Hurricane Maria Mortality Study (NEJM online, Kishore et al., May 29 2018)

Available online: https://www.nejm.org/doi/full/10.1056/NEJMsa1803972

Last updated: June 15, 2018

Technical FAQ

In the paper, we provide an estimate and confidence interval for excess deaths. To make this calculation we use three quantities: the average observed death rate for the period in question, a baseline death rate representing mortality for the period in question had the hurricane not occurred, and the effective population size (the average during the period). We will refer to these quantities as r, r_0, and N, respectively. If we denote the excess death count with Y, then we can write the formula used in the paper as:

$$Y = (r - r_0) \times N \times K$$

with K a constant to account for units of person-time — because we report rates in deaths per year per 1000 people, $K = (102/365)/1000$.

We used survey data to estimate r, which we denote here with $\hat{r}$, as well as a confidence interval. For the other two quantities we inserted plug-in estimators which we will denote with $\hat{r}_0$ and $\hat{N}$. So our estimate of excess counts, denoted with $\hat{Y}$, will be:

$$\hat{Y} = (\hat{r} - \hat{r}_0) \times \hat{N} \times K$$

Surveys rarely produce completely unbiased estimates. Many sources of bias are difficult to detect, and even after identifying them, often have competing approaches for addressing these biases. Each approach for estimating $\hat{Y}$ has strengths and weaknesses, and thus no approach should be considered final or perfect. As stated in our paper, for transparency, collaboration, and future analyses we have made our data publicly available and encourage others to try their own approaches for attempting to address these biases.

In this document, we describe some of the alternative approaches to estimating $\hat{Y}$ in addition to the analyses shown in the paper. This document includes the `R` code that can be used reproduce all of our analysis with the publicly available data.

Setting up

We begin by loading libraries and setting some options.

```
library("tidyverse")
library("magrittr")
library("lubridate")
library("survey")
options(digits = 3)
```

Why did you not use a weighted approach to estimate the rate and confidence interval since you used a cluster sample?

We did consider a weighted approach. The code is included below. You can see that the resulting estimate and confidence interval for $\hat{r}$ does not substantially differ from the ones presented in the paper.

```r
days_before_hurricane <-
  difftime(ymd("2017-09-20"), ymd("2016-12-31"), units = "days") %>%
  as.numeric()
years_before <- days_before_hurricane / 365
years_after <- (365 - days_before_hurricane) / 365

source("../ref/base.R")

#loading resources
hh_main <- readRDS("../data/rdata/hh_main.RDS")
deaths <- readRDS("../data/rdata/deaths.RDS")
deaths_official <- readRDS("../data/rdata/deaths_official.RDS")
weighted_pop_est <- readRDS("../data/rdata/pop_est.RDS")
adj_rates <- readRDS("../data/rdata/adj_rates.RDS")
individuals <- readRDS("../data/rdata/individuals.RDS")
weights <- readRDS("../data/rdata/final_weights.RDS")

weights$sel_strata_w <- weights$barrio_w/13 # weight for each barrio
hh_main$id <- as.character(hh_main$id)
weights$id <- as.character(weights$id)
hh_main <- left_join(hh_main, weights, by = c("id", "strata"))

# create single final weight for each household
hh_main$hh_w_f <- as.numeric(hh_main$hh_w) *
    as.numeric(hh_main$sel_strata_w)

# add data on age and gender of individuals
ind_main <- left_join(individuals, hh_main, by = "hh_id")
ind_main$count <- years_after

# now merge deaths
deaths_main <- left_join(ind_main, deaths, by = c("hh_id", "hh_id__1"))
deaths_main$deaths_after_hurricane <-
    ifelse(as.numeric(deaths_main$died_month) >= 10 |
               (deaths_main$died_month == 9 &
                    deaths_main$died_b_p_hurricane == 2), 1, 0)

deaths_main$deaths_after_hurricane[
    is.na(deaths_main$deaths_after_hurricane)] <- 0

# subtract deaths that occured before the hurricane from the population count
deaths_main$deaths_before_hurricane <- ifelse(as.numeric(deaths_main$mo) < 9.2, 1, 0)
deaths_main <- deaths_main %>% mutate(count = case_when(deaths_before_hurricane == 1 ~ 0,
                                       TRUE ~ count))

# create survey design object
id.form <- ~ strata + id
wt.form <- ~ 1 + hh_w_f
dsvy <- svydesign(id = id.form, weights = wt.form,
                  data = deaths_main, nest = TRUE)

rate <- svyratio(~deaths_after_hurricane, ~count, dsvy, na.rm = TRUE)
```

```
print(
    data.frame(rate_after = unlist(rate[1]) * 1000,
           lower_rate = (unlist(rate[1]) - 1.96*SE(rate)) * 1000,
           upper_rate = (unlist(rate[1]) + 1.96*SE(rate)) * 1000)
)
```

```
##       rate_after lower_rate upper_rate
## ratio       15.6       11.7       19.5
```

However, our data exploration did not reveal strong evidence of clustering as demonstrated by this code:

```
deaths_by_location <- deaths %>%
  mutate(death_after = mo >= 9.2 ) %>%
  group_by(hh_id) %>%
  summarize(tot_before = sum(!death_after), tot_after = sum(death_after)) %>%
  ungroup() %>%
  left_join(hh_main, by = "hh_id")

deaths_by_location %>%
  group_by(id, mun_id) %>%
  summarize(tot_before = sum(tot_before),
            tot_after=sum(tot_after)) %>%
  arrange(desc(tot_after)) %>%
  print(n = 10)
```

```
## # A tibble: 44 x 4
## # Groups:   id [43]
##    id    mun_id tot_before tot_after
##    <chr>  <dbl>      <int>     <int>
##  1 134      48.          0         2
##  2 14       30.          0         2
##  3 171      11.          0         2
##  4 259      29.          0         2
##  5 426      37.          0         2
##  6 752      31.          0         2
##  7 152      53.          0         1
##  8 237       8.          0         1
##  9 269      52.          0         1
## 10 291       5.          0         1
## # ... with 34 more rows
```

Note that no barrio had more than two deaths.

We also explored the data by municipality. Due to privacy concerns, we cannot publicly release the names of municipalities; rather we generate a muninicipality ID for this analysis. Although two of the three municipalities showing three or more deaths were on the east coast, we did not observe a strong geographical clustering among those with more than one death.

```
deaths_by_location %>%
  group_by(mun_id) %>%
  summarize(tot_before = sum(tot_before),
            tot_after=sum(tot_after)) %>%
  arrange(desc(tot_after)) %>%
  filter(tot_after > 1) %>%
  print()
```

```
## # A tibble: 9 x 3
```

```
##   mun_id tot_before tot_after
##    <dbl>      <int>     <int>
## 1    29.          0         4
## 2    19.          3         3
## 3    30.          0         3
## 4    11.          0         2
## 5    20.          0         2
## 6    31.          0         2
## 7    37.          0         2
## 8    39.          0         2
## 9    48.          0         2
```

Furthermore, we did not see evidence of an effect across remoteness strata. Specifically, we saw no strong evidence for the number of deaths changing by remoteness:

```
deaths_by_location %>%
  group_by(strata) %>%
  summarize(before = sum(tot_before),
            after = sum(tot_after)) %>%
  arrange(strata) %>%
  print()
```

```
## # A tibble: 8 x 3
##   strata before after
##    <dbl>  <int> <int>
## 1     1.      2     4
## 2     2.      2     7
## 3     3.      4     1
## 4     4.      2     3
## 5     5.      4     8
## 6     6.      1     7
## 7     7.      1     5
## 8     8.      2     3
```

In conclusion, we did not find strong evidence of clustering in our data. To maintain simplicity and avoid additional noise, we decided not to use a weighted approach.

The pre-hurricane death rate is estimated as 2.6 per 1000. Why is it so much lower than 2016?

To see the low rates you can run this code:

```
N <- sum(hh_main$hh_size, na.rm = TRUE)

deaths %>%
  mutate(death_after = mo >= 9.2 ) %>%
  summarize(deaths_before = sum(!death_after),
            deaths_after = sum(death_after)) %>%
  mutate(rate_before = deaths_before/N*1000 / years_before,
         rate_after = deaths_after/(N - deaths_before)*1000 / years_after)
```

```
##   deaths_before deaths_after rate_before rate_after
## 1            18           38        2.62       14.3
```

As mentioned in the paper, our survey is unable to capture deaths in single-person households. This introduces a bias that results in underestimation of both before and after the hurricane. The following code can be used

to see this:

```
## compute median age of each household
hh_median_age <- individuals %>%
  group_by(hh_id) %>%
  summarize(median_age = median(age, na.rm=TRUE))

## comput household size and median age table
hh_stats <- hh_main %>%
  dplyr::select(hh_id, hh_size) %>%
  filter(!is.na(hh_size) & hh_size>0) %>%
  left_join(hh_median_age, by="hh_id")

## create table for before and after deaths for each household
hh_deaths <- deaths %>%
  mutate(death_after = mo >= 9.2 ) %>%
  group_by(hh_id) %>%
  summarize(tot_before = sum(!death_after),
            tot_after = sum(death_after)) %>%
  ungroup()

## Divide households by size and compute a total, strata size,
## and before and after rates for each one
cuts_hh_size <- c(0, 1, 2, 4, Inf)
cuts_hh_size_labels <- c("1", "2", "3-4", "5+")
rates_by_hh_size <- left_join(hh_stats, hh_deaths, by = "hh_id") %>%
  mutate(household_size = cut(hh_size, cuts_hh_size,
                              labels = cuts_hh_size_labels)) %>%
  group_by(household_size) %>%
  summarize(median_age = median(median_age),
            total_households = n(),
            N = sum(hh_size),
            deaths_before = sum(tot_before, na.rm = TRUE),
            deaths_after = sum(tot_after, na.rm = TRUE)) %>%
  mutate(rate_before = deaths_before/N*1000 / years_before,
         rate_after = deaths_after/(N - deaths_before)*1000 / years_after)
names(rates_by_hh_size)[1] <- "hh_size"

rates_by_hh_size %>%
    print(width = 100)
```

```
## # A tibble: 4 x 8
##   hh_size median_age total_households     N deaths_before deaths_after
##   <fct>        <dbl>            <int> <dbl>         <int>        <int>
## 1 1             69.0              534  534.             0            0
## 2 2             64.2             1074 2148.             9           15
## 3 3-4           44.0             1255 4290.             8           16
## 4 5+            27.0              436 2550.             1            7
##   rate_before rate_after
##         <dbl>      <dbl>
## 1       0.          0.
## 2       5.81       25.1
## 3       2.59       13.4
## 4       0.544       9.83
```

The death rate in single-person households is found to be 0 — this cannot be true. First, the median age in this single-person households is 69. Second, this death rate is expected because we missed (i.e., could not count) the single-person households where there had been a death (as there was no one to answer the door!). Further evidence that our survey undercounted these deaths is provided through the change in household size distributions as compared to the ACS in 2016. (Figure 1B in the paper shows this as well.)

```
load("../data/rdata/ACS2016.Rdata")

household_dist <- acs.hh_size %>%
  mutate(hh_size = cut(hh_size, cuts_hh_size, labels = cuts_hh_size_labels)) %>%
  group_by(hh_size)%>%
  summarize(count = sum(count)) %>%
  ungroup() %>%
  mutate(pop_freq = count / sum(count)) %>%
  dplyr::select(-count)

rates_by_hh_size %>%
  dplyr::select(hh_size, total_households) %>%
  mutate(surevy_freq = total_households/sum(total_households)) %>%
  dplyr::select(-total_households) %>%
  left_join(household_dist, by = "hh_size") %>%
  print()
```

```
## # A tibble: 4 x 3
##   hh_size surevy_freq pop_freq
##   <fct>         <dbl>    <dbl>
## 1 1             0.162   0.261
## 2 2             0.326   0.314
## 3 3-4           0.380   0.348
## 4 5+            0.132   0.0772
```

A simple approach to adjust for the problem posed by the lack of deaths in single-person households is to remove all households of size one from our study.

```
N <- hh_main %>%
  filter(hh_size > 1) %>%
  summarize(n = sum(hh_size)) %>%
  .$n

deaths %>%
  left_join(hh_main, by = "hh_id") %>%
  filter(hh_size > 1) %>%
  mutate(death_after = mo >= 9.2 ) %>%
  summarize(deaths_before = sum(!death_after),
            deaths_after = sum(death_after)) %>%
  mutate(rate_before = deaths_before/N*1000 / years_before,
         rate_after = deaths_after/(N-deaths_before)*1000 / years_after) %>%
  print()
```

```
##   deaths_before deaths_after rate_before rate_after
## 1            18           38        2.78       15.2
```

Removing all households with the size of one increases the "before estimate" but not by much. However, it is important to note that the households of size one appeared to be older, on average, than other households which likely results in an underestimate of the mortality rate if we simply remove these from our analysis.

An alternative approach would be to plug in a more realistic rate for these households and then adjust for

the bias favoring larger households. We took the conservative approach and plugged in the same rate for both before and after the hurricane. To compute a confidence interval we use Keyfitz's approximation. Here is the function that computes the adjustment for any given plug-in:

```
adjust <- function(rates_by_hh_size, plugin){
  cuts_hh_size <- c(0, 1, 2, 4, Inf)
  cuts_hh_size_labels <- c("1", "2", "3-4", "5+")

  rates_by_hh_size_adj <- rates_by_hh_size
  rates_by_hh_size_adj$rate_before[1] <-
      rates_by_hh_size_adj$rate_after[1] <- plugin

  rates_by_hh_size_adj$deaths_before[1] <-
    rates_by_hh_size_adj$rate_before[1] *
      rates_by_hh_size_adj$N[1]/1000*years_before

  rates_by_hh_size_adj$deaths_after[1] <-
      rates_by_hh_size_adj$rate_after[1] *
      rates_by_hh_size_adj$N[1]/1000*years_after

  household_dist <- acs.hh_size %>%
    mutate(hh_size = cut(hh_size, cuts_hh_size,
                         labels = cuts_hh_size_labels)) %>%
    group_by(hh_size)%>%
    summarize(count = sum(count)) %>%
    ungroup() %>%
    mutate(pop_freq = count / sum(count))

  rates_by_hh_size %>%
    dplyr::select(hh_size, total_households) %>%
    mutate(survey_freq = total_households/sum(total_households)) %>%
    left_join(household_dist, by = "hh_size") %>%
    dplyr::select(-count, -total_households)

  adjusted_rates <- rates_by_hh_size_adj %>%
    left_join(household_dist, by = "hh_size") %>%
    summarize(rate_before = sum(rate_before*pop_freq),
              rate_after = sum(rate_after*pop_freq))

  se_before <- adjusted_rates$rate_before/sqrt(sum(hh_deaths$tot_before))
  se_after <- adjusted_rates$rate_after/sqrt(sum(hh_deaths$tot_after))

  adjusted_rates$before_lower <- adjusted_rates$rate_before - 1.96*se_before
  adjusted_rates$before_upper <- adjusted_rates$rate_before + 1.96*se_before

  adjusted_rates$after_lower <- adjusted_rates$rate_after - 1.96*se_after
  adjusted_rates$after_upper <- adjusted_rates$rate_after + 1.96*se_after

  adjusted_rates <- dplyr::select(adjusted_rates,
                                  c("rate_before", "before_lower",
                                    "before_upper", "rate_after",
                                    "after_lower", "after_upper"))
  adjusted_rates
}
```

Now we can see how these estimated rates change when we plug in a value for the households of size one. The plug-in we used in the paper was simply the death rate before the hurricane:

```
population_by_year <- readRDS("../data/rdata/deaths_official.RDS") %>%
  dplyr::select(Year, Popv17) %>%
  setNames(c("year", "pop"))

population_by_year <- readxl::read_excel("../data/pr_popest_2010_17.xlsx")
official <- readRDS("../data/rdata/official_long.RDS")
days_in_month <- c(31, 28, 31, 30, 31, 30, 31, 31, 30, 31, 30, 31)
official_rates <- official %>% left_join(population_by_year, by = "year") %>%
  mutate(days = days_in_month[month]) %>%
  mutate(days = ifelse(year %% 4 == 0 & month ==2, days + 1, days)) %>%
  mutate(rate = deaths / pop * 365/days * 1000) %>%
  group_by(year) %>%
  summarize(before_rate = sum(deaths*(month<9) +
                                deaths*(month==9)*2/3) / pop[1] * 1000 *
              365/(sum(days*(month<9))+20),
            after_rate = sum(deaths*(month>9) +
                               deaths*(month==9)*1/3) / pop[1] * 1000 *
              365/(sum(days*(month>9))+10))

plugin <- official_rates %>%
  filter(year == 2017) %>%
  .$before_rate
```

Here are the adjusted rates using that rate:

```
adjust(rates_by_hh_size, plugin)
```

```
## # A tibble: 1 x 6
##   rate_before before_lower before_upper rate_after after_lower after_upper
##         <dbl>        <dbl>        <dbl>      <dbl>       <dbl>       <dbl>
## 1        5.08         2.73         7.42       15.6        10.6        20.6
```

Note that the before rate goes up substantially. However, the value we plugged in is probably an underestimate for those households given that the median age is 69. If we plug in the death rate for a 69 year old, which in the US is about 19 per 1000 we get the following:

```
adjust(rates_by_hh_size, 19)
```

```
## # A tibble: 1 x 6
##   rate_before before_lower before_upper rate_after after_lower after_upper
##         <dbl>        <dbl>        <dbl>      <dbl>       <dbl>       <dbl>
## 1        7.72         4.16         11.3       18.2        12.4        24.0
```

It is also possible that the deaths before the hurricane were affected by other biases such as recall bias. We therefore did not use these data in the calculation of excess deaths. Specifically, we did not use these data to inform the choice of the baseline rate $\hat{r}_0$. We think further research is needed in understanding potential biases.

The age distribution of your survey was different from the distribution reported by ACS. Why did you not adjust the baseline death rate for age?

We cannot know for sure if the difference in distribution we see is due to the survey bias, an actual change in distribution due to out-migration, or a combination of both. For this reason we estimated the rate $\hat{r}$ without

an age adjustment. However, the plug-in estimate $\hat{r}_0$ was obtained from historical data which is based on a population with a different population distribution. This implies that some of the different $\hat{r} - \hat{r}_0$ may be explained by age differences, which makes $\hat{Y}$ an overestimate.

Below we include code that performs an age adjustment that assumes that post-hurricane age distribution was the same as in 2016.

```
cuts <- c(0, 55, 65, 75, 85, Inf)

total <- sum(acs.age$count)
acs <- acs.age %>%
  mutate(upper = ifelse(is.na(upper), 125, upper)) %>%
  mutate(age_strata = cut(upper, cuts,
                          include.lowest = TRUE, right = FALSE)) %>%
  group_by(age_strata) %>%
  summarize(pop_2016_freq = sum(count)/total)

n_by_age <- individuals %>%
  left_join(hh_stats, by = "hh_id") %>%
  filter(hh_size>1) %>%
  filter(!is.na(gender)) %>%
  mutate(age_strata = cut(age, cuts,
                          include.lowest = TRUE, right = FALSE)) %>%
  group_by(age_strata) %>%
  summarize(n = n())  %>%
  ungroup()

n_by_age %>%
  mutate(survey_freq = n/sum(n)) %>%
  dplyr::select(-n) %>%
  right_join(acs, by = "age_strata") %>%
  print()
```

```
## # A tibble: 5 x 3
##   age_strata survey_freq pop_2016_freq
##   <fct>            <dbl>         <dbl>
## 1 [0,55)          0.617         0.703
## 2 [55,65)         0.143         0.123
## 3 [65,75)         0.141         0.100
## 4 [75,85)         0.0735        0.0531
## 5 [85,Inf]        0.0259        0.0210
```

```
r <- deaths %>%
  mutate(death_after = mo >= 9.2 ) %>%
  mutate(age_strata = cut(age, cuts,
                          include.lowest = TRUE, right = FALSE)) %>%
  group_by(age_strata) %>%
  summarize(tot_before = sum(!death_after), tot_after = sum(death_after)) %>%
  ungroup() %>%
  left_join(n_by_age, by = "age_strata") %>%
  mutate(rate_before = tot_before/n*1000/years_before,
         rate_after = tot_after/(n-tot_before)*1000/years_after,
         ratio = rate_after / rate_before)

adjusted_rates <- r %>%
    left_join(acs, by = "age_strata") %>%
```

```
    summarize(rate_after = sum(rate_after*pop_2016_freq))

se_after <- adjusted_rates$rate_after/sqrt(sum(hh_deaths$tot_after))
adjusted_rates$after_lower <- adjusted_rates$rate_after - 1.96*se_after
adjusted_rates$after_upper <- adjusted_rates$rate_after + 1.96*se_after

print(adjusted_rates)
```

```
## # A tibble: 1 x 3
##   rate_after after_lower after_upper
##        <dbl>       <dbl>       <dbl>
## 1       12.5        8.50        16.4
```

We can see that once we adjust to the 2016 ACS age distribution the estimated rate decreases.

We see that adjusting for household size increases the estimated rate, while adjusting for age decreases the estimated rate. We are currently working on an adjustment that does both.

Did you consider the variability of the baseline estimate?

In the paper we used historical data to construct a plug-in $\hat{r}_0$. Specifically we used the following code:

```
official_deaths <- readRDS("../data/rdata/deaths_official.RDS")

baseline_deaths <- official_deaths %>%
  subset(Year == 2016) %>%
  dplyr::select(Sep,Oct,Nov,Dec) %>%
  mutate(Sep = Sep*(1/3)) %>% sum

baseline_pop <- official_deaths %>%
  subset(Year == 2016) %>%
  .$Popv17 * years_after

print(baseline_deaths/baseline_pop * 1000)
```

```
## [1] 8.82
```

We can also estimate the standard error of this estimated rate:

```
print(sqrt(baseline_deaths/(baseline_pop^2)) * 1000)
```

```
## [1] 0.0963
```

Note that this is less than 5% of the variability associated with the survey estimate:

```
survey_deaths <- deaths %>%
  subset(as.numeric(died_month) >= 10 |
           (died_month == 9 & died_b_p_hurricane == 2)) %>%
  nrow
survey_pop <- (sum(hh_main$hh_size) - survey_deaths) * (102/365)

print(sqrt(baseline_deaths/(baseline_pop^2)) /
         sqrt(survey_deaths/(survey_pop^2)))
```

```
## [1] 0.0414
```

We therefore did not include it in the reported confidence interval.

In the next section we describe different approaches to estimating $\hat{r}_0$. We note that the variability across approaches is larger than the statistical variability of the estimate.

Did you consider other values for $\hat{r}_0$?

In the previous section we describe how we compute the $\hat{r}_0$ used in the paper.

We consider other possibilities which we describe here.

The historical average death rate

There is year to year variability in death rates which we can see from the table we computed earlier:

```
official_rates %>%
    filter(year < 2017) %>%
    print()
```

```
## # A tibble: 7 x 3
##    year before_rate after_rate
##   <dbl>       <dbl>      <dbl>
## 1 2010.        7.74       8.19
## 2 2011.        8.17       8.03
## 3 2012.        8.20       8.15
## 4 2013.        8.18       7.92
## 5 2014.        8.20       9.28
## 6 2015.        8.10       8.09
## 7 2016.        8.50       8.82
```

We considered using the average:

```
official_rates %>%
    filter(year < 2017) %>%
    summarize(ave=mean(after_rate)) %>%
    print()
```

```
## # A tibble: 1 x 1
##     ave
##   <dbl>
## 1  8.36
```

But because there appears to be an upward trend, we decided using the average death rate could result in an underestimate.

The historical average death rate in the barrios included in the survey

To the best of our knowledge, barrio-level death rates are not publicly available.

The historical average death rate in the municipalities included in the survey

To the best of our knowledge, the municipality level data is not publicly available. We examined municipality level data for 2005-2015 using restricted-access multiple cause of death data from the National Center for Health Statistics, which is a similar but not identical dataset. Unfortunately, these data are not public-use and cannot be shared.

However, in our analyses, we noticed that the municipality-to-municipality variability was not large. We also noted that some municipalities had very small populations which could introduce another level of variability. We therefore decided against using this approach.