A Bayesian Approach to Covid-19 Excess vs Official Deaths in the Latin America Region

Timothy Lee

Introduction

Inspired from The Economist article on "Tracking covid-19 excess deaths across countries" (The Economist, 2021), this report aims to investigate the proportion of overall excess deaths vs the official reported death counts due to the covid-19 pandemic. We will primarily focus on 4 Latin America countries (Chile, Ecuador, Mexico, and Peru) as a representation of the Latin America region, which have mostly "experienced a devastating first wave from April to July 2020" (The Economist, 2021). It is hypothesized that reported covid-19 deaths will be a much smaller proportion of overall excess deaths for Latin American countries due to various reasons including underreporting, misclassification, or having a higher number of total covid-19 cases in general (hospital overload).

Definition of Terms

First, we will define the term "excess deaths" as the "difference between the observed numbers of deaths in specific time periods and expected numbers of deaths in the same time periods" (CDC, 2022). This distinction is especially important in the context of the covid-19 pandemic as it is crucial to recognize that causes of deaths are hard to be attributed to specific factors since "some deaths due to covid-19 may be assigned to other causes of deaths, [hence] tracking all-cause mortality can provide information about whether an excess number of deaths is observed, even when COVID-19 mortality may be undercounted." (CDC, 2022).

Methodology and Modelling Assumptions

Next, we will move on to our methodology of how to calculate excess deaths. For simplicity reason, we will assume the starting date of the covid-19 pandemic for all countries to be on March 1st, 2020 since the official date when the WHO made the assessment that covid-19 can be characterized as a pandemic is on March 11, 2020 (WHO, 2020). Hence for each country, a baseline model of total weekly mortatility counts will be fitted on all the pre-covid years and out-of-sample forecasts will be generated for the covid years (March 1st onwards). The rationale here is that seasonal patterns and the trend would be capture by this baseline model based on the pre-covid mortatility counts. Using this baseline model, we can then compare the excess deaths by subtracting the actual number of mortatility counts in the covid years with this historical baseline. This way, our model will include people who died from covid-19 but the cause of death wasn't correctly attributed to the official covid mortatiltiy numbers due to the reasons aforementioned. Finally, we will then compare this excess death against the official reported covid-19 deaths for each country.

The Data

First, for consistency considerations, the total mortatility counts and the official covid death counts for each country will be retrieved from The Economist's publicly available data repository on GitHub (https://github.com/TheEconomist/covid-19-excess-deaths-tracker/tree/master/output-data/historical-deaths). The original source includes "the Human Mortality Database, a collaboration between UC Berkeley and the Max Planck Institute in Germany, and the World Mortality Dataset, created by Ariel Karlinsky and Dmitry Kobak." (The Economist, 2021). We will first plot the total deaths counts against the official covid-19 death counts for each Latin America country. Note that both death counts are reported in a weekly time frame.

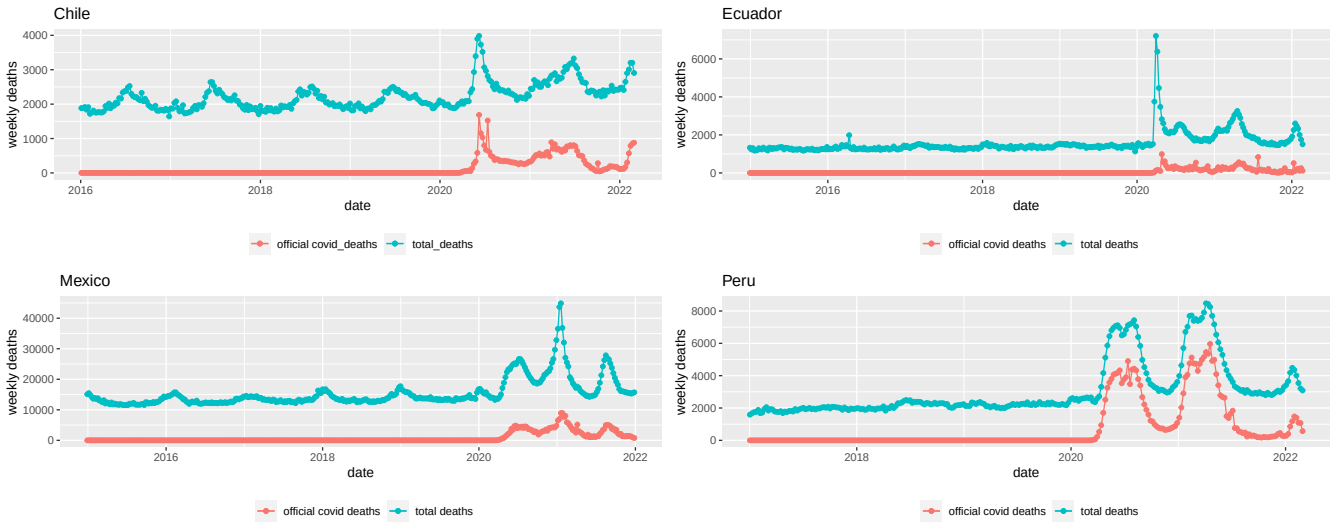

Figure 1: Plot of official covid-19 vs total deaths counts for each country

Modelling

In the original article, The Economist had first used "a five-year average of deaths in a given region to calculate a baseline for excess deaths", followed by "a statistical model for each region, which predicts the number of deaths we might normally have expected in 2020. The model fits a linear trend to years, to adjust from long-term increases or decreases in deaths, and a fixed effect for each week or month." (The Economist, 2021)

However, I believe both models might lack the complexity to successfully capture the underlying pattern of the mortality data, which might not necessarily be "linear" in nature. I hypothesize that general mortality data could be further decomposed into both a seasonal/cyclical and trend component, suggesting a time-series model might be a better choice (Hyndman, 2018). Furthermore, the general death rate over the years might be increasing (non-stationary) as well, which is unrelated to covid-related factors but is rather attributed to trends such as the increasing elderly population globally (Cheng et al, 2020) or due to factors such as "drug overdoses, alcohol, suicides, and cardiometabolic condition" for the younger and middle-aged population (National Academies of Sciences, 2021).

Hence, a Bayesian model using INLA or integrated nested Laplace approximation (Rue et al, 2009) will be fitted for each country and will be trained only using the data with dates before "2020-03-01". This will act as our historical baseline which will used to calculate the excess deaths during the pandemic. Predicted values along will their corresponding 95% credible intervals (CI) will also be generated. Finally, the actual weekly deaths will be overlayed as well.

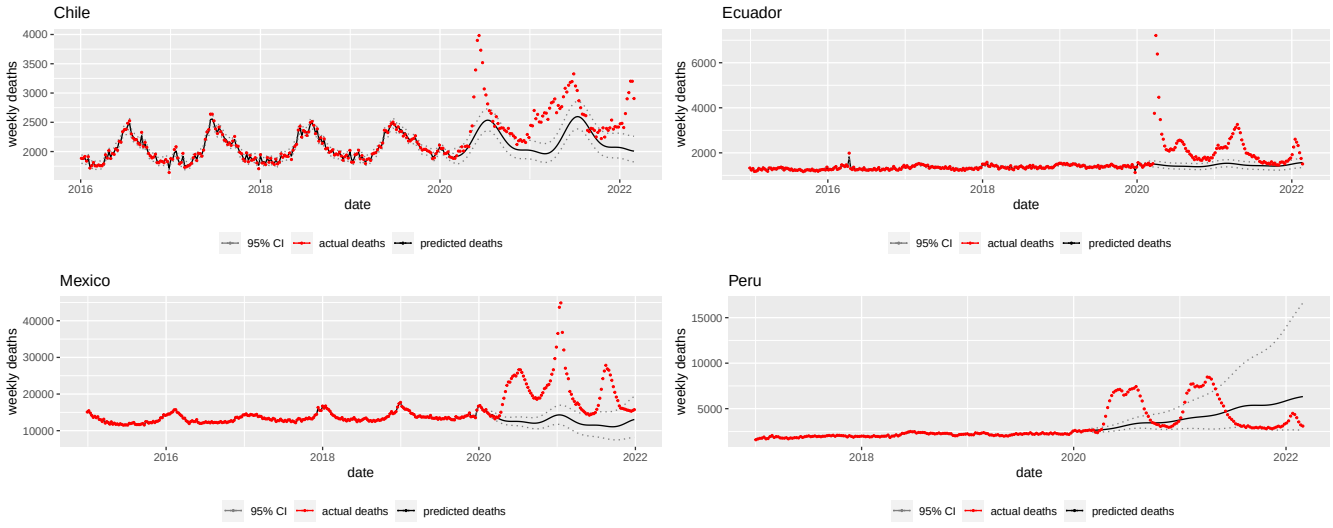

Figure 2: Plot of predicted vs actual deaths

Sampling Predicted Deaths from the Posterior Distributions

Next, we will sample 10 simulations from the corresponding posterior distributions for each country along with the actual deaths after accounting for the covid-19 pandemic.

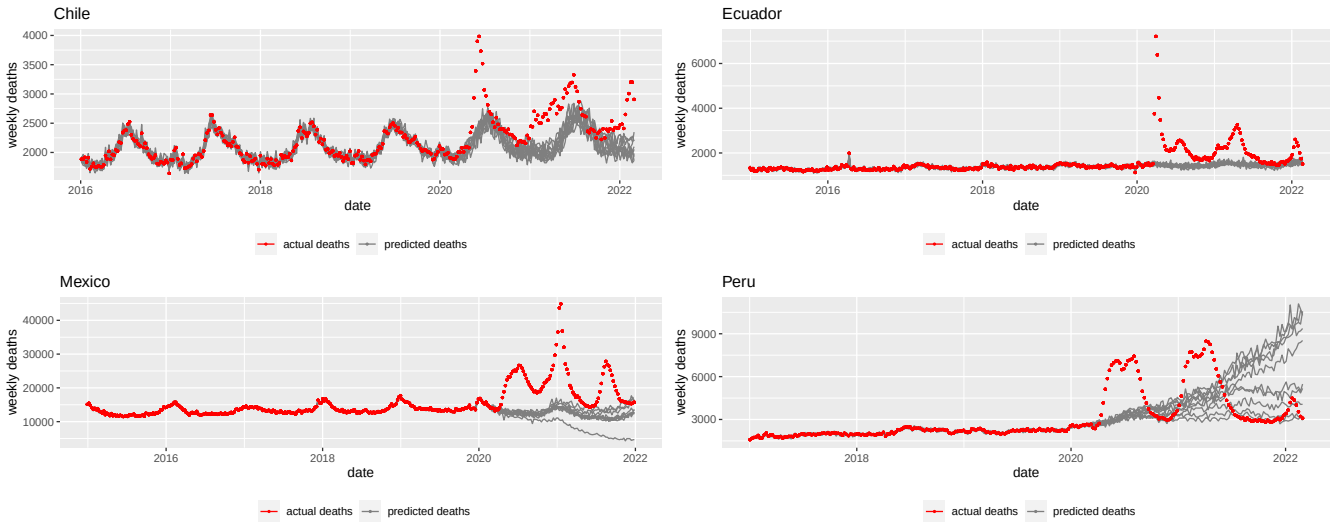

Figure 3: Plot of sampled predicted vs actual deaths from 10 simulations

Excess Deaths vs Official COVID-19 Deaths

Finally, we will now calculate the excess deaths by subtracting the estimated predicted deaths from the actual total deaths for each simulation. We will also overlay the official reported covid-19 deaths for each country.

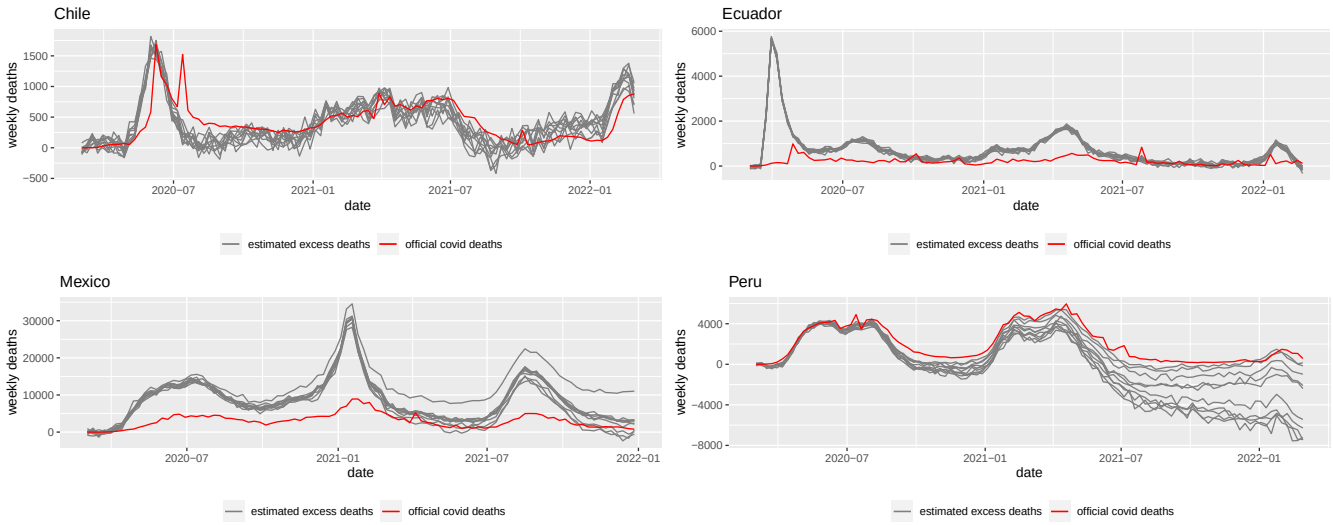

Figure 4: Plot of sampled predicted vs official covid-19 deaths

Discussion

Based on Figure 2, we can see that INLA was able to successfully capture the trend and the seasonal patterns of the pre-covid death trends and was able to generate reasonable predictions (posterior medians) using only the pre-covid data going into the pandemic. The seasonality is mainly generated by the linear combination of the 6-months and 12-months sines and cos-sines cycles specified by the parameters of `sin12 + sin6 + cos12 + cos6`. The priors chosen for both models (random walk 2 and independent/iid) are both `pc.prec` or Precision Penalized Complexity prior, which is able to "penalize departure from the base model" (Virgilio, 2020). These priors are also "invariant to reparameterisations, have a natural connection to Jeffreys' priors, are designed to support Occam's razor" (Daniel et al, 2017). The standard deviation of the prior is set to be 0.01 with a median quantile of 0.5, which denotes that the

slope of the death rate can change by about 0.01 from one week to the next. Note that the scale of this prior is only an estimate as we do not know the exact death rate for each country.

We can also observe that the 95% C.I. gets wider the further away our prediction gets as we are now less confident about the predictions. This is especially the case for Peru (Figure 2), as INLA has suggested a rather steep and increasing trend for death rates with a very high uncertainty as well.

On Figure 3, we generated 10 samples from the posterior distributions instead of just plotting the posterior medians and quantiles. We noticed that the samples are pretty much consistent for Chile and Ecuador, but experienced some variations for Mexico and Peru, which is also consistent to Figure 2. This suggests that perhaps we need to choose a more sensitive prior for these two countries or tune other hyperparameters to adjust for a better model fit.

Finally on Figure 4, we have subtracted the predicted deaths from total deaths to generate the "excess deaths" for each simulation and also plotted the official covid-19 death counts from each country. Theoretically, we should be able to observe that the official reported covid-19 death counts to be the same as the excess deaths we calculated, which would mean that the majority of the cause of deaths during the pandemic period have all been caused by and attributed to the virus. However, this happens to be only the case of Chile. For Ecuador and Mexico, we can observe our estimated excess deaths have all been higher than the official reported covid death counts. Finally for Peru, the estimated excess death counts are actually lower than that of the official covid death count, which suggests that our historical baseline model is under-fitting leading to our excess deaths counts to be extremely noisy with a large variance.

Conclusion

In conclusion, we don't have much evidence that the Latin America region had a much lower official reported covid-19 deaths compared to the estimated excess deaths. Out of the four chosen countries (Chile, Ecuador, Mexico, and Peru), there is only evidence that Ecuador and Mexico has estimated excess deaths to be greater than the reported covid-19 deaths. Chile, on the other hand, has excess deaths to be more or less the same as its reported covid deaths. These results are also consistent with the Economist's article (The Economist, 2021) which uses a linear fixed effects model. Finally, our model has not fitted the Peru data well, resulting in an inconclusive result.

Some potential limitations include the four countries chosen randomly as a representation for the Latin American region which could introduce some selection biases. Furthermore, one could also argue that each country should really be treated as independent samples and the regional fixed effect might not exist at all whereas factors such as GDP per capita, health infrastructure scores, or corruption indices might be more useful indicators rather than just subsetting by region. This could also be an area for exploration as next steps.

Finally, future recommendations could include a more clever and stronger choice of informative priors (based on domain knowledge or prior researches) or the better tuning of the other hyper-parameters which could avoid underfitting and the highly volatile results of the Peru dataset. Other models and methods such as GAMs or time series models such as ARIMA or exponential smoothing could also be explored for comparison purposes. It would also be worthwhile to compare the results of calculated excess deaths vs reported covid deaths for OECD countries such as Canada as a baseline for comparison.

References

- CDC, Centers for Disease Control and Prevention (2022) *Excess Deaths Associated with COVID-19*. Retrieved: https://www.cdc.gov/nchs/nvss/vsrr/covid19/excess_deaths.htm#references
- Cheng X, Yang Y, Schwebel DC, Liu Z, Li L, Cheng P, et al. (2020) *Population ageing and mortality during 1990–2017: A global decomposition analysis*. PLoS Med 17(6): e1003138. https://doi.org/10.1371/journal.pmed.1003138
- Daniel Simpson, Håvard Rue, Andrea Riebler, Thiago G. Martins, Sigrunn H. Sørbye. "Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors." Statistical Science, 32(1) 1-28 February 2017. https://doi.org/10.1214/16-STS576
- Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice.* (2nd ed.) OTexts. https://otexts.org/fpp2/
- National Academies of Sciences, Engineering, and Medicine. 2021. *High and Rising Mortality Rates Among Working-Age Adults.* Washington, DC: The National Academies Press.https://doi.org/10.17226/25976.
- H. Rue, S. Martino, and N. Chopin. *Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). Journal of the Royal Statistical Society, Series B, 71(2):319{392, 2009.*
- The Economist (2021) *Tracking covid-19 excess deaths across countries.* Retrieved: https://www.economist.com/graphic-detail/coronavirus-excess-deaths-tracker
- Gómez-Rubio, Virgilio (2020). Bayesian Inference with INLA. Chapman & Hall/CRC Press. Boca Raton, FL. Retrieved: https://becarioprecario.bitbucket.io/inla-gitbook/index.html
- WHO (2020) *WHO Timeline - COVID-19.* Retrieved: https://www.who.int/news/item/27-04-2020-who-timeline---covid-19

Appendix

```
# importing libraries
library(tidyverse)
library(INLA)
library(Pmisc)

chile <- read.csv("https://raw.githubusercontent.com/TheEconomist/covid-19-excess-deaths-tracker/master/ou
ecuador <- read.csv("https://raw.githubusercontent.com/TheEconomist/covid-19-excess-deaths-tracker/master/
peru <- read.csv("https://raw.githubusercontent.com/TheEconomist/covid-19-excess-deaths-tracker/master/out
mexico <- read.csv("https://raw.githubusercontent.com/TheEconomist/covid-19-excess-deaths-tracker/master/o

chile_plt <- chile %>% ggplot(aes(x= as.Date(start_date)) ) +
          geom_point(aes(y= total_deaths, color="total_deaths")) +
          geom_line(aes(y= total_deaths, color="total_deaths")) +
          geom_point(aes(y= covid_deaths, color="official covid_deaths")) +
          geom_line(aes(y= covid_deaths, color="official covid_deaths")) +
          xlab("date") + ylab("weekly deaths") + labs(title="Chile", color="") +
          theme(legend.position="bottom")

ecuador_plt <- ecuador %>% ggplot(aes(x= as.Date(start_date)) ) +
          geom_point(aes(y= total_deaths, color="total deaths")) +
          geom_line(aes(y= total_deaths, color="total deaths")) +
          geom_point(aes(y= covid_deaths, color="official covid deaths")) +
          geom_line(aes(y= covid_deaths, color="official covid deaths")) +
          xlab("date") + ylab("weekly deaths") + labs(title="Ecuador", color="") +
          theme(legend.position="bottom")

mexico_plt <- mexico %>% ggplot(aes(x= as.Date(start_date)) ) +
          geom_point(aes(y= total_deaths, color="total deaths")) +
          geom_line(aes(y= total_deaths, color="total deaths")) +
          geom_point(aes(y= covid_deaths, color="official covid deaths")) +
          geom_line(aes(y= covid_deaths, color="official covid deaths")) +
          xlab("date") + ylab("weekly deaths") + labs(title="Mexico", color="")+
          theme(legend.position="bottom")

peru_plt <- peru %>% ggplot(aes(x= as.Date(start_date)) ) +
          geom_point(aes(y= total_deaths, color="total deaths")) +
          geom_line(aes(y= total_deaths, color="total deaths")) +
          geom_point(aes(y= covid_deaths, color="official covid deaths")) +
          geom_line(aes(y= covid_deaths, color="official covid deaths")) +
          xlab("date") + ylab("weekly deaths") + labs(title="Peru", color="")+
          theme(legend.position="bottom")

cowplot::plot_grid(chile_plt, ecuador_plt, mexico_plt, peru_plt)

generate_predicted_deaths <- function(x){
  x$time <- as.Date(x$start_date)
  x$dead <- x$total_deaths

  dateCutoff = as.Date('2020/3/1')
  xPreCovid = x[x$time < dateCutoff, ]
  xPostCovid = x[x$time >= dateCutoff, ]
  toForecast = expand.grid(time = unique(xPostCovid$time), dead = NA)

  xForInla = rbind(xPreCovid[,colnames(toForecast)], toForecast)
  xForInla= xForInla[order(xForInla$time), ]

  xForInla$timeNumeric = as.numeric(xForInla$time)
```

```
  xForInla$timeForInla = (xForInla$timeNumeric)/365.25
  xForInla$timeIid = xForInla$timeNumeric
  xForInla$sin12 = sin(2*pi*xForInla$timeNumeric/365.25)
  xForInla$sin6 = sin(2*pi*xForInla$timeNumeric*2/365.25)
  xForInla$cos12 = cos(2*pi*xForInla$timeNumeric/365.25)
  xForInla$cos6 = cos(2*pi*xForInla$timeNumeric*2/365.25)

  res = inla(dead ~ sin12 + sin6 + cos12 + cos6 +
      f(timeIid, model='iid', prior='pc.prec', param= c(0.01, 0.5)) +
      f(timeForInla, model = 'rw2', scale.model=FALSE,
        prior='pc.prec', param= c(0.01, 0.5)),
    data=xForInla,
    control.predictor = list(compute=TRUE, link=1),
    control.compute = list(config=TRUE),
  #  control.inla = list(fast=FALSE, strategy='laplace'),
    family='poisson')
  qCols = paste0(c('0.5', '0.025', '0.975'), 'quant')
  res_bind = rbind(res$summary.fixed[,qCols], Pmisc::priorPostSd(res)$summary[,qCols])
  res_bind_inla = cbind(xForInla, res$summary.fitted.values[,qCols],
                        total_deaths=x$total_deaths, covid_deaths = x$covid_deaths)
  return(list(res, res_bind_inla, res_bind, xForInla))
}

chile_predicted_deaths = generate_predicted_deaths(chile)
ecuador_predicted_deaths = generate_predicted_deaths(ecuador)
mexico_predicted_deaths = generate_predicted_deaths(mexico)
peru_predicted_deaths = generate_predicted_deaths(peru)


chile_inla_plt = chile_predicted_deaths[[2]] %>%
  ggplot(aes(x=time)) +
  geom_line(aes(y=`0.5quant`, color='predicted deaths')) +
  geom_line(aes(y=`0.025quant`, color='95% CI'), linetype="dotted") +
  geom_line(aes(y=`0.975quant`, color='95% CI'), linetype="dotted") +
  geom_point(aes(y=total_deaths, color='actual deaths'), size=0.5) +
  scale_color_manual(values=c("grey50", "red", "black"))+
  xlab("date") + ylab("weekly deaths") + labs(title="Chile", color="")+
  theme(legend.position="bottom")

ecuador_inla_plt = ecuador_predicted_deaths[[2]] %>%
  ggplot(aes(x=time)) +
  geom_line(aes(y=`0.5quant`, color='predicted deaths')) +
  geom_line(aes(y=`0.025quant`, color='95% CI'), linetype="dotted") +
  geom_line(aes(y=`0.975quant`, color='95% CI'), linetype="dotted") +
  geom_point(aes(y=total_deaths, color='actual deaths'), size=0.5) +
  scale_color_manual(values=c("grey50", "red", "black"))+
  xlab("date") + ylab("weekly deaths") + labs(title="Ecuador", color="")+
  theme(legend.position="bottom")

mexico_inla_plt = mexico_predicted_deaths[[2]] %>%
  ggplot(aes(x=time)) +
  geom_line(aes(y=`0.5quant`, color='predicted deaths')) +
  geom_line(aes(y=`0.025quant`, color='95% CI'), linetype="dotted") +
  geom_line(aes(y=`0.975quant`, color='95% CI'), linetype="dotted") +
  geom_point(aes(y=total_deaths, color='actual deaths'), size=0.5) +
  scale_color_manual(values=c("grey50", "red", "black"))+
  xlab("date") + ylab("weekly deaths") + labs(title="Mexico", color="")+
  theme(legend.position="bottom")
```

```
peru_inla_plt = peru_predicted_deaths[[2]] %>%
  ggplot(aes(x=time)) +
  geom_line(aes(y=`0.5quant`, color='predicted deaths')) +
  geom_line(aes(y=`0.025quant`, color='95% CI'), linetype="dotted") +
  geom_line(aes(y=`0.975quant`, color='95% CI'), linetype="dotted") +
  geom_point(aes(y=total_deaths, color='actual deaths'), size=0.5) +
  scale_color_manual(values=c("grey50", "red", "black"))+
  xlab("date") + ylab("weekly deaths") + labs(title="Peru", color="")+
  theme(legend.position="bottom")
cowplot::plot_grid(chile_inla_plt, ecuador_inla_plt, mexico_inla_plt, peru_inla_plt)

sample_deaths_from_posterior <- function(res, res_bind_inla, n_sim=10){
  sampleList = INLA::inla.posterior.sample(n_sim, res, selection = list(Predictor=0))
  sampleIntensity = exp(do.call(cbind,
    Biobase::subListExtract(sampleList, 'latent')))
  sampleDeaths = matrix(rpois(length(sampleIntensity), sampleIntensity),
    nrow(sampleIntensity), ncol(sampleIntensity))
  sample_deaths_df = cbind(sampleDeaths, res_bind_inla)
  sample_deaths_df_wide = sample_deaths_df %>%
    pivot_longer(1:n_sim, names_to = "simulation", values_to = "predicted_deaths")
  return(sample_deaths_df_wide)
}

sample_deaths_chile = sample_deaths_from_posterior(chile_predicted_deaths[[1]], chile_predicted_deaths[[2]]
sample_deaths_ecuador = sample_deaths_from_posterior(ecuador_predicted_deaths[[1]], ecuador_predicted_deat
sample_deaths_mexico = sample_deaths_from_posterior(mexico_predicted_deaths[[1]], mexico_predicted_deaths[
sample_deaths_peru = sample_deaths_from_posterior(peru_predicted_deaths[[1]], peru_predicted_deaths[[2]])

chile_posterior_plt = sample_deaths_chile %>%
  ggplot(aes(x=time)) +
  geom_line(aes(y=predicted_deaths, group=simulation, color="predicted deaths"))  +
  geom_point(aes(y=total_deaths, color='actual deaths'), size=0.5) +
  scale_color_manual(values=c("red", "grey50"))+
  xlab("date") + ylab("weekly deaths") + labs(title="Chile", color="")+
  theme(legend.position="bottom")

ecuador_posterior_plt = sample_deaths_ecuador %>%
  ggplot(aes(x=time)) +
  geom_line(aes(y=predicted_deaths, group=simulation, color="predicted deaths"))  +
  geom_point(aes(y=total_deaths, color='actual deaths'), size=0.5) +
  scale_color_manual(values=c("red", "grey50"))+
  xlab("date") + ylab("weekly deaths") + labs(title="Ecuador", color="")+
  theme(legend.position="bottom")

mexico_posterior_plt = sample_deaths_mexico %>%
  ggplot(aes(x=time)) +
  geom_line(aes(y=predicted_deaths, group=simulation, color="predicted deaths"))  +
  geom_point(aes(y=total_deaths, color='actual deaths'), size=0.5) +
  scale_color_manual(values=c("red", "grey50"))+
  xlab("date") + ylab("weekly deaths") + labs(title="Mexico", color="")+
  theme(legend.position="bottom")

peru_posterior_plt = sample_deaths_peru %>%
  ggplot(aes(x=time)) +
  geom_line(aes(y=predicted_deaths, group=simulation, color="predicted deaths"))  +
  geom_point(aes(y=total_deaths, color='actual deaths'), size=0.5) +
  scale_color_manual(values=c("red", "grey50"))+
  xlab("date") + ylab("weekly deaths") + labs(title="Peru", color="")+
```

```
  theme(legend.position="bottom")
cowplot::plot_grid(chile_posterior_plt, ecuador_posterior_plt, mexico_posterior_plt, peru_posterior_plt)

generate_excess_deaths_from_posterior <- function(sample_deaths){
  excess_deaths = sample_deaths %>% filter(time > as.Date('2020/3/1')) %>%
  mutate(excess_deaths=total_deaths - predicted_deaths)
  return(excess_deaths)
}

excess_deaths_chile = generate_excess_deaths_from_posterior(sample_deaths_chile)
excess_deaths_ecuador = generate_excess_deaths_from_posterior(sample_deaths_ecuador)
excess_deaths_mexico = generate_excess_deaths_from_posterior(sample_deaths_mexico)
excess_deaths_peru = generate_excess_deaths_from_posterior(sample_deaths_peru)

excess_deaths_chile_plt = excess_deaths_chile %>%
  ggplot(aes(x=time)) +
  geom_line(aes(y=excess_deaths, group=simulation, color="estimated excess deaths"))  +
  geom_line(aes(y=covid_deaths, color="official covid deaths"))  +
  scale_color_manual(values=c("grey50", "red"))+
  xlab("date") + ylab("weekly deaths") + labs(title="Chile", color="")+
  theme(legend.position="bottom")

excess_deaths_ecuador_plt = excess_deaths_ecuador %>%
  ggplot(aes(x=time)) +
  geom_line(aes(y=excess_deaths, group=simulation, color="estimated excess deaths"))  +
  geom_line(aes(y=covid_deaths, color="official covid deaths"))  +
  scale_color_manual(values=c("grey50", "red"))+
  xlab("date") + ylab("weekly deaths") + labs(title="Ecuador", color="")+
  theme(legend.position="bottom")

excess_deaths_mexico_plt = excess_deaths_mexico %>%
  ggplot(aes(x=time)) +
  geom_line(aes(y=excess_deaths, group=simulation, color="estimated excess deaths"))  +
  geom_line(aes(y=covid_deaths, color="official covid deaths"))  +
  scale_color_manual(values=c("grey50", "red"))+
  xlab("date") + ylab("weekly deaths") + labs(title="Mexico", color="")+
  theme(legend.position="bottom")

excess_deaths_peru_plt = excess_deaths_peru %>%
  ggplot(aes(x=time)) +
  geom_line(aes(y=excess_deaths, group=simulation, color="estimated excess deaths"))  +
  geom_line(aes(y=covid_deaths, color="official covid deaths"))  +
  scale_color_manual(values=c("grey50", "red"))+
  xlab("date") + ylab("weekly deaths") + labs(title="Peru", color="")+
  theme(legend.position="bottom")

cowplot::plot_grid(excess_deaths_chile_plt, excess_deaths_ecuador_plt,
                   excess_deaths_mexico_plt, excess_deaths_peru_plt)
```