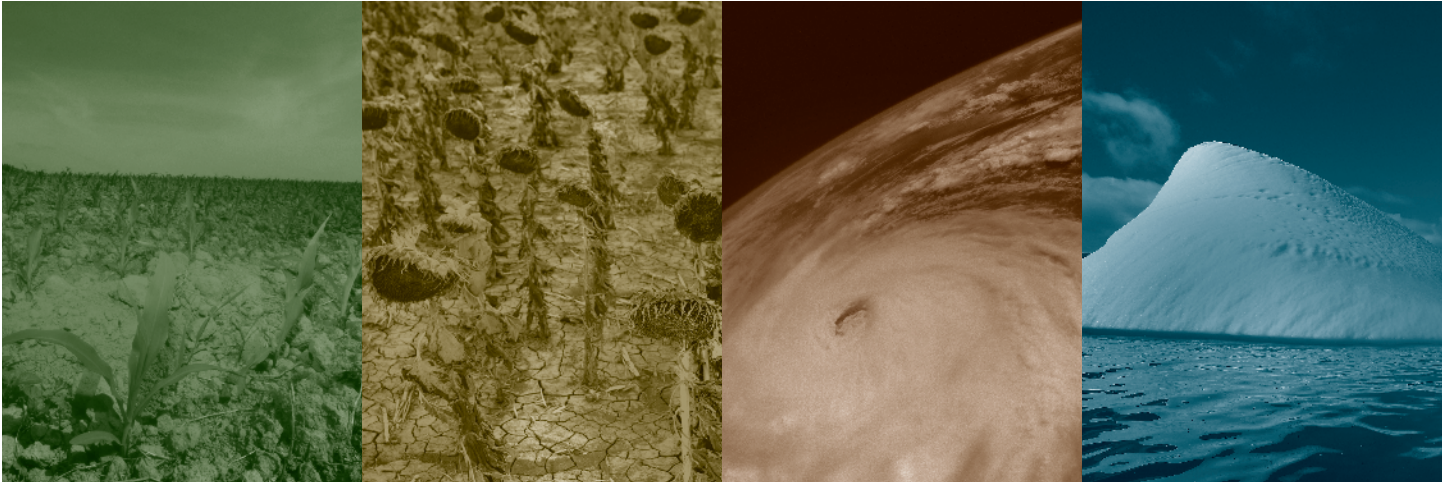# Projecting Climate-Related Disease Burden:

A Case Study on Methods for Projecting Respiratory Health Impacts

Climate and Health Technical Report Series

Climate and Health Program, Centers for Disease Control and Prevention

Wan-Hsiang Hsu[1], Syni-An Hwang[1,2], Shubhayu Saha[3], Paul J. Schramm[3], Alissa R. Van Zutphen[1,2], Kathryn Conlon[3], Shao Lin[4]

[1]Bureau of Environmental and Occupational Epidemiology, Center for Environmental Health, New York State Department of Health.

[2]Department of Epidemiology and Biostatistics, School of Public Health, University at Albany, State University of New York.

[3]Climate and Health Program, Division of Environmental Hazards and Health Effects, National Center for Environmental Health, Centers for Disease Control and Prevention.

[4]Department of Environmental Health Science, School of Public Health, University at Albany, State University of New York.

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

National Center for Environmental Health
Division of Environmental Hazards and Health Effects

CDC

1. Background

This document is a follow up to the "Projecting Climate-Related Disease Burden" technical report that was developed by the Climate and Health Program at the Centers for Disease Control and Prevention (CDC).[1] Using case studies (completed and in development) from Climate Ready States and Cities Initiative (CRSCI) grantees utilizing the Building Resilience Against Climate Effects (BRACE) framework, these guidance documents are examples of specific methodology health departments can utilize to project climate-related disease burden.

The example in this document provides one approach to estimate the excess current and future public health burden related to respiratory hospitalizations attributable to extreme heat in summer in New York State (NYS). A data set with daily observations for NYS from 1991–2004 is used as the baseline; disease burden is projected for the years 2046–2065 (midcentury) and 2080–2099 (end of century). These dates were chosen based on available climate projections and a time frame in which changes in disease burden would be evident (see sections 4.1 and 4.2). The methodology in this example is similar to that in a paper published in *Environmental Health Perspectives,*[2] but uses updated climate projections. The document is organized as follows: Section 2 describes the required data sets and their organization and linkage for the analysis; Section 3 describes the derivation of the exposure-outcome function (i.e., the estimation of the change in the number of adverse health effects in a given population for a given change in concentration for a given exposure); and Section 4 presents the approach for projecting the disease burden. Note that all data used here is for demonstration purposes only and not the actual data.

2. Assembling the data elements and estimating the baseline disease burden

To determine the future impact of climate change on respiratory diseases, we must first estimate the baseline respiratory disease burden (Section 2) and develop the exposure-outcome function between respiratory diseases and temperature (Section 3). This activity requires two primary data elements: 1) health endpoint data, such as respiratory hospitalizations and 2) meteorological data as our exposure measurement. In this section, we describe how to extract and organize health and exposure data, conduct the data linkage, and estimate the baseline respiratory disease burden.

2.1. Extraction and organization of the health data

The health data set used is the New York State Department of Health's Statewide Planning and Research Cooperative System (NYSDOH SPARCS), a hospital discharge data set which includes principal diagnosis, hospital admission date, sources of payment, date of birth, sex, race, ethnicity, total charges, length of stay, and patient residential street address. We first extracted daily summer respiratory hospitalization records at baseline (June-August, 1991–2004) according to the *International Classification of Disease, 9th Revision, Clinical Modification* (ICD-9-CM) (respiratory diseases: 491–493, 496, and 490 and 466 for child age < 5 years) and geocoded patient residential address to the street level. For geocoding, ArcGIS (ESRI) online geocoding services or other standard geocoding packages can be used. Next, we aggregated individual admissions to daily hospitalization counts spatially for each of the 14 NYS weather regions described by Chinery and Walker (2009).[3] These 14 NYS weather regions were thus the spatial unit of analysis in this example. We also summed daily hospitalizations for selected socio-demographic characteristics (e.g., sex, race, ethnicity) for the purpose of stratified analyses.

As an example, five randomly selected observations from the initial health data set (named ***RESP***) are displayed in table 1 below with selected variables and geocoded results.

- DATE: hospital admission date

- REGION: NYS weather region

- ICD9: ICD-9-CM codes

- ETHN: patient's ethnicity

- LENSTAY: days hospitalized

- RACE: patient's race

- SEX: patient's immutable biological classification as either male or female

- TCHARG: total charges

OBS	DATE	REGION	ICD9	DOB	ETHN	LENSTAY	RACE	SEX	TCHARG
1	06/01/91	1	496	1930/01/01	2	7	01	M	7907
2	07/13/91	1	496	1956/12/15	2	5	01	M	3536
3	08/20/91	1	496	1942/01/20	2	7	01	M	7359
4	07/15/91	2	493	1965/02/05	2	5	01	M	2708
5	08/15/91	1	493	1940/01/02	1	52	88	M	51740

Table 1: Example of Initial Health Data Set

Next, to obtain the daily counts of respiratory hospitalizations for each region, we sorted this health data set *(RESP)* by DATE and REGION and summed the observations for each date and each region, creating the data set named *DAILY_RESP.* SAS code for creating this health data set is in Appendix SAS code 1. As an example, Table 2 below displays daily counts of respiratory hospitalizations for five selected days. Note that the individual–level information (e.g., sex, race, ethnicity) was lost (Table 1 vs. Table 2) when we aggregated individual observations into the daily counts by DATE and REGION. The *DAILY_RESP* data set is composed of complete counts **(DAILY_TOTAL)** for each day in June-August, 1991–2004. This data set has the required format for conducting time series analyses. **SUM_TCHARG** are the daily sum of the total charges and **SUM_LENSTAY** is days hospitalized on a single date for each region. These two variables are used as economic costs to translate estimated excess respiratory admissions due to climate change into economic burden in Section 4. For example, for **Obs 1** there were a total of 80 respiratory hospitalizations on June 1, 1991 in region 1 with total charges of $370,078 and 336 days of hospitalization for the 80 individuals.

To perform stratified analysis by an additional variable (e.g., patient's sex) or variables, the data set must be aggregated by the additional variable(s). Table 3 is an example of daily counts of respiratory hospitalizations sorted and summed by **DATE, REGION,** and **SEX.**

OBS	DATE	REGION	DAILY_TOTAL	SUM_TCHARG	SUM_LENSTAY
1	06/01/91	1	80	370078	336
2	06/02/91	1	15	90967	94
3	06/03/91	1	2	4621	5
4	06/04/91	1	17	97679	121
5	06/05/91	1	13	79397	82

Table 2: Example of Sorted and Summed Health Data Set

OBS	DATE	REGION	SEX	DAILY_TOTAL	SUM_TCHARG	SUM_LENSTAY
1	06/01/91	1	M	20	150024	136
2	06/01/91	1	F	60	220054	200
3	06/02/91	1	M	7	50425	44
4	06/02/91	1	F	8	40542	50
5	06/03/91	1	M	2	4621	5
6	06/03/91	1	F	0	0	0
7	06/04/91	1	M	10	60139	68
8	06/04/91	1	F	7	37540	53
9	06/05/91	1	M	8	50855	46
10	06/05/91	1	F	5	28542	36

Table 3: Example of Sorted and Summed Health Data Set for An Additional Variable (Patient's Sex)

2.2. Estimation of the baseline disease burden

The next step is to estimate the baseline disease burden from the previously established health data set. There are several possible estimates for the baseline disease burden. In our example, we estimate the baseline respiratory disease burden as the mean number of daily respiratory hospitalizations during June–August, 1991–2004. In general, the choice of estimate is based on the study design and available data; for instance, disease prevalence is often one of the chosen metrics. For heat-related mortality due to climate change, mortality rates for specific administrative units (like county, state) could be used to estimate the baseline burden.

2.3. Extraction and organization of the climate/weather data

This subsection describes the climate/weather data sources and their extraction and organization. Meteorological data can come from either monitoring sites (including ground-based monitors or satellite data) or climate models. When analyses will be conducted on large geographic areas, but with smaller analytic geographic units, such as census tracts or other sub-county areas, modeled data may be more suitable since it can provide higher resolution. If a study targets metropolitan areas, which usually have more monitoring sites, the weather data from ground-based monitors may provide comprehensive coverage. However, the choice of monitored or modeled data still depends on resources and study design.

As mentioned above, this study examined 14 NYS weather regions with relatively homogeneous weather and ozone exposures. We used the monitored meteorological data since these 14 NYS weather regions are well covered by weather stations. The meteorological data (temperature and dew point) was obtained from the Data Support Section of the Computational and Information System Laboratory at the National Center for Atmospheric Research (NCAR) (https://www2.cisl.ucar.edu/cisl-computing). Researchers can submit an online data request based on the study geography and time period. The meteorological data includes hourly temperature, barometric pressure, and dew point. In addition, measures of air pollution may need

to be considered as confounders. High ozone concentration is related to elevated temperature, and ozone exposure has been associated with respiratory disease and reduction in lung function. Hourly ambient ozone levels were obtained from NYS Department of Environmental Conservation (DEC) ambient air monitors. Data on air pollutants can also be obtained from the US Environmental Protection Agency (EPA). We controlled for an 8-hr maximum average ozone concentration limited to 10:00–18:00 as a confounder in our example; this period represents the most likely time for outdoor exposure. Both meteorological and ambient air monitoring sites were geocoded to street level. The daily average exposure was calculated for each region. If there were multiple monitoring sites in one region, we used the average value among the monitoring sites.

We used daily mean apparent temperature (AT [°C] = -2.653 + 0.994*(daily mean temperature [°C]) + 0.0153*(dew point [°C])^2) as our temperature indicator since AT is an index of human discomfort resulting from the combined effects of heat and humidity[4, 5]. The exposure data set (named **CLIMATE**) is similar in structure to the **DAILY_RESP** health data set, where each observation is a regional daily measurement of AT and ozone. The units of AT and ozone are in degrees Fahrenheit (°F) and parts per million (ppm), respectively. It is important to make note of the units of the data being used, for example °F vs °C. The exposure data are shown in Table 4 below for the same days shown in the previous table.

DATE	REGION	AT	O3
06/01/91	1	81.335	0.038
06/02/91	1	74.997	0.055
06/03/91	1	76.569	0.068
06/04/91	1	64.557	0.033
06/05/91	1	57.426	0.021

Table 4: Example Data Set with Regional Daily Apparent Temperature and 8-hr Maximum Average Ozone Concentration.

2.4. Merging the health and climate data sets

We used the common variables **(DATE** & **REGION)** in both data sets to link the health outcome data to the meteorological data. Note that both data sets **(*DAILY_RESP* &** ***CLIMATE*)** must be sorted by **DATE** and **REGION** before merging. The merged data set is named ***ALL***; SAS code for the data linkage is in Appendix SAS code 2. The merged data for the five selected days are shown below.

OBS	DATE	REGION	DAILY_TOTAL	SUM_TCHARG	SUM_LENSTAY	AT	O3
1	06/01/91	1	80	370078	336	81.335	0.038
2	06/02/91	1	15	90967	94	74.997	0.055
3	06/03/91	1	2	4621	5	76.569	0.068
4	06/04/91	1	17	97679	121	64.557	0.033
5	06/05/91	1	13	79397	82	57.426	0.021

Table 5: Example of Merged Data Set Containing Health and Meteorological Variables.

3. Derivation of the exposure-outcome function

Time series and case-crossover approaches are the most commonly used techniques for developing exposure-outcome functions to assess the change in an adverse health outcome associated with a given environmental exposure change. Exposure-outcome functions are also known as health impact functions or concentration-response functions. The health outcome in this example is the daily total count of respiratory hospitalizations. We applied time series analysis using a two-stage Bayesian model that includes a regional analysis and a statewide estimate adjusted for regional confounders. Because of the technical nature of the material, the process was broken down into steps; SAS and R code are provided in the Appendix. An alternative to using this method is to obtain exposure-outcome functions from the existing epidemiological literature.

3.1. Stage 1: regional analysis

In stage 1, we estimated the association between extreme heat and respiratory disease hospitalization for each of the 14 NYS weather regions using generalized additive models (GAMs) with Poisson distributional errors and a log link function for each of the 14 NYS regions, while controlling for possible confounders.[6] Since a U-, V-, or J-shaped relationship between temperature and health endpoints is usually found, we used a linear-threshold model to quantify the effect of high temperature.[7,8] That is, we assumed a log-linear increase in health risk above a temperature threshold (T_0), which was determined by comparing the maximum likelihood estimates over all possible threshold values in the range of data and using the value with the lowest deviance.[8,9]

We used a linear term, $AT>T_0$, to approximate the right part of the curve (extreme heat effect). The model is:

$$\log(\textbf{DAILY\_TOTAL}) = \beta_0 + \beta_1(\textbf{AT}>T_0) + s(\textbf{AT}{\leq}T_0, DF=12) + s(\textbf{DATE}, DF) + s(\textbf{O3}, DF=8) + \beta2 + \ldots + \beta9 + \varepsilon$$

where T_0 is the threshold value of AT, β_0 is the intercept term, ε is the error term, β_1 is the slope parameter for $\textbf{AT}>T_0$, and DF equals degrees of freedom. Smoothing splines, indicated by s($\textbf{AT}{\leq}T_0$, DF=12), s($\textbf{DATE}$, DF), and s($\textbf{O3}$, DF=8), are used to model the non-linear effects of $\textbf{AT}{\leq}T_0$, long-term trends and seasonal variation (**DATE**), and ozone (**O3**). DF for **DATE** was determined using an automatic procedure based on minimizing the sum of absolute values for the first 30 items of the partial autocorrelation function (PACF) of the model residuals. We also controlled for the effect of day of the week (Monday–Saturday represented by β_2–β_7), and Northeast blackout events on 08/14/06–08/15/06 (β_8, β_9). Detailed procedures are outlined in Box 1 below; the corresponding SAS code is presented in Appendix SAS codes 3-10.

1) Create sub-data sets for each region and their corresponding categorical variables for the day of the week and Northeast blackout events.

2) Search for the optimal DF for **DATE** (SAS codes 3-4).

   a. Fit GAMs with a possible range of DF for the smoothing spline of **DATE** while adjusting for above-mentioned confounders. Note that we model the smoothing spline of the full range of AT here before we identify the AT threshold.

   b. Calculate the partial autocorrelation function (PACF) of the model residuals resulting from GAMs in Step 2.a, up to 30 lags.

   c. Determine **DATE**'s DF based on minimizing the sum of absolute values for the first 30 items of the PACF.

3) Calculate the marginal prediction of **AT** and the linear prediction values of daily summer respiratory admissions given the optimal DF for **DATE** derived from Step 2 (SAS code 5).

4) Search for the optimal AT threshold (T_0) (SAS codes 6-9).

   a. Under each AT threshold, for those with **AT**>T_0, calculate the modified marginal prediction of **AT** (the linear effect of AT after threshold) for a possible range of a slope of **AT**>T_0, the linear prediction values of daily summer respiratory admissions, the deviances and log-likelihood estimates.

   b. Select a slope and its corresponding 95% CI based on maximum log-likelihood estimate.

   c. For those with **AT**≤T_0, calculate deviances and log-likelihood estimates. Note that because of **AT**≤T_0, it is not necessary to calculate the modified marginal prediction of **AT** and linear prediction values of daily summer respiratory admissions.

   d. Sum the total deviances and log-likelihood estimates from those with **AT**>T_0 and **AT**≤T_0.

   e. Determine AT threshold with minimum deviance.

5) Estimate the AT effects on respiratory hospitalization (SAS code 10).

   a. Create new variables, **HT** for **AT**>T_0 and **LT** for **AT**≤T_0, given an identified AT threshold (T_0).

   b. Fit GAMs with new variables given an optimal **DATE**'s DF.

   c. Report the regional modeled estimates as the percent change in respiratory hospitalizations per 1°F AT increase above threshold.

Box 1: Procedure Used to Perform Regional Analysis of Association Between Extreme Heat and Respiratory Disease Hospitalization

3.2. Stage 2: statewide analysis

In stage 2, we pooled region-specific estimates to generate a statewide estimate using a Bayesian hierarchical model. We controlled for region-level covariates using yearly data estimated from the 1990 and 2000 U.S. Census, including population density, health-care access (minimum distance to clinics), race and ethnicity (percent of black and Hispanic residents), percent of residents with ≤ high school education, mean AT during June–August, percent living below the poverty level, and percent of the regional population that was elderly (age ≥ 75 years) and living alone. The pooling of information across regions improves statistical power and generalizability of results and controls for geographic heterogeneity in effects.[10, 11] All stage analyses were conducted using the *tlnise* function in R.[12] Detailed procedures follow; the corresponding R code is presented in Appendix R code 1.

1) Link regional modeled estimates and their corresponding standard errors from the Stage 1 analysis with region-level covariates. Note that region-level covariates are mean centered. The partial data (named **STAGE2**) for this example are as follows:

REGION	ESTIMATE	SE	DIS_CLINIC	PCT_ELDERLY	POP_DEN	…	PCT_POVERTY
1	0.0989	0.0128	-2.49	-2.25	32308.71		8.92
2	0.0911	0.0229	-2.16	-0.46	14913.16		6.02
3	0.0953	0.0188	-1.77	-1.94	2499.28		-1.77
4	0.0355	0.0299	0.07	-0.16	-2753.32		-6.04
…							
13	0.1014	0.0353	1.47	-2.24	-4706.85		-3.93
14	0.0059	0.0478	-0.45	0.71	-4799.81		-2.39

- ■ REGION: NYS weather region
- ■ SE: standard error of β_1 from stage 1
- ■ PCT_ELDERLY: % elderly
- ■ PCT_POVERTY: % poverty
- ■ DIS_CLINIC: minimum distance to clinics
- ■ ESTIMATE: (β_1) from stage 1
- ■ POP_DEN: population density

2) Run *tlnise* function in R *tlnise* package.

3) Calculate the pooled estimate as the percent change in respiratory hospitalizations per 1°F AT increase above threshold and its corresponding 95% confidence interval (CI).

Box 2: Procedure Used to Generate a Statewide Estimate of Association Between Extreme Heat and Respiratory Disease Hospitalization

4. Projection of the disease burden

The final step of the process is to use climate projection data to project future disease burden. A climate projection shows how climate might change several decades to centuries in the future if certain scenarios develop. We first describe the rationale for the choice of climate projections, the acquisition and organization for climate projection data sets, and how to incorporate population growth, and finally provide an approach to estimate future disease burden, including hospital admissions and monetary value.

4.1. Rationale for the choice of climate projections

Analysis of the potential impacts of climate change on communities requires information about the changing climate. Global climate models (GCMs) generate climate projections based on estimates of future greenhouse gas emissions. An ensemble of multiple GCMs has generally been found to provide a better forecast than any individual model.[6] The Intergovernmental Panel on Climate Change (IPCC) has defined a series of scenarios that represent a range of possible future trends in carbon emissions for the whole of the 21st century. Each scenario is based on a set of assumptions about socioeconomic growth, technology development, land-use change, and energy demand. The fourth assessment report of the IPCC (AR4 IPCC) followed the method described in the Special Report on Emission Scenarios (SRES).[13] The method called for sequentially developing socioeconomic and emissions scenarios and then carrying out climate change projections based on those scenarios.[13-15] SRES scenarios have four storylines and a total of 40 scenarios. Recently, the fifth assessment report of the IPCC (AR5 IPCC) followed a slightly different approach for scenario development. Four representative concentration pathways (RCP8.5, RCP6, RCP4.5 and RCP2.6) were derived in a parallel process with emission and socioeconomic scenarios.[14, 15] Furthermore, in a relatively new climate modeling development, there are downscaled climate models with smaller-scale outputs achieved by using statistical techniques or dynamic weather modeling. Each type of model has advantages and disadvantages. GCMs simulate how the earth's climate might change, whereas downscaled climate models generate more locally relevant projections for communities and cities. Ideally, the choice of GCMs or downscaled outputs depends upon the analytic geographic unit, time periods, meteorological variables of interest, and resources. For example, if one is interested in estimating public health burden due to extreme weather events, such as heavy rains, droughts, and hurricanes, which occur at the local or regional level, downscaled climate models would be preferable since these models simulate these weather patterns on a finer resolution by incorporating region-specific environmental processes.

4.2. Obtain and organize the climate projection data for the exposure of interest

The Data Distribution Centre of the IPCC and the Research Data Archive of NCAR provide freely available climate projections with a variety of meteorological variables, temporal coverage, and emission scenarios.[16, 17] These climate projection data sets have been used in many well-documented studies/reports. For example, the AR4 and AR5 of the IPCC and the study by Lin et al. (2012).[2, 13, 15]

In this example, we estimated future AT by using temperature and precipitation obtained from Downscaled CMIP3 and CMIP5 Climate and Hydrology Projections.[18] The downscaled outputs of the CMIP5 (Coupled Model Intercomparison Project Phase 5) multi-model ensemble was structured with daily meteorological variables by longitude and latitude (grid size~12 km x 12 km).[19] Similar to what was done in section 2.3 with daily climate/weather ("exposure") data, we utilized the longitude and latitude in climate projections to estimate spatially averaged mean summer daily AT for each region during 2046–2065 and 2080–2099 under the selected climate scenarios. We chose RCP 4.5 and 8.5 since they represent the intermediate emissions and the highest emissions scenarios, respectively. Future AT data sets for 2046–2065 and 2080–2099 (named **AT65** and **AT99**, respectively) with five selected days are shown in Table 6 below, where the first row represents the AT values in Region 1 on 6/1/46 under scenarios of RCP 4.5 and RCP 8.5.

DATE	REGION	RCP4.5_AT	RCP8.5_AT
06/01/46	1	81.335	85.375
06/02/46	1	74.997	82.937
06/03/46	1	76.569	80.869
06/04/46	1	64.557	69.587
06/05/46	1	57.426	62.436
06/01/80	1	85.335	87.445
06/02/80	1	76.955	84.955
06/03/80	1	79.578	83.877
06/04/80	1	66.537	70.555
06/05/80	1	59.226	64.411

Table 6: Example of Future Apparent Temperature (AT) under RCP4.5 and RCP8.5 in 2046 and 2080.

4.3.   Obtain projection information on the population of concern

This case study assumed that the population remains constant into 2100; however, researchers could easily incorporate population growth into the projections. For example, we could use the population for New York from the 2000 Census (18,976,457) as our baseline population (1991–2004).[20] Martin et al. estimated the 2050 New York State population as 29,301,275.[21] The population growth rate is 1.5. This growth rate can be applied to the estimated disease burden as shown in the next section.

4.4.   Derive estimates of the future burden

We estimated the excess admissions attributable to extreme heat at baseline (1991–2004) and project for 2046–2065 and 2080–2099 as: $A = R \times \Delta T \times M$, where A represents the estimated number of daily admissions attributable to $AT > T_0$ for each time period; R is the estimated percentage increase in admissions per 1°F increase in $AT > T_0$ at baseline based on the exposure-outcome function in Section 3 [i.e., $R = \exp(\beta_1) - 1$]; M is the mean number of daily respiratory admissions during June–August at baseline; and ΔT is the difference between the mean daily AT for each time period and T_0 on days when $AT > T_0$. We used the thresholds identified from the baseline data (1991–2004) for each region to project future impacts.

We also estimated excess hospitalizations due to extreme heat using medical charges for treatment of illness and hospitalization days, which were estimated as: $C = A \times D$, where C represents either daily temperature-attributable hospitalization charges or days hospitalized; A is as defined in the previous paragraph; and D is the average number of days hospitalized per hospitalization at baseline or the average charge per hospitalization adjusted for inflation for each time period and standardized to 2004 U.S. dollars to ensure that charges are comparable across the different years in this case study.[22] Because a dollar in the future is considered to be of less value than a current dollar, it is common practice in health cost estimation to discount for future costs incurred across different time periods to derive the current worth of all future amounts. We used an annual discount rate of 3% as recommended by the U.S. Panel on Cost-Effectiveness in Health and Medicine to estimate the 2004 dollar value of the future stream of costs.[23, 24] The detailed procedures follow; the corresponding SAS code is in Appendix SAS codes 11–13.

1) Calculate the mean number of daily respiratory hospitalizations, the average number of days hospitalized per respiratory admission, and the average cost per respiratory admission with inflation adjustment during June-August at baseline for each region.

2) Compute the AT difference between the mean daily AT and threshold for days with $AT>T_0$ and present it as a yearly average.

3) Link data sets resulting from Step 1) and Step 2) and risk estimates from Section 3.

4) Compute excess admissions, days hospitalized, and charges per year and adjust future charges by a discount rate of 3%.

5) To incorporate the population growth, multiply the estimated disease burden by the population growth rate.

Box 3: Procedure Used to Estimate Future Disease Burden and Future Health Costs

References

1. Hess JJ, Saha S, Schramm PJ, Conlon KC, Uejio CK, Luber G. Projecting Climate-Related Disease Burden: A Guide for Health Departments. 2015.

2. Lin S, Hsu W-H, Van Zutphen AR, Saha S, Luber G, Hwang S-A. Excessive heat and respiratory hospitalizations in New York State: estimating current and future public health burden related to climate change. Environmental health perspectives. 2012;120(1):1571-7.

3. Chinery R, Walker R. Development of exposure characterization regions for priority ambient air pollutants. Human and Ecological Risk Assessment. 2009;15(5):876-89.

4. Kalkstein LS, Valimont KM. An evaluation of summer discomfort in the United State using a relative climatological index. Bulletin of the American Meteorological Society. 1986;67(7):842-8.

5. Steadman RG. The assessment of sultriness. Part I: A temperature-humidity index based on human physiology and clothing science. Journal of Applied Meteorology. 1979;18(7):861-73.

6. Tebaldi C, Knutti R. The use of the multi-model ensemble in probabilistic climate projections. Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences. 2007;365(1857):2053-75.

7. Linares C, Diaz J. Impact of high temperatures on hospital admissions: comparative analysis with previous studies about mortality (Madrid). The European Journal of Public Health. 2008;18(3):317-22.

8. McMichael AJ, Wilkinson P, Kovats RS, Pattenden S, Hajat S, Armstrong B, et al. International study of temperature, heat and urban mortality: the 'ISOTHURM' project. International journal of epidemiology. 2008;37(5):1121-31.

9. Armstrong B. Models for the relationship between ambient temperature and daily mortality. Epidemiology. 2006;17(6):624-31.

10. Dominici F. Invited Commentary: Air Pollution and Health—What Can We Learn from a Hierarchical Approach? American journal of epidemiology. 2002;155(1):11-5.

11. Dominici F, Daniels M, Zeger SL, Samet JM. Air pollution and mortality: estimating regional and national dose-response relationships. Journal of the American Statistical Association. 2002;97(457):100-11.

12. Everson P. Two-level normal independent sampling estimation (TLNise). Swarthmore, PA: Swarthmore College. 2000.

13. Change IPOC. Climate change 2007: The physical science basis. Agenda. 2007;6(07):333.

14. Moss RH, Babiker M, Brinkman S, Calvo E, Carter T, Edmonds JA, et al. Towards new scenarios for analysis of emissions, climate change, impacts, and response strategies. Pacific Northwest National Laboratory (PNNL), Richland, WA (US), 2008.

15. Pachauri RK, Allen M, Barros V, Broome J, Cramer W, Christ R, et al. Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. 2014.

16. Data Distribution Centre (DDC) [Internet]. Available from: http://www.ipcc-data.org/

17. Research Data Archive [Internet]. National Center for Atmospheric Research. Available from: http://rda.ucar.edu/

18. Eccel E. Estimating air humidity from temperature and precipitation measures for modelling applications. Meteorological Applications. 2012;19(1):118-28.

19. Brekke L, Thrasher B, Maurer E, Pruitt T. Downscaled CMIP3 and CMIP5 climate and hydrology projections: Release of downscaled CMIP5 climate projections, comparison with preceding information, and summary of user needs. US Dept of the Interior, Bureau of Reclamation, Technical Services Center, Denver. 2013.

20. Administration USDoCEaS. New York: 2000. 2002.

21. Martin J, Fogel S. Projecting the US Population to 2050. FAIR March. 2006.

22. CPI Inflation Calculator [Internet]. [cited 12/22/2015]. Available from: http://www.bls.gov/data/inflation_calculator.htm

23. Goodman CS. Introduction to health technology assessment. Retrieved from United States National Library of Medicine, National Institutes of Health website: https://www.nlm.nih.gov/nichsr/hta101/HTA_101_FINAL_7-23-14.pdf. 2004.

24. Phillips KA, Chen JL. Impact of the US panel on cost-effectiveness in health and medicine. American journal of preventive medicine. 2002;22(2):98-105.

Appendix

CODE	DESCRIPTION	SECTION
SAS code 1	Create the daily summer hospitalization data set	Table 2
SAS code 2	Link and merge daily summer hospitalization and climate data sets	Table 4
SAS code 3	Calculate the sum of the absolute values of the PACF up to 30 lags for each DATE's DF	Section 3.1, Box 1
SAS code 4	Determine DATE's DF based on the minimum of the sum of the absolute values of the PACF up to 30 lags	Section 3.1, Box 1
SAS code 5	Calculate the marginal prediction of AT and the linear prediction values of daily summer hospitalization	Section 3.1, Box 1
SAS code 6	Calculate log of factorial for the upcoming calculation of log-likelihood estimates	Section 3.1, Box 1
SAS code 7	Calculate deviances and log-likelihood estimates	Section 3.1, Box 1
SAS code 8	Repeat SAS code 7 with threshold (60°F−100°F) and slope (-5% to +5%)	Section 3.1, Box 1
SAS code 9	Select threshold with the minimum deviance within 95% interval	Section 3.1, Box 1
SAS code 10	Run final models to estimate the effects of AT > threshold	Section 3.1, Box 1
SAS code 11	Estimate the mean number of daily summer respiratory admissions, days hospitalized per admission and the average charge per admission adjusted for yearly inflation rate at baseline.	Section 4.4, Box 3
SAS code 12	Calculate the yearly averaged difference between AT and threshold during the baseline, 2046-2065, and 2080-2099	Section 4.4, Box 3
SAS code 13	Project burden including excess hospital admission, charges, and length of stay	Section 4.4, Box 3
R code 1	Conduct stage 2 analysis	Section 3.2, Box 2

SAS code 1

```
*CREATE DAILY SUMMER (FOR MONTHS OF JUNE, JULY AND AUGUST) RESPIRATORY ADMISSIONS FOR
NYS DURING 1991-2004 (this is what produces Table 2);
proc sql;
create table DAILY_RESP as
select DATE, REGION, count(DATE) as DAILY_TOTAL, sum(TCHARG) as SUM_TCHARG,
sum(LENSTAY) as SUM_LENSTAY
from RESP
where month(DATE) between 6 and 8
group by DATE, REGION;
quit;
```

SAS code 2

```
*MERGE DATA SETS OF DAILY_RESP AND CLIMATE (this is what produces Table 4);
proc sort data=DAILY_RESP; by DATE REGION; run;
proc sort data=CLIMATE; by DATE REGION; run;
data ALL;
merge DAILY_RESP CLIMATE(in=r);
by DATE REGION;
if r;
run;
```

SAS code 3

```
%macro pacf;
      %do ID=1 %to 14;
      *CREATE SUB-DATA SETS FOR EACH REGION AND THEIR CORRESPONDING CATEGORICAL
      VARIABLES;
      data ALL_REG;
      set ALL;
      where REGION=&ID;
      DOW=weekday(DATE);
      if DATE='14Aug2003'd then B8=1; else B8=0;
      if DATE='15Aug2003'd then B9=1; else B9=0;
      run;

      %do DATE_DF=10 %to 120 %by 2; *A POSSIBLE RANGE OF DATE'S DF;
          proc datasets library=work nolist; delete EST_&ID DF_PACF_&ID;
          quit;
          *FIT GAMS FOR EACH POSSIBLE DATE'S DF;
          proc gam data=ALL_REG;
          class DOW;
          model DAILY_TOTAL=param(DOW B8 B9) spline(DATE, df=&DATE_DF)
          spline(AT, df=12) spline(O3, df=8) / link=log dist=poisson notest;
          output out=EST_&ID residual;
          run;
          *CALCULATE THE PACF;
          proc sort data=EST_&ID; by DATE; run;
          proc arima data=EST_&ID;
          identify var=R_DAILY_TOTAL outcov=PCORR nlag=30 noprint;
          run;
          quit;
          *SUM THE ABSOLUTE VALUES OF THE PACF UP TO 30 LAGS;
          proc sql;
          create table DF_PACF_&ID as
          select &DATE_DF as DATE_DF, sum(abs(PARTCORR)) as PACF
          from PCORR
          where 1 le lag le 30;
          quit;
```

```
                proc append base=DF_&ID. data=DF_PACF_&ID.; run;
        %end;
        dm log 'clear';
        dm output 'clear';
%end;
%mend;
```

SAS code 4

```
*DETERMINE DATE'S DF BASED ON THE MINIMUM OF THE SUM OF ABSOLUTE VALUES OF THE PACF
WITH 30 LAGS;
%macro define_min;
%do ID=1 %to 14;
proc sql;
create table DATA&ID as
select &ID as REGION, DATE_DF, PACF
from DF_&ID
where PACF=(SELECT min(PACF) from DF_&ID);
quit;
proc append base=MIN_PACF data=DATA&ID; run;
%end;
%mend;
%define_min;
```

SAS code 5

```
*CALCULATE THE MARGINAL PREDICTION OF AT AND THE LINEAR PREDICTION VALUES OF DAILY
SUMMER RESPIRATORY ADMISSIONS;
%macro AT_effect;
%do ID=1 %to 14;
        data _null_;
              set MIN_PACF;
              where REGION=&ID;
              call symput('DATE_DF',DATE_DF);
        run;

        data ALL_REG;
        set ALL;
        where REGION=&ID;
        DOW=weekday(DATE);
        if DATE='14Aug2003'd then B8=1; else B8=0;
        if DATE='15Aug2003'd then B9=1; else B9=0;
        run;

        ods output parameterEstimates=LINEAR;
        proc gam data=ALL_REG;
        class DOW;
        model DAILY_TOTAL=param(DOW B8 B9) spline(DATE, df=&DATE_DF)
        spline(AT, df=12) spline(O3, df=8) / link=log dist=poisson notest;
        output out=OUT P LINP;
        run;

        data _null_;
        set LINEAR;
        where parameter="Linear(at)";
        call symput('Linear',ESTIMATE);
        run;

        data P_&ID;
        set OUT;
        *THE MARGINAL PREDICTION OF AT=LINEAR + SPLINE;
```

```
        EFFECT=&linear.*AT + P_AT;
        run;
%end;
%mend;
%AT_effect;
```

SAS code 6

```
*CALCULATE LOG OF FACTORIAL FOR THE BELOW CAUCULATION OF LOG-LIKELIHOOD ESTIMATES;
%macro fact;
data test;
do i=1 to 240;
        total+log(i);
        FACT='FACT'||compress(put(i,3.));
        output;
end;
run;
data _null_; set test; call symput(FACT,TOTAL); run;
%mend;
%fact;
```

SAS code 7

```
%macro gam_dev_cutp(EST, T_MIN, T_MAX, SL_MIN, SL_MAX, SL_INCR, SEG);
proc datasets library=work nolist; delete &EST._DEVIANCE_DIRECT; run; quit;

%do THRESHOLD =&T_MIN %to &T_MAX; *A POSSIBLE RANGE OF THRESHOLD;
        *SELECT THE MARGINAL PREDICTION OF AT~THRESHOLD AS THE LINEAR-ADJUSTED
        NONPARAMETRIC PREDICTION;
        proc sql;
        create table CLOSEAT as
        select EFFECT as CLOSEAT
        from &EST
        where AT=(select max(AT) from &EST where AT<&THRESHOLD);
        quit;

        data _null_; set CLOSEAT; call symput('CLOSEAT', CLOSEAT); run;

        proc datasets library=work nolist; delete DEV2; run; quit;

        %do SL=&SL_MIN %to &SL_MAX %by &SL_INCR; *A POSSIBLE RANGE OF SLOPE;
        %let SLOPE=%sysevalf(&SL/&SEG);
        %put THRESHOLD=&THRESHOLD SLOPE=&SLOPE;
                *FOR THOSE WITH AT > THRESHOLD, CALCULATE THE MODIFIED MARGINAL PREDICTION
                OF AT AND THE LINEAR PREDICTION VALUES OF DAILY SUMMER RESPIRATORY
                ADMISSIONS;
                data test2;
                set &EST;
                where AT > &THRESHOLD;
                EFFECT_AT=&CLOSEAT + &SLOPE.*(AT-&THRESHOLD);
                P_DAILY_TOTAL = P_DAILY_TOTAL - EFFECT + EFFECT_AT;
                run;
                *FOR THOSE WITH AT > THRESHOLD, CALCULATE DEVIANCES AND LOG-LIKELIHOOD
                ESTIMATES;
                data test2;
                set test2;
                S2=deviance('poisson', DAILY_TOTAL, exp(LINP_DAILY_TOTAL));
                if DAILY_TOTAL le 0 then L2=0;
                else do;
                        call symput('FACTN', 'FACT'||compress(put(DAILY_TOTAL, 3.)));
                        L2=DAILY_TOTAL*LINP_DAILY_TOTAL-exp(LINP_DAILY_TOTAL)-
                        symget(symget('FACTN'));
```

```
    end;
    run;
    *SUM DEVIANCES AND LOG-LIKELIHOOD ESTIMATES AMONG OBSERVATIONS WITH AT >
    THRESHOLD;
    proc means data=TEST2 noprint;
    var S2 L2;
    output out=DEVIANCE2(drop_type_ _freq_) sum=;
    run;

    data DEVIANCE2; set DEVIANCE2; SLOPE=&SLOPE; run;

    proc append base=DEV2 data=DEVIANCE2; run;

dm "out;clear;log;clear;";
%end;

*UNDER EACH THRESHOLD, SELECT SLOPE WITH THE MAXIMUM LOG-LIKELIHOOD ESTIMATE;
proc sql;
create table DEV2_THRESHOLD as
select &THRESHOLD as THRESHOLD, *
from dev2
where L2= (select max(L2) from DEV2);
quit;
*THE 95% CI FOR SLOP;
proc sql;
create table CI as
select SLOPE
from DEV2
where L2+1.92 > (select max(L2) from DEV2);
quit;
proc means data=CI noprint;
var SLOPE;
output out=SLOPE_CI(drop_type_ _freq_) min=L_SLOPE max=U_SLOPE;
run;
data DEV2_THRESHOLD;
merge DEV2_THRESHOLD SLOPE_CI;
if L_SLOPE gt SLOPE then L_SLOPE=SLOPE;
if U_SLOPE lt SLOPE then U_SLOPE=SLOPE;
run;

*FOR THOSE WITH AT <= THRESHOLD, CALCULATE DEVIANCE AND LOG-LIKELIHOOD
ESTIMATES;
data test1;
set &EST;
where AT le &THRESHOLD;
S1=deviance('poisson', DAILY_TOTAL, exp(LINP_Daily_TOTAL));
if DAILY_TOTAL le 0 then L1=0;
else do;
call symput('FACTN', 'FACT'||compress(put(DAILY_TOTAL, 3.)));
L1=DAILY_TOTAL*LINP_DAILY_TOTAL-exp(LINP_DAILY_TOTAL)-symget(symget('FACTN'));
end;
run;
*SUM DEVIANCES AND LOG-LIKELIHOOD ESTIMATES AMONG OBSERVATIONS WITH AT <=
THRESHOLD;
proc means data=TEST1 noprint;
var S1 L1;
output out=DEVIANCE1(drop_type_ _freq_) sum=;
run;
*SUM DEVIANCE AND LOG-LIKELIHOOD ESTIMATES FOR OBSERVATIONS WITH AT > THRESHOLD
AND AT <= THRESHOLD;
data DEV_THRESHOLD;
merge DEVIANCE1 DEV2_THRESHOLD;
```

```
            L_DEV=S1;
            R_DEV=S2;
            T_DEV=L_DEV+R_DEV;
            L_LL=L1;
            R_LL=L2;
            T_LL=L1+L2;
            drop L1 L2 S1 S2;
            run;

            proc append base=&EST._DEVIANCE_DIRECT data=DEV_THRESHOLD; run;

    %end;
    %mend;
```

SAS code 8

```
*REPEAT SAS code 7 FOR 14 NYS WEATHER REGIONS WITH THRESHOLD (60F-100F) AND SLOPE (-5%
to +5%);
%macro search;
%do ID=1 %to 14;
%gam_dev_cutp(P_&ID.,60,120,-50,50,1,1000);
%end;
%mend;
%search;
```

SAS code 9

```
*SELECT THRESHOLD WITH THE MINIMUM DEVIANCE WITHIN 95% INTERVAL;
%macro threshold;
%do ID=1 %to 14;
proc sql;
create table THRESHOLD_&ID. as
select &ID. as REGION, min(THRESHOLD) as THRESHOLD
from P_&ID._DEVIANCE_DIRECT
where T_DEV le 3.84 + (select min(T_DEV) from P_&ID._DEVIANCE_DIRECT);
quit;

proc append base=THRESHOLD data=THRESHOLD_&ID.; run;
%end;
%mend;
%threshold;
```

SAS code 10

```
*ESTIMATE THE EFFECTS OF AT > THRESHOLD;
%macro estimate;
proc datasets library=work; delete linear param; run;

%do ID=1 %to 14;
        data _null_;
        set THRESHOLD;
        where REGION=&ID;
        call symput('THRSHLD',THRESHOLD);
        run;

        data _null_;
        set min_pacf;
        where REGION=&ID;
        call symput('DATE_DF',DATE_DF);
        run;
```

```
data ALL_REG;
set ALL;
where REGION=&ID;
DOW=weekday(DATE);
if DATE='14Aug2003'd then B8=1; else B8=0;
if DATE='15Aug2003'd then B9=1; else B9=0;
*DEFINE EXTREME AT;
HT=(AT-&THRSHLD)*(AT>&THRSHLD);
LT=AT*(AT<=&THRSHLD);
run;

ods output parameterEstimates=LINEAR;
proc gam data=ALL_REG;
class DOW;
model DAILY_TOTAL=param(DOW B8 B9 HT) spline(date, df=&DATE_DF)
spline(LT, df=12) spline(O3, df=8) / link=log dist=poisson notest;
run;

data _null_;
set LINEAR;
if trim(Parameter)="HT" then do;
    call symput('HT',ESTIMATE);
    call symput('SE',STDERR);
    call symput('P_VALUE',PROBT);
end;
run;

data PARAM;
REGION=&ID;
THRESHOLD=&THRSHLD;
HT=&HT;
EFFECT=exp(&HT)-1;
P_VALUE=&P_VALUE.;
L_CI=exp(&HT-&SE*1.96)-1;
H_CI=exp(&HT+&SE*1.96)-1;
run;

proc append base=REGION_EST data=PARAM; run;
%end;
%mend;
%estimate;
```

SAS code 11

```
*CALCULATE THE MEAN NUMBER OF DAILY SUMMER RESPIRATORY ADMISSIONS,
DAYS HOSPITALIZED PER HOSPITALIZATION AND THE AVERAGE CHARGE PER HOSPITALIZATION
ADJUSTED FOR INFLATION FOR EACH TIME PERIOD AT BASELINE;
proc sql;
create table M_B as
select REGION, mean(DAILY_TOTAL) as M, sum(SUM_LENSTAY)/sum(DAILY_TOTAL) as M_LEN,
sum(case
    when year(DATE)=1991 then SUM_TCHARG*1.39
    when year(DATE)=1992 then SUM_TCHARG*1.35
    when year(DATE)=1993 then SUM_TCHARG*1.31
    when year(DATE)=1994 then SUM_TCHARG*1.27
    when year(DATE)=1995 then SUM_TCHARG*1.24
    when year(DATE)=1996 then SUM_TCHARG*1.20
    when year(DATE)=1997 then SUM_TCHARG*1.18
    when year(DATE)=1998 then SUM_TCHARG*1.16
    when year(DATE)=1999 then SUM_TCHARG*1.13
    when year(DATE)=2000 then SUM_TCHARG*1.10
    when year(DATE)=2001 then SUM_TCHARG*1.07
```

```
         when year(DATE)=2002 then SUM_TCHARG*1.05
         when year(DATE)=2003 then SUM_TCHARG*1.03
          else SUM_TCHARG
         end)/sum(DAILY_TOTAL) as M_COST
from ALL
group by REGION;
quit;
```

SAS code 12

```
*CALCULATE THE YEARLY AVERAGED DIFFERENCE BEWTEEN AT AND THRESHOLD DURING THE
BASELINE, 2046-2065, AND 2080-2099;
%macro delta;
%do ID=1 %to 14;
      data _null_;
      set REGION_EST;
      where REGION=&ID;
      call symput('THRESHOLD',THRESHOLD);
      run;
      data DELTA_BASELINE;
      set ALL(where=(REGION=&ID));
      if AT > &THRESHOLD then DELTA=AT-&THRESHOLD;
      else DELTA=0;
      run;
      data DELTA_65;
      set AT65(where=(REGION=&ID));
      if AT > &THRESHOLD then DELTA_45_65=AT-&THRESHOLD;
      else DELTA_45_65=0;
      if AT > &THRESHOLD then DELTA_85_65=AT-&THRESHOLD;
      else DELTA_85_65=0;
      run;
      data DELTA_99;
      set AT99(where=(REGION=&I));
      if AT > &THRESHOLD then DELTA_45_99=AT-&THRESHOLD;
      else DELTA_45_99=0;
      if AT > &THRESHOLD then DELTA_85_99=AT-&THRESHOLD;
      else DELTA_85_99=0;
      run;
      proc sql;
      create table D_B as
      select &ID as REGION, sum(DELTA)/14 as D_B
      from DELTA_BASELINE
      group by REGION;
      quit;
      proc sql;
      create table D_65 as
      select &ID as REGION, sum(DELTA_45_65)/20 as D_45_65, sum(DELTA_85_65)/20 as
      D_85_65
      from DELTA_65
      group by REGION;
      quit;
      proc sql;
      create table D_99 as
      select &ID as REGION, sum(DELTA_45_99)/20 as D_45_99, sum(DELTA_85_99)/20 as
      D_85_99
      from DELTA_99
      group by REGION;
      quit;
      data DELTA; merge D_B D_65 D_99; by REGION; run;
      proc append base=DELTA_ALL data=DELTA; run;
%end;
%mend;
%delta;
```

SAS code 13

```
*CALCULATE BURDEN;
data FINAL_PH; merge REGION_EST(keep=REGION EFFECT) M_B DELTA_ALL; by REGION; run;
data EST_BURDEN;
set FINAL_PH;
AD_B=EFFECT*D_B*M;
C_B=AD_B*M_COST;
L_B=AD_B*M_LEN;
AD_RCP45_65=EFFECT*D_45_65*M;
C_RCP45_65=AD_RCP45_65*M_COST*((1+0.03)**50);
L_RCP45_65=AD_RCP45_65*M_LEN;
AD_RCP85_65=EFFECT*D_85_65*M;
C_RCP85_65=AD_RCP45_85*M_COST*((1+0.03)**50);
L_RCP85_65=AD_RCP45_85*M_LEN;
AD_RCP45_99=EFFECT*D_45_99*M;
C_RCP45_99=AD_RCP45_99*M_COST*((1+0.03)**100);
L_RCP45_99=AD_RCP45_99*M_LEN;
AD_RCP85_99=EFFECT*D_85_99*M;
C_RCP85_99=AD_RCP45_99*M_COST*((1+0.03)**100);
L_RCP85_99=AD_RCP45_99*M_LEN;
run;
```

R code 1

```
library(tlnise)
# DEFINE ESTIMATOR, VARIANCE, REGIONAL COVARIATES
y= STAGE2[,2]
v= STAGE2[,3]
w=as.matrix(cbind(rep(1,14), STAGE2[,-c(1:3)]))

# USING TLNISE PACKAGE TO CONDUCT STAGE 2 ANALYSIS
out=tlnise(y, v, w, prior=2, maxiter=200)

# STAGE 2 RESULT
pooled=(exp(out$gamma[1,1])-1)*100
CI_L=(exp(out$gamma[1,1]-1.96*out$gamma[1,2])-1)*100
CI_H=(exp(out$gamm a[1,1]+1.96*out$gamma[1,2])-1)*100
```