Intro:

Power outage incidence is increasing. As climate change increases the frequency and intensity of extreme heat events, wildfires, hurricanes, and other severe weather, the United States electrical grid is aging. On average, electrical customers experienced 8 hours without power in 2020, the longest duration on record.

Power outages have substantial health effects in vulnerable populations such as older adults, people who use life-sustaining electricity-dependent medical equipment, and children. Loss of electricity can be life-threatening for people who use electricity-dependent medical equipment such as at-home ventilators or oxygen tanks. Older adults may be susceptible to cardiorespiratory disease-related health effects from heat or cold exposure, when heating and air conditioning is unavailable during power outages. Outages increase pediatric injuries, because of accidents related to increased generator and natural gas use. They also increase pediatric asthma emergencies, as power outages increase heat and humidity exposure when air conditioning is unavailable.

Despite the health risks of power outage, especially to vulnerable populations, data describing power outage exposure is extremely limited. Only one New York State-wide dataset describes outage exposure across space and time. Other studies of power outage have relied on large-scale events (e.g., hurricanes) as a surrogate of outages across a broad area (e.g., an entire city) at a specific time (e.g., August 2003)*.* These studies cannot disentangle health effects from disaster exposure and power outage exposure, or estimate exposure-response relationships.

In our previous work, we created a new national dataset of 10-minute resolution power outage exposure in sub-county spatial units for the continental United States. We used this dataset to describe power outage exposure by region and social vulnerability, finding that outages were more common in the southeast and northeast US, and with high outage incidence and high social vulnerability co-occurring most frequently in the southeastern US. This dataset will allow the characterization of exposure-response relationships between power outage and health outcomes nationally.

However, there are still major challenges with exposure assessment using this newly available power outage data. First, there is no standard definition of power outage exposure in the literature, which would allow comparison and aggregation of study results. There is also no literature on the clinically relevant length of power outage. In order to conduct an epidemiological study of power outage exposure using these newly available data, power outage needs to be defined and quantified. Any definition of power outage exposure would depend on the clinically relevant length of outage exposure. An incorrect assumption about the clinically relevant length of power outage could substantially bias the results of an epidemiological study. Second, both the national dataset we developed and the existing New York State dataset are missing large amounts of data. This missingness could also substantially bias results of an epidemiological study of power outage and any health outcome.

In this paper we will address these issues with power outage exposure assessment by running simulations. We will test how incorrect assumptions about the clinically relevant length of power outage could bias the results of an epidemiological study of the health effects of power outage. We will also test by how much missing power outage data (as it appears in the new national dataset we have created and in the New York State data) could bias the results of a study. Our results will allow us and other researchers to define power outage exposure in studies of power outage exposure and any health outcome, using the datasets currently available and in other power outage datasets available in the future.

Methods:

Simulation set up:

We designed a simulation representing an epidemiological study measuring the association between power outage exposure and all-cause hospitalization rates by county-day, though this outcome is standing in for many potential outcomes of interest and could be anything. In this hypothetical study, daily binary power outage exposure is measured in 100 US counties for 1 year. Daily county-level hospitalization rates are also measured over 1 year, and the study aims to estimate the effect of county-level binary power outage exposure on county-level hospitalization rates.

Real power outage data:

We wanted to simulate power outage data resembling the real power outage data (which we used in Do et al. and will use in future studies) as closely as possible. In our previous work, we purchased raw power outage data covering the continental US for the years 2018-2020 from poweroutages.us. These data came from the public websites maintained by utility companies, designed to be used by utility customers to check if there is a power outage in their area. Poweroutages.us scraped counts of customers without power from these websites in real time, every ten minutes, from 2018-2020. The resulting data contains 10-minute resolution counts of customers without power all served by the same utility in a sub-county unit. Utilities define a ‘customer’ as a grid connection, which can correspond to a household, apartment building, or business. A sub-county unit can be an entire county, city (where there are possibly multiple cities in a county), or neighbourhood (where there are possibly multiple neighbourhoods in a city, in turn nested in a county). All sub-county units were nested inside counties. Many utilities can serve the same location, so these sub-county units were not necessarily geographically distinct. Two houses next to each other might be in two different spatial units in the power outage data if they were served by different utilities.

The sub-county unit level data is a time series of ten-minute intervals covering 3 years, where there is an estimate of the number of customers without power in a sub-county unit in every 10 minute interval. The counts of customers out do not necessarily track the same customers: if 10 people are reported without power in two subsequent 10-minute periods in a subcounty unit, the data do not contain information about whether the same 10 households were out. The data only show that 10 households were out in each interval.

Because all subcounty units were nested inside counties, we were able to aggregate 10-minute counts of customers without power to the county level. We also aggregated these counts to the hourly level.

Previous definitions of power outage exposure in the literature:

Other studies have defined binary power outage exposure over a spatial unit (whether there was an outage, Y/N, in a county or zip code etc.) by considering an outage “on” if the percentage of customers without power in a spatial unit exceeds a threshold. There is exposure misclassification inherent in this definition: when the spatial unit is ‘exposed’, some households in the spatial unit will be without power and others will not. Studies have dealt with this by conducting sensitivity analyses varying the threshold after which a unit is considered exposed for power outage. For example, Northrop et al. considered a spatial unit exposed to power outage if more than 10% of the customers served in that unit were without power, and conducted two sensitivity analyses where they considered a spatial unit exposed to power outage if more than 20% and 30% of the customers served in that unit were without power. As the threshold increases, the specificity of this type of definition of power outage increases.

The length of power outages matter. Studies have defined daily binary power outage exposure over a spatial unit by considering a unit exposed if there were 8+ consecutive hours where the percentage of customers out exceeded a threshold such as 10% within a 24-hour period. However, there is no literature on the clinically relevant length of power outage. 8+ hours has been used because batteries for most electricity-dependent medical equipment last 8 hours. During a power outage, electricity-dependent medical device users might experience adverse health effects without their equipment immediately after losing power to the equipment. Without air conditioning or heat, indoor temperatures may also begin to change over the course of 8 or more hours.

The clinically relevant length of outage depends on the health outcome being studied. In reality, for medical-device related emergencies, pediatric asthma and injuries, and cardiorespiratory health effects in older adults, some individuals may experience health effects as soon as a power outage begins. As an outage lasts longer, the likelihood of adverse health effects increases, and we hypothesize that there are threshold effects for some outcomes (ex: 8+ hours for medical equipment related emergencies, or some number of hours for heat).

There is a difference between the clinically relevant length of power outage for an individual, and the clinically relevant ‘length of outage’ in a spatial unit. If the number of customers out in a spatial unit exceeds 10% for 8 hours, this does not mean that 10% of customers are without power for 8 hours consecutively. Some customers could have power restored, and others could experience an outage 4 hours into the 8-hour window, and contribute to the percentage of customers without power. If the number of customers out in a spatial unit exceeds 10% for 8 hours, this may indicate that many individuals in the spatial unit are without power for approximately 8 hours, which may in turn produce health effects. Again, in reality, as outages (as measured at the spatial unit level) last longer, the likelihood of individuals in that spatial unit experiencing adverse health effects increases, and we hypothesize that there are threshold effects at certain outage lengths for certain outcomes.

Simulation on the clinically relevant length of power outage:

In this simulation, we aimed to test how much incorrect assumptions about the length of clinically relevant power outage would bias the results of an epidemiological study. At the spatial unit level, if ‘power outage exposure’ as defined above (at least X% of customers without power for Y number of hours) produces the most health effects when Y >= 8, how much would misidentifying the length of outage bias the results of an epidemiological study? If shorter or longer outages produced health effects, and we assumed 8+ hour outages produced health effects, this would bias study results. Said differently, if it was instead 4+ hour outages or 12+ hour outages (as measured at the spatial unit level) 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 non-differential exposure misclassification.

To do this, we modeled two scenarios where the ground truth was that 4+ or 12+ power outages caused health effects, but the researchers conducting the study (us lol) incorrectly assumed that 8+ outages caused health effects. In this simulation, we created outcome data – hospitalization data – where we increased the hospitalization rate when a county-day was exposed to a 12-hour outage or a 4-hour outage. Then, we modelled the relationship between the outcome data based on the 12-hour exposures and 4-hour exposures, and the exposure data based on the 8-hour power outage definition. Finally, we evaluated by how much this exposure misclassification biased the results of these studies.

Data preparation for the simulation:

We generated 100 simulated counties. Then, we assigned each county a number of nested sub-county areas. We determined the number of sub-county areas by drawing from the empirical distribution of sub-county areas in the poweroutages.us dataset. We populated each sub-county area with a total number of simulated customers, again drawn from the empirical distribution of customers in sub-county areas in the real poweroutages.us dataset. We then generated ten-minute counts of customers without power for each sub-county 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 sub-county unit. We constructed our simulated data from the proportion of customers without power because of very different counts of customers served by sub-county unit nationwide.

We aggregated these sub-county 10-minute counts of customers without power to the hourly level. We also aggregated hourly counts of customers out to the county level, by summing all the customers served in each sub-county area and the customers out in each hour in each sub-county area. Figure 1 is a flowchart of the data structure in the simulated and real data and how we aggregated it.

We used a “0.5% out” definition to develop a county time series of hours exposed to power outage based on hourly-county counts of customers without power. We marked a county-hour exposed if the percentage of customers without power in county *i* during hour *j* exceeded 0.5% of the customers served in county *i*.[[1]](#footnote-1) Finally, we aggregated to the daily level: we considered a county-day as exposed if there were 8 consecutive hours of ‘power outage on’ (customers without power percentages > 0.5% of county) in that county on that day. Power outages could last more than 24 hours. We also considered a county-day as exposed if a power outage lasting longer than 8 hours ended on that county-day. This process produced a one-year time series indicating if there was an 8+ hour power outage on each day for each of the 100 simulated counties.

First, to establish a ground truth case representing an unbiased scenario where exposure was measured correctly, we generated simulated outcome data of all-cause hospitalization counts by day and county based on the simulated 8+ hour exposure data. We drew cardiorespiratory hospitalization counts for each county-day based on the total number of customers living in a county from a Poisson distribution with a base rate of 0.1%. County-days that met the 8-hour outage definition received a 1% rate increase (for a total hospitalization rate of 0.101%).

We then created two additional exposure datasets for each county, marking a county-day as exposed if there was either a 4+ hour outage, or 12+ hour outage (customers without power counts > 0.5% of total customers for 4 or 12 consecutive hours). We generated two additional datasets of outcome data (simulated ‘all-cause hospitalization data) based on the same hospitalization rate of 0.1%, and a 1% rate increase on days with 4+ hour and 12+ hour power outages. This represented exposure misclassification, where the misclassification was happening because we identified the length of a clinically relevant power outage incorrectly.

To generate a ground-truth estimate for the effect of an 8-hour power outage on daily county-level hospitalization counts, for each day exposed in a county, we chose control days that were more than 1 week away from the exposed day, but within 4 weeks of the exposed day, matching on the day of week. This process was meant to approximate what could be done in an augmented difference-in-differences design that could be used to compare hospitalization rates in counties exposed to power outages with those not exposed in a real study. We used these exposed and control dates in Poisson models. We used one Poisson model per 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. These models represented an unbiased scenario where exposure was measured correctly – 8 hr exposure caused an increase in hospitalizations, and this was reflected correctly in the simulation.

To represent the exposure misclassification scenarios, we paired the 8+ hour exposure data with outcome data generated based on 4+ and 12+ hour exposure data. We used two Poisson model per county (for a total of 100 models) to model the relationship between the binary 8+ hour power outage exposure and daily hospitalization counts based on the 4+ hour exposure data, and 12+ hour exposure data, again with an offset for the number of customers in a county. In the case where 4 hours was the relevant exposure, we would have classified many exposed county-days as unexposed, whereas if 12 hours was the relevant exposure, we would have classified many unexposed county-days as exposed when we used an 8-hour duration for analysis.

Simulation on missing data:

Many counties in the poweroutages.us dataset are missing substantial amounts of data. Some are missing up to 70% of observations. This missingness has the potential to bias effect estimates from an epidemiological study of power outage exposure and a health outcome. Counties with no or little missing data contribute information towards the overall effect estimate, while including counties with substantial amounts of missing data in an analysis could introduce bias.

Substantial missingness in the poweroutages.us dataset happens in two ways. In some cases, data are missing for an entire sub-county unit: the unit is never included in the dataset. This may happen because a sub-county unit is located in a rural area, and is served by a small co-operative utility without a website. Because poweroutages.us data is scraped from websites, no data – no average value of customers without power, or even the number of customers served by the utility in the subcounty unit – is included in the dataset in these cases. Because there is no information at all included in the dataset about these missing areas, interpolating the number of customers without power in missing subcounty units is near impossible to do with any accuracy.

Data may also be missing from the poweroutages.us website because utility websites may be offline or inaccessible for long periods of time (months or years). People living in the areas served by utilities without active websites are not represented in the dataset during the time periods when websites are not accessible. In this case, interpolating missing values is also near impossible, since values in a months or years-long period of missingness can’t be interpolated from prior observations with any accuracy.

In this simulation, we treat missing data as if there are no customers without power at any time in the missing subcounty units and missing time periods. Zero is also by far the most common value, and the average value, of customers out in the non-missing subcounty data in poweroutages.us. If data are missing at random, this means any bias from missing data would be towards the null.

To reduce bias due to missing data in an epidemiological study of power outage exposure and a health outcome using the poweroutages.us dataset, researchers could exclude counties that are missing more than a threshold percentage of observations. Within counties with large amounts of missing data, treating missing data as if it indicates no exposure, effect estimates are likely to be very biased towards the null. Counties with no or little missing data contribute information towards the overall effect estimate, and within these counties, bias is likely small. Excluding counties with high missingness could result in less biased overall effect estimates. To do this, researchers must identify the threshold at which a missing data in a county begins to severely bias effect estimates, and use this threshold to determine which counties to include in an analysis.

In this simulation, we aimed to find this threshold. We simulated an epidemiological study of daily county-level power outage exposure and daily county-level hospitalizations, as in the previous simulation. We aimed to estimate how conducting this analysis with an exposure dataset missing increasing amounts data would bias the effect estimates of the study. We created exposure datasets missing 10%, 30%, 50%, and 70% of observations, and repeated the analysis with a complete dataset, and then with each of these increasingly incomplete datasets. We assessed bias in the results in each of these five cases.

aimed to estimate how the bias introduced into an analysis of daily county-level power outage exposure and daily county-level hospitalizations increased, as more exposure data was missing from the

model a scenario where 8+ hour power outages at the county level resulted in increased hospitalizations, as in the previous simulation but the researchers (us) were missing increasing percentages of exposure data. We created power outage exposure data describing 8+ hour outages, and created hospitalization data based on this exposure data where power outages resulted in increased hospitalizations. Then, we removed increasing percentages of observations from the exposure dataset, creating datasets missing 10%-70% of observations, and modeled the relationship between exposure data with this missingness and the outcome (hospitalizations). Finally, we assessed by how much missing data biased the results of each of these studies, hoping to identify at which percentage of missing data results were biased enough to justify removing counties from a larger study.

Data preparation for simulation on missing data:

We generated power outage exposure for 100 simulated counties as in the previous simulation. We aggregated the simulated data from sub-county units to the county level, and identified days exposed and unexposed to 8+ power outages. This process produced a one-year time series indicating if there was an 8+ hour power outage on each day for each of the 100 simulated counties.

As in the previous simulation, we generated simulated outcome data of all-cause hospitalization counts by day and county based on the simulated 8+ hour exposure data. We drew cardiorespiratory hospitalization counts for each county-day based on the total number of customers living in a county from a Poisson distribution with a base rate of 0.1%. County-days that met the 8-hour outage definition received a 1% rate increase (for a total hospitalization rate of 0.101%).

To test by how much missing data biased the results of a study relating daily county-level power outage exposure to daily county-level hospitalization rate, we created four additional exposure datasets, each with increasing percentages of missing observations (10%, 30%, 50%, 70%). To do this, we randomly removed observations from the complete exposure dataset until there were 10%, 30%, 50%, or 70% of observations missing. We treated missing observations in these datasets as if they indicated no power outage exposure.

We then modeled the relationship between exposure in each dataset (assessed as well as possible given the missing data) and all-cause hospitalization counts generated based on a complete dataset. We used five Poisson model per county (one for no missingness, and 10%, 30%, 50%, and 70% missingness respectively for a total of 100 models per dataset) to model the relationship between the binary 8+ hour power outage exposure based on the datasets with increasing missingness, and daily hospitalization counts based on the complete 8+ hour exposure data, again with an offset for the number of customers in a county.

We quantified the percentage of missing data in the following way. If no power outage exposure data was missing from a county, then the data should have information on all the customers served for all of the hours in a year. The total amount of data missing from a county can be represented as a percentage of that total customer-hours in county *i*. Customer-hours present in county *i* should be:

[customers served in county *i*] \* [number of hours in the year] = N customer-hours

When we removed data, we removed

Caveat: it’s likely that not all the data missing from poweroutages.us are missing at random. Some may be – some utilities may not have websites simply because they haven’t developed them, or utility websites may be undergoing maintenance or development, which could lead to random missingness. However, a lot of missingness is likely patterned: for example, small, rural utilities are much less likely to have a website or API, meaning that many of those utilities may not be in the poweroutages.us dataset. It’s also possible that utility websites and APIs crash during extremely large outages, meaning that there may be missing data when counts of customers out are extremely high. Additionally, in both these cases, the missingness may be related to health outcomes influenced by power outage. For the outcome of hospitalization, rural counties may have different hospitalization rates than urban ones. Large power outages are often caused by natural disasters, which may also affect hospitalization rates. However, to evaluate the potential effects of missingness not at random, we would have to speculate about so many different simulation parameters with almost no information to inform our assumptions that such a simulation would not be informative. We therefore limit ourselves to an investigation of missingness that we assume is random.

RIP

Many counties in the poweroutages.us dataset are missing substantial amounts of data. Some are missing up to 70% of observations. This missingness has the potential to bias effect estimates of the relationship between power outage exposure and a health outcome. In some cases, data is missing from an entire sub-county unit in the poweroutages.us dataset. This may happen because a rural area is served by a small utility without a website. When that happens, the dataset contains no information on that unit – no average value for customers out, and sometimes not even an estimate of the number of customers served by the utility in the sub-county unit area. For this reason, it’s very hard to infer or interpolate missing values. Data may also be missing from the poweroutages.us website because utility website may be offline or inaccessible for a long period of time (months or years), and because data is collected into this dataset in real time off utility websites, people living in the area served by that utility are not represented in the dataset during that period of time. Interpolating missing values in this case is also challenging, since values in a months or years-long period of missingness can’t be interpolated from prior observations with any accuracy.

It’s possible to treat missing values in the poweroutages.us dataset as if there is no exposure during the missing periods. Zero is also by far the most common value, and the average value, of customers out in the non-missing subcounty data in poweroutages.us. By treating missing values as non-exposure, bias from missing data in the dataset will be towards the null.

However, we still need a strategy for dealing with this bias. A strategy would be to exclude counties after a certain amount of their observations are missing. If no power outage exposure data was missing from a county, then the data should have information on all the customers served for all of the hours in a year. The total amount of data missing from a county can be represented as a percentage of that total customer-hours in county *i*. Customer-hours present in county *i* should be:

[customers served in county *i*] \* [number of hours in the year] = N customer-hours

To reduce bias in an epidemiological study of power outage exposure using the poweroutages.us dataset or another dataset with similar missingness issues, we could exclude counties with more than say 50% of observations missing. We want to figure out at what threshold missing data biases the results enough that a county with that percentage of missing data is biases the effect estimates instead of contributing information.

The data missing from poweroutages.us are likely not all missing at random. Some may be – some utilities may not have websites simply because they haven’t developed them, or some websites may be undergoing maintenance or development which would lead to random missingness. However, a lot of missingness is patterned: for example, small, rural utilities are much less likely to have a website or API, meaning that many of those utilities may not be in the poweroutages.us dataset. It’s also possible that utility websites and APIs crash during extremely large outages, meaning that there may be missing data when counts of customers out are extremely high. Additionally, in both these cases, the missingness may be related to the outcome of hospitalization. Rural counties may have different hospitalization rates than urban ones. Large power outages are often caused by natural disasters, which may also affect hospitalization rates. However, to evaluate the potential effects of missingness not at random, we would have to speculate about so many different parameters in a simulation that we don’t feel that a simulation would be informative. So, we’ve limited ourselves to an investigation of missingness that we assume is random.

When counties are missing such large amounts of data, interpolating missing values would involve very strong assumptions about power outage exposure. For example, if a county is missing data in one sub-county area for an entire year, there is really no information on which to base assumptions about power outage exposure during this year.

* Many counties missing a lot of data
* Need a plan for dealing with these data
* What happens to the effect estimates when there is a ton of missing data
* Could choose to exclude units where the data is unreliable, but we’d need a threshold at which to do that
* Chunks and long periods of time can be missing
* Very hard to interpolate what power outage should be during these times because when there are long periods of time missing or whole chunks missing, any assumptions to fill in these periods would have to be very strong assumptions
* Treating these regions as unexposed seems like the only reasonable option, especially since the vast majority of power outage counts are 0 most of the time anyway and that is the mean power outage value.
* Going to do a simulation where we remove increasing percentages of data from all the units in the analysis, and then see by how much this biases the results of the study
* Assume data is missing at random even though probably not true
* But impossible to really do anything else

Simulation set up:

1. This is smaller than the percentage thresholds described above, because counties are larger than the spatial units used in Northrop et al., meaning that the percentage of customers that need to be out for it to be a ‘substantial’ power outage is a smaller proportion of the population. [↑](#footnote-ref-1)