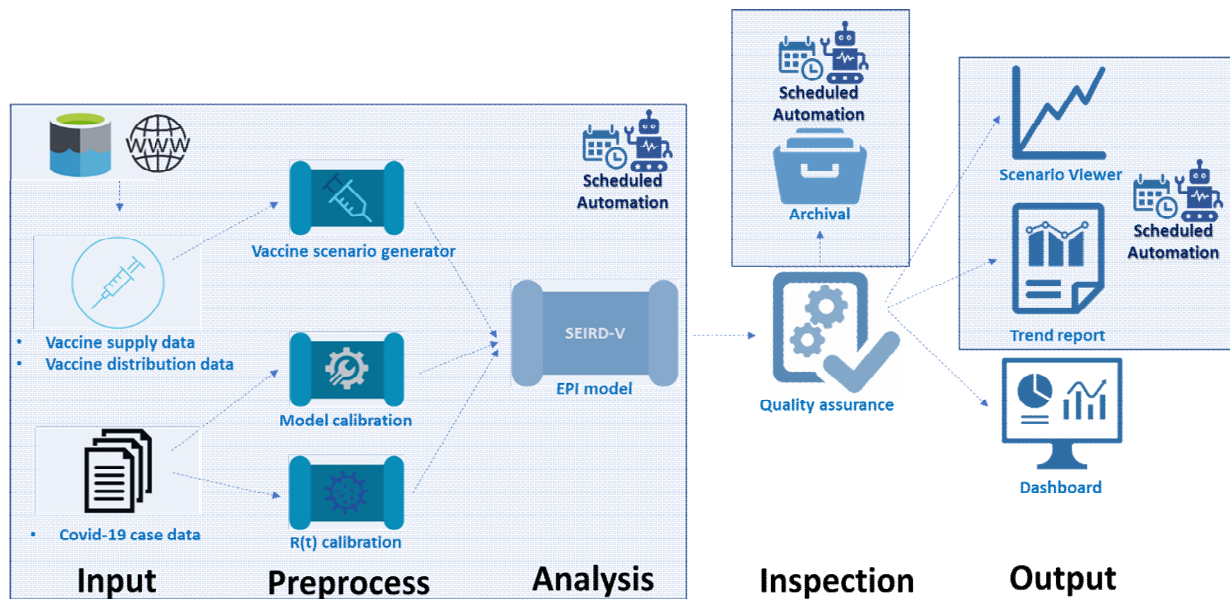
**Figure 3.3-1**  
**Sample visualization of the SEIRDV epidemiological curves**



### 3.4 Pipeline automation

A significant amount of time and effort was spent automating the model pipeline, because we were required to calibrate, run, and update the dashboard on a weekly cycle during the peak periods of the pandemic. To improve productivity and efficiency, we automated almost all pipelines in our model using a scheduler in Azure Data Factory. This has significantly reduced our workload and cost for executing the model and helped quicken the release cycle of our results and dashboard refreshes.

**Figure 3.4-1**  
**Automation of pipeline**



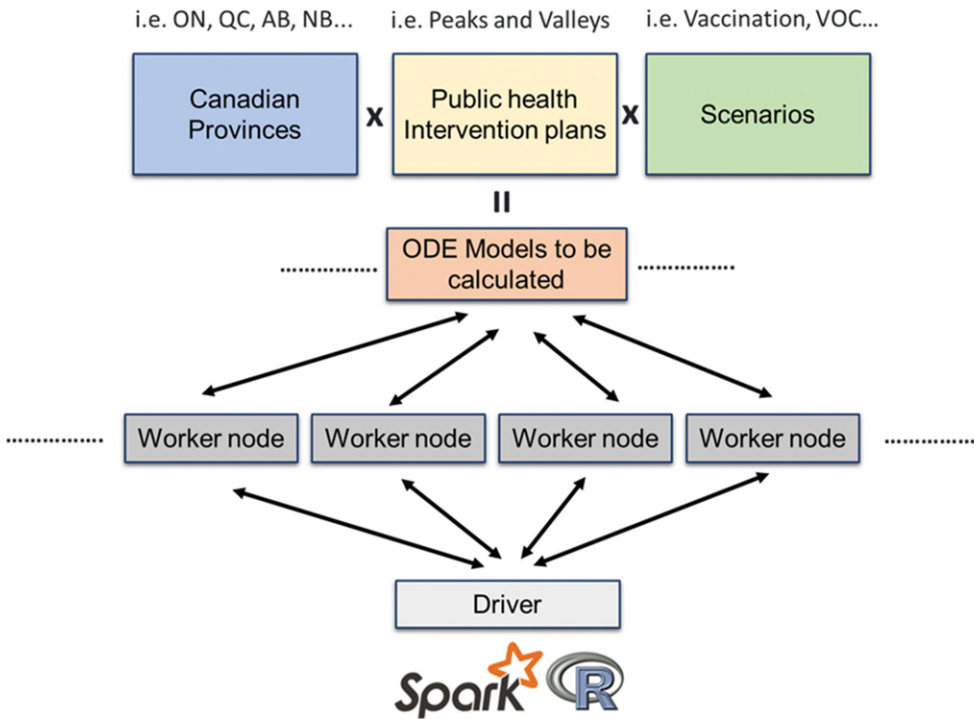
### 3.5 Optimization

A number of different optimization efforts were made in our model pipeline. One example of optimization that significantly reduced the time-complexity of our model was the parallelization of computation. This development greatly increased our efficiency as we were required to run thousands of models during the weekly calibration cycle, in addition to hundreds of model runs during the weekly production cycle.

In order to parallelize computation, we split our large epidemiological model into smaller model subsets, each covering a portion of the province and scenario combinations. Then, an Apache Spark platform (Venkataraman et al., 2016) was used in conjunction with Azure Databricks to: (1) submit each model subset to the driver node of the computing cluster; (2) distribute the required computation across the executor nodes; (3) retrieve the results from the distributed worker nodes; and (4) combine the output into a single data frame form.

Through this implementation, we were able to reduce the computation time by nearly 90%, which made it possible for us to have both our calibration and production run on a weekly basis during the critical periods of the pandemic.

**Figure 3.5-1**  
**Distributed computing**



**Note:** VOC= Variant of concern; ODE=ordinary differential equations

#### 4. Summary and future work

Through applications of the data science approach, we were able to substantially improve the efficiency of our Pan-Canadian Demand and Supply model pipeline, in particular the SEIRDV epidemiological model. Our optimized production work-flow allowed us to respond to the rapidly changing conditions of the pandemic in a timely manner. Overall, the Pan-Canadian Demand and Supply model has made a valuable contribution to modelling the SARS-CoV-2 pandemic in Canada, specifically allowing the estimation and projection of PPE demand and supply sufficiency, in order to inform policy decisions around procurement and deployment.

#### References

Copeland, M., Soh, J., Puca, A., Manning, M., & Gollob, D. (2015), Microsoft Azure. *Microsoft Azure*. <https://doi.org/10.1007/978-1-4842-1043-7>

Eggertson, L. (2020), "Canadian primary care doctors face shortage of protective equipment", *CMAJ*, 192(14), E380–E381. <https://doi.org/10.1503/CMAJ.1095856>

Ghanam, R., Boone, E. L., & Abdel-Salam, A. S. G. (2020), "SEIRD Model for Qatar Covid-19 Outbreak: A Case Study", *Letters in Biomathematics*, 8(1), pp. 19–28. <https://arxiv.org/abs/2005.12777v1>

Korolev, I. (2021), "Identification and estimation of the SEIRD epidemic model for COVID-19", *Journal of Econometrics*, 220(1), 63. <https://doi.org/10.1016/J.JECONOM.2020.07.038>

Ludwig, A., Berthiaume, P., Orpana, H., Nadeau, C., Diasparra, M., Barnes, J., Hennessy, D., Otten, A., & Ogden, N. (2020), "Assessing the impact of varying levels of case detection and contact tracing on COVID-19

transmission in Canada during lifting of restrictive closures using a dynamic compartmental model", *Canada Communicable Disease Report*, 46(1112), pp. 409–421. <https://doi.org/10.14745/CCDR.V46I1112A08>

Piccolomini, E. L., & Zama, F. (2020), "Monitoring Italian COVID-19 spread by an adaptive SEIRD model", *MedRxiv*, 2020.04.03.20049734. <https://doi.org/10.1101/2020.04.03.20049734>

Turner, M., Budgen, D., & Brereton, P. (2003), "Turning software into a service", *Computer*, 36(10), pp. 38–44. <https://doi.org/10.1109/MC.2003.1236470>

Venkataraman, S., Yang, Z., Liu, D., Liang, E., Falaki, H., Meng, X., Xin, R., Ghodsi, A., Franklin, M., Stoica, I., & Zaharia, M. (2016), "SparkR: Scaling R programs with spark", *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 1099–1104. <https://doi.org/10.1145/2882903.2903740>