Introduction (ish):

* With the aging US electrical grid and climate change underway, power outages are becoming increasingly common, and they pose substantial health risks to vulnerable populations like older adults and people who use durable medical equipment.
* To study the health effects of outages, we assembled a dataset of nation-wide county-level power outage data, used in Do et al. “Spatiotemporal distribution of power outages with climate events and social vulnerability in the USA”.
* However, the study of power outage as a health-related exposure is new, and it’s unclear how to define power outage exposure in part because we don’t know how long an outage needs to be to influence health.
* It’s also unclear how missing data could influence the results of an epidemiologic study of the effect of outages on health when using power outage data.
* In this paper, we present the results of simulations that we designed to test how missing data and incorrect definitions of power outage exposure (leading to exposure misclassification) could influence the results of a case-crossover study we are conducting now, testing the effect of county-level power outage exposure on county-level Medicare hospitalization rates nationwide.
* Authors working with several kinds of environmental exposure data similar to power outage data could face challenges in defining exposure and handling missing data like the ones we addressed with simulations here.
* In particular, other data on climate-related events, such as precipitation data, or data describing extreme weather events and disasters, follow a similar structure to power outage data. Data are available in a time series (containing the of amount of precipitation or number of people affected by a disaster) for a spatial unit, and data cleaners have to decide on a spatial-unit-level threshold for exposure. Here, we test how defining an exposure power outage incorrectly could bias results of our epidemiologic study testing the effect of power outage exposure on the outcome of hospitalization counts by people insured with Medicare.
* Water quality data at the county level is made up of several smaller spatial units of local water systems, and may be missing water systems that cover a certain proportion of people in the county, or time series for certain water systems may be incomplete. Here, we also test how missing data of these kind could bias results of our study.
* We hope other environmental epidemiologists can design simulations similar to what we did here, to test questions about exposure definition and data coverage with datasets that follow a similar structure. Simulations like these could be used to assess potential bias in similar contexts.

Methodssss:

Data:

* We got raw power outage data covering the continental US for the years 2018 and 2019 from poweroutages.us, a site that aggregates power outage data by collecting it from utility company APIs.
* The raw data came to us at the spatial unit of ‘neighbourhood-utility’ level – the data contained counts of customers without power all served by the same utility in a neighbourhood or city. Many utilities can serve the same place, so these units were not geographically distinct. Two houses next to each other might have been in two different spatial units if they were served by different utilities. This neighbourhood-utility level data came in a time series of ten-minute intervals covering 2 years.
* Neighbourhood-utilities were nested inside of counties, so it was possible for us to aggregate 10-minute counts of customers without power to the county level. We also aggregated these counts to the hourly level.

Simulation on exposure misclassification:

Why?

* We were interested in defining a county-level power outage exposure. Other studies have defined binary power outage exposures in a spatial unit (whether there is an outage, yes/no) by deciding that there is an “outage ON” when the proportion of customers without power goes beyond a certain threshold, for example, above the 95th percentile of counts of customers without power in a spatial unit. We reasoned that if counts of customers without power are this high, many individual households and therefore people are likely to be without power continuously while the counts remain high.
* We followed (cite the papers – ours and dominianni etc.) and used the 95th percentile-based definition to develop a time series of hours exposed or unexposed based on hourly-county counts of customers without power.
* We initially hypothesized that outages that were 8 hours or longer could produce health effects, since the batteries of most durable medical equipment last 8 hours. When batteries die, vulnerable people may be without life-sustaining medical devices like oxygen tanks or at-home ventilators.
* We initially marked days as exposed or unexposed only if there were 8 continuous hours of ‘power outage on’ (customers without power counts > 95th percentile) on that day.
* However, we don’t have any evidence to support the claim that 8 hours is the true threshold after which a power outage becomes clinically relevant. So, we were concerned that if it was in fact shorter or longer outages that mattered for health, we could get really biased results in our study if we cleaned the data and marked days as exposed or unexposed based on 8 hour outages
* Said differently, if it was instead 4 hour outages or 12 or outages that were clinically relevant and mattered for health, and we cleaned the data as if it were 8 hour outages that mattered, we would be introducing serious exposure misclassification.

How we did the simulation:

* To test how much we would bias our results if we defined a power outage incorrectly, we generated 100 counties with an average of N (have to go get this number- around 700) neighbourhood-utility units nested in each of them. We randomly drew the number of neighbourhood-utility units from the real empirical distribution of neighbourhood-utility units within counties in the real poweroutage.us data.
* We populated each neighbourhood-utility with a total number of customers “living or doing business there”, drawn from the empirical distribution of number of customers served in each neighbourhood-utility unit in the real poweroutage.us data.
* We then generated ten-minute counts of customers without power for each neighbourhood-utility unit for one year. The counts of customers without power at each 10-minute interval was also drawn from the empirical distribution of the proportion of customers without power in the poweroutage.us data, then multiplied by the customers served in the neighbourhood-utility unit.
* We aggregated these neighbourhood-utility 10-minute counts of customers without power to the hourly-county level.
* Again, following dominianni etc, we used the 95th percentile-based definition to develop a county time series of hours exposed to power outage or unexposed to power outage based on hourly-county counts of customers without power. We marked an hour exposed if the count of customers without power in a county in that hour was above the 95th percentile of counts of customers without power in that county for the year’s worth of hourly data in each county.
* First, to establish a base case representing an unbiased scenario where exposure was defined correctly, we generated simulated outcome data of hospitalization counts by day and county based on a definition of 8-hour power outages.
* We initially marked days as exposed or unexposed only if there were 8 consecutive hours of ‘power outage on’ (customers without power counts > 95th percentile) ending on that day.
* Then we created outcome data: we drew ‘hospitalization counts’ for each day in each county from a Poisson distribution based on the total number of customers living in that county, with a base hospitalization rate of 0.1%, and a 10% rate increase (for a total hospitalization rate of 0.11%) on days marked as exposed to a power outage based on the “8 hour” definition. This base hospitalization rate *is the actual hospitalization rate in medicare data.*  – it’s actually not right now but when I get access to the data it will be.
* To generate an effect estimate for the effect of an 8-hour power outage on daily county-level hospitalization rate, for each day exposed in a county, we chose a control day 2 weeks before the exposed day and 2 weeks after, if such days were available and not before the study period began or after the simulated study period ended. We used these cases and controls in a Poisson model, one for each county (for a total of 100 models) to model the relationship between the binary 8-hour power outage exposure and daily hospitalization counts, with an offset for the number of customers in a county. This model represented an unbiased scenario where exposure was defined correctly.
* To test how defining an outage incorrectly could bias results, we first changed how we defined our power outage exposure.
* We created two additional exposure dataset for each county, marking a day in a county exposed if there was either a 4-hour outage (customers without power counts > 95th percentile for 4 consecutive hours) ending on that day, or a 12 hour outage (defined the same but for 12 hours) ending on that day.
* We generated two additional datasets of outcome data (daily county hospitalization counts) based on the same base hospitalization rate of 0.1%, and a 10% rate increase (for a total hospitalization rate of 0.11%) on days with a power outage, but we based these counts on 4-hour outage exposure data and 12-hour outage exposure data.
* We then modelled the relationship between the binary 8-hour power outage exposure and hospitalization counts generated based on the 4-hour exposure, and then on the 12-hour exposure.
* This represented exposure misclassification, happening because we defined the length of a clinically relevant power outage incorrectly. In the 12-hour case, there were many days marked as exposed (because there was an 8-hour power outage ending on that day) but were not actually exposed (because the hospitalization rate in this simulations was only increased if there was a 12-hour power outage ending on that day). In the case of the model using outcome data based on the 4-hour exposure and exposure data based on the 8-hour exposure, there were many days in the dataset that were actually exposed but were not marked as exposed.
* For each day marked as exposed given the 8-hour definition in a county, we chose 2 control days as above. We used these cases and controls in 100 Poisson models for the 4 hour exposure and 100 for the 12 hour exposure like the one described above.
* We assessed the bias in these models by plotting the estimated rate increases with power outage exposure from each model and coverage of 95% confidence intervals (will be Figure 1). We also repeated the simulations for 2 other effect sizes of power outage on hospitalization (will be Figure 1).

Simulation on missing data:

Why?

* The poweroutages.us dataset is incomplete. We estimated the number of utility customers in each county using EIA data as part of our paper Do et al., and there were a lot of counties where the poweroutages.us data only covers a fraction of the people in that county. There are neighbourhood-utility units missing from the dataset.
* and substantial chunks of time (such as several weeks or months) were missing from some time series. Some neighbourhood-utility units were missing from the datasets entirely. These sources of msisingness affected the total percentage of person-time that should be covered in a county.
* Describe power outage data
* Describe exposure misclassification because of defining the exposure wrong
* Describe the simulation we designed to assess bias
* Describe missing data issues
* Describe the simulation we used to assess bias
* have a similar structure to the power outage data we worked with.
* Counties are split up by electrical service provider and neighbourhood – so the data came to us at the neighbourhood-utility level. Many utilities can serve the same place, so these units were not geographically distinct. The neighbourhood-utility level data came in a time series of ten-minute intervals covering 2 years, and substantial chunks of time (such as several weeks or months) were missing from some time series. Some neighbourhood-utility units were missing from the datasets entirely. These sources of msisingness affected the total percentage of person-time that should be covered in a county.
* Other sources of data on climate-related events, such as precipitation data, or data on extreme weather events, follow a similar structure in that data are available in a time series of amount of precipitation or ??? that may have missingness, and you have to make choices about how to define exposure from a continuous time series. Water system data at the county level is made up of several smaller spatial units and may
* because this is a new exposure that hasn’t been studied a lot at all besides in new York data and in the case of extreme events, and nobody really knows how long a clinically significant power outage is, we weren’t sure how to define exposure based on the data we got and how this could bias results of a study we did
* We also didn’t know how missing data would influence any epidemiologic study that we did
* we wanted, in particular, to look at how county-level power outages influenced hospitalizations among people with medicare, using a case-control setup while accounting for lagged effects.
* We decided to run simulations.
* We ran simulations where we decreased the person-time coverage in a county so that progressively more of the power outage data was missing, and poisson regression models with a case-control setup to test whether missingness would bias effect estimates of the effect of power outages on county-level hospitalization counts.
* We also ran simulations where we created outcome data as if 4 hour and 12 hour power outages were clinically significant, but then cleaned exposure data under the assumption that it was 8 hour outages that were significant, to test how this exposure misclassification would bias the same model results.
* Many kinds of epidemiologic data have a similar structure to the power outage data we worked with.
* Counties are split up by electrical service provider and neighbourhood – so the data came to us at the neighbourhood-utility level. Many utilities can serve the same place, so these units were not geographically distinct. The neighbourhood-utility level data came in a time series of ten-minute intervals covering 2 years, and substantial chunks of time (such as several weeks or months) were missing from some time series. Some neighbourhood-utility units were missing from the datasets entirely. These sources of msisingness affected the total percentage of person-time that should be covered in a county.
* Other sources of data on climate-related events, such as precipitation data, or data on extreme weather events, follow a similar structure in that data are available in a time series of amount of precipitation or ??? that may have missingness, and you have to make choices about how to define exposure from a continuous time series. Water system data at the county level is made up of several smaller spatial units and may