# COVID-19 感染予測 (日本版): ユーザーガイド

英語版はこちら

## 概要

COVID-19 感染予測 (日本版) は、COVID-19（新型コロナウイルス感染症）の日本全国での感染の広がりに関する予測データを提供しています。このデータは都道府県別に、対象期間である将来 28 日間のあいだに予測される死亡者数、陽性者数、入院・療養等患者数[1]を表しています。これらの項目に関する重要な指標として累計や日別のデータ、95 %予測区間等も掲載しています。また、全国の予測値は都道府県の予測値を足し合わせることで表示しています。

このモデルは、医療機関や公的機関を始めとする COVID-19 の影響を受ける組織が、今後に向けてより適切な対処を検討・準備する上で手がかりとなる情報の一つとして利用されることを目的に公開しています。なんらかの決定を下すためにこの予測データを使用する際は、必ず他の情報源も併せて参照してください。また、モデルの予測がその用途に適しているかどうかは、ご利用者様ご自身で判断していただく必要があり、Google はその結果について一切の責任を負いません。

このユーザーガイドでは、モデルの概要、予測データ出力形式、このデータへのアクセス方法と使用方法について説明します。予測データの使用方法に関する情報とモデルカードの内容を、次の 4 つのセクションで記載します。（1）出力される予測データの使用、（2）モデルの概要、（3）追加情報とリソース、（4）公平性に関する重要な考察。

## セクション 1: ユーザーガイド - 出力される予測データの使用

**データへのアクセス方法**

COVID-19 感染予測 (日本版) は、Google Cloud の一般公開データセット プログラムの一環で公開される BigQuery テーブルとして、次のテーブル名で提供されます。

- bigquery-public-data.covid19_public_forecasts.japan_prefecture_28d

データは、こちらから .CSV 形式でダウンロードすることもできます。可視化されたデータは一般公開されているデータポータル ダッシュボードでご確認ください。予測データをダウンロードまたは使用するには、Google の利用規約に同意する必要があります。

[1] 入院、宿泊療養、自宅療養、入院・療養調整中の患者（死亡・回復を除く）

### 予測データ出力形式

出力のスキーマは、各列の簡単な説明とともに次のページに掲載します。詳細については、以下のセクションをご覧ください。

### 予測データ出力テーブル スキーマ

| 列名 | データ型 | 説明 |
|---|---|---|
| **japan_prefecture_code** | STRING | 都道府県コード。たとえば、北海道は「JP-01」。 |
| **prefecture_name** | STRING | 予測対象の都道府県名。 |
| **prediction_date** | DATE | 予測データまたは過去のデータが参照する日付（YYYY-MM-DD）。 |
| **cumulative_confirmed** | FLOAT | 予測される COVID-19 の累計陽性者数。 |
| **cumulative_confirmed_q0025** | FLOAT | 累計陽性者数の 95% 予測区間の下限（2.5% 分位数）。 |
| **cumulative_confirmed_q0975** | FLOAT | 累計陽性者数の 95% 予測区間の上限（97.5% 分位数）。 |
| **cumulative_deaths** | FLOAT | **prediction_date** 以前（当日を含む）の、予測される COVID-19 による累計死亡者数。 |
| **cumulative_deaths_q0975** | FLOAT | 累計死亡者数の 95% 予測区間の上限（97.5% 分位数）。 |
| **cumulative_deaths_q0025** | FLOAT | 累計死亡者数の 95% 予測区間の下限（2.5% 分位数）。 |
| **hospitalized_patients** | FLOAT | **prediction_date** における、予測される COVID-19 による入院・療養等患者数。 |
| **hospitalized_patients_q0975** | FLOAT | 入院・療養等患者数の 95% 予測区間の上限（97.5% 分位数）。 |
| **hospitalized_patients_q0025** | FLOAT | 入院・療養等患者数の 95% 予測区間の下限（2.5% 分位数）。 |
| **recovered** | FLOAT | 予測される COVID-19 からの累計回復者数。 |
| **recovered_q0975** | FLOAT | 累計回復者の 95% 予測区間の上限（97.5% 分位数）。 |
| **recovered_q0025** | FLOAT | 累計回復者の 95% 予測区間の下限（2.5% 分位数）。 |
| **cumulative_confirmed_ ground_truth** | FLOAT | COVID-19 陽性者数の報告された実績値の累計。 |
| **cumulative_death_ground_ truth** | FLOAT | COVID-19 による死亡者数の報告された実績値の累計。 |
| **hospitalized_patients_ground_truth** | FLOAT | COVID-19 による入院・療養等患者数の報告された実績値。 |
| **recovered_ground_truth** | FLOAT | COVID-19 からの回復者数の報告された実績値の累計。 |
| **forecast_date** | DATE | 予測作成日（YYYY-MM-DD）。 |
| **new_deaths** | FLOAT | **prediction_date** における、予測される COVID-19 による新規死亡者数。 |
| **new_confirmed** | FLOAT | **prediction_date** における、予測される COVID-19 の新規陽性者数。 |
| **new_deaths_ground_truth** | FLOAT | 当該日に発生したCOVID-19 による死亡者数の報告された実績値。 |
| **new_confirmed_ground_truth** | FLOAT | 当該日に発生したCOVID-19 による陽性者数の報告された実績値。 |

**列の定義 - 詳細説明**

- **japan_prefecture_code**: 都道府県「geo_id」は日本の都道府県の ISO 3166-2 の定義に対応し、「JP-XX」の形式で表記します。たとえば、「埼玉」は「JP-11」です。
- **prefecture_name**: 予測対象の都道府県名。
- **prediction_date**: 関連する値が真であると予測される日付（YYYY-MM-DD 形式）。Ground Truth データを格納する行の場合は、値が真であると報告された日付を指します。
- **cumulative_confirmed**: 当該日に予測される、COVID-19 の累計陽性者数。厚生労働省により発表される累計陽性者数（例: 国内の発生状況など）に対応します。厚生労働省は、新規陽性者数について、各自治体がプレスリリースしている個別の事例数（再陽性例を含む）を積み上げて算出していると述べています。CDC（アメリカ疾病予防管理センター）とは異なり、推定陽性の症例はこの値に含まれていないことに注意してください（すべての陽性者数はPCR検査または抗原検査等が陽性であることを示しています）。
- **cumulative_confirmed_q0025**: COVID-19 累計陽性者数の 95% 予測区間[2]の下限（2.5% 分位数）。
- **cumulative_confirmed_q0975**: COVID-19 累計陽性者数の 95% 予測区間[1]の上限（97.5% 分位数）。
- **cumulative_deaths**: **prediction_date** 以前（当日を含む）の、COVID-19 による累計死亡者数（厚生労働省の説明に従って PCR 検査または抗原検査等で確認された陽性者のみを数える）。
- **cumulative_deaths_q0025**: COVID-19 による累計死亡者数の 95% 予測区間[1]の下限（2.5% 分位数）。
- **cumulative_deaths_q0975**: COVID-19 による累計死亡者数の 95% 予測区間[1]の上限（97.5% 分位数）。
- **hospitalized_patients**: **prediction_date** における、COVID-19 による入院・療養等患者数の 1 日あたりの有効数。
- **hospitalized_patients_q0025**: 入院・療養等患者数の95% 予測区間[1]の下限（2.5% 分位数）。
- **hospitalized_patients_q0975**: 入院・療養等患者数の95% 予測区間[1]の上限（97.5% 分位数）。
- **recovered**: 予測される COVID-19 からの累計回復者数。
- **recovered_q0025**: 累計回復者数の 95% 予測区間[1]の下限（2.5% 分位数）。
- **recovered_q0975**: 累計回復者数の 95% 予測区間[1]の上限（97.5% 分位数）。
- **cumulative_confirmed_ground_truth**: 当該日までに発生した COVID-19 陽性者数の報告された実績値の累計。
- **cumulative_death_ground_truth**: 当該日における COVID-19 による死亡者数の報告された実績値の累計。
- **hospitalized_patients_ground_truth**: 当該日における、1 日あたりのCOVID-19 による入院・療養等患者数の報告された実績値。
- **recovered_ground_truth**: 当該日における COVID-19 からの回復者数の報告された実績値の累計。
- **forecast_date**: 予測作成日（YYYY-MM-DD 形式）。このモデルは、予測作成日を含むその日付までに利用可能なデータを使用しています。
- **new_deaths**: ***prediction_date*** における、予測される COVID-19 による新規死亡者数。これは、前日と当日の特定の都道府県の ***cumulative_deaths*** 値の差に相当する計算値です。
- **new_confirmed**: ***prediction_date*** における、予測される COVID-19 の新規陽性者数。これは、前日と当日の特定の都道府県の ***cumulative_confirmed*** 値の差に相当する計算値です。
- **new_deaths_ground_truth**: 当該日に発生したCOVID-19 による死亡者数の報告された実績値。

[2] 予測区間を生成する際には、最初にコンパートメントモデルを使用してポイント推定値を算出します。次に、ポイント推定値をベクトル推定値に変換するpost-hoc処理を行い、ピンボールロス (参照)に基づき学習可能な変数を最適化します。このベクトル推定値を予測分位点としています。ベクトルから0.025、0.5、 0.975分位点を選定し、予測と予測区間を算出します。

- **new_confirmed_ground_truth**: 当該日に発生したCOVID-19 による陽性者数の報告された実績値。

**Ground Truth データ**

出力テーブルには、厚生労働省や各都道府県等より一般公開されているデータからの過去の日付の Ground Truth 値も含まれています。入力されているデータソースのリストは、このユーザーガイドの最後にある「トレーニング データ ソース」セクションに含まれています。

陽性者数と死亡者数の Ground Truth値は、厚生労働省が公開するオープンデータから取得しています。

これらの過去の Ground Truth 値は、予測データ出力テーブルの、「ground_truth」を含む名前の列に含まれています。Ground Truth が含まれる列は次のとおりです。

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

**予測データ出力に関する追加注意事項**

累計数・新規数・ 1 日の有効数の予測

**cumulative_confirmed** 列と **cumulative_deaths** 列は累計値です。たとえば、**prediction_date** までの COVID-19 による死亡者総数などです。特定の日における死亡者数と陽性者数の 1 日の発生数は、**new_deaths** 列と **new_confirmed** 列に格納されます。「**new**」列は、前日との累計値の差から計算された 1 日の増分値です。

**hospitalized_patients** 列に指定された値は、1 日の有効数の値を反映しています。これは、特定の日に COVID-19 による入院・療養等患者数です。入院・療養等患者数の値は、COVID-19 の陽性診断を受けた人のうち、死亡者・回復者を除く入院、宿泊療養、自宅療養、入院・療養調整中の患者の人数を反映しています。

予測値は最も近い整数に丸められない

このモデルは、実際には値（死亡数など）が常に整数である場合でも、予測値を浮動小数点として出力します。出力テーブルには予測された浮動小数点値が表示されますが、ダッシュボードにはこれらの値が最も近い整数に丸められて表示されます。

## セクション 2: モデルカード - モデルの概要

**モデルの説明**

このモデルは集団内の各個人を病状に基づいて「区画」に割り当てる「 SEIR 」（ Susceptible（感染前の状態、免疫なし）- Exposed（ウイルスに曝露したものの他者への感染性を有さない状態、曝露）- Infected（他者への感染性を有する状態、発症）- Recovered（回復して免疫を獲得あるいは死亡した状態、回復）モデルを拡張したものです。このモデルでは、入院・療養等患者数などの区画が追加されています。モデルの区画の詳細は、ホワイトペーパーでも確認できます。

本モデルでは機械学習を使用して、過去のデータに基づき、区画間の遷移率に影響を与える他の関連要因（共変量）を考慮に入れて、個人が区画間をどのような確率で遷移するかを推定します。たとえば、モデルでは移動指数が異なる場所にいる個人が Exposed（曝露）区画から Infected（発症）に遷移率に影響を与える共変量であると判断する場合があります。遷移率は、都道府県ごとに判断されます。

モデルは、過去の陽性者数や医療システム情報などの一般公開データを使用してトレーニングされています。予測が可能な限り最新かつ正確であることを保証するため、モデルは定期的に再トレーニングされます。モデルのトレーニングに使用されるデータセットは、このガイドの最後にある「トレーニングデータソース」セクションに記載されています。また、 COVID-19 感染拡大が続く限り、この予測データを継続的に提供する予定です。

**モデルのパフォーマンス**

モデルのパフォーマンスは、バックテストを使用して継続的に評価されます。この評価では、モデルの予測値を過去の実績データと比較し、モデルが過去の期間の陽性者数をどれだけ正確に予測したか精査します。予測は 28 日間の値の予測で、パフォーマンスは実際の Ground Truth データとの差を計算することによって評価されます。詳細なモデル パフォーマンス分析は、ホワイトペーパーに記載されています。

**予測出力頻度**

新しい予測は、陽性者数、死亡者数、コミュニティ モビリティ レポートなどの新しいデータが利用可能になるタイミングに応じて、定期的に出力されます。新しい予測の出力後、公開されている図表は、新しい予測値と過去の日付の公式ソースから新たに利用可能になった「 Ground Truth 」データに基づき更新されます。

## セクション 3: モデルカード - 追加の情報とリソース

**利用制限とトレードオフ**

- 入力データのタイムラグ: トレーニングに用いられている一部のデータソースでは、最新の状態よりも 1 ～ 3 日遅れてデータが更新されます。つまり、予測は定期的に更新されますが、予測出力時には入力データのすべてに最新状況が含まれていない可能性があります。
- 急激なトレンドの変化: 特定の場所で陽性者数が突然変化したような場合（たとえば、検査の報告方針の変更や、モデルに含まれていない他の共変量の変化が原因となるようなこと）、これらの変化はタイムリーにデータソースや予測値に反映されない場合があります。
- Ground Truth データの精度: モデルが使用している Ground Truth データは、完全に正確ではない可能性があります。たとえば、公式のデータソースで使用される陽性者数を判断するための方法は、地域によって異なる場合があります。
- サードパーティデータとサンプリングの分散: モデルのトレーニング データには、サードパーティソースからの正確でない、一貫性がない、あるいは最新でない可能性のあるデータが含まれます。サードパーティはそれぞれに異なるアプローチを採用する可能性があり、サンプリングバイアスが存在する可能性があります（たとえば、一部の都道府県または部分母集団に関して過少報告があるなど）。

**トレーニング データソース**

モデルは、次の一般公開データでトレーニングされています。
GitHub と Google Cloud の BigQuery 一般公開データセット プログラムでホストされている一般公開データ:

- 『東洋経済オンライン「新型コロナウイルス 国内感染の状況」制作：荻原和樹』データセット。データセットの概要については、こちらをご覧ください。
- Google コミュニティ モビリティ レポート
- Covid-19 World Symptom Survey

その他のデータ:

- 厚生労働省オープンデータ
- 日本政府の非常事態宣言の発表、2020 年（首相官邸が発表した通知）
- 日本統計年鑑
- 国勢調査
- Handbook of Health and Welfare Statistics
- 病床オープンデータ
- 医師・歯科医師・薬剤師統計
- 国立感染研究所感染症情報センター
- 国民生活基礎調査
- 国税庁　統計情報・各種資料
- 国民健康・栄養調査報告

## セクション 4: モデルカード - 公平性に関する重要な考察

Google は AI 利用における基本方針を遵守しています。COVID-19 感染予測(日本版)の開発にあたり、Google では COVID-19 による影響の不均衡性と、特にこの原則の第 2 項「不公平なバイアスの発生、助長を防ぐ」への準拠において及ぼす影響について熟慮しました。

COVID-19 は、特定の人種やサブグループに強い影響を与えています。（例：CDCの研究結果）アメリカでの感染予測モデルの公開に先立ち、Google の担当チームは、公平性に関する分析を包括的に行い、そうした不均衡性が予測の正確性に与える影響とその適切な解釈について考察しました。日本のモデルのパフォーマンスについても同様に、公平性に関する同様の分析を実施しています。具体的には年齢、性別、民族性、収入を調査した上で、陽性者の総数を考慮に入れても、モデルのパフォーマンスは人口統計グループ全体で概ね一貫していることがわかりました。各都道府県別のパフォーマンスを見ても、陽性者数が多く人口密度が高い都道府県で見られるパフォーマンスと一致していました。手がかりの一つとして COVID-19 感染予測(日本版)の活用を検討される際には、公平性に関する分析とホワイトペーパーを精読されることをおすすめします。

## ホワイトペーパー

モデル、モデルの開発に使用された方法論、モデルのパフォーマンスの詳細については、公開されているホワイトペーパーをご確認ください。

## 謝辞

*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.*

Google Cloud

詳しくは、**google**.com/cloud をご覧ください