Estimating a mortality baseline from limited data:

The Spanish Influenza in Madrid

Laura Cilek, Gerardo Chowell, Diego Ramiro Fariñas *

Abstract

Quantifying the strength and timing of epidemics requires a reasonable expectation of seasonal baseline mortality. However, in historical and some subgroups of contemporary populations, it is difficult to find this information at weekly or daily intervals. Using several data sources of varying temporal aggregations (individual death records, weekly, and monthly aggregated death counts primarily related to the Spanish flu), we explore traditional methods of baseline and excess mortality calculations as well as some adaptations. We then propose additional ways to calculate and refine the seasonality in yearly mortality baseline using both Metropolis-hastings MCMC and interpolation. We present baseline and excess mortality estimates from all models and conclude with a discussion of the merits and practicality of each method.

1 Background

Contemporary estimations claim the influenza pandemic events between 1918 and 1921, the so-called "Spanish" flu, account for the deaths of more than 50 million people throughout the world [1]. The series of successive influenza virus outbreaks gripped the world beginning in early 1918, however results of various phylogenetic and molecular-clock analyses indicate the initial circulation of the virus from avian or swine and other mammal species to humans may have occurred as early as 1911 [2, 3]. Moreover, the symptoms and mortality patterns associated with this particular flu pandemic are particularly unique. For example, young-adults often exhibited the highest excess mortality rates, in contrast to seasonal influenza epidemics, which primarily affect the very young and elderly [4, 5]. Also, death from the virus was often labeled a result of respiratory illnesses, such as pneumonia or bronchitis rather than directly from "influenza." For this reason, the study of all respiratory related mortality is desirable by researchers when such data exists. The difference in influenza-related mortality by year and flu sub-type is often examined by using age specific mortality rates as both an absolute value and as a ratio of excess mortality rates between vulnerable (young and old) and lesser affected populations [5]. Ergo, a reliable estimation of baseline overall and baseline mortality is essential.

The first major world-wide influenza pandemic during the modern age of transportation was that of the Russian flu in 1889-1890, spreading to every corner of Europe in only 6 weeks and throughout the world as the winter progressed [6, 7]. Due to this shift in the speed of which influenza pandemics spread and their increase in scope, the study of pandemic timing has also become a topic of interest. However, within the context of the 1918 Spanish flu, quantitative research about the specific timing of each wave is difficult, as at smaller geographic levels, most longitudinal mortality data is aggregated. Thus, while some areas collected daily information about the flu as it progressed, it is difficult to ascertain how its specific timing may have differed from baseline flares of seasonal influenza. Smaller intervals of baseline mortality data in these areas may provide a better indication of the specific time-frame from which the virus spread not only into Spain, but

*laura.cilek@cchs.csic.es

Center for Humanities and Social Sciences, Spanish National Research Council, Spain; Division of Epidemiology & Biostatistics, School of Public Health, Georgia State University

through the rest the world. While the location of the first human infection remains unclear, the virus likely moved to Spain via Spanish and Portuguese labor migrants returning to the Iberia peninsula from Southern France during the World War I [8].

In fact, the actual timeline and progression leading to the virus's emergence is debated, though likely, the H1N1 strains responsible for the Spanish flu are related to those which caused the "Russian" pandemic influenza events at the end of the 19th century and may have been present in both swine and humans more than 5 years before the first waves in 1918 [2]. While strains of the H1N1 virus continue to circulate in the form of seasonal influenza viruses, biological remnants of the particularly deadly 1918 strains are still found in avian species via the presence of specific encoded proteins [9]. In this manner, continued research into the timing of the Spanish flu, as well as its health and mortality impacts on different populations is essential to understanding the potential effects that a virulent influenza strain could have on the global population today. To quantify these impacts at a refined level, a reasonable and reliable mortality baseline (at small time intervals of time) in non-epidemic years from which excess mortality may be determined is vital.

Moreover, because waves often affected places more than once and at different times of the year, calculations of excess mortality depend heavily on the estimated underlying level of mortality. For example, in the 1918 herald wave in Madrid, some weeks had excess levels well above the seasonal level, but others were barely–or not at all–above what would have been a normal seasonal peak in the fall or winter [10].

More specifically, the effect of a reliable mortality baseline can shape the debate surrounding the specific topics of interest in the flu. In the wealth of information and research on Spanish and other influenza outbreaks, much examination has focused on how age-specific mortality differs from seasonal outbreaks. Often, analyses featuring little-to-no baseline mortality information show high rates of excess mortality for young adults, which in the past led to some discussion of a "w"-shaped curve of excess mortality. However, other research calls into question this large peak; while the standard mortality ratio (SMR) continues to follow the shape of an inverted "V"–that is, the probability of dying in the epidemic period relative to a non-epidemic period is higher in the young adult ages–, the actual amount of excess mortality peaks only at the lowest and highest ages. While this is not to say only one shape of age-specific mortality curve is possible, the extent to which a baseline is calculated and implemented can have a large impact on results. Given the ongoing debate about total mortality related to the Spanish flu and its associated pattern, we found it fitting, one hundred (and one) years following the pandemic, to reexamine traditional baseline estimation methods and the application of an interpolation technique to refine aggregate data for timing analyses on a smaller-scale.

As such, the rest of this paper is structured as follows. First, we briefly outline the two sets of data used in our calculations, whose peculiarities inspired this examination. Next, we quickly introduce the "standard" Serfling Regression model used to estimate seasonal mortality patterns, then provide some potential adaptations to the method to ensure a better fit and quantify uncertainty when faced with with limited pre-epidemic mortality data. We then propose using a Metropolis-hastings MCMC approach to estimate the distribution of deaths throughout the year by optimizing the Serfling parameters. Finally, we switch gears to explore how monthly-aggregated data may be interpolated in order to provide a "best-guess" of weekly mortality patterns throughout the year. After providing the results of these methods when applied to our Madrid data, we calculate excess mortality from each method and discuss the similarities and differences of each completed baseline and the applications of these methods in the future.

2 Data

The Madrid Civil Register of Deaths provides excellent, detailed information about the deceased, including age, sex, civil status, location, data, and cause of death [11]. However, the death records are only available for the years 1917-1922; this is enough to cover the full period of the Spanish-flu related outbreaks in Madrid but only one year of data from which to calculated a baseline.

Other mortality data sources exist, including the Boletín Mensual Estadístico Sanitario-Demográfico, which provides monthly mortality information by selected causes for the entire Madrid province (both the city proper and surrounding rural area in the administrative region) [12]. While the data does not provide individual level information, these monthly death counts published by the Ministry of Government in Spain for the years 1915-1919 encapsulate the first three waves of influenza in Spain. In addition to total deaths, this data also provides counts for several causes; thus, we can create and compare our baseline estimations for both overall and influenza-specific mortality in the city and province of Madrid.

Finally, we gain population information city the from the Yearly Statistical Books of Madrid. At the time of the influenza outbreaks, the evolution of the city's population was recorded via a quasi-register based system, of which a census-equivalent was taken every five years. Published on an annual basis, the volumes we use provide population by district both for the city and the region of Madrid given reported and estimated changes. Because our mortality data describes two different geographic areas and thus, different population numbers, we calibrate our mortality estimates according to the at risk population of both the city and the region it encompassed.

3 Methods

Several methods exist to infer a seasonal mortality baseline using the single year of 1917 data, but the monthly counts from 1915-1917 also allow a way to see if mortality in 1917 is significantly different from prior years, and thus, using only this year mis-estimates overall and cause specific excess morality during the influenza epidemic waves. We first briefly review the Serfling regression model which incorporates parameters for time and seasonal trends in mortality. Then, we outline some additional modifications to the traditional model in order to better fit Madrid's non-traditional mortality pattern and overcome our own data limitations before explaining how the application of a Metropolis-Hastings Markov Chain Monte Carlo model can also estimate the mortality baseline and its upper bound. Lastly, we explore how interpolating monthly-aggregated data can provide additional insight into mortality patterns at the weekly level.

While for epidemiologists, the "gold standard" of mortality baseline estimation remains the noted Serfling regression model and "current model" for count data, additional methods exists to calculate changing mortality patterns across the year. Several recent studies have begun to estimate seasonal mortality through the use of other methods, such as with Poisson counts or cubic splines (i.e. [13, 14, 15, 16]), which also allows for baseline changes during the observation period, such for the introduction of a vaccine. We do not specifically review these methods in this paper. We also expand this paper to discuss the use of Markov Cs to model seasonal variations in mortality and determine the presence of epidemic waves through examining the likelihood that an observed point of weekly mortality occurs based on the parameterized baseline distribution [17].

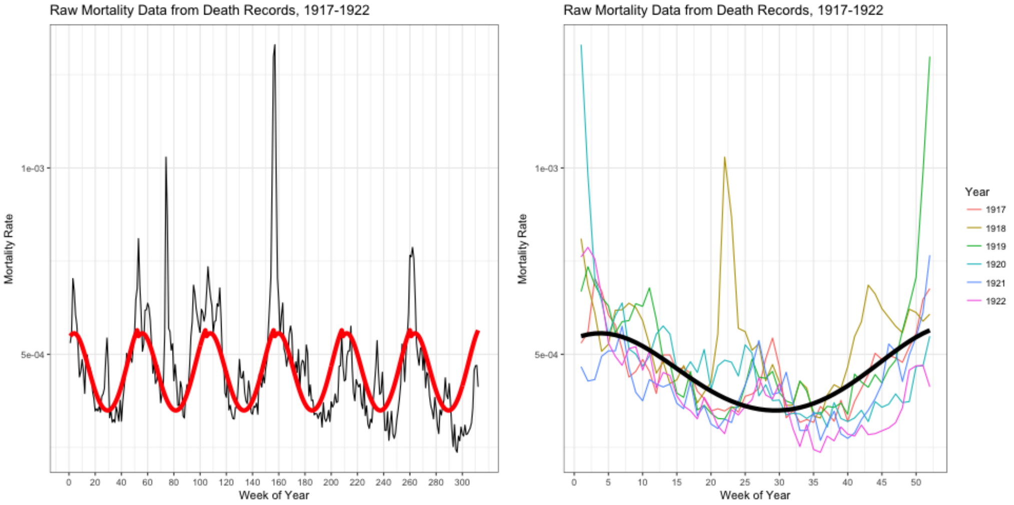
3.1 Serfling Regression

Often, in an effort to quantify seasonal mortality and smooth a baseline across several years of data, researcher employ a Serfling cyclical regression model [18], which provides an average mortality level incorporating time and seasonal peaks through cosine and sine parameters

$$\frac{Deaths_{x_t}}{Population_{x_t}} = u + \alpha * (t) + \beta * sin(\frac{2\pi}{52.17} * t) + \beta * cos(\frac{2\pi}{52.17} * t)$$

After calculation of the baseline, observed deaths during the period of analysis above the upper 95% confidence interval bound of the expected values are defined as an epidemic period, from which excess mortality is calculated. However, the traditional Serfling baseline approach requires several years of pre-epidemic mortality data–normally at least three years–due to the fluidity of seasonal mortality. Additionally, shorter time periods could result in a loss of continuity between the beginning and end of the year. In the case of our 1917 data, this is displayed in Figure 1, where due to the change in the timing of the seasonal peak between the winter of 1916-1917 and 1917-1918, the baseline suggests there is significantly higher mortality in week fifty-two of the year than in week one.

Figure 1: The black line shows the real weekly mortality rates from 1917-1922, while the solid red line displays the predicted mortality values based on a simple Serfling regression (left). Lines show the real weekly mortality rates from 1917-1922, while the solid black line displays the predicted mortality value based on a simple Serfling regression (right).



Other issues can arise when mortality does not follow the parameter patterns of a traditional Serfling model. For example, Madrid experiences a small but noticeable summer mortality peak which the basic Serfling model does not account for (see Figure 1). However, this may be rectified through the addition of more time and seasonality parameters to better fit the mortality pattern [19]. While much statistical research avoids overfitting the data, if the mortality data used to create the baseline is trusted to be a correct representation of actual mortality, and that actual mortality is assumed to follow a normal pattern in the area of study, than the regression model with additional parameters used to create the baseline can be better in providing a realistic expectation of normal mortality than the simple approach. In the case of the 1917 mortality records, the visible mortality peak in the summer can be represented through the addition of parameters such that the baseline equation is written as

$$\begin{aligned}\frac{Deaths_{x_t}}{Population_{x_t}} = \quad & u + \alpha * (t) + \alpha * (\frac{100}{t})^2 + \\ & \beta * sin(\frac{2\pi}{52.17} * t) + \beta * sin(\frac{4\pi}{52.17} * t) + \\ & \beta * sin(\frac{8\pi}{52.17} * t) + \gamma * cos(\frac{2\pi}{52.17} * t) + \\ & \gamma * cos(\frac{4\pi}{52.17} * t) + \gamma * cos(\frac{8\pi}{52.17} * t)\end{aligned}$$

The added coefficients in the model account for both linear and non-linear time (α) and seasonal (β & γ) variations in normal mortality activity that create the oscillations present in the data.

3.1.1 Serfling Regression with Parametric Bootstrapping

While visual analysis reveals similarity in our yearly pre- and post-Spanish flu epidemic mortality data, we only use mortality information from before the epidemic to construct the baseline, as post-outbreak mortality may be influenced by the increased number of deaths during the flu onslaught. We felt that using only the one year of available death records (1917) may ultimately provide an incorrect estimation of the baseline, as it forced the assumption that mortality in 1917 followed a normal pattern at all ages. Thus, we used parametric bootstrapping to generate some uncertainty and account for the potential variability of the 1917 data from typical mortality patterns [20].

We first simulated data before fitting the above regression model, accounting for the possibility of aforementioned fluctuation in the annual timing of winter and summer mortality peaks. A single set of mortality points from which the bootstrapped points were estimated consisted of six consecutive iterations of total weekly deaths in 1917, to mimic the six years of mortality data used in our analysis.[1] For each of these 312 week sets of weekly death counts, we simulated a number of expected deaths, assuming a Poisson count distribution. Our Poisson estimations assumed the mean and variance of a week were equal the observed total number of deaths in that week of 1917.

From each of simulated six-year datasets, α, β, and γ parameters are estimated according to the modified seasonal regression model above. We calculate our five year baseline from the mean values of the coefficients from the models and compute the upper baseline from the upper quartile value of the 95% confidence interval of coefficients. As in previous literature, we define weeks with mortality above the upper baseline as "epidemic" [19, 21, 22, 23, 14].

3.2 Metropolis-Hastings Markov Chain Monte Carlo

In the example above, numbers of weekly deaths are simulated according to parameters of their observed values in 1917 to account for the lack of multi-year baseline data. From these simulated vectors of weekly mortality data, we create a baseline and upper 95% certainty epidemic threshold according with Serfling regression. However, other ways exist to estimate these baseline parameters that may better take into account how each parameter affects mortality at different times. Here, we explore an application of the Metropolis-Hastings algorithm via Monte Carlo Markov Chains to recreate the distribution of deaths throughout the year [24].

The application of these methods are particularly appropriate when considering the inter-year seasonality in mortality. As modeled in Serfling regression, mortality throughout the year depends heavily on parameters both the time and the extent to which mortality is seasonal. Yet, rather than maximizing the likelihood of the regression function, the Metropolis-Hastings algorithm approaches the vector of weekly deaths as a probability

[1] We consider each year to have 52 weeks, and calculate the "total" deaths in the 52nd week of the year as the $\frac{7}{8}$ or $\frac{7}{9}$ of deaths in the final week and associated excess day(s) of the year (1920 was a leap year).

distribution of deaths throughout the year. That is to say, the distribution can be expressed mathematically as the integral of the parameters from time 0 to the end of the year, or, in the case of Madrid, as:

$$\int_0^{52} t + (\frac{100}{t})^2 + sin(\frac{2\pi t}{52.14}) + sin(\frac{4\pi t}{52.14}) + sin(\frac{8\pi t}{52.14}) + cos(\frac{2\pi t}{52.14}) + cos(\frac{4\pi t}{52.14}) + cos(\frac{8\pi t}{52.14}) + \varepsilon$$

By sampling the distribution parameters within the state space, defined as the space of one year, the algorithm converges to a specific distribution by comparing the log-likelihoods of the functions for current and randomly selected proposed values [25]. After the distribution and its parameters have been optimized, the 95% probability of the distribution can be calculated from the parameter quantiles of accepted values during the random walk. As in the other methods, weeks in which total deaths are higher than this 95% threshold are assumed to be "epidemic."

Thinking of the yearly baseline mortality pattern as a distribution is quite useful. For example, we can consider that the mortality rate in a year is relatively constant–i.e., from year to year, the total number of deaths in a stable population (akin to the total density of the distribution) will not change. However, the distribution of these deaths changes throughout the year, resulting in a non-normal distribution that generally has two modes at the beginning and end of the year. In the case of Madrid, the deaths appear to have a distinct tri-modal distribution with peaks at the beginning (winter), middle (small summer peak), and end of the year. This general distribution, for which the Metropolis-hastings algorithm seeks to define parameters, follows the same general pattern from year to year. While the peaks are defined mostly according to the observed points in 1917, the uncertainty allows for the possibility that the exact timing of these peaks may vary from year to year, as does the onset of seasonal diseases and weather patterns.

3.2.1 Implementation

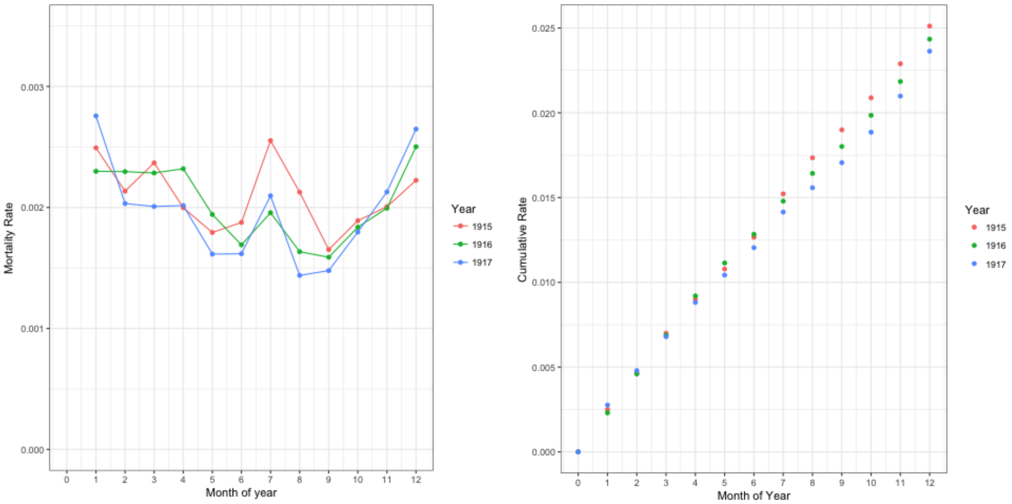
To implement the Metropolis-Hastings procedure, the target distribution is defined according to the adapted Serfling regression parameters explored in section 3.1, such that 9 parameters are proposed and tested with each iteration. This includes a value for the intercept, two for the time coefficients, and three-each for sine and cosine coefficients. The initial proposed distribution gives a value of average mortality to the intercept and 9 to the other seasonal and time-varying parameters such that the mortality distribution is represented as a straight line, equal throughout the state space (year). A Markov Chain samples new parameters selected from a random uniform distribution, then accepts or rejects these new parameters based on the change in log-likelihood of the function [25]. The log-likelihood value is preferred over the normal likelihood as it is more numerically stable. If the log-likelihood of the proposed parameters is better than that of the previous ones by a randomly specified amount, the proposed parameters become the new values from which a new set of randomly selected proposed values are generated. Over time, the parameters convene to an optimal distribution based on the 1917 weekly mortality. This method can be generalized for all seasonal mortality distributions by optimized the log-likelihood of an appropriate set of parameters.

3.3 Interpolation of Monthly Data

The above methods explored involve mathematically determining a baseline from data at smaller intervals in order to quantify seasonal patterns in *weekly* mortality. Especially in the context of both historical and population- and geographic-specific subgroups, mortality data is not always available in such refined time intervals. Thus, here we explore the idea of interpolating data such that weekly death counts can be inferred from monthly aggregated mortality information [16]. Several methods exist to interpolate values form smaller intervals of time from aggregated values. Within the realm of fertility research, this often involves determining single-year age specific fertility rates from grouped five-year intervals. Here, we adapt similar methods to those outlined in the Human Fertility Database [26] to interpolate yearly age-specific rates from aggregated data.

Monotonicity is a requirement of most interpolation techniques, but both yearly mortality and the age-specific fertility curve do not follow a pattern of strictly increasing or decreasing values through time (see Figure 2). However, by using the aggregate amount of expected births (or deaths, in our case) across the time period of study, a strictly increasing number of total deaths can be observed; that is, there will never be *fewer* total deaths during a year on one day than on the day before.

Figure 2: Aggregate mortality data by month (left) and cumulative (right)



To interpolate, we first take the logit of the cumulative (year-to-date) monthly mortality rates according to $Y(t) = log[\frac{M_t}{M_{t_{max}}} - M_t]$, where Y is equal to the logit of the cumulative mortality rate (M) at time t. T ranges from 0, interpreted as the beginning of the year (before the nonoccurence of any deaths), to 12 (the end of December), when all deaths have occurred, equal to M_t. At these extreme t points, we replace the logit value with reasonable values. As according to the method in the Human Fertility Database, we then perform cubic spline interpolation according to the *"Interp1"* function of the *"signal"* package in *R*.[2] We interpolate for 52 points along the logit function of cumulative mortality in order to represent the weeks of the year.

Inverse logit transformation is used to return the interpolated values to cumulative mortality rates, using the formula $M(x)_{hat} = [\frac{e^{Y_{hat}(x)}}{1+e^{Y_{hat}(x)}}] * M_{x_{max}}$. From the difference of these cumulative points, we determine weekly mortality estimates for three years of the monthly aggregated mortality data. The baseline and its 95% estimate are created by taking the upper bound of a Poisson distribution containing each the three interpolated points.

4 Results and Discussion

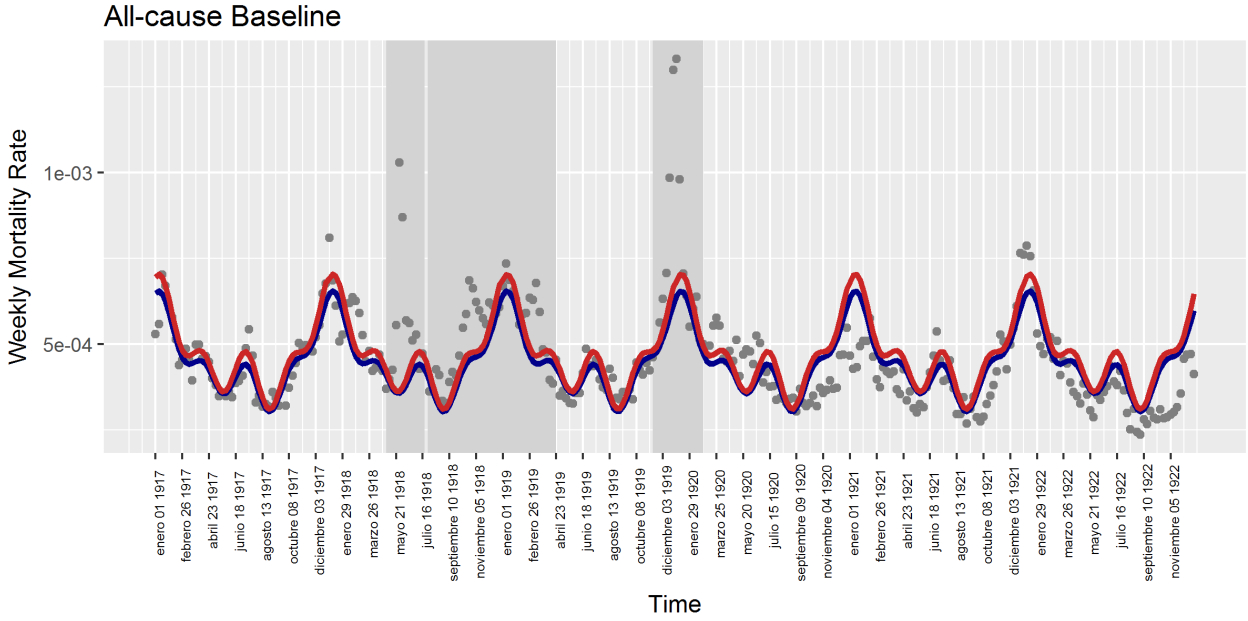
4.1 Baseline Results from Serfling Regression with Parametric Bootstrapping

The estimation technique for the mortality baseline created a general linear pattern mimicking mortality patterns for the year 1917. In the years following the main epidemic waves, the all-cause baseline continues to mimic the general shape and pattern of the baseline, but the overall mortality rate falls relative to 1917.

[2] HFD uses Hermite spline interpolation, but we choose to use cubic splines.

Given the rich information present in our data, we can also use this method to account for mortality baseline variations by age and estimate baseline mortality accordingly.

Figure 3: Grey points show the real weekly mortality rates from 1917-1922, while the blue and red lines display mean and upper 95% bound baseline from simulated 1917 deaths data. Shaded gray blocks represent the three epidemic wave periods.

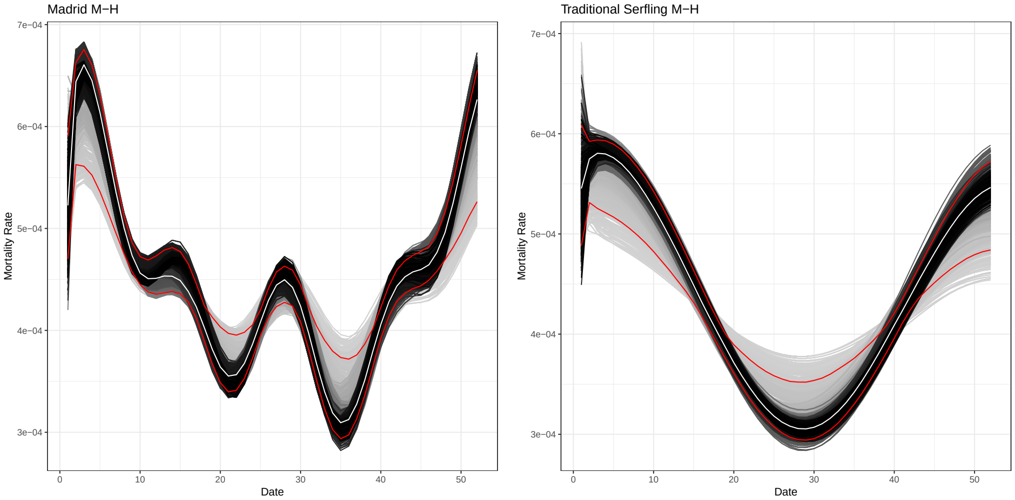


4.2 M-H MCMC optimization

Figure 4 shows graphical results of the Metropolis-Hastings MCMC for two potential distributions. The gray-scale shows, after the a series of "warm up steps, the accepted distribution parameters during the optimization process from start (gray) to end (black). The white line shows the mean distribution, and the red lines represent the upper and lower 95% confidence intervals. The first figure demonstrates the random walk of the optimization process using the Madrid 1917 data. The second figure shows the results of this algorithm to a different distribution of deaths, similar to the standard Serfling regression parameters, such that:

$$\int_0^{52} t + (\frac{100}{t})^2 + sin(\frac{2\pi t}{52.14}) + cos(\frac{2\pi t}{52.14}) + \varepsilon$$

Figure 4: MH MCMC results for 1917 Madrid data (left) and other distribution (right)



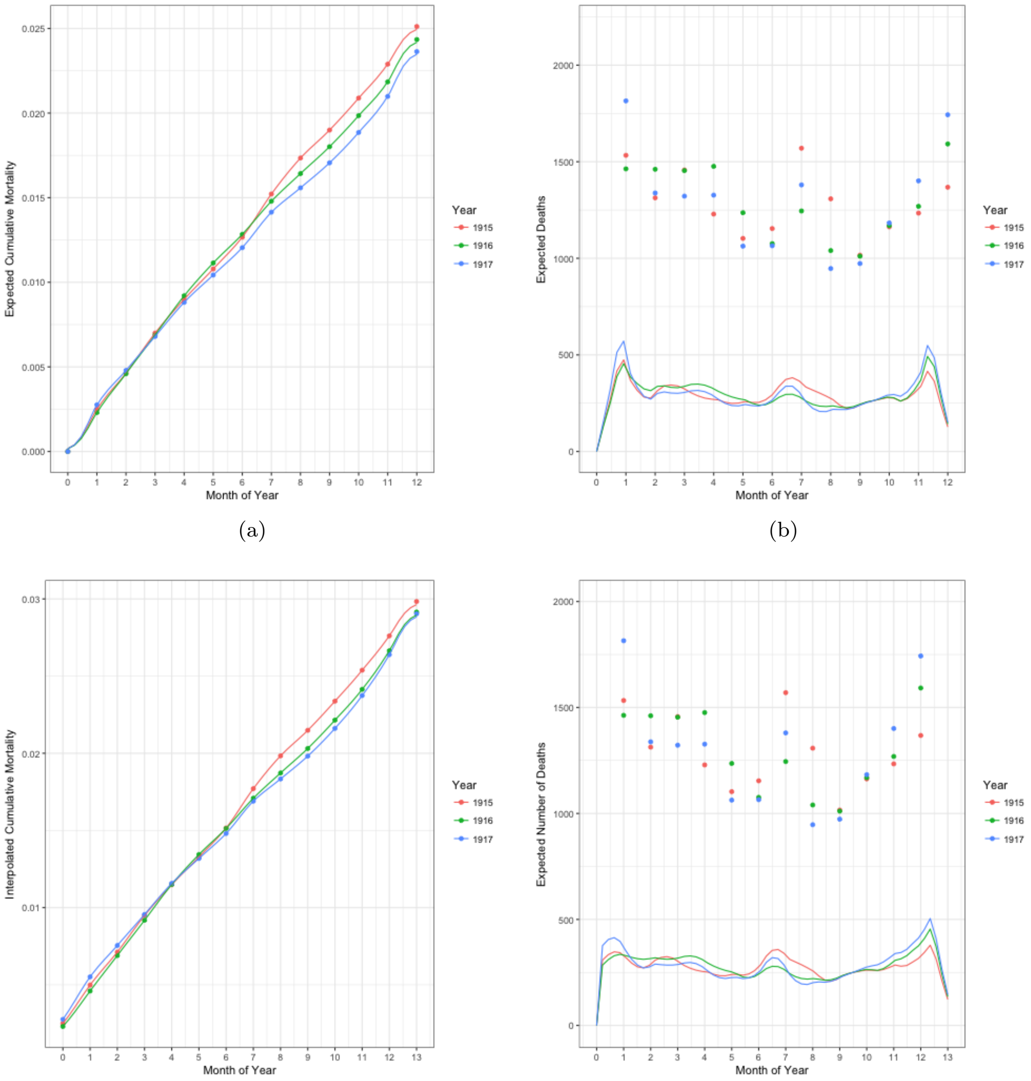
In both cases, the algorithm slowly optimizes the distribution by sampling and accepting or rejecting new parameter values. In outbreak periods, observed weekly rates higher than the red 95% confidence interval of the accepted parameters will be considered epidemic weeks, and the level of excess is the difference between the observed value and mean expected value (white line).

4.3 Baseline According to Interpolation Method

Initial results obtained by strictly adhering to the HFD methods revealed realistic estimations of the baseline for all months, with the exception of the first and last several month of the year. In the case of time $t = 0$ to time $t = 4$, the cumulative mortality rate is equal to zero at the beginning of the year and some value of January mortality at $t = 4$, but the interpolation function interprets this as zero deaths occurred at time 0, rather than simply that zero deaths during the year had occurred. Likewise, the end of the year shows a flattening of mortality rates towards zero as the values of the logits move towards an asymptote of the function where cumulative mortality remains the same. This is not realistic, as cumulative mortality will continue to grow even after the end of our early observation period.

To rectify this, we add "phantom deaths" to the beginning and end of the year before calculating cumulative mortality, the logits, and performing interpolation. This can be interpreted as providing information about mortality patterns before and after the end of the year (i.e. December mortality in the year prior to our baseline and January of the following year). We randomly generate these six (two each for 1915, 1916, and 1917) values according to a Poisson distribution fit from the number of monthly deaths during the months of January and December present in our data. Thus, the initial problems of our interpolated values remain at the extremes of our interpolated function, but all values of cumulative mortality during the observation year fall within a part of the increasing logit function such that the resulting values are realistic.

Both cases of interpolation results are visually depicted in Figure 5, where the quick rise of mortality in the beginning of the year and plateau at the end are easily visible. The right side of the figure also depicts the change in the expected number of deaths at the beginning and end of the year without and with the addition of the phantom deaths.



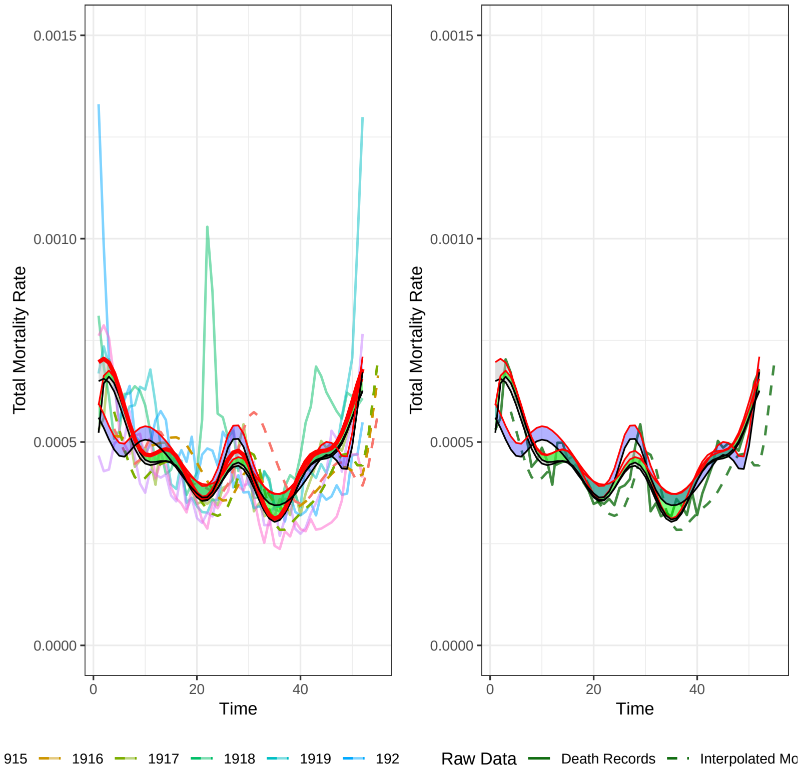
(a) (b) (c) (d)

Figure 5: **Interpolation issues and adaptation:** Intitial technique–(a) estimated weekly values of cumulative yearly mortality and (b) mis-estimation of deaths at the beginning and end of the year. Phantom deaths–(c) interpolated cumulative mortality including additional December-January deaths and (d) realistic estimation of deaths from January to December of observed year

4.4 Comparison of Baselines

Figure 6 shows both baselines overlain on all mortality data as well as for the year 1917 according to both monthly and death record data. There are many similarities between the three; all share the same basic shape and take into account the summer mortality peak, and the baselines follow nearly identical paths and values in the second half of the year. Overall, the baselines reinforce each other's legitimacy, as the three years of interpolated data validate the relative shape of the curve determined by only 1917 data. Moreover, the thickness of the confidence intervals show that neither our simulated data and the number of simulations involved in our bootstrapping nor the iterations of the MH MCMC optimization overfit the variability in the baseline estimation.

Figure 6: Blue shaded region shows baseline and upper 95% of interpolated monthly data, gray shaded region shows the same for parametric bootstrapping of death records, green shaded region shows same for the m-h mcmc estimated baseline; mortality rates for all years of data plotted on left, two years of 1917 (death records and interpolated provincial counts) on right



Outside of its calculation method, the MH MCMC baseline largely differs from the parametric bootstrapping technique in that it predicts the mean and 95% uncertainty interval of the summer and winter peaks with greater certainty and the spring and fall mortality troughs with less uncertainty. This is due to the selection mechanism for the proposed parameters; it should be noted that a greater number of simulations and increase in the specified “burnout period would further decrease the level of uncertainty and lower the confidence intervals. Overall, the two share very similar mean values of predicted mortality levels, though this is expected given they are derived from the dame data and are calculated using the same parameter values, even if via different methods.

Nonetheless, there are also noticeable differences in the the baselines. Because the data used to construct the single-year baseline follows the distribution of deaths in 1917, the resulting shape of the mortality curve, despite the introduced uncertainty, adheres to the timing and strength of the peaks in 1917. That is to say, the winter seasonal peak in 1917 occurred mid-way through January, and this is reflected through the early peak while the monthly data declines from the beginning of the year. Likewise, the summer peak was lower in 1917 than what was observed in the aggregated monthly data. However, this could also possibly be attributed to the timing of the summer peak. Whereas the individual death dates can pinpoint an exact peak, the monthly data may show a muted rise if the peak occurred during the end of one month and the beginning of another.

In general, the methods using the 1917 data only highlight the importance of using more than one year of baseline mortality data; both place winter mortality peaks in the middle of January *and* at the end of December. This seemingly indicates a winter “trough in the beginning of January, however, as a general pattern of seasonal mortality, this valley is unrealistic. However, we also cannot be sure that the monthly data is interpolated correctly such that the location of the peaks within a month is where it actually occurred. The interpolation infers a smooth rate of increase towards and away from the peak according to the monthly aggregates, and without more precise information, we must rely on this assumption.

Several additional interpolation methods exist that should be considered in order to best quantify a weekly baseline from aggregated data. We briefly considered a type of polynomial regression [27], but found that due to the summer mortality peak, the high-order function produced results that were quite complex for our simple baseline. Another option to determine our weekly baseline involves using a calibrated splines method to estimate four segments of the mortality curve across the year [28]. Though this technique would require estimation of several segments of the baseline, its applications to fertility data (if the age-specific fertility curve is thought of in a similar manner to the time-of-year-specific mortality curve) appear to produce at least comparable, if not better results than the current Human Fertility Database technique [29] and thus should be considered as a possible additional procedure to interpolate mortality data.

4.5 Excess Mortality

Table 1 shows five estimates of excess mortality in Madrid for the influenza waves in the spring of 1918 and the combined fall/winter waves in 1918-1919. The first estimates are derived from our technique of parametric bootstrapping our single year of 1917 mortality data. We calculate the second column from the 95% probability interval of the MH MCMC derived distribution of deaths, and we estimate the third set from our interpolated data, using the upper 95% interval of an assumed Poisson distribution. The fourth set applies the MH MCMC method to the interpolated data, and the final set of estimates come from a previously published paper on excess mortality in Spain during the pandemic [30]. These excess mortality rates are based on a simple three year Serfling model using the monthly data in its aggregated form.

Table 1: Estimates of Excess Mortality with Different Baseline Estimations

Wave	Serfling with Param Bootstrapping	MH MCMC	Interp Data & Parm Bootstrapping	Interp Data with MCMC	Monthly Data
Spring 1918	18.19	17.46	14.92	16.58	11.7
Fall/Winter 1918-1919	22.44	22.15	25.36	35.61	55
Overall	40.63	39.61	40.28	52.19	66.7

Despite the differences in the overall baseline discussed above, both of our baseline estimations reveal very similar numerical results of excess mortality during the two waves in Madrid but with varying degrees of relative difference in excess mortality. The spring waves are calculated from only three months of data; thus, the differences in the upper baselines are different enough that the results are about 12.5% different. Our estimated excess mortality estimates for the fall wave differ by about 6.25%. This period encompasses a longer time period, and the variations in the epidemic threshold of the mortality baseline appear to even each other out over time.

Our findings of overall excess mortality do differ from those of the previous study of excess monthly mortality in all provinces of Spain [30]. For example, the excess mortality from both of our baselines is much lower in the fall/winter wave, and higher in the spring. When the interpolated data is used to calculate a baseline with the MH MCMC method, we also find higher results of excess mortality than with the other methods, most notably in the combined fall/winter wave. Part of this could come from the lower expected mortality of interpolated data at the begining of the year, but it is difficult to ascertain what specifically contributes to these differences. Also, it should be mentioned the previous estimates use additional monthly data from 1918 and 1919 to estimate the excess, while we exclusively use our death records to estimate the excess from calibrated baselines in the first four methods of table 1.

Additionally, the overall timing of the baseline mortality rate for the city of Madrid may vary compared to the province due to varying levels of influenza immunity. Previous exposure to influenza virus and a person's triggered immune defense changes the probability of (a) contracting a specific strain of influenza from exposure and (b) the body's reaction to fighting such a virus [31]. Moreover, the connectedness of the city itself resulted in the presence of a strong spring wave and this *is* present in the provincial data due to its inclusion of the city. However, the population in the countryside of the surrounding region likely did not face the same exposure to the spring wave, meaning the monthly provincial data may not perfectly reflect the mortality conditions in the city during the outbreaks.

In many urban centers that experienced a stronger herald wave in the spring or summer of 1918, the severity of the succeeding fall wave is less pronounced relative to the rest of the world. Often, the total effect of the epidemic was lower in cities with herald waves. While the spring wave lacked the virulence of the successive fall outbreaks, its overall transmission rates were quite high, and where present, the virus tended to spread through much of the population [32, 33]. However, excess mortality rates specific to the spring waves are generally lower than in the fall. Thus, the differences in the province-wide mortality from our city estimates may be due to a lack of exposure to the spring wave, followed by a strong fall wave to which the provincial population had no mortality. While further sensitivity analysis will be done to look at the differences in the city and surrounding populations, the current results reveal a stark contrast of more than 50% higher mortality in the province itself.

4.6 Final Notes

While here, we explore novel ways to calibrate a mortality baseline, caution should be taken when deciding the best approach to take. When producing a mortality baseline, it is most essential that the seasonal rates are a reasonable expectation of "normal" mortality patterns for the population of analysis. As mentioned, the monthly data used in the paper covers the entire province of Madrid, whereas the individual-level death records provide information from the city only. While the provincial-level data takes into account the population of the city as well, the differences in mortality during the baseline period of 1917 are stark, specifically during the winter peaks. Both baselines accurately reflect their input data–that is, according to available data, they both paint a believable picture of mortality–, but there are large, significant differences in the epidemic threshold, particularly in the first two thirds of the year.

The content of this paper currently focuses on the city of Madrid during several waves of Spanish influenza, but as noted, these estimation techniques are applicable to all types of aggregated mortality data. To further understand the benefits and limitations of the interpolation and baseline re-calculation process, future analyses should incorporate data from other locations and time periods. Yet, in this analysis, we find that reasonable mortality time-series at finer time intervals can be created through interpolating higher-level aggregated data. For example, the severity of the 1918 Influenza Pandemic throughout the world led many cities, counties, and other administrative areas to collect daily and/or weekly surveillance and mortality data during the epidemic periods, especially in the fall of 1918. However, the usefulness of this data relies heavily on the estimation of baseline mortality during the time and the discernment of how much larger an outbreak was compared to seasonal flu. In these same areas with detailed 1918 information, often, previous years of mortality data are available only in a larger aggregate scale. Thus, interpolating this information, as shown through our estimations, provides a viable method through which a finer baseline may be calculated. From this, researchers can better understand some of the intricacies of the influenza pandemic and other events still debated today.

References

[1] N. P. Johnson and J. Mueller, “Updating the accounts: global mortality of the 1918-1920 ”Spanish” influenza pandemic,” *Bulletin of the History of Medicine*, vol. 76, no. 1, pp. 105–115, 2002.

[2] G. J. D. Smith, J. Bahl, D. Vijaykrishna, J. Zhang, L. L. M. Poon, H. Chen, R. G. Webster, J. S. M. Peiris, and Y. Guan, “Dating the emergence of pandemic influenza viruses,” *Proceedings of the National Academy of Sciences*, vol. 106, no. 28, pp. 11709–11712, 2009.

[3] M. Worobey, G.-Z. Han, and A. Rambaut, “Genesis and pathogenesis of the 1918 pandemic H1N1 influenza A virus,” *Proceedings of the National Academy of Sciences*, vol. 111, no. 22, pp. 8107–8112, 2014.

[4] M. A. Miller, C. Viboud, M. Balinska, and L. Simonsen, “The signature features of influenza pandemics—implications for policy,” *New England Journal of Medicine*, vol. 360, no. 25, pp. 2595–2598, 2009.

[5] L. Simonsen, M. J. Clarke, L. B. Schonberger, N. H. Arden, N. J. Cox, and K. Fukuda, “Pandemic versus epidemic influenza mortality: a pattern of changing age distribution,” *Journal of infectious diseases*, vol. 178, no. 1, pp. 53–60, 1998.

[6] D. Ramiro, S. Garcia, Y. Casado, L. Cilek, and G. Chowell, “Age-specific excess mortality patterns and transmissibility during the 18891890 influenza pandemic in madrid, spain,” *Annals of Epidemiology*, 2017.

[7] G.-F. Sara, *La gripe de 1889-1890 en Madrid.* PhD dissertation, Universidad Complutense de Madrid, 2017.

[8] A. Trilla, G. Trilla, and C. Daer, “The 1918 ”Spanish flu” in Spain,” *Clinical infectious diseases*, vol. 47, no. 5, pp. 668–673, 2008.

[9] T. Watanabe, G. Zhong, C. A. Russell, N. Nakajima, M. Hatta, A. Hanson, R. McBride, D. F. Burke, K. Takahashi, S. Fukuyama, *et al.*, “Circulating avian influenza viruses closely related to the 1918 virus have pandemic potential,” *Cell host & microbe*, vol. 15, no. 6, pp. 692–705, 2014.

[10] D. Ramiro Farias, L. Cilek, and G. Chowell, “Age-Specific Excess Mortality Patterns During the 19181920 Influenza Pandemic in Madrid, Spain,” *American Journal of Epidemiology*, vol. 187, pp. 2511–2523, 08 2018.

[11] M. C. Registry, “Civil register records on deaths,” 1917-1922.

[12] I. G. y Estadstico, “Resumen (mensual) del movimiento natural de la población de españa y de las capitales de provincia,” 1915-1919.

[13] W. W. Thompson, D. K. Shay, E. Weintraub, L. Brammer, N. Cox, L. J. Anderson, and K. Fukuda, “Mortality associated with influenza and respiratory syncytial virus in the united states,” *Jama*, vol. 289, no. 2, pp. 179–186, 2003.

[14] W. W. Thompson, E. Weintraub, P. Dhankhar, P.-Y. Cheng, L. Brammer, M. I. Meltzer, J. S. Bresee, and D. K. Shay, “Estimates of us influenza-associated deaths made using four different methods,” *Influenza and other respiratory viruses*, vol. 3, no. 1, pp. 37–49, 2009.

[15] E. Goldstein, C. Viboud, V. Charu, and M. Lipsitch, “Improving the estimation of influenza-related mortality over a seasonal baseline,” *Epidemiology (Cambridge, Mass.)*, vol. 23, no. 6, p. 829, 2012.

[16] C. Warren-Gash, K. Bhaskaran, A. Hayward, G. M. Leung, S.-V. Lo, C.-M. Wong, J. Ellis, R. Pebody, L. Smeeth, and B. J. Cowling, “Circulating influenza virus, climatic factors, and acute myocardial infarction: a time series study in england and wales and hong kong,” *Journal of Infectious Diseases*, vol. 203, no. 12, pp. 1710–1718, 2011.

[17] W. K. Hastings, “Monte carlo sampling methods using markov chains and their applications,” 1970.

[18] R. E. Serfling, “Methods for current statistical analysis of excess pneumonia-influenza deaths,” *Public health reports*, vol. 78, no. 6, pp. 494–506, 1963.

[19] A. J. Cobos, C. G. Nelson, M. Jehn, C. Viboud, and G. Chowell, “Mortality and transmissibility patterns of the 1957 influenza pandemic in Maricopa County, Arizona,” *BMC infectious diseases*, vol. 16, no. 1, p. 405, 2016.

[20] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap.* CRC press, 1994.

[21] G. Chowell, C. Viboud, L. Simonsen, M. A. Miller, R. Acuna-Soto, J. M. O. Díaz, and A. F. Martínez-Martín, "The 1918–19 influenza pandemic in Boyaca, Colombia," *Emerging infectious diseases*, vol. 18, no. 1, pp. 48–56, 2012.

[22] S.-E. Mamelund, "A socially neutral disease? individual social class, household wealth and mortality from Spanish influenza in two socially contrasting parishes in Kristiania 1918–19," *Social Science & Medicine*, vol. 62, no. 4, pp. 923–940, 2006.

[23] L. Simonsen, T. A. Reichert, C. Viboud, W. C. Blackwelder, R. J. Taylor, and M. A. Miller, "Impact of influenza vaccination on seasonal mortality in the us elderly population," *Archives of internal medicine*, vol. 165, no. 3, pp. 265–272, 2005.

[24] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, "Simulated annealing," *Journal of Chemical Physics*, vol. 21, pp. 1087–1092, 1953.

[25] C. P. Robert, *The MetropolisHastings Algorithm*, pp. 1–15. American Cancer Society, 2015.

[26] A. Jasilioniene, D. Jdanov, T. Sobotka, E. Andreev, K. Zeman, V. Shkolnikov, J. Goldstein, D. Philipov, and G. Rodriguez, "Methods protocol for the human fertility database," *Rostock, MPIDR, 56p*, 2012.

[27] S. Krenk, "On the use of the interpolation polynomial for solutions of singular integral equations," *Quarterly of Applied Mathematics*, vol. 32, no. 4, pp. 479–484, 1975.

[28] C. P. Schmertmann, "Calibrated spline estimation of detailed fertility schedules from abridged data^1," *Revista Brasileira de Estudos de População*, vol. 31, no. 2, pp. 291–307, 2014.

[29] O. Grigorieva, A. Jasilioniene, D. Jdanov, P. Grigoriev, T. Sobotka, K. Zeman, and V. Shkolnikov, "Methods protocol for the human fertility collection," 2015.

[30] G. Chowell, A. Erkoreka, C. Viboud, and B. Echeverri-Dávila, "Spatial-temporal excess mortality patterns of the 1918–1919 influenza pandemic in Spain," *BMC infectious diseases*, vol. 14, no. 1, p. 371, 2014.

[31] A. Gagnon, J. E. Acosta, J. Madrenas, and M. S. Miller, "Is antigenic sin always "original?" re-examining the evidence regarding circulation of a human H1 influenza virus immediately prior to the 1918 Spanish flu," *PLOS Pathogens*, vol. 11, pp. 1–6, 03 2015.

[32] D. R. Olson, L. Simonsen, P. J. Edelson, and S. S. Morse, "Epidemiological evidence of an early wave of the 1918 influenza pandemic in New York City," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 31, pp. 11059–11063, 2005.

[33] V. Andreasen, C. Viboud, and L. Simonsen, "Epidemiologic characterization of the 1918 influenza pandemic summer wave in Copenhagen: implications for pandemic control strategies," *The Journal of infectious diseases*, vol. 197, no. 2, pp. 270–278, 2008.