Bioinformatics Summer School Final Report, 2025

Beyond Swabs:

A Mortality-anchored Method to Estimate True COVID-19 Infections

Lamis Saeed Alghamdi, University of Cambridge

QUESTION

Deaths are harder to miss than infections. Can we use confirmed deaths to determine how many cases we missed during the pandemic?

IMPORTANT ABBREVIATIONS

CFR: Case Fatality Ratio – the risk of death among reported cases

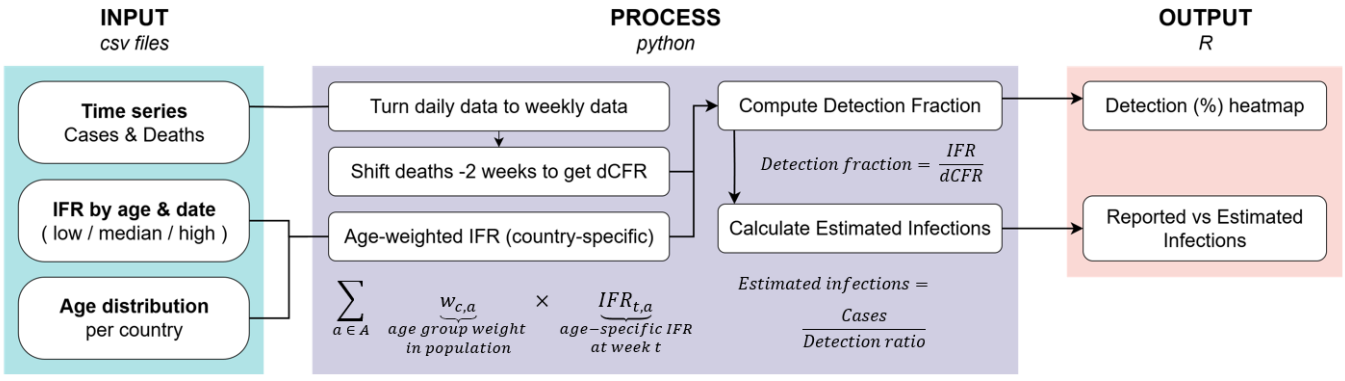
$$CFR = \frac{Deaths}{Reported\ infections}$$

IFR: Infection Fatality Ratio – the risk of death per infection (includes undiagnosed)

$$IFR = \frac{Deaths}{Total\ infections}$$

METHODOLOGY AND APPROACH

Main flowchart:

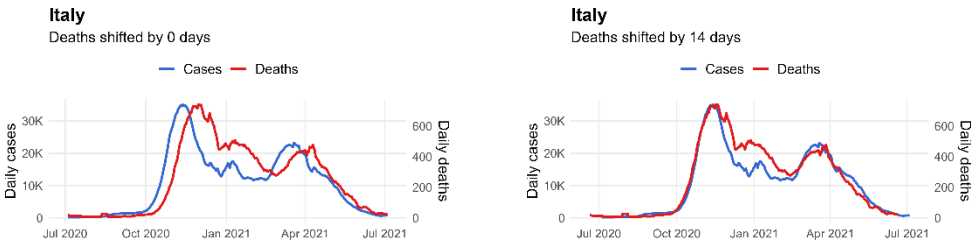


a *— age bin (e.g., 6–24, 25–44, 45–54, 55–64, 65–74, 75+).*
A *— the set of all age bins:* $A = \{6-24, 25-44, 45-54, 55-64, 65-74, 75+\}$
c *— country (Saudi Arabia, France, Italy, United Kingdom)*
t *— epidemiological week*

Raw inputs

- **JHU CSSE COVID-19 time series** — *Johns Hopkins University CSSE. COVID-19 Data Repository.* GitHub (global confirmed/deaths time-series).[1]
- **UN WPP age structure** — *United Nations, Department of Economic and Social Affairs, Population Division. World Population Prospects (2024 revision), population by 5-year age groups (CSV).*[2]
- **IFR by age & date (low/median/high):** Ward et al., *Nature Communications* (2024), England IFR time series, Source Data. Used to build weekly age-weighted IFR per country.[3]
- **Variant context** — *CoVariants.org variant frequency explorer.*[4]

I estimated weekly infections by tying mortality to case counts. First, I converted daily cases and deaths to weekly totals and computed a delayed CFR (dCFR). Deaths mostly come from infections that happened in the past, so I overlaid daily cases and deaths and shifted deaths backward to align peaks (visual check across countries). So, the dCFR pairs deaths two weeks later with the current week's cases to reflect the infection-to-death lag. Here is a visual example:



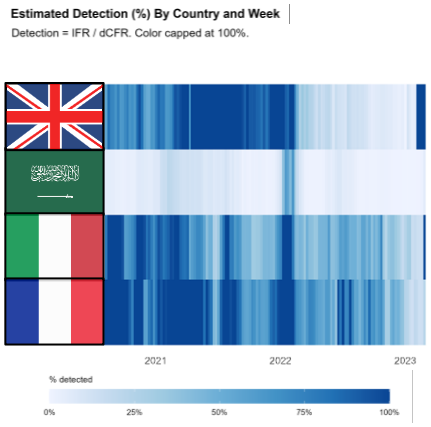
Next, I built a country-specific weekly IFR by weighting age-specific IFR values with each country's population shares (6–24, 25–44, 45–54, 55–64, 65–74, 75+), keeping low/median/high versions for uncertainty. Comparing this IFR to the delayed CFR gives a detection fraction, the share of infections surveillance that was likely captured. I then derived estimated infections primarily from cases divided by the detection.

Variant eras are based on **CoVariants.org** weekly sequence frequencies. An era begins in the first week a variant accounts for >50% of sequences. This was only used for visualization and had nothing to do with the analysis.

RESULTS & REASONING

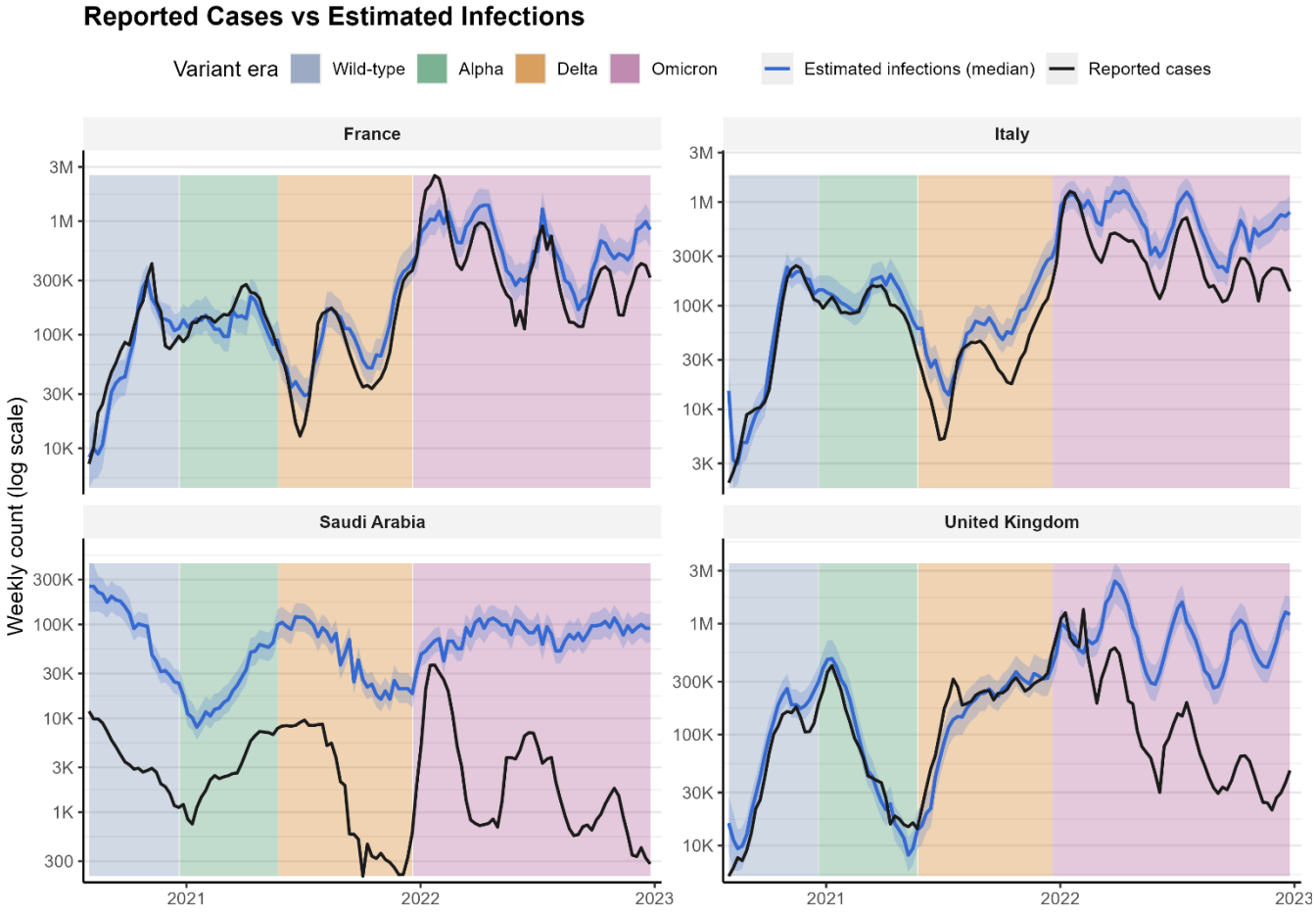
Detection fraction over time

I mapped the weekly detection fraction (≈ IFR / dCFR) as a heatmap. The first weeks show elevated detection because the analysis window starts mid-pandemic, not at the very beginning. As we move into late-2021/2022, detection declines. This is consistent with the emergence of and changes in testing/reporting (next plot).



Reported Cases vs Estimated Infections

Each panel compares reported cases (black) with estimated infections (blue; median line with an uncertainty ribbon) on a log scale. Colored bands mark variant eras (Wild-type, Alpha, Delta, Omicron). Estimated infections are computed case-based as cases × dCFR / age-weighted IFR; the ribbon reflects low/high IFR assumptions.



Some countries, like the UK, Italy and France, show good alignment through Alpha/Delta. During Omicron, estimated infections consistently exceed reported cases, suggesting under-reporting as testing practices changed, and self-testing rose. This aligns with the nature of Omicron, where it spreads quickly yet the infections were not severe, so people were not getting tested by testing centers.

But why does Saudi Arabia look so different?

1. **IFR prior and why it matters**

I use an IFR prior from England (Ward et al., *Nature Communications*, 2024). It reflects an older, high-income population with time-varying, age-specific risks. If I apply those values to Saudi Arabia, they likely **overstate Saudi's true IFR**. In a death-anchored approach that leads me to overestimate infections for Saudi.[3]

2. **Why Saudi may have a lower IFR**

Saudi Arabia and the wider GCC have younger populations, and IFR rises sharply with age. Regional evidence, for example from Qatar, reports lower severity and fatality, which fits a lower all-age IFR across the Gulf. Pair that with Saudi's early, large-scale testing and swift public-health measures, and the result is a smaller mortality burden. This also explains why my estimated infections can sit higher than reported cases. Part of the "low detection" signal likely comes from using an England prior rather than a GCC-specific IFR. [3, 5, 6]

3. **Response context**

Saudi acted early, drew on MERS experience, restricted travel and gatherings, and scaled testing quickly. Roughly 43 million tests were completed by mid-2022. These steps likely raised detection and reduced transmission and deaths, which supports the view that England's IFR is a conservative choice for Saudi, meaning higher than Saudi's true IFR.[6]

CHALLENGES

- The nature of Cambridge's version of the bioinformatics summer school only allowed us to work on this project for 3 days. I later spent around a week refining the code and the model when I returned to Saudi.
- IFR generalization. My estimates rely on age-specific, time-varying IFRs reported by Ward et al. (the study was done in England). I age-standardize these IFRs to each country's population structure and apply them globally. This implicitly assumes that infection severity conditional on age is similar across settings, which may be violated by differences in vaccination coverage, comorbidities, healthcare capacity, reporting practices, and variant mix. I adopted this source because it is the only publicly available longitudinal IFR-by-age series I was able to find.
- Given the time constraint, the infection-to-death delay was set empirically by visually aligning case and death peaks (≈13 days for most countries) rather than by fitting a formal delay model. I didn't have time to experiment with the code and was struggling with this step.

NEXT STEPS

Using deaths to estimate infections works, but several refinements will make the results more reliable, like using country-specific IFR, account for vaccination, prior infection, treatment changes, and variant mix when building weekly IFR. Then validate estimates against seroprevalence studies

RESOURCES AND TOOLS

Analysis: Python (pandas, numpy) in DataSpell and Jupyter
Plotting: R and RStudio (tidyverse, ggplot2)

Data & code can be found here: https://github.com/L4M315/beyond_swabs.git

REFERENCES

1. **JHU CSSE COVID-19 time series.**
   Johns Hopkins University Center for Systems Science and Engineering (CSSE). *COVID-19 Data Repository* (global confirmed/deaths time series), GitHub. https://github.com/CSSEGISandData/COVID-19
2. **UN WPP age structure.**
   United Nations, DESA, Population Division. *World Population Prospects 2024 (WPP2024)* population by 5-year age groups (CSV). https://population.un.org/wpp/
3. **IFR by age & date (low/median/high).**
   Ward, T., Fyles, M., Glaser, A., Paton, R. S., Ferguson, W., & Overton, C. E. (2024). *The real-*

time infection hospitalisation and fatality risk across the COVID-19 pandemic in England. **Nature Communications, 15**(1), 4633. https://doi.org/10.1038/s41467-024-47199-3

4. **Variant context.**
   Hodcroft, E. B. *CoVariants: SARS-CoV-2 Mutations and Variants of Interest*. https://covariants.org (variant frequency explorer).
5. **Lower severity evidence in the Gulf (Qatar).**
   Chemaitelly, H., Ayoub, H. H., Faust, J. S., et al. (2023). *Turning point in COVID-19 severity and fatality during the pandemic: a national cohort study in Qatar.* **BMJ Public Health, 1**(1):e000479. PMCID: PMC11812731. https://pmc.ncbi.nlm.nih.gov/articles/PMC11812731/
6. **Context on Saudi testing and response.**
   Arab News. *The Kingdom vs. COVID-19* (feature on early response and large-scale testing). https://www.arabnews.com/TheKingdomVsCOVID19