COVID-19 Public Forecasts for Japan: User Guide & Model Card

A Japanese version of this guide is available here.

Summary

Google Cloud's COVID-19 Public Forecasts provides data on the estimated spread of COVID-19 throughout Japan. The forecasts predict projected death toll, confirmed case counts, active cases[1], and other important values for tracking and projecting the spread of COVID-19. The forecasts predict data for the next 28 days. A prediction is given for each separate prefecture.

The model is intended to help decision makers in healthcare, the public sector, and other impacted organizations be better prepared for what lies ahead. A user should use this forecast alongside other sources of information for decision making, and each user should evaluate whether the model's forecast is appropriate for their specific use.

This guide provides more information on the model itself, the format of forecast outputs, and how to access and use this data. The document provides information on how to use the forecasts as well as information found in Model Cards in the following four sections: (1) Using the Forecast Output, (2) Overview of the Model, (3) Additional Information and Resources, and (4) An Important Look at Fairness.

Section 1: User Guide - Using the Forecast Output

How to Access the Data

The COVID-19 Public Forecasts are provided as publicly accessible BigQuery tables as part of Google Cloud's Public Dataset Program, with the following table names:

- bigquery-public-data.covid19_public_forecasts.japan_prefecture_28d

Users can also download the data in the .CSV format here. Data visualizations are also available in a public Data Studio Dashboard, here. In order to download or use the forecasts, you must agree to the Google Terms of Service.

Forecast Output Format

The schema for the output is shown on the next page, as well as a brief description of each column. More details are available in the subsequent section.

[1] Includes confirmed cases that are inpatient hospitalized, outpatient home/lodging facilities, and pending facility designation. Does not include deaths and recovered patients.

Forecast Output Table Schema

Column Name	Type	Description
japan_prefecture_code	STRING	The prefecture code, for example 'JP-01' for Hokkaido.
prefecture_name	STRING	Full name of the prefecture being forecast.
prediction_date	DATE	Date (YYYY-MM-DD) that the forecast or historical data refers to.
cumulative_confirmed	FLOAT	Cumulative number of people forecast to have confirmed COVID-19 cases.
cumulative_confirmed_q0025	FLOAT	Lower bound of the 95% prediction interval for cases (2.5% quantile).
cumulative_confirmed_q0975	FLOAT	Upper bound of the 95% prediction interval for cases (97.5% quantile).
cumulative_deaths	FLOAT	Cumulative number of people forecast to have died due to COVID-19 on and before the **prediction_date**.
cumulative_deaths_q0975	FLOAT	Upper bound of the 95% prediction interval for deaths (97.5% quantile).
cumulative_deaths_q0025	FLOAT	Lower bound of the 95% prediction interval for deaths (2.5% quantile).
hospitalized_patients	FLOAT	Number of people forecast to be actively receiving medical treatment (not necessarily in a hospital) due to COVID-19 on this **prediction_date**.
hospitalized_patients_q0975	FLOAT	Upper bound of the 95% prediction interval for hospitalized_patients (97.5% quantile).
hospitalized_patients_q0025	FLOAT	Lower bound of the 95% prediction interval for hospitalized_patients (2.5% quantile).
recovered	FLOAT	Cumulative number of people who are forecast to have recovered from COVID-19.
recovered_q0975	FLOAT	Upper bound of the 95% prediction interval for recovered (97.5% quantile).
recovered_q0025	FLOAT	Lower bound of the 95% prediction interval for recovered (2.5% quantile).
cumulative_confirmed_ ground_truth	FLOAT	Reported cumulative confirmed cases of COVID-19.
cumulative_death_ground_ truth	FLOAT	Reported cumulative deaths due to COVID-19.
hospitalized_patients_ground _truth	FLOAT	Reported number of people actively receiving medical treatment (not necessarily in a hospital) due to COVID-19.
recovered_ground_truth	FLOAT	Reported cumulative number of people who have recovered from COVID-19.
forecast_date	DATE	Date (YYYY-MM-DD) when the forecast was created.
new_deaths	FLOAT	New deaths due to COVID-19 forecast on **prediction_date**.
new_confirmed	FLOAT	New confirmed case due COVID-19 forecast on **prediction_date**.
new_deaths_ground_truth	FLOAT	New deaths due COVID-19 that occurred on **prediction_date**.
new_confirmed_ground_truth	FLOAT	New confirmed cases of COVID-19 that occurred on **prediction_date**.

Column Definitions - Additional Details

- **japan_prefecture_code:** the prefecture 'geo_id' corresponds to the ISO 3166-2 designation for Japan prefectures. The designations are all of the form 'JP-XX'. For example, 'Saitama' is 'JP-11'.
- **prefecture_name:** the English name of the prefecture.
- **target_prediction_date:** date (in YYYY-MM-DD format) on which the associated values are forecast to be true. For rows containing ground truth data, this refers to the date that the values were reported to have been true.
- **cumulative_confirmed:** cumulative total number of confirmed COVID-19 cases forecast on that date. This corresponds to the cumulative total number of cases expected to be reported by MHLW (e.g., on these daily briefing pages). MHLW states that they aggregate and report these case counts as published by individual prefectures. Note that unlike the US CDC, presumptive/probable positive cases are *not* included in this figure. (All confirmed cases reflect a positive nucleotide/antigen test.)
- **cumulative_confirmed_q0025:** lower bound of the 95% prediction interval[2] for cumulative number of confirmed COVID-19 cases (2.5% quantile).
- **cumulative_confirmed_q0975:** upper bound of the 95% prediction interval for cumulative number of confirmed COVID-19 cases (97.5% quantile).
- **cumulative_deaths:** cumulative total number of people who have died from COVID-19, counting only nucleotide/antigen-testing confirmed cases, in accordance with the MHLW description.
- **cumulative_deaths_q0025:** lower bound of the 95% prediction interval for the cumulative number of people who have died from COVID-19 (2.5% quantile).
- **cumulative_deaths_q0975:** upper bound of the 95% prediction interval for the cumulative number of people who have died from COVID-19 (97.5% quantile).
- **hospitalized_patients:** daily active number of people hospitalized or receiving medical treatment due to COVID-19.
- **hospitalized_patients_q0025:** lower bound of the 95% prediction interval for the (2.5% quantile).
- **hospitalized_patients_q0975:** upper bound of the 95% prediction interval for the (97.5% quantile).
- **recovered:** cumulative number of people expected to have recovered from COVID-19.
- **recovered_q0025:** lower bound of the 95% prediction interval for the (2.5% quantile).
- **recovered_q0975:** upper bound of the 95% prediction interval for the (97.5% quantile).
- **cumulative_confirmed_ground_truth:** actual cumulative total number of confirmed COVID-19 cases on that date.
- **cumulative_death_ground_truth:** actual cumulative total number of people who have died from COVID-19 on that date.
- **hospitalized_patients_ground_truth:** actual daily active number of people hospitalized or receiving medical treatment due to COVID-19 on that date.
- **recovered_ground_truth:** actual cumulative number of people who have recovered from COVID-19 on that date.
- **forecast_date:** date (in YYYY-MM-DD format) on which the forecast was made. The underlying model used data available up to and including this date in order to develop the forecast.

[2]To generate the **prediction intervals** we first generate the point estimates using our compartmental model. We then apply post-hoc processing that first transforms the point estimates into vector estimates and then optimizes the learnable variables based on the pinball loss (see references here). These vector estimates are the forecast quantiles we output. Finally, we pick out the 0.025, 0.5 and 0.975 quantiles from the vector to report the prediction and prediction interval.

- **new_deaths:** new deaths forecast on the **prediction_date** due to COVID-19. This is a calculated value equivalent to the difference in the **cumulative_deaths** value for a given location on the previous date and the current date.
- **new_confirmed:** new confirmed cases of COVID-19 forecast on the **prediction_date**. This is a calculated value equivalent to the difference in the **cumulative_confirmed** value for a given location on the previous date and the current date.
- **new_deaths_ground_truth:** actual daily number of people who have died from COVID-19 on that date.
- **new_confirmed_ground_truth:** actual daily number of people confirmed Covid-19 cases on that date.

Ground Truth Data

The output tables also include the ground truth values for previous days, from official sources. The full list of input sources is included in the 'Training Data Sources' section at the end of this document.

Ground truth values for confirmed cases and deaths are retrieved from announcements by Japan's Ministry of Health, Labor, and Welfare (MHLW).

These historical ground truth values are included in the forecast output tables, in columns with a name that includes 'ground_truth'. The following columns contain ground truth information::

- **cumulative_confirmed_ground_truth**
- **cumulative_deaths_ground_truth**
- **hospitalized_patients_ground_truth**
- **recovered_ground_truth**
- **new_deaths_ground_truth**
- **new_confirmed_ground_truth**

Additional Notes on the Forecast Output

Cumulative vs. New vs. Daily Active forecasts

The **cumulative_confirmed** and **cumulative_deaths** columns are cumulative values, i.e., the total number of deaths due to COVID-19 by the **prediction_date**. The daily increase in deaths and confirmed cases on a given day are provided in the **new_deaths** and **new_confirmed** columns. Columns prefixed by 'new' are daily incremental values calculated via the difference in the cumulative values from the previous day.

The provided values for the **hospitalized_patients** column reflects the daily active values. This is the number of active cases due to COVID-19 on a given day. This includes confirmed cases that are inpatient hospitalized, outpatient home/lodging facilities, and pending facility designation. Does not include deaths and recovered patients.

Forecast values are not rounded to the nearest integer

The model outputs forecast values as floats, even when the values (e.g. deaths) are always an integer in practice. The output tables give the forecast float values, whereas the dashboard presents these values rounded to the nearest integer.

Section 2: Model Card - Model Overview

Model Description

The model expands upon the 'SEIR' (Susceptible - Exposed - Infected - Recovered) model, which assigns each individual in the population to a 'compartment' based on their disease state. The model adds additional compartments, such as hospitalizations. More details on the model's compartments can also be found in our White Paper.

The model uses machine learning to estimate the transition rate between compartments based on historical data and taking into account other relevant factors ('covariates') that influence the transition rates. For example, the model may determine that mobility indices are covariates that influence the rates at which individuals in different locations move from the Exposed compartment to the Infected compartment. Transition rates are determined for each prefecture.

The model was trained with public data, including historical case numbers and healthcare system information. To ensure the forecast is as current and accurate as possible, the model is retrained regularly. The datasets used to train the model are listed in the 'Training Data Sources' section towards the end of this guide. The model outputs will be provided as long as they are relevant to the COVID-19 pandemic in Japan.

Model Performance

The model's performance is evaluated on an ongoing basis, using backtesting. This evaluation compares how accurately the model would have predicted case numbers for a past time period and compares the model's predictions to historical data. The forecasts predict values for a 28-day time period, and the performance is evaluated by calculating the difference to the actual ground truth data. A detailed model performance analysis is available in our White Paper.

Forecast Output Frequency

New forecasts are produced regularly as new data for confirmed cases, deaths, mobility, and more become available. After a new forecast is created, the public-facing tables are updated with the forecast new values, as well as the newly available 'ground truth' data from official sources for previous days.

Section 3: Model Card - Additional information and resources

Limitations and tradeoffs

- *Input data time lag:* Some of the training data sources update their data with a lag of ~1-3 days. This means that while the forecasts are updated on a regular basis, the latest developments might not be included in all input data at the time of prediction.
- *Rapid trend changes may not be captured:* When the case count in a particular location changes suddenly (e.g., due to modifications in reporting policies or due to other covariates that are not included in the model), these dynamics may not be captured in a timely manner.

- *Ground truth data accuracy:* The ground truth data that the model is using might not be completely accurate. For example, the methods used by official sources to determine the number of confirmed cases may vary across locations.
- *Third party data and sampling variances:* The model's training data includes data from third party sources that may not be accurate, consistent, and/or up-to-date. These third parties may use different approaches, and there could be sampling biases (e.g. underreporting from some prefectures or with respect to some subpopulations).

Training Data Sources

The model is trained on the following public data:

Public data which is hosted on Github and Google Cloud's BigQuery Public Datasets Program:

- Toyo Keizai Online Dataset. See here for an overview of the dataset.
- Google Mobility reports
- Covid-19 World Symptom Survey

Other data:

- Japanese Govt MHLW data
- Japanese Government State-of-Emergency intervention, 2020 (notices announced by PM Cabinet)
- Japan statistical yearbook
- National census
- Handbook of Health and Welfare Statistics
- Open data beds
- Survey of Physicians, Dentists and Pharmacists
- Infectious Disease Surveillance Center
- Comprehensive Survey of Living Conditions
- Statistics on Imposition of Liquor Tax
- National Health and Nutrition Survey Report

Section 4: Model Card - An Important Look at Fairness

Google is committed to a core set of AI principles. In developing the COVID-19 Public Forecasts, we paid close attention to the disproportionate impact the disease has had and how that would impact our adherence to these principles, particularly principle #2: 'Avoid creating or reinforcing unfair bias.'

The COVID-19 pandemic has impacted certain communities and demographic subgroups more than others (e.g., CDC research). We conducted a comprehensive Fairness Analysis to investigate how the forecasts differ across subgroups, similar to the Fairness Analysis for the US launch. The analysis looked at age, sex, ethnicity and income, finding that once the total number of cases had been taken into account model performance was generally consistent across demographic groups. When looking at individual prefectures, performance was again consistent with the best performance seen in prefectures with more cases and greater population density. We encourage all users who intend to make decisions in part based

on the COVID-19 Public Forecasts to closely review the Fairness Analysis as well as our White Paper.

White Paper

More details on the model, the methodology used to develop the model, and the model's performance can be found in our published White Paper.

Acknowledgements

Special thanks to those on the Google Cloud, Google Japan and AI Fairness teams who worked on this project, including Andrew Max, Ben Hutchinson, Emilio Garcia, Fergal Daly, Hiroki Kayama, Ivor Horn, Joe Ledsam, Joel Shor, Jinsung Yoon, Junichi Kawai, Kaho Kobayashi, Karen Ouk, Kris Popendorf, Madeleine Elish, Mike Dusenberry, Nanako Yamaguchi, Natalie Wei, Nate Yoder, Peter Fitzgerald, Raj Sinha, Ryu Hirayama, Sakura Tominaga, Sercan Arik, Takahide Kato, Timnit Gebru, Tomas Pfister, Tomohiko Kikuchi, Vik Menon.