BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Using logic regression to characterize extreme heat exposures and their health associations: a time-series study of emergency department visits in Atlanta

Shan Jiang[1], Joshua L. Warren[2], Noah Scovronick[3], Shannon E. Moss[1], Lyndsey A. Darrow[4], Matthew J. Strickland[4], Andrew J. Newman[5], Yong Chen[6], Stefanie T. Ebelt[3] and Howard H. Chang[1*]

Abstract

Background: Short-term associations between extreme heat events and adverse health outcomes are well-established in epidemiologic studies. However, the use of different exposure definitions across studies has limited our understanding of extreme heat characteristics that are most important for specific health outcomes or subpopulations.

Methods: Logic regression is a statistical learning method for constructing decision trees based on Boolean combinations of binary predictors. We describe how logic regression can be utilized as a data-driven approach to identify extreme heat exposure definitions using health outcome data. We evaluated the performance of the proposed algorithm in a simulation study, as well as in a 20-year time-series analysis of extreme heat and emergency department visits for 12 outcomes in the Atlanta metropolitan area.

Results: For the Atlanta case study, our novel application of logic regression identified extreme heat exposure definitions that were associated with several heat-sensitive disease outcomes (e.g., fluid and electrolyte imbalance, renal diseases, ischemic stroke, and hypertension). Exposures were often characterized by extreme apparent minimum temperature or maximum temperature over multiple days. The simulation study also demonstrated that logic regression can successfully identify exposures of different lags and duration structures when statistical power is sufficient.

Conclusion: Logic regression is a useful tool for identifying important characteristics of extreme heat exposures for adverse health outcomes, which may help improve future heat warning systems and response plans.

* Correspondence: howard.chang@emory.edu
[1]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, USA
Full list of author information is available at the end of the article

Introduction

Extreme heat events have significant public health impacts as demonstrated, for example, by historical heat waves in Chicago [1, 2] and Europe [3, 4]. Recent epidemiology studies have also found consistent short-term associations between extreme heat events and various cause-specific mortality [5–9] and morbidity outcomes [10–12]. However, synthesizing existing evidence has been challenging because of the use of various exposure definitions in previous studies [13, 14]. How extreme temperatures become hazardous can vary across locations due to differences in societal and biological adaptations [15], as well as across outcomes due to differences in physiological mechanisms [16]. Identifying extreme heat characteristics that are most important for specific health outcomes or vulnerable subpopulations may help improve heat warning systems and response plans.

Extreme heat events are typically characterized by their exceedance over an intensity *threshold* and their sustained *duration*. For example, previous studies have defined heat waves as a period where temperature exceeds the 98th percentile over two or more consecutive days [17–19]. Many other relative (or absolute) thresholds and durations have been used to define extreme heat in health studies [20]. Furthermore, the choice of *heat metric* represents another source of variation across studies. Daily maximum and average temperatures have been the most commonly used heat metrics. But there is increasing interest in investigating apparent temperature or wet-bulb temperature [21] that may better reflect human discomfort, and minimum temperature that reflects night-time exposure [22–24].

The current approach of assessing effect heterogeneity due to exposure definitions within a study involves examining different heat metrics, different extreme thresholds, and different durations one-at-a-time. Studies often also need to take into account statistical power in exposure definition due to the low frequency of extreme heat days [25]. There has been limited work in leveraging health data directly to develop exposure definitions for extreme heat. Recently, machine learning methods have been applied to predict adverse health outcomes using meteorological variables [26]. However, the resulting algorithms can be difficult to interpret in terms of the two key characteristics of duration and intensity threshold. These approaches may also suffer from the lack of rigorous control for confounders, making results more difficult to translate into causal effects for subsequent intervention and impact analysis.

In this paper, we examine the use of *logic regression* [27–29], a machine learning method, to help identify characteristics of extreme heat events that are associated with adverse health outcomes. Logic regression estimates a decision tree constructed using Boolean combinations of binary predictors. Logic regression has been utilized extensively in genetic association studies for identifying high-dimensional interactions [30, 31], and has recently been extended to other exposures [32–34]. We show how logic regression provides a data-driven approach to construct a daily *extreme heat exposure indicator* that is binary (i.e., presence versus absence of the exposure) and can capture impacts of *sustained* extreme exposure over several days (i.e., heat waves). We evaluated the performance of the approach in simulation studies and applied the method to a 20-year time-series analysis of daily emergency department (ED) visits in Atlanta, Georgia.

Materials and methods

Atlanta emergency department visit and meteorology data, 1993–2012

Patient-level ED visits data were obtained directly from hospitals within the 20-county Atlanta metropolitan area from 1993 to 2004 and then from the Georgia Hospital Association from 2005 to 2012. For some outcomes, secondary diagnoses were also included because they showed stronger associations with temperature in a previous Atlanta analysis [35]. The selected health outcomes are internal causes (INTERN), heat illness (HEAT), ischemic stroke (STK), fluid and electrolyte imbalances (FLEL), all renal disease (RENAL), acute renal failure (ARF), all circulatory system disease (CIRC), hypertension (HT), myocardial infarction (MI), congestive heart failure (CHF), ischemic heart disease (IHD), and diabetes (DIA). Table 1 provides summary statistics of daily ED visits and ICD-9 codes for each outcome. Only admissions during the warm seasons (May 1st to September 30th) were used in this analysis because of our interest in comparing extreme heat events versus non-event warm days.

Hourly ambient air (dry-bulb) temperature, dew-point temperature, and apparent temperature were obtained at the Atlanta Hartsfield International Airport weather station from the National Climatic Data Center from 1993 to 2012. We used airport monitor due to its high-quality, complete temporal observations and central location in the study area, which has little variation in elevation. Apparent temperature in °C was defined as $-1.3 + 0.92 T + 2.2e$, where T is ambient air temperature (°C) and e is water vapor pressure (kPa) [36]. We considered six heat metrics: daily maximum (MX), minimum (MN), and average (Avg) of either dry-bulb temperature or apparent temperature (AT). For each daily temperature variable, we created binary *extreme indicators* at the 95th, the 98th or the 99th percentile thresholds based on observations over the 20-year study period. Specifically, the extreme indicator takes the value 1 when the temperature value exceeds the percentile threshold.

Table 1 Descriptive statistics for daily emergency department visits during May to September in the 20-county Atlanta Metropolitan area, 1993–2012. Outcomes are ordered by decreasing total counts

Disease	ICD-9 Code(s)	Mean Daily ED Visits	Standard Deviation	Min	Max
All Internal Causes	001–799	2553	1328	192	5557
Diabetes	249, 250	215	171	0	630
Fluid and Electrolyte Imbalance[a]	276	162	116	0	413
All Circulatory System Diseases	390–459	623	467	3	1652
Hypertension	401–405	491	408	0	1407
Ischemic Heart Disease	410–414	120	87	0	327
Myocardial Infarction[a]	410	14	8	0	40
Congestive Heart Failure	428	74	58	0	245
Ischemic Stroke	433–437	23	15	0	67
All Renal Diseases	580–593	139	118	0	447
Acute Renal Failure[a]	584	36	42	0	155
Heat-related Illnesses[a]	992.5	4	6	0	64

[a] Diagnosis based on primary ICD code only

Time-series model for ED visits and temperature

We first describe the quasi-Poisson log-linear model used to estimate short-term associations between daily ED visit counts and extreme temperature [36]. Following our previous analysis of heat waves and ED visits in Atlanta, [37], denote μ_t^a the mean ED visit count for adverse health outcome a on day t. The time-series model is given by

$$\begin{aligned}\ln\left(\mu_t^a\right) &= \beta_0+\beta H(\boldsymbol{X}_t)+ns(T_t)+ns(\overline{T}_t)+ns(DPT_t)+\alpha_{0,year_t}\\ &+f(DATE_t)\times\alpha_{1,year_t}+\sum_{i=1}^{6}\lambda_i DOW_{ti}\\ &+\sum_{j=1}^{2}\delta_j HOLIDAY_{tj}+\sum_{k=1}^{32}\gamma_k HOSPITAL_{tk}.\end{aligned}$$

Our parameter of interest β is the log relative risk (RR) associated with a binary extreme heat exposure $H(\boldsymbol{X}_t)$. The use of logic regression to define $H(\boldsymbol{X}_t)$ is described in detail later in Section 2.3. The above time-series model adjusts for *non-extreme* continuous same-day temperature (T_t), average temperature over the last 3 days (lag 1, lag 2 and lag 3) ($\overline{T}_t$), and same-day maximum dew-point temperature (DPT_t) to reflect the discomfort level due to the humidity. Specifically, in primary analyses, we defined the continuous temperature T_t and $\overline{T}_t$ by *truncating* the value at the extreme heat threshold (the 95th, 98th, or 99th percentile), i.e., setting any daily temperature value above the threshold to be the threshold value. In a sensitivity analysis, we also examined the use of the entire range of observed (i.e., non-truncated) temperature; in this case, $H(\boldsymbol{X}_t)$ can be interpreted as the "added" impact of extreme heat beyond the continuous temperature effect. The use of truncated continuous exposure is to provide better interpretation of β because $H(\boldsymbol{X}_t)$ only reflects the temperature effect beyond the threshold. Moreover, the tail of exposure-response function is often difficult to estimate due to sparse data and can be more sensitivity to influential observations. Hence, adjusting for non-truncated temperature may result in $\beta H(\boldsymbol{X}_t)$ accommodating for mis-specification of the exposure-response functions $ns(T_t)$ and $ns(\overline{T}_t)$ at the extreme tail. When apparent temperature is the exposure of interest, we did not include dew-point temperature in the model. To model the possible non-linear effect of meteorology, natural cubic splines, denoted by *ns (.),* were used for $T_t, \overline{T}_t$, and DPT_t with 2 equidistant internal knots.

We adjusted for long-term temporal trend in the time-series model as follows. Within a year, seasonal variation (from May to September), denoted by $f(DATE_t)$, was modeled smoothly with natural cubic splines and monthly knots. We included year-specific indicators $\alpha_{1,year_t}$ and their interactions with the seasonal pattern in our model to allow for between-year variation. We also adjusted for day of week effect using indicators DOW_{ti} and for federal or state holiday using indicators $HOLIDAY_{tk}$. Finally, $HOSPITAL_{tk}$ represents indicator variables for whether hospital k contributes to the ED visits counts on day *t;* these indicators were used to account for potential temporal drops in ED counts due to missing data from individual hospitals.

Logic regression

Logic regression is an adaptive regression method that attempts to construct predictors as Boolean combinations of binary covariates. It constructs a simple decision tree or a set of decision trees (multiple trees) with binary predictors connected by *and* (∧), *or* (∨), and *not* (^{c}) operators. For example, consider the following four binary extreme heat indicators based on minimum apparent temperature (ATMN),

$$X_1 = \mathrm{lag0}(\mathrm{today}^{'}\mathrm{s})\ \mathrm{ATMN} > 98\mathrm{thpercentile},$$

$$X_2 = \mathrm{lag1}(\mathrm{yesterday}^{'}\mathrm{s})\ \mathrm{ATMN} > 98\mathrm{thpercentile},$$

$$X_3 = \mathrm{lag2\ ATMN} > 98\mathrm{thpercentile},$$

and

$$X_4 = \mathrm{lag3\ ATMN} > 98\mathrm{thpercentile},$$

where we suppress the subscript t for presentation purposes. A simple extreme heat definition may be $H(\boldsymbol{X}_t) = X_1 \wedge X_2 \wedge X_3$, which describes a period of 3-consecutive days with high temperature. Hence, using these set of indicators, we can potentially capture both lagged and sustained effects of extreme heat. In our application, we focus on the use of the single tree model to estimate $H(\boldsymbol{X}_t)$. In another example, $H(\boldsymbol{X}_t) = X_1^c \wedge X_2 \wedge X_3$ describes a period of 2-consecutive days with high temperature, excluding the lag-0 day. This is different from $X_2 \wedge X_{3,}$ which does not place a restriction on whether X_1 being 0 or 1. We also note that $H(\boldsymbol{X}_t)$ is a binary exposure variable and offers better interpretation for risk associations compared to a more naïve approach of including X_1, X_2, $X_{3,}$ and X_4, as well as their interactions, jointly in a model.

Different logic trees may result in the same classification of days. For example, the two trees identified by logic regressions:

$$H_1(\boldsymbol{X}_t) = (X_1^c \wedge X_2 \wedge X_3) \vee (X_2 \wedge X_3 \wedge X_4) \mathrm{and} H_2(\boldsymbol{X}_t) = (X_1^c \vee X_4) \wedge (X_2 \wedge X_3).$$

give the same classification according to the distributive law (i.e., $A \wedge (B \vee C) = (A \wedge B) \vee (A \wedge C)$). To aid in interpretation of logic regression results, we used the following scheme to describe $H(\boldsymbol{X}_t)$. The Boolean combinations of heat indicators generated by logic regression can always be expressed by a series of logic statements joined by the $\vee$ (or) operators. Hence in the above example, $H_1(\boldsymbol{X}_t) = (\mathrm{X}_1^{\mathrm{c}} \wedge \mathrm{X}_2 \wedge \mathrm{X}_3) \vee (\mathrm{X}_2 \wedge \mathrm{X}_3 \wedge \mathrm{X}_4)$ is the preferred form with logic statements $X_1^c \wedge X_2 \wedge X_3$ and $X_2 \wedge X_3 \wedge X_4$.

The '*Logicreg*' package in R was used to fit logic regression models. Estimation was based on simulated annealing as the optimization algorithm to stochastically explore all 2^{2^k} Boolean combinations for k predictors. Ten-fold cross validation was used to select the size of the tree (i.e., number of leaves) and reduce issues related to overfitting.

A multi-stage estimation approach

We applied logic regression to time-series analysis of ED visit data in a three-stage approach. In the first stage, for each temperature variable (i.e., ambient air temperature and apparent temperature) and metric (i.e., maximum, minimum, or average), a quasi-Poisson log-linear model *without* $H(\boldsymbol{X}_t)$ was fit.

In the second stage, the Pearson residuals from the first-stage model were used to identify the Boolean combination of different extreme heat indicators at various lags. The Pearson residuals were calculated as

$$r_t^a = \frac{Y_t^a - \hat{\mu}_t^a}{\sqrt{V(\hat{\mu}_t^a)}}$$

where Y_t^a and $\hat{\mu}_t^a$ are, respectively, the observed number of daily ED visits and the predicted mean number of daily ED visits for outcome of interest a on day t, and $V(\hat{\mu}_t^a)$ is the product of $\hat{\mu}_t^a$ and the dispersion parameter. In the first stage we removed effects of continuous same-day and lagged temperatures, as well as other temporal trends. Because Pearson residual represents a scaled difference between the observed and expected counts, the logic regression tree $H(\boldsymbol{X}_t)$ estimated using Pearson residuals aims to captures additional lagged and sustained associations due to extreme temperature not explained by the base model. Finally, in the third stage, we refit the full time-series model with $H(\boldsymbol{X}_t)$ and all other covariates.

Simulation study

Simulation setup

We performed a simulation study to assess the performance of logic regression a our multi-stage procedure in detecting the structure of extreme heat exposures and in estimating associations with health outcomes. Let X_1, X_2, and X_3 be binary indicators for the minimum apparent temperature exceeding the 98th percentile threshold on lag 0, lag 1, and lag 2 day, respectively. We considered three different true $H(\boldsymbol{X}_t)$ exposures:

E1 Same-day effect: $H(\boldsymbol{X}_t) = X_1$,
E2 Sustained 2-day effect: $H(\boldsymbol{X}_t) = X_1 \wedge X_2$, and
E3 Sustained 2-day lagged-only effect: $H(\boldsymbol{X}_t) = X_1^c$⊠$X_2$⊠ X_3.

We considered three different health outcomes with different sample sizes, temporal patterns, and overdispersion (CIRC, RENAL, and HEAT). For each disease, we first fit the time-series model with all confounders as described in Section 2.1 with the Atlanta ED and meteorology data during 1993–2012 to obtain the baseline mean daily ED visits. We assumed a true relative risk (RR) of 1.01 or 1.05 for $H(\boldsymbol{X}_t)$ and simulated daily ED visit counts from a negative-binomial distribution and observed meteorology data using the time-series model given in Section 2.2. Regression coefficients and overdispersion were based on estimated values from models fitted with real data. We then applied the three-stage algorithm described in Section 2.4 to the simulated data and estimated the log RR of interest.

We ran the simulation 100 times for each scenario. The relative bias and relative root mean squared error (RRMSE) were used to examine the performance of the proposed approach. Relative bias and RRMSE were calculated as

$$Relative\ bias = \frac{\frac{1}{100}\sum_{i=1}^{100}\left(\widehat{RR_i} - RR_{true}\right)}{RR_{true}},$$

$$RRMSE = \frac{\sqrt{\frac{1}{100}\sum_{i=1}^{100}\left(\widehat{RR_i} - RR_{true}\right)^2}}{RR_{true}}.$$

Here, $\widehat{RR_i}$ is the estimated RR of $\hat{H}(\boldsymbol{X}_t)$ based on logic regression from the i$^{\text{th}}$ simulation and RR_{true} is the true RR for each scenario. We also estimated the sensitivity and specificity by comparing days indicated by $\hat{H}(\boldsymbol{X}_t)$ to be exposed/unexposed to the true exposure status given by E1, E2, or E3.

Simulation study results

Results from the simulation study are summarized in Table 2. We found that performance of the proposed method was better with larger RRs and for outcomes with larger daily event counts (e.g., comparing all renal diseases versus heat-related illnesses). We also ran the same simulation using the non-truncated continuous temperature and found similar results. Among the three different exposure scenarios, logic regression performed best for scenario E1 (single-lag, same-day) and worst in E3 (sustained two-day lagged consecutive exposure). The frequency of the extreme heat event may explain the different performances; the frequency of occurrence in our Atlanta study for E1, E2, and E3 were 146, 73, and 29 days, respectively. Importantly, we also found that in our Atlanta ED visits application, the average bias was negative, indicating that the effect estimate was attenuated towards the null. This is likely due the presence of exposure misclassification: when sensitivity/specificity is not 100%, some days are classified incorrectly as exposed/unexposed.

Application to the Atlanta ED visit data

Primary analyses

We applied the multi-stage algorithm described in Section 2.4 separately for each heat metric: MX, MN, and Avg of temperature (T) or apparent temperature (AT), and separately for each threshold (95th, 98th, 99th percentile). Extreme binary indicators were also defined for

Table 2 Summarized simulation study results of the performance of logic regression

Disease	True Relative Risk	Exposure Scenario	Sensitivity[a]	Specificity[a]	Relative bias		Relative root mean square error (RRMSE)	
					Logic regression	Known $H(X_t)$	Logic regression	Known $H(X_t)$
All Circulatory System Diseases	1.01	E1	50%	99%	−0.56%	−0.01%	1.09%	0.55%
		E2	35%	98%	−0.84%	− 0.04%	1.59%	0.69%
		E3	23%	98%	−1.13%	−0.28%	2.02%	0.85%
	1.05	E1	100%	100%	−0.11%	− 0.11%	0.69%	0.69%
		E2	99%	100%	−0.03%	0.03%	0.99%	0.71%
		E3	98%	100%	−0.09%	−0.02%	1.23%	1.00%
All Renal Diseases	1.01	E1	34%	99%	−0.40%	0.36%	1.39%	1.10%
		E2	36%	98%	−0.27%	0.10%	2.03%	1.26%
		E3	26%	98%	−0.92%	−0.22%	3.03%	1.61%
	1.05	E1	93%	100%	−0.14%	0.12%	1.75%	1.21%
		E2	82%	100%	−1.00%	0.18%	3.29%	1.16%
		E3	73%	99%	−1.37%	−0.16%	3.31%	1.57%
Heat-Related Illnesses	1.01	E1	19%	99%	−0.57%	0.35%	6.58%	5.61%
		E2	29%	98%	0.30%	1.05%	12.02%	6.93%
		E3	17%	98%	0.09%	0.91%	16.21%	9.03%
	1.05	E1	33%	98%	−2.53%	−0.23%	8.43%	6.44%
		E2	30%	98%	−0.83%	0.45%	11.66%	6.40%
		E3	23%	98%	−3.07%	−0.92%	16.42%	9.10%

[a] Sensitivity is defined as the proportions of days assigned as exposed using the exposure metric estimated from logic regression among days assigned as exposed using the true exposure metric (E1, E2, or E3) for simulating health data. Specificity is defined similarly for days assigned as unexposed. For each simulation scenario, the sensitivities and specificities reported are averaged across 100 simulations

up to three lagged days. The model with the smallest quasi-AIC was selected.

Table 3 summarizes the structure of the extreme heat exposures $H(X_t)$ and their associations with ED visits. Supplementary Figure 1 shows the structure of the logic regression tree $H(X_t)$ for selected outcomes. Logic trees can be read "bottom-up" to construct a corresponding logic statements where the tree split is given by the Boolean statement and/or. Given a logic regression tree, individual days in our 20-year study period data were divided into two groups and we defined the reference (i.e., $H(X_t) = 0$) as the type of days more frequently observed.

Overall, we found several positive associations with exposures. For example, for heat-related diseases (HEAT), extreme heat exposure was defined as days where (1) lag 0 and lag 1 ATMIN are above the 95% percentile or (2) lag 1 and lag 3 ATMIN is above the 95% percentile. This exposure was associated with an increase in mean HEAT ED visits by 40% (95% CI: 27–54%). For acute renal failure (ARF) ED visits, the exposure identified (i.e., recent two days' ATMIN is above the 95% percentile) was associated with an increase of 5% (95% CI: 2–9%). For all renal diseases, the exposures identified were more complicated and was associated with an increase of ED visits by 2% (95% CI: 1–4%). Associations with circulatory disease were generally null, except for ischemic stroke and hypertension, for which we found a negative association with extreme heat exposure

Sensitivity analyses

We also conducted three additional sensitivity analyses. First, we examined the more conventional heat wave definition where duration is defined as at least two consecutive days exceeding the threshold. Result for HEAT, STK, and RENAL outcomes are given in Supplementary Table S1. We found that the conventional heat wave definitions also indicated positive associations, but the magnitude can be attenuated. Second, because logic regression is a machine learning algorithm that optimizes predictability, the lag structure may contain holes (e.g., exceeding the temperature threshold on lag-1 and lag-3, but not lag-2). Based on the exposure lags identified by logic regression as a guide, we defined alternative exposure metrics by filling in gaps and removing the "not" statement. These alternative metrics are defined over consecutive days and may be more interpretable. Results are given in Table S2. For HEAT and RENAL, we found that using alternate exposures that are more extreme (less frequent) continues to estimate positive associations. However, in the case of STK, where the alternate exposures are less extreme, the associations with ED visits were positive, but confidence intervals included the null.

In a second analysis, we replaced the truncated same-day and 3-day moving-average of temperature heat metric with the original variable without truncation. Here the application of logic regression attempts to

Table 3 Summary of extreme heat exposure estimated by logic regression and their short-term associations with warm-season emergency department visits in Atlanta, 1993 to 2012. Relative risk estimates and 95% confidence intervals (CI) were from time-series models adjusting for truncated continuous temperature. Within each outcome, each row of the extreme heat exposure corresponds to an "or" statement derived from the logic regression tree. "Not" statement is indicated by a superscript *c*

ED Visit Outcome	Heat Metric and Quantile	Extreme Heat Exposure	lag0	lag1	lag2	lag3	Frequency (days)	Relative Risk (95% CI)
HEAT	ATMN95	(lag0 and lag1)	Y	Y			237	**1.400 (1.269, 1.544)**
		OR (lag1 and lag3)		Y		Y		
FLEL	ATMN95	lag1		Y			335	**1.030 (1.016, 1.045)**
RENAL	ATMN95	(lag0 and lag3)	Y			Y	255	**1.025 (1.011, 1.039)**
		OR (lag2^{C} and lag3)			N	Y		
ARF	ATMN95	lag0 and lag1	Y	Y			190	**1.052 (1.021, 1.085)**
STK	TMX98	lag0 and lag1 and lag2^{C} and lag3	Y	Y	N	Y	5	**1.257 (1.061, 1.477)**
CIRC	ATMX95	lag1		Y			363	0.996 (0.987, 1.004)
HT	ATMX95	lag0 and lag1 and lag3^{C}	Y	Y		N	128	**0.988 (0.977, 0.999)**
		OR (lag1 and lag2 and lag3^{C})		Y	Y	N		
IHD	TMN99	lag0	Y				63	1.013 (0.986, 1.041)
MI	TMN99	lag0	Y				63	1.044 (0.964, 1.131)
CHF	ATMN98	lag0	Y				146	0.976 (0.947, 1.007)
DIA	TMN95	lag2			Y		238	1.009 (0.996, 1.021)
INTERN	ATMN95	lag3				Y	335	**1.008 (1.002, 1.015)**

HEAT heat illness, *STK* ischemic stroke, *ARF* Acute renal failure, *RENAL* all renal disease, *FLEL* Fluid and electrolyte imbalance, *CIRC* all circulatory system disease, *CHF* Congestive heart failure, *IHD* Ischemic heart disease, *MI* Myocardial infarction, *DIA* Diabetes, *HT* Hypertension, *ITERN* all internal causes) identify *additional risks* beyond that conferred by the continuous exposure-response function. Results are given in Supplementary Table S2. In general, we found that logic regression identified extreme heat exposure based on similar heat metrics but RR with smaller magnitude. However, the use of truncated continuous temperature, as in our main analysis, was generally associated with better model fit and stronger associations with ED visits. This may be attributed to data sparsity in the extreme tail of the temperature distribution such that when using non-truncated temperature, the tail of the continuous exposure-response function may be misspecified and has high uncertainty.

Discussion

In this paper, we propose the use of logic regression to help identify characteristics of extreme heat exposure that are associated with short-term adverse health risks. Our 20-year time-series analysis shows that ED visits for various disease outcomes were associated with exposure identified by logic regression using a multi-stage algorithm. Our motivating hypothesis is that most harmful characteristics of heat exposures are likely to vary between outcomes because of the different vulnerable subpopulations and different pathophysiological mechanisms they may impact. While the strength of association of different exposure definitions varied by outcome, the most health-relevant exposures were generally those characterized by temperatures exceeding a threshold over multiple lags. We also found evidence that apparent temperature and daily minimum temperature gave better model fit, providing support that humidity and night-time exposure are important considerations for quantifying adverse health effects of extreme heat events. However, Armstrong et al. [38] found that adding relative humidity or dewpoint temperature to a model with temperature does not improve model fit.

Though studies on heat waves and cause-specific ED visits are limited compared to cause-specific mortality, our results are consistent with much of the previous research that utilized different definitions. Petitti et al. [39] used three pre-specified temperature trigger points (minimum risk temperature, increasing risk temperature, and excess risk temperature) in Maricopa County Arizona. This study found significant associations with all three trigger points and ED visits for heat related diagnosis but found no association with CVD related outcomes and total ED visits. Another study that looked at high ambient temperature and ED visits in California found that same day ambient temperature was positively associated with heat illness and ARF, but was negatively associated with hypertension [40], as were our results.

For hypertension, we found that the exposures estimated from logic regression tended to be negatively associated with ED visits. Previous studies on heat waves and morbidity outcomes have found similar results. Lim et al. [9] and Sherbakov et al. [41] found that hospital admissions due to hypertension and other cardiovascular related outcomes decreased with increased temperatures. Similarly, a study by Michelozzi et al. [42] found that cardiovascular related morbidity was reduced with higher temperatures, but cardiovascular mortality increased. A possible mechanism for these findings is that blood pressure can decrease from vasodilation and sweating in the summer, thus potentially reducing hypertension related hospital admissions [41, 43].

Our extreme heat exposures did not always identify consecutive days of high temperature as being the most harmful. For example, the *not* logic statement and *missing* lags were selected for several disease outcomes (Table 3 and S2). This may be due to the lack of statistical power as periods with consecutive days of high temperature were less frequent. It is also possible that the risk of consecutive days of high temperature is less harmful compared to periods with more *variable* but high temperature due to awareness of extreme heat events.

We found that different specifications of the base model (i.e., adjusting for truncated versus non-untruncated continuous temperature) can have an impact on the extreme heat exposure identified and its estimated risk ratio. While both types of analyses are common in the literature, associations of extreme heat events from these two approaches should be interpreted differently (i.e., total risks above a certain threshold versus risks in addition to the continuous exposure-response function). This motivated our multi-stage estimation approach such that the base model is specified a priori and the data-driven logic regression is only utilized to explore potential excess risks not explained by the base model. We note that when performing risk assessment of high temperature, both extreme temperature events and the tail of the exposure-response function should be considered.

The main advantage of logic regression is its supervised learning approach for deriving study-specific and outcome-specific exposure definitions that are flexibly constructed by indicators of different extreme heat characteristics. Study population, geographical region, and the set of exposures being considered may contribute to the observed heterogeneity in extreme heat health effects within and across studies. The algorithm also has a user-friendly software package that can efficiently evaluate a large suite of possible exposure definitions to identify those that are most important for individual health outcomes. Compared to regression tree methods, logic regression has two important advantages: (1) the ability to incorporate “and”, “or” and “not” statement between predictors, and (2) the focus on binary classification of a continuous outcome. In contrast, Classification And Regression Trees (CART) will give multiple terminal nodes and the resulting heat metric will likely be very complex because only "and" statement is allowed as the decision tree is split.

Finally, logic regression has several limitations that warrant further investigations. First, during estimation, logic regression can become trapped in a local minimum when many binary predictors are being considered. Hence, model fitting requires a comprehensive evaluation of tuning parameters (e.g., starting temperature, finishing temperature, and cooling schemes) and initial values, which increases computational burden. In our application, the number of binary predictors is considerably smaller than the typical genetic association studies and we were able to evaluate different control parameters for simulated annealing. Second, we only utilized logic regression to select the exposure lag structure given a heat metric and threshold. We found that the current sample size cannot accommodate including all possible heat exposure indicators that are highly correlated. One future direction is to consider additional penalization within logic regression. Third, our multi-stage estimation approach, while allowing us to work with established time-series analysis methodology, does not account for estimation uncertainty associated with the extreme heat metric, which may be important for quantifying associations for rare exposure events. Statistical inference such as pseudolikelihood [44, 45] could be integrated to the current method to account for the ignored uncertainty. Moreover, recent advances in fitting logic regression under a Bayesian framework [46] allows for direct quantification of uncertainties via posterior samples. How to incorporate these uncertainties in a multi-stage health analysis, similar to an exposure measurement error framework, warrants further investigations.

Abbreviations

ARF: Acute renal failure; AT: Apparent temperature; CHF: Congestive heart failure; CIRC: Circulatory system diseases; ED: Emergency department; DIA: Diabetes; FLEL: Fluid and electrolyte imbalances; HEAT: Heat illness; HT: Hypertension; IHD: Ischemic heart disease; INTERN: Internal causes; MI: Myocardial infarction; MN: Minimum; MX: Maximum; RENAL: Renal diseases; RR: Relative risk; STK: Ischemic stroke; T: Temperature

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-021-01278-x.

Additional file 1: Figure S1. Structure of logic regression tree of extreme heat exposures for selected warm-season ED visit outcomes in Atlanta, Georgia, 1993–2012. **Table S1.** Summary of alternative extreme temperature metrics and their short-term associations with warm-season emergency department visits in Atlanta, 1993 to 2012. **Table S2.** Summary of alternative extreme temperature metrics with consecutive lags and their short-term associations with warm-season emergency department visits in Atlanta, 1993 to 2012. **Table S3.** Summary of extreme heat metrics from truncated continuous versus continuous temperature metric and their short-term associations with warm-season emergency department visits in Atlanta, 1993 to 2012.

Authors' contributions

SJ contributed to data analysis, data interpretation and drafting the manuscript. JWL, NS, LAD, MJS, YC, AJN, STE and HHC assisted with the writing and interpretation of findings. STE and SM contributed to health and exposure data acquisition and management. HHC contributed to the conceptual framework of the study. All authors critically reviewed and approved the final manuscript.

Funding

This research was supported by the National Institute of Environmental Health Sciences (NIEHS) of the National Institutes of Health (NIH) under Award R01 ES027892, R01 ES028346, and R21 ES023763. The content of this manuscript is solely the responsibility of the authors and does not necessarily represent the official views of the NIH and does not endorse the purchase of any commercial products or services mentioned in this manuscript.

Availability of data and materials

The daily emergency department data used in this study contain protected health information and cannot be shared due to existing data use agreements. Example R code for carrying out the simulation study and analysis are available from the corresponding author upon request.

Declarations

Ethics approval and consent to participate

The study is approved by the Emory Institutional Review Board. Informed consent was not obtained because this study involved secondary analysis of administrative databases. All analyses were carried out in accordance with relevant guidelines and regulations.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, USA. [2]Department of Biostatistics, Yale University, New Haven, USA. [3]Gangarosa Department of Environmental Health, Emory University, Atlanta, USA. [4]School of Community Health Sciences, University of Nevada Reno, Reno, USA. [5]Research Applications Laboratory, National Center for Atmospheric Research, Boulder, USA. [6]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, USA.

Received: 19 January 2021 Accepted: 12 April 2021
Published online: 26 April 2021

References

1. Semenza JC, Rubin CH, Falter KH, Selanikio JD, Flanders WD, Howe HL, et al. Heat-related deaths during the July 1995 heat wave in Chicago. N Engl J Med. 1996;335(2):84–90. https://doi.org/10.1056/NEJM199607113350203.
2. Semenza JC, McCullough JE, Flanders WD, McGeehin MA, Lumpkin JR. Excess hospital admissions during the July 1995 heat wave in Chicago. Am J Prev Med. 1999;16(4):269–77. https://doi.org/10.1016/S0749-3797(99)00025-2.
3. Dhainaut JF, Claessens YE, Ginsburg C, Riou B. Unprecedented heat-related deaths during the 2003 heat wave in Paris: consequences on emergency departments. Crit Care. 2003;8(1):1.
4. Pirard P, Vandentorren S, Pascal M, Laaidi K, Le Tertre A, Cassadou S, et al. Summary of the mortality impact assessment of the 2003 heat wave in France. Eurosurveillance. 2005;10(7):7–8. https://doi.org/10.2807/esm.10.07.00554-en.
5. Cheng J, Xu Z, Bambrick H, Prescott V, Wang N, Zhang Y, et al. Cardiorespiratory effects of heatwaves: a systematic review and meta-analysis of global epidemiological evidence. Environ Res. 2019;177:108610. https://doi.org/10.1016/j.envres.2019.108610.

6. Guo Y, Gasparrini A, Armstrong BG, Tawatsupa B, Tobias A, Lavigne E, et al. Heat wave and mortality: a multicountry, multicommunity study. Environ Health Perspect. 2017;125(8):087006. https://doi.org/10.1289/EHP1026.
7. Lim Y, Lee K, Bae H, Kim D, Yoo H, Park S, et al. Estimation of heat-related deaths during heat wave episodes in South Korea. Environ Epidemiol. 2019;3:241.
8. Williams S, Nitschke M, Weinstein P, Pisaniello DL, Parton KA, Bi P. The impact of summer temperatures and heatwaves on mortality and morbidity in Perth, Australia 1994–2008. Environ Int. 2012;40:33–8. https://doi.org/10.1016/j.envint.2011.11.011.
9. Zhang Y, Feng R, Wu R, Zhong P, Tan X, Wu K, et al. Global climate change: impact of heat waves under different definitions on daily mortality in Wuhan, China. Global Health Res Policy. 2017;2(1):10.
10. Wang XY, Barnett A, Guo YM, Yu WW, Shen XM, Tong SL. Increased risk of emergency hospital admissions for children with renal diseases during heatwaves in Brisbane, Australia. World J Pediatr. 2014;10(4):330–5. https://doi.org/10.1007/s12519-014-0469-x.
11. Xu Z, Tong S, Cheng J, Crooks JL, Xiang H, Li X, et al. Heatwaves and diabetes in Brisbane, Australia: a population-based retrospective cohort study. Int J Epidemiol. 2019;48(4):1091–100. https://doi.org/10.1093/ije/dyz048.
12. Yin Q, Wang J. The association between consecutive days' heat wave and cardiovascular disease mortality in Beijing, China. BMC Public Health. 2017; 17(1):223. https://doi.org/10.1186/s12889-017-4129-7.
13. Kent ST, McClure LA, Zaitchik BF, Smith TT, Gohlke JM. Heat waves and health outcomes in Alabama (USA): the importance of heat wave definition. Environ Health Perspect. 2014;122(2):151–8. https://doi.org/10.1289/ehp.1307262.
14. Poumadere M, Mays C, Le Mer S, Blong R. The 2003 heat wave in France: Dangerous climate change here and now. Risk Anal. 2003;25(6):1483–94.
15. Guo Y, Gasparrini A, Armstrong B, Li S, Tawatsupa B, Tobias A, et al. Global variation in the effects of ambient temperature on mortality: a systematic evaluation. Epidemiology. 2014;25(6):781–9. https://doi.org/10.1097/EDE.0000000000000165.
16. Åström DO, Ebi KL, Vicedo-Cabrera AM, Gasparrini A. Investigating changes in mortality attributable to heat and cold in Stockholm, Sweden. Int J Biometeorol. 2018;62(9):1777–80. https://doi.org/10.1007/s00484-018-1556-9.
17. Anderson GB, Bell ML. Heat waves in the United States: mortality risk during heat waves and effect modification by heat wave characteristics in 43 US communities. Environ Health Perspect. 2011;119(2):210–8. https://doi.org/10.1289/ehp.1002313.
18. Bobb JF, Obermeyer Z, Wang Y, Dominici F. Cause-specific risk of hospital admission related to extreme heat in older adults. JAMA. 2014;312(24):2659–67. https://doi.org/10.1001/jama.2014.15715.
19. Chen T, Sarnat SE, Grundstein AJ, Winquist A, Chang HH. Time-series analysis of heat waves and emergency department visits in Atlanta, 1993 to 2012. Environ Health Perspect. 2017;125(5):057009. https://doi.org/10.1289/EHP44.
20. Vaidyanathan A, Kegler SR, Saha SS, Mulholland JA. A statistical framework to evaluate extreme weather definitions from a health perspective: a demonstration based on extreme heat events. Bull Am Meteorol Soc. 2016; 97(10):1817–30. https://doi.org/10.1175/BAMS-D-15-00181.1.
21. Heo S, Bell ML, Lee JT. Comparison of health risks by heat wave definition: applicability of wet-bulb globe temperature for heat wave criteria. Environ Res. 2019;168:158–70. https://doi.org/10.1016/j.envres.2018.09.032.
22. Hattis D, Ogneva-Himmelberger Y, Ratick S. The spatial variability of heat-related mortality in Massachusetts. Appl Geogr. 2012;33:45–52. https://doi.org/10.1016/j.apgeog.2011.07.008.
23. Loughnan ME, Nicholls N, Tapper NJ. The effects of summer temperature, age and socioeconomic circumstance on acute myocardial infarction admissions in Melbourne, Australia. Int J Health Geogr. 2010;9(1):1–1.
24. Murage P, Hajat S, Kovats RS. Effect of night-time temperatures on cause and age-specific mortality in London. Environ Epidemiol. 2017;1(2):e005. https://doi.org/10.1097/EE9.0000000000000005.
25. Hajat S, Armstrong B, Baccini M, Biggeri A, Bisanti L, Russo A, et al. Impact of high temperatures on mortality: is there an added heat wave effect? Epidemiology. 2006;1:632–8.
26. Park J, Kim J. Defining heatwave thresholds using an inductive machine learning approach. PLoS One. 2018;13(11):e0206872. https://doi.org/10.1371/journal.pone.0206872.
27. Ruczinski I, Kooperberg C, LeBlanc M. Logic regression. J Comput Graph Stat. 2003;12(3):475–511. https://doi.org/10.1198/1061860032238.
28. Ruczinski I, Kooperberg C, LeBlanc ML. Exploring interactions in high-dimensional genomic data: an overview of logic regression, with applications. J Multivar Anal. 2004;90:178–95.
29. Schwender H, Ruczinski I. Logic regression and its extensions. Adv Genet. 2010;72:25–45. https://doi.org/10.1016/B978-0-12-380862-2.00002-3.
30. Schwender H, Ickstadt K. Identification of SNP interactions using logic regression. Biostatistics. 2008;9(1):187–98. https://doi.org/10.1093/biostatistics/kxm024.
31. Yoo W, Ference BA, Cote ML, Schwartz A. A comparison of logistic regression, logic regression, classification tree, and random forests to identify effective gene-gene and gene-environmental interactions. Int J Appl Sci Technol. 2012;2(7):268.
32. Rathod SD, Li T, Klausner JD, Hubbard A, Reingold AL, Madhivanan P. Logic regression-derived algorithms for syndromic management of vaginal infections. BMC Med Inform Decis Mak. 2015;15(1):1–7.
33. Lorenz MW, Abdi NA, Scheckenbach F, Pflug A, Bülbül A, Catapano AL, et al. Automatic identification of variables in epidemiological datasets using logic regression. BMC Med Inform Decis Mak. 2017;17(1):1–1.
34. Bellavia A, Rotem RS, Dickerson AS, Hansen J, Gredal O, Weisskopf MG. The use of logic regression in epidemiologic studies to investigate multiple binary exposures: an example of occupation history and amyotrophic lateral sclerosis. Epidemiol Methods. 2020;25(1) open issue. https://doi.org/10.1515/em-2019-0032. in press.
35. Winquist A, Grundstein A, Chang HH, Hess J, Sarnat SE. Warm season temperatures and emergency department visits in Atlanta, Georgia. Environ Res. 2016;147:314–23. https://doi.org/10.1016/j.envres.2016.02.022.
36. Steadman RG. A universal scale of apparent temperature. J Climate Appl Meteor. 1984;23(12):1674–87. https://doi.org/10.1175/1520-0450(1984)023<1674:AUSOAT>2.0.CO;2.
37. Bhaskaran K, Gasparrini A, Hajat S, Smeeth L, Armstrong B. Time series regression studies in environmental epidemiology. Int J Epidemiol. 2013; 42(4):1187–95. https://doi.org/10.1093/ije/dyt092.
38. Armstrong B, Sera F, Vicedo-Cabrera AM, Abrutzky R, Åström DO, Bell ML, et al. The role of humidity in associations of high temperature with mortality: a multicountry, multicity study. Environ Health Perspect. 2019;127(9):097007.
39. Petitti DB, Hondula DM, Yang S, Harlan SL, Chowell G. Multiple trigger points for quantifying heat-health impacts: new evidence from a hot climate. Environ Health Perspect. 2016;124(2):176–83. https://doi.org/10.1289/ehp.1409119.
40. Basu R, Pearson D, Malig B, Broadwin R, Green R. The effect of high ambient temperature on emergency room visits. Epidemiology. 2012;23(6):813–20. https://doi.org/10.1097/EDE.0b013e31826b7f97.
41. Sherbakov T, Malig B, Guirguis K, Gershunov A, Basu R. Ambient temperature and added heat wave effects on hospitalizations in California from 1999 to 2009. Environ Res. 2018;60:83–90.
42. Michelozzi P, Accetta G, De Sario M, D'Ippoliti D, Marino C, Baccini M, et al. High temperature and hospitalizations for cardiovascular and respiratory causes in 12 European cities. Am J Respir Crit Care Med. 2009;179(5):383–9. https://doi.org/10.1164/rccm.200802-217OC.
43. Abdulla K, Taka M. Climatic effects on blood pressure in normotensive and hypertensive subjects. Postgrad Med J. 1988;64(747):23–6. https://doi.org/10.1136/pgmj.64.747.23.
44. Gong G, Samaniego FJ. Pseudo maximum likelihood estimation: theory and applications. Ann Stat. 1981;9(4):861–9.
45. Liang KY, Self SG. On the asymptotic behaviour of the pseudolikelihood ratio test statistic. J R Stat Soc Ser B Methodol. 1996;58(4):785–96.
46. Hubin A, Storvik G, Frommlet F. A novel algorithmic approach to Bayesian logic regression (with discussion). Bayesian Anal. 2020;15(1):263–333.

Publisher's Note