Jurisdictional Vulnerability Assessment
Frequently Asked Questions

The questions in this document have been received from various jurisdictions. Although the responses may not capture the specific details relevant for each question, we are providing the responses from a variety of subject matter experts as they may be informative to others. Please follow-up with Cailyn at CSTE (clingwall@cste.org) or your Science Officer for additional questions.

Step 1: Identify and Prioritize Indicators

Acute hepatitis C may not be a reliable outcome of injection drug use due to the small number of counties recording cases. Is there another option such as combining acute and chronic cases together or adding HCV and HIV cases together?

- You can consider using (either alone or in combination) chronic hepatitis C.
- Other states are using combined acute and chronic hepatitis C cases and, particularly for lower burden or less populated states, this is a supported option. For chronic hepatitis C cases, consider age range options that might help focus on more recent infections, such as <baby boomer, <40 or <30, depending on your case counts.
- We would not recommend adding HIV cases into the proxy measure since there is a fairly substantial delay in diagnosis and therefore transmission may not have occurred in the location of the diagnosis.

On average, how many acute HCV incidences did the National Assessment report for each county?

- In 2012 and 2013, 1,778 and 2,138 cases of acute HCV infection were reported to CDC. Of the reported cases, 1710 and 2074, respectively, were reported from 2,970 counties with valid county FIPS codes. The mean rate was 0.97 per 10,000 population, ranging from 0 to 100.4 per 10,000 population.

On average, how many acute HCV incidences did TN have for each county?

- For 2012, there was a range of 0-15 cases per county in the state of TN (n=144 total) and for 2013, there was a range of 0-14 cases per county (n=138 total). These included both confirmed and probable acute HCV.

Should HIV and STDs be used as indicators or primary outcome?

- The National Assessment did not use either of these outcomes in the analysis. HIV has a substantial diagnosis delay and thus transmission may not have occurred in the location of the diagnosis. STDs are usually diagnosed more timely, but are not as directly associated with unsterile injection drug use. However, consider looking at both of these for contextual information to inform where HIV and STD prevention services are and may be needed within the identified communities (e.g., counties).

Should I combine skin infections and endocarditis together?

- Some states are considering using endocarditis for the proxy outcome (Schranz AJ et al. Trends in drug use–associated infective endocarditis and heart valve surgery, 2007 to 2017: A study of statewide discharge data. Ann Intern Med 2018 Dec 4; [e-pub]. (https://doi.org/10.7326/M18-2124)
- In the event that skin infections and endocarditis are NOT both specifically defined as drug-use-associated, we wouldn't recommend combining, since they would likely represent various causes, and would be cautious about using either if not limited to drug-use-associated.

Which one of the indicators in EMS dataset should be used for opioid overdose? Overdoses reported by dispatch, or provider impression, or cause of injury?

- Case ascertainment can be difficult when using EMS data. Here are some considerations on the variables that may be available:
    - Overdoses reported by dispatch – Only about 20% of calls to a 911 service correctly identify the patient as having a drug overdose.
    - Provider impression – While this is the best variable to select, it is not always complete, nor is it always accurate. Sometimes EMS will put down heart issues for drug overdose patients going into cardiac arrest.
    - Cause of injury – Many people associate injuries with blood and guts, not poisoning. In the old ICD-9-CM system any ICD code between 800-999 was an injury. Poisoning is in the upper 900s mostly. It is an injury. This is not a very good variable to select.
- Below is information on a paper using the Provider's impression and a combined 911 call (an either or approach). However, we found that it omits too many records to be a good approach. Here is an example of a study using that approach: https://www.ncbi.nlm.nih.gov/pubmed/25905856
    - The best approach that we found was to look at whether naloxone was administered or not. If you are using the NEMSIS data, there is a table showing what kinds of medications were dispensed. Sometimes EMS will administer naloxone when they don't know what is wrong with a person. However, EMS personnel can size up a scene very quickly and have environmental cues that suggest if the injury is a drug overdose.
      1.) https://www.ncbi.nlm.nih.gov/pubmed/28481656
      2.) https://www.ncbi.nlm.nih.gov/pubmed/30091966

How did the National Assessment identify indicators?

- The National Assessment started with a framework similar to the social vulnerability index to determine domains: sociodemographic, unsterile injection drug use proxies, drug-related behaviors, drug-related outcomes, and drug-related health care access. The National Assessment initially began with 40 combinations of indicators and potential sources and narrowed down to 15 indicators with complete county-level data. **See Table 1 for the 15 indicators assessed for association with acute Hepatitis C virus infection as proxy for unsafe injection drug use.*
- While considered relevant, drug related arrest data, non-fatal overdoses, and EMS overdose were not included because they were not available at the county-level nationally.
- Other indicators were excluded because they were deemed not as proximally associated with the outcome of interest, unsterile injection drug use: HBV, HIV diagnosis, age.

Step 2: Compile Data and Calculate Indicators

What years of data are appropriate to use for the assessment?

- We suggest aiming to use data from the year 2015 and newer so that it is newer than the data used in the national assessment. This will also help to account for changes in the opioid epidemic that may affect the outcomes of interest.

Will CDC or CTSE make available or provide assistance to get updated data from the original models (e.g., DEA data)?

- At this time, NCHHSTP does not have data more recent than 2014 and in most cases 2012 or 2013. Additionally, we have restrictive data use agreements that do not allow sharing of even the aggregate county-level data from DEA, HSIP 2012 Gold Database and National Vital Statistics.
- As an alternative to the DEA data, we recommend states consider state Prescription Drug Monitoring Program data. If that is not an option, please contact your Science Officer and we can discuss additional alternative options.
- CDC does not have an alternative to the HSIP data at this time, but please contact your Science Officer if you are interested in this data and need assistance.
- For the National Vital Statistics data, please consider the small area estimates the NCHS has developed as an alternative if in-state data are not available: https://www.cdc.gov/nchs/data-visualization/drug-poisoning-mortality/index.htm

Are there data sources for the social determinants of health?

- The NCHHSTP AtlasPlus includes 5 social determinants of health indicators: poverty, uninsured, less than a high school education and vacant housing nationally and by state and county; percentage of population living in rural areas nationally and by state; and county urbanization level.
    - Access this resource at: https://www.cdc.gov/nchhstp/atlas/index.htm

- The Robert Wood Johnson Foundation County Health Rankings includes data on the social determinants of health related to clinical care, social & economic factors, quality of life, and health behaviors.
    - Access this resource at: http://www.countyhealthrankings.org

For indicators available in two different datasets, should I choose the one with higher counts?

- Generally, we'd suggest using the dataset that is most reliable / highest data quality. For example if one dataset regularly counts preliminary test results and the other is based on confirmed (thus lower counts), we suggest using the dataset with confirmed cases only.

We have data for 2015-2017 for many variables. Do you recommend averaging the numbers for these three years or using the highest value?

- It is not recommended to use the highest value. Under the assumption that there is enough data from each year for the outcome (hep c cases) and indicator variables, it is not recommended to use the highest value or average. Instead, it is recommended that you use data 'year by year'. However, sometimes aggregating (averaging) the data is necessary.

For the Admission for IDU treatment, the state's Mental Health & Substance Abuse Service data can be used. How is that data obtained? Are ICD codes used? And if yes, what ICD codes specifically?

- In the case of the state of TN, The TN DMHSAS collects claims data from sites that they fund for mental health services. The claims data that we were able to get from them only contains these state-funded claims. At the time of the study, there wasn't a way to collect medical claims directly from the facilities, so they leveraged inter-agency partnerships. Another note about this variable is that state claims were looked at for admissions around categories of drugs that could be injected (heroin, opioids, prescriptions) -- there wasn't a specific category of 'admissions for injection drug use'.

For the MME rate for analgesics, the Tennessee PDMP database is an option. What NDC codes were used to distinguish the prescriptions?

- One option is the most current version of the CDC's MME conversion table (https://www.cdc.gov/drugoverdose/resources/data.html). This table includes the majority of the opioid NDCs that appear in the CSMD. This was supplemented with the NDCs that have been identified in past versions of the table that are no longer included in the most recent version.

Step 3: Develop Vulnerability Assessment

Statistical Analysis

When a variable is reported in percentage, is it ok to use the population as the offset for that variable?

- Yes, and please make sure that you are modeling a count and not a percent. For example, let's say that the rate is 60% and the population is 1000. The estimated number of events (i.e., the rate's numerator) is 600. So this is the count that you are modeling using Poisson or NB. The denominator (or offset) is 1000.

Can we set the offset for a subset of variables? For instance, if I have an indicator reported in rates, should I use the population as the offset again for that variable?

- If the rates are all based on the same population, you should use the same offset.

When should an offset variable be incorporated in count-based (i.e., Poisson/Negative Binomial) regression models?

- This depends on your data. In general, if you are strictly modeling counts, then an offset is not required; however, if you are modeling event rates, then a population offset should be included. If you're unsure, think of the population denominator for each of your sampling units (e.g., counties). Is the population denominator constant between sampling units over study follow-up? If so, then an offset is not required, but if the population denominators are different, an offset is recommend. For instance, if you were modeling the number of incident HCV infections over the past 12 months across 5 counties, and each county had starkly different population denominators, then the estimated counts alone would not be meaningful; instead, the number of people living in each county (i.e., the offset) should be incorporated and rates reported. We typically recommend researchers use a population offset if these data are available. When incorporating a population offset in your regression model, the value must be transformed via natural log (not $\log_{10}$). The natural log corresponds to the link function used in count-based models.

After checking the assumptions, did all the included indicators have a positive/negative correlation with HCV cases for the TN vulnerability assessment? Were none of them were intercorrelated with each other?

- Yes. The PCA, Factor Analysis, and Correlation Matrix steps of TN's dimension reduction "algorithm" eliminated collinear/highly correlated covariates, so the final Poisson model did not contain highly correlated variables. In addition, TN used a step-wise insertion procedure (as Van Handel, et al. did) to retain only statistically significant predictors of our acute HCV outcome in the model (so all indicators in the final model did have some significant association with the outcome).

How do you select your final regression variables while accounting for potential multicollinearity?

- When dealing with multicollinearity, we advise the reader to review the *JVA Regression Model Taskflow* tool. Generally, we first recommend running correlations on all predictor variables, irrespective of the count outcome. Contingent on data types, these may include Pearson, Spearman, or Bi-serial correlations. While there is no standard cut-off, take note of any predictor variables that are highly correlated with each other (e.g., >0.7). Then, run bivariable regression models and retain predictors with $p<0.1$. If any two predictors are significant with the outcome at $p<0.1$, but also correlate strongly with each other, retain the predictor most significant with the outcome and eliminate the other. After redundant variables have been eliminated, build the final model using forward, backward, or stepwise selection. If the reader wants to go beyond simple correlation tests, they may employ the methods implemented by Tennessee, specifically, PCA and Factor Analysis. As a side note for general linear regression problems, variance inflation factors (VIFs) and condition indices are available, but equivalents are not easily calculable for count-based regression models.

If a state has one very large city with many smaller counties, should we exclude that city from analysis?

- One suggestion is to perform the assessment with and without that city to compare how results might differ. Also consider if more granular data (e.g, census tract) is available for the city and a mini-assessment can be done within the city.

How do you decide between using a Poisson regression model versus a Negative Binomial Regression Model?

- We advise the reader to review the *JVA Regression Model Taskflow* tool and the *Jurisdiction Level Vulnerability Assessment Technical Assistance Webinar* slides for tips, but briefly, Poisson regression is appropriate when the mean and variance for the count outcome are roughly equal. If the outcome variance is much higher than the mean (i.e., the data are highly right skewed), Negative Binomial regression is more appropriate. Beyond this simple summary check, a Chi-square goodness of fit (GOF) test for overdispersion may be used. Slide 8 of the *Jurisdiction Level Vulnerability Assessment Technical Assistance Webinar* details SAS code for this test. A significant GOF test indicates the data <u>do not</u> well fit the Poisson model, and a Negative Binomial model should be employed.

Which of the following methods is more appropriate? Poisson regression or Inflated-Poisson regression, as we might get many zeros in different areas.

- This depends on your data. In the case of excessive zeroes, a zero-inflated model is preferred. A general rule of thumb for excess zeroes is 30% to 40% or higher of the outcome data are 0. A formal test for zero-inflation exists and is called the Vuong's Non-Nested Hypothesis test. SAS has provided a macro, available at: http://support.sas.com/kb/42/514.html. Vuong's Hypothesis test can also be implemented in R using the *mpath* package. Overdispersion is a different issue related to the underlying assumptions of the Poisson model. The Poisson model assumes that the mean and variance of the count distribution are equal. If they are not, one may choose to include a dispersion parameter in the Poisson model or switch to the negative binomial model which estimates mean and variance separately.

We expect to get many zeros for data from many locations within the state. In that case, should we use a Zero-inflated Poisson Model or should we just remove them from our analysis?

- We advise against removing zeros for data from analysis because the zeros provide important information.

Regarding the census data, how did you consider the uncertainty in the income data or the % of population under poverty? Was the reported margin of error for the simulation used or did you use normal distribution to produce the data?

- We used normal distribution for each indicator in the simulation to calculate the confidence intervals for consistency, since the reported margin of error for not available for all indicators.

For the dataset that we do not have detailed data (such as drug-related crimes that are only reported based on counties and only in some of the counties in our state). Should we ignore them for the regression modeling and just illustrate them on the map in case they inform our healthcare officials?

- In general, it's better to use the same level of data for all areas in the modeling and ranking. However, rather than including an incomplete variable in the model, you could develop a map to add contextual information using data available in those counties.

In case you needed to convert a zip code based indicator to a county-based or census tract based indicator when a zip code crosses multiple borders, how would you do this allocation?

- While working at the census boundary level, over ZIP is much preferred, in those cases when only zip data are available, GRASP has used the HUD crosswalk files here: https://www.huduser.gov/portal/datasets/usps_crosswalk.html. Each ZIP would map to several CTs and vice versa. ZIP codes are not designed to fit census boundaries so they are difficult to go between. The nice thing about the HUD files is that because they have access to proprietary residential data, the crosswalk files are supposed to more accurately represent geographic locations of populations than a more basic areal proportion method. The down side of using these files is that you are ultimately introducing error into your dataset that is unmeasurable, and thus unreportable. If you choose to go this route, consider including this as a limitation.
- The National Vulnerability Assessment cross-walked zip code to county based on residential proportion.

When do you transform data to log10? Should we transform the response variable or only dependent variables? Do you check the assumptions on the log transform or the count data?

- Regarding log transformation, when there is an indicator or variable with a high value and wide range (e.g., per capita income, population), log transformation can be done to improve model fit.

Based on the data, which variables do you suggest to be transformed using Log10?

- Log transformation is usually performed to improve model fit. So, the choice of which variables to transform is based on the behavior of each individual variable. The results can be compared graphically but, as a start, I would recommend that you run both models (with transformed and non-transformed variables) and compare the model fits using the AIC or likelihood ratio test. The variable with the best fit is the preferred one.

How do you recommend handling missing data?

- Missing data problems can be complex, but there are a few general ways to handle this circumstance:
    - The most straightforward option is to employ list-wise deletion and only analyze complete cases. For SAS users, this is the default in many procedures, including PROC GENMOD.
    - The second option is to use mixed models. Mixed models are useful when outcome data are missing, but predictor data are available. With mixed models, the missing outcome data are assumed to be at random (MAR), and estimates are calculated via maximum likelihood functions derived separately for subjects with and without complete data. These values are then maximized together to obtain the final model parameters. This method gives unbiased estimates and standard errors when data are not systematically missing. If data are systematically missing, we term this missing not at random (MNAR), and this problem requires more complicated, yet imperfect procedures. For SAS users, mixed models for count data can be performed in PROC GLIMMIX; in R, users may use the *glmm* or *lme4* packages.
    - The final option is to use imputation. Imputation can be performed simply using summary measures (e.g., mean, median) or single imputation regression models, or with more sophisticated methods, such as multiple imputation. When performed correctly, multiple imputation produces the least biased results. SAS users can perform multiple imputation with PROC MI; R users should consider the *MICE* and *caret* packages. The type of imputation performed is at the discretion of the user, but broadly, if few data are missing (<10%), then summary or single imputation procedures are generally appropriate. For larger amounts of missing data, we recommend the use of multiple imputation.
- In general, we recommend a combination of the methods above, to act as sensitivity checks for the results; moreover, we generally do not recommend imputation for variables that have over 50% of information missing (although this can be circumstance-specific).

What do you do when your regression model fails to converge?

- We have created a document dedicated to convergence issues for mixed effects regression models (*JVA Convergence Questions*). This document is geared towards SAS users, but the broad tips can be implemented in any software. In general, convergence issues are resolved through trial and error, and employ quality checks on the raw data, evaluation for model misspecification, and consideration of the optimization algorithms. We recommend the reader investigate the referenced document.

How do you account for correlated observations in your dataset?

- We recommend the use of mixed models when your data are correlated or longitudinal (e.g., repeated observations over multiple time points). These models allow the user to specify a variance-covariance matrix to properly adjust for observation-to-observation correlations within sampling units. Under a mixed model framework, the user will typically specify fixed effect covariates (i.e., variables that are constant across sampling units), such as your independent predictor variables, and random effect covariates (i.e., variables that randomly vary across sampling units). For purposes of the HCV study, states and counties were modeled as random effects, to both account for both year-to-year correlations in the data and inherent clustering (i.e., counties nested within states). Fixed effects were the predictors: opioid death rates, opioid sale rates, etcetera. For SAS users, we recommend PROC GLIMMIX, and for R users, we recommend *glmm* or *lme4* packages.

How does the offset variable factor-in to predicted counts/rates from the regression models?

- When a user manually calculates predicted values from a count-based regression equation, the raw output is in log-counts per person. We can exponentiate these values and get predicted counts per person (i.e., person-rates). This rate is not population-specific, and can be comparable across other sampling units (e.g., counties). Many times, it's reasonable to multiply this person-rate by a constant to get an epidemiological rate (e.g., rate per 100,000). For SAS users, the procedure default is not to give this person-rate, but a predicted count, based on the offset population. In PROC GLIMMIX, the workaround is to take the predicted count value and divide by the offset population, then multiply by a constant to get the epidemiological rate. These steps are detailed beginning in slide 15 of the *Jurisdiction Level Vulnerability Assessment Technical Assistance Webinar*.

How do you interpret your final regression coefficients in your multivariable count models?

- If you wish to interpret the raw coefficients, the interpretation is in terms of log-counts. For instance, if you have a continuous predictor variable and the beta coefficient is 0.5, we would say for every one-unit increase in our predictor variable, we expect the log-count to increase by 0.5. Since the link function for count-models is the natural-log, we can exponentiate our betas and interpret them in the form of an incidence rate. Exponentiating 0.5, we get 1.65 and can say that for every one-unit increase in our predictor variable, the incidence rate for our outcome will increase by 1.65 times, or 65%. For multivariable models, it is assumed all other covariates are held constant.

<u>Social Vulnerability Index and Hot-Spot Analysis</u>

Is it appropriate to start with an optimized hot spot analysis and then use a Getis-Ord Gi as a more sophisticated hot spot approach?*

- The optimized hot spot analysis uses the Getis-Ord Gi*statistic, so just run the hot spot analysis once.

How should we use the Social Vulnerability Index (SVI) to inform our analysis? Should we incorporate our state's existing SVI or make our own based on the 15 indicators?

- You can start with the overall SVI and themes and see if any fit extremely well and other themes/variables fit poorly. If there are many poorly fitting themes/variables, maybe you exclude them from your analysis. You can find the data to download and a detailed description of all variables here: svi.cdc.gov. However if you create your own index, follow steps in the handbook on constructing composite indicators: https://www.oecd.org/sdd/42495745.pdf

What classification method do you recommend for the Social Vulnerability Index?

- On our SVI we use quartiles. (We also have an additional approach – of those tracts in the quarter of highest vulnerability, we flag those in the most vulnerable ten percent.) For much data I don't recommend going beyond quintiles (my preference, as it provides an 'average' category of observations in the middle). Beyond that, I often find it difficult to grasp the meaning of a category. But there are certainly data where more quantiles would be beneficial, or data that lends itself to natural breaks. See the link below for additional guidance (pages 12 and 13).

https://www.cdc.gov/dhdsp/maps/gisx/resources/cartographic_guidelines.pdf. See pages 12 & 13

Using the Hot Spot and Outlier Analysis, absolute hot posts and vulnerable areas to opioid overdoses and HIV/HCV outbreak can be identified. Any suggestions?

- Our geospatial analyst suggests looking into bivariate Moran's I using GeoDa software (https://geodacenter.github.io/workbook/5b_global_adv/lab5b.html) to identify areas with high vulnerability and high presence of disease outcome. You can compare time periods with a differential Moran's I which could help you assess progress and changes between years. Additionally drive time analyses or service area analyses could be very applicable. I am linking to an article Elaine was the lead author on relating to service areas and estimating populations within those regions which may be helpful (https://ij-healthgeographics.biomedcentral.com/articles/10.1186/s12942-017-0102-z which details the methodology behind https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6125785/). Another way to approach this research might be to look at treatment centers and compare that information to vulnerable populations information. What is the population density in those service areas and which areas are remote (including lack of access to public transportation).

After the hot spot analysis, is it an appropriate next step to complete a binary logistic regression to identify appropriate demographic indicators?

- Yes.

Are there SVI indicators that you recommend we prioritize, given the limited assessment timeline?

- Start with overall SVI rank and themes prior to prioritizing specific indicators.

Qualitative Methods

Should we consider integrating more qualitative work in our assessment methods?

- In addition to the vulnerability assessment, qualitative work that can identify prevention and intervention needs and gaps, in particular, it would be useful to informing a plan to address those needs. Other activities within the scope of the project should be discussed with your Science Officer.

Step 4: Identify Gaps in Services in Vulnerable Areas

Is there a particular distance from a resource that is considered a 'gap'?

- The National Assessment considered average daily commute as a distance for which to consider for 'HIV proximity' but didn't apply it to service distances. The average daily commute to access that service could be one idea.

General: Jurisdictional Vulnerability Assessment

Is IRB approval necessary for the jurisdictional vulnerability assessment?

- As required in the NOFO, this funding opportunity does not support research. Determination of research was conducted and based on the activities submitted in the initial workplans were deemed non-research. Please contact your Science Officer if a determination of research is needed.

*The vulnerability assessment project and associated technical assistance was funded through CDC's National Center for Injury Prevention and Control.