Supplementary Materials

Details of Dashboard Features

Dynamic documentation is available via links from the Dashboard landing page (https://covid19pvi.niehs.nih.gov/) in the form of a Quick Start Guide for users (https://www.niehs.nih.gov/research/programs/coronavirus/covid19pvi) and a Details page (https://www.niehs.nih.gov/research/programs/coronavirus/covid19pvi/details/index.cfm). Here, we summarize features of the current Dashboard interface.

On loading, the Dashboard displays a map of the continental United States, overlaid with a choropleth layer of overall PVI scores for the current day. Users can navigate the map by dragging and/or by zooming in and out with the PLUS/MINUS icons or keyboard shortcuts. Clicking on a county brings up the current daily scorecard. The scorecard shows the graphic representation of the PVI, the overall score, and the rank and score for each data stream. Selection of a county also populates the surrounding panels with county-specific information. Scrollable panels on the left of the Dashboard include plots of vulnerability drivers relative to their nationwide distribution across all U.S. counties, with the location of the selected county delineated. The panels across the bottom of the Dashboard report cumulative county numbers of cases and deaths; timelines of cumulative cases, deaths, PVI scores, and PVI ranks; daily changes in cases and deaths for the most recent 14-day period (a measure commonly used in reopening guidelines); and predicted cases and deaths for the given forecast horizon.

The main menu bar provides numerous options for visualizing the map that are detailed in the Quick Start Guide. The base map can be changed from the default gray to options that include satellite imagery and topology (Change Map menu option). Users can choose a WebMap by inserting a WebMap Portal Item ID. Details of the WebMap Extent options are available through Esri [1]. Cumulative case and death counts can be changed or removed through the COVID-19 Legend menu option. The size and opacity of the PVIs displayed on the map can be changed through the PVI Model Legend option. Interactive options to display data from different days are available in the 'Covid-19 Layers' panel for cumulative case and death data and 'PVI Model Layer' panel for the historical PVI.

To generate the display in Figure S1, the PVI Model Filter menu option was used to select all counties in the state of Alabama. The county of Autauga was highlighted by mouse-click, bringing up the Scorecard and populating info panels along the left and bottom of the Dashboard. Scrolling down the left side under the collapsible 'PVI Slice Legend' displays the PVI slice legend and histograms of the overall and slice-wise PVI across all U.S. counties. For each histogram, a black line indicates the location of the selected county within this distribution. For all distributions, lower values (to the left) indicate lower relative vulnerability (i.e. shorter slices) whereas higher values (to the right) reflect higher vulnerability (i.e. longer slices).

In addition to the Change Map, PVI Model Legend, and Covid-19 Legend options, additional display options are available through the main drop-down menu (see the Quick Start Guide). The PVI Model Filter option allows the interactive filtering of display options, including the number of profiles displayed. The top-ranked counties, which are those with the highest vulnerability according to the overall PVI, are displayed by default. The PVI Model Filter option also allows users to display the most vulnerable counties based on individual data streams, with options for multi-level filters. These options for restricting the ranges of one or more slices enabling the identification of counties with similar profiles. The PVI Model Clusters option provides additional opportunities for data-driven contextualization. The two options for clustering PVI profiles are labeled KMeans (agglomerative with k=10) and HClust (divisive, displaying the top-10 splits), after the algorithm used. Both options identify counties with a similar PVI profile "shape". The distance matrix is calculated from the integrated profile, which considers all slice scores and weights. For KMeans, the clustering is implemented as a Java port of the *kmeans* R function using the Hartigan-Wong algorithm [2]. For HClust, the clustering is implemented as a Java port of the *hclust* R function using the "complete" method [2]. By filtering the Dashboard to display only specific clusters, users can identify clusters with multivariate profile similarity. This enables the detection of clusters that may be geographically adjacent as well as geographically distinct to reflect the diversity of risk profiles on a national scale. The existence of geographically disparate areas with similar drivers highlights the need for integrated profiles to encourage effective local-level policies targeted at communities at risk.

Forecasting

To predict COVID-19 case and death counts, we model the observed log-counts for new cases and deaths in each U.S. county using a Bayesian functional model. We assume deaths are a proportion of active COVID cases, and we jointly model both cases and deaths so these data are shared between these observations. In this model, the expected case count is the case rate offset by the observed number of deaths and is spatially modeled across the United States. Currently, the model assumes a normal distribution on the logs of the observed counts.

Let C_{tj} and D_{tj} be the observed case and death count for day t, $1 \leq t \leq T$, and county j, $1 \leq j \leq J$. For county j, we also observe a vector $x_j = \left(1, x_{1j}, \ldots, x_{mj}\right)'$ of m explanatory variables that are static over time (e.g., population density) and a vector $z_{tj} = \left(z_{1tj}, \ldots, z_{ktj}\right)'$ of k explanatory variables that are observed each day (e.g., social distancing metrics). Let X_j be the matrix of observed explanatory variables, both static and dynamic, for the T time points in county *j*. Conditional on knowing the case and death rates λ_{Cij} and λ_{Dij} , we assume that C_{ij} and D_{ij} are Poisson variates where

$$\lambda_{ctj} = exp(X_j\beta \ + \gamma_c(s_j,t) + \ \log[\bar{C}_{tj}] + \epsilon_{tj}^{C}), \qquad (1)$$

and

$$\lambda_{Dtj} = exp\big(X_j\alpha \ + \gamma_D(s_j,t) + \ \log[\bar{\lambda}_{ctj}] + \epsilon_{tj}^{D}\big). \ (2)$$

To define these rates, $\log[\bar{C}_{ij}]$ is the log geometric-mean of the previously observed count; $\log[\bar{\lambda}_{ctj}]$ is the log of the new case rate for the last *m* days (i.e., $\log[\bar{\lambda}_{ctj}] = \frac{1}{m}\sum_{i=t-m}^{t} X_j\beta \ + \gamma_c(s_j,i) + \ \bar{C}_{ij}$); $\mu_c(s_j,t)$ and $\mu_D(s_j,t)$ are spatial process accounting for unobserved heterogeneity in the response at time *t* and county location s_j; and $\epsilon_{tj}^{C} \sim N(0,\tau_{cj}^{-1})$ and $\epsilon_{tj}^{D} \sim N(0,\tau_{Dj}^{-1})$. This defines a Poisson-lognormal model over C_{ij} and D_{ij}, which is an over-dispersed count model that allows for efficient sampling using Bayesian computation with conditional Gibbs updates.

The random fields borrow information from nearby counties under the assumption that geographically proximate counties have public health departments with similar testing strategies and testing resources and likely have similar responses to the pandemic, which would account for heterogeneity not captured by the covariates. Time is included as testing strategies and testing resources are expected to change over time as well as by region.

Modeling the Spatial-Temporal Power Term

Let $\gamma_C\big(s_j,t\big) = f_C\big(s_j\big)g_C(t)$ and $\gamma_D\big(s_j,t\big) = f_D\big(s_j\big)g_D(t)$, where $f_C \sim GP(0,\sigma_C[\cdot,\cdot])$ and $f_D \sim GP(0,\sigma_D[\cdot,\cdot])$, which are Gaussian processes [3] with a 0 mean and squared exponential covariance kernel functions $\sigma_C[s,s'] = \exp[-\tau_c(s-s')^2]$ and $\sigma_C[s,s'] = \exp[-\tau_D(s-s')^2]$, where τ_c and $\tau_D$are length-scale parameters controlling the amount of correlation between spatial locations s and s'. For f_C and f_D, it is generally assumed that the covariance kernel has an unknown variance component. In our case, we fix this parameter to 1 because it is unidentifiable given g_c and g_D.

For g_c and g_D, we use first-order B-splines [4], which are local linear piecewise splines. That is,

$$g_c(t) = \sum_{k=1}^{K_c} \zeta_{ck} b_{ck}(t),$$

and

$$g_D(t) = \sum_{k=1}^{K_D} \zeta_{Dk} b_{Dk}(t),$$

where $b_{ck}(t)$ and $b_{Dk}(t)$ are defined on K_c and K_D evenly spaced knots, respectively. We chose linear splines to minimize end-knot variability. This formulation is a tensor product formulation [4,5], which allows modeling the three-dimensional surface as the product of a two-dimensional surface and a one-dimensional surface.

Bayesian Specification and Computation

We conducted inference and prediction using Bayes' rule. As such, all parameters in the models described in Equations (1) and (2) are given proper priors, and inference is completed using Markov chain Monte Carlo (MCMC) methods. All coefficients on the explanatory covariates are given normal(0,10) priors. The precision terms, namely τ_{Cj}^{-1} and τ_{Dj}^{-1} , are given Gamma(10,1) priors. For the Gaussian processes, the length-scale terms τ_c and τ_D are given discrete uniform priors over a range of equally spaced values that cover a variety of covariance functions. Finally, the ζ_{ck} and ζ_{Dk} terms, which specify the spline coefficients over the time component in the tensor product, are given Bayesian P-spline priors [6]. This allows flexible smooth modeling of the time component and defines the variance of the tensor product Gaussian processes, conditional on the time component.

All priors are conditionally conjugate, with inference conducted using Gibbs sampling. In total, 11,000 MCMC samples were taken, with the first 1,000 disregarded as burn-in. We examined trace plots for convergence and mixing from separate chains. These indicated convergence in the chain, typically within 100 iterations, and reasonable mixing. We took posterior predictive observations (i.e., future cases and deaths) every 20th sample, for a total of 500 posterior predictive observations. Dashboard predictions are made by taking the mean and standard deviation of these observations. Because 1 is added to the original count, 1 is subtracted from the estimate. In the rare case that the estimated data point is negative, a 0 average case count or death count is recorded. Otherwise, non-integer values are used for the forecast.

A major issue involved in this construction is the dramatic increase in the computational time required for the Gaussian processes with more observations (i.e., algorithms on covariance matrices are intrinsically $O(n^3)$). Bayesian computation using the exact Gaussian process is not feasible for 3,142 distinct geographical locations, so, as an alternative, we use the method developed by Moran and Wheeler [7], which employs compressed covariance matrices that are nearly exact to the original covariance matrix (i.e., constructed such that the max norm of the two matrices is approximately 1e-14), but has the added benefit of the computational complexity of $O(nlog^2n)$. Utilizing these algorithms, the computational time required is decreased by a factor of 40, from two weeks to 6 to 8 hours. Unlike Moran and Wheeler [7], we do not use Ambikasaran and Darve's [8] HODLR compression technique due to its inability to scale to more than one dimension. Instead, we utilize Börm's [9] H2 matrix compression method that utilizes the H2lib matrix library.

While the model produces reliable estimates, rapid increases in infection rate can occasionally generate outlier extrapolations. For reporting within the Dashboard, counties with < 100 cases were examined for the largest count increase week over week nationwide and multiplied by 1.5 a single time to derive a cases-increase threshold. For counties with the prior week of actuals < 100, if the predicted cases from the aforementioned Bayesian models were capped at this value. For counties with cases >= 100 in the prior period, no such threshold was applied. Similar logic was applied to deaths < 5 counts.

Epidemiological Modeling

To provide context and ensure that the data streams provide conclusions and priority rankings that are broadly consistent with other epidemiological models, we performed cross-sectional analysis of cumulative (i) cases, (ii) deaths, (iii) deaths as a proportion of the population, and (iv) deaths as a proportion of reported cases. We emphasize that the PVI is not intended to be an epidemiological modeling tool per se as it does not explicitly distinguish between factors of vulnerability for cases vs. deaths. Our modeling described here is intended to anchor the components of the PVI and provide context within the larger field of COVID-19-related epidemiological modeling. Additionally, this modeling is not intended to provide forecasts, which are the primary focus of projection models, as discussed in the previous section.

As the initial analyses displayed evidence of count overdispersion, we performed generalized linear modeling in R version 3.5 with the gam() procedure using a negative binomial model with observed cumulative counts as the response [2]. For analyses (i), (ii), and (iv), we used log(population size) values as predictors with estimated coefficients. For analysis (iii), we used the "offset" command to model the death rate. Similarly, for analysis (iv), we used log(cumulative cases) as an offset to model the death rate among cases, which may produce biased results due to regional variation in reporting rates. It should be noted that a constant underreporting bias across counties would be absorbed into the intercept and would otherwise produce valid coefficient estimates for the predictors. Analysis (iv) may provide important clues about the death risk as including cases in the denominator removes a large portion of the stochastic variation. Moreover, for all analyses, we used the proportion of the state population that has been tested as a predictor to account for additional sources of bias.

To anchor our efforts to previous work, we included as additional fixed predictors those from Wu et al. [10], who focused primarily on the effects of a $PM_{2.5}$ air pollution index using an analysis analogous to our model (iii). Before analysis, we removed predictors with pairwise correlation with any other predictor greater than 0.85 and predictors that would be collinear with a series of predictors, such as the overall proportion of minority residents. For pairs exceeding the correlation threshold, we favored predictors with the lower missingness rate (if any) or those that are reported in other work. Dynamic predictors (i.e., those that changed substantially over the modeled period) were incorporated using simple county averages over the March-August period covered by the PVI. With over 3,100 counties (according to FIPS codes), most with >0 cases and deaths, the analysis can easily support the 27 to 28 final predictors used. To facilitate comparison with previous sources, we used predictors as they are given in their source. Accordingly, in some instances, predictors are represented as proportions and, in other instances, they are represented as percentages.

To provide additional context, we also performed negative binomial modeling (R version 3.5 bam() with “REML” fitting) [2] of daily cases, using the fixed county predictors as well as unaveraged dynamic predictors. Due to the nature of the model, we included the two-weeklagged cumulative number of cases as an additional predictor, as well as a smoothing spline time-dependent term to reflect a nationwide component of risk. Although it is formally a fixed-effects model, we refer to this model as dynamic and treat each day outcome as an independent realization, with the rate determined by the predictors. To account for potential time-dependent latent correlation structures, we determined standard errors for the coefficients by bootstrapping, treating each county across all dates as an observational unit for bootstrap resampling. We also built a dynamic version of the generalized linear model for cases and deaths as a proportion of the population to further investigate the effects of social distancing and other predictors that change daily. Final significance testing was based on bootstrapping to account for potential time-dependent correlation structures.

References

1. Esri. ArcGIS Javascript API, version 4.13. Esri; 2020 [cited on 31 August 2020]. Available from: https://www.esri.com/en-us/home

2. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2019 [cited on 31 August 2020]. Available from: https://www.R-project.org/

3. Rasmussen CE, Williams CKI. Gaussian processes for machine learning, vol. 2. Cambridge, MA: MIT Press; 2006.

4. De Boor C. A practical guide to splines. Applied Mathematical Sciences, vol. 27. New York: Springer; 1978.

5. Wheeler MW. Bayesian additive adaptive basis tensor product models for modeling high dimensional surfaces: an application to high-throughput toxicity testing. Biometrics. 2019 Mar;75(1):193-201.

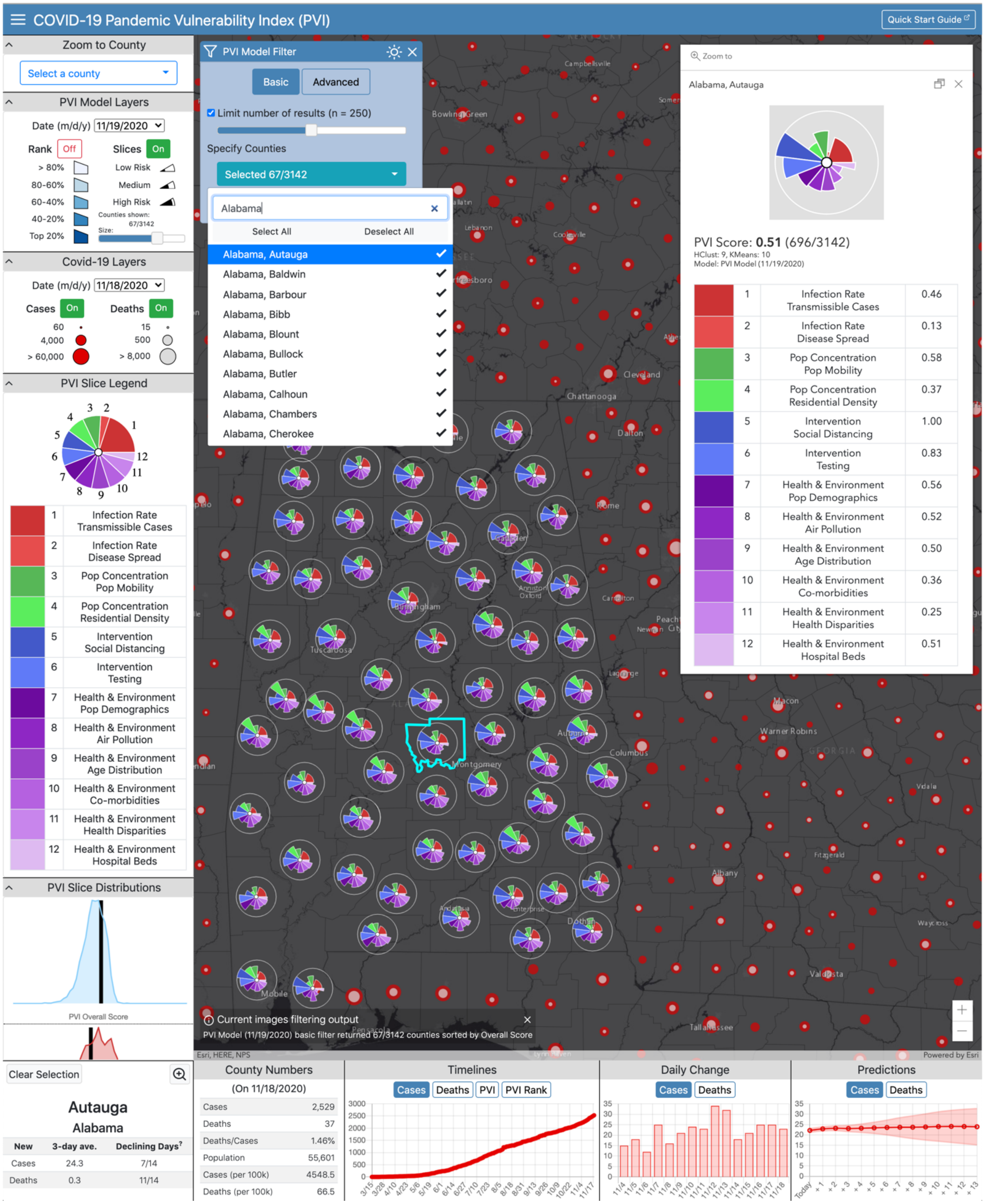
6. Lang S, Brezger A. Bayesian P-splines. J Comput Graph Stat 2004 Mar 1;13(1):183-212.

7. Moran KR, Wheeler MW. 2020. Fast increased fidelity approximate Gibbs samplers for Bayesian Gaussian process regression. arXiv:2006.06537. 2020 Jun 11 [cited on 31 August 2020]. Available from: https://arxiv.org/abs/2006.06537

8. Ambikasaran S, Darve E. An O(NlogN) Fast Direct Solver for Partial Hierarchically Semi-Separable Matrices. J Sci Comput. 2013 Dec 1;57(3):477-501.

9. Börm S. Efficient numerical methods for non-local operators: H2-matrix compression, algorithms and analysis. Zurich: European Mathematical Society; 2010.

10. Wu X, Nethery RC, Sabath BM, Braun D, Dominici F. 2020b. Exposure to air pollution and COVID-19 mortality in the United States. 2020 Apr 27; medRxiv preprint. Available from: https://doi.org/10.1101/2020.04.05.20054502.



PVI Model Layers

Date (m/d/y) 11/19/2020

Rank Off — Slices On

Rank	Slices
> 80%	Low Risk
80-60%	Medium
60-40%	High Risk
40-20%	Counties shown: 67/3142
Top 20%	Size:

Covid-19 Layers

Date (m/d/y) 11/18/2020

Cases On — Deaths On

Cases	Deaths
60	15
4,000	500
> 60,000	> 8,000

PVI Slice Legend

🟥	1	Infection Rate Transmissible Cases
🟥	2	Infection Rate Disease Spread
🟩	3	Pop Concentration Pop Mobility
🟩	4	Pop Concentration Residential Density
🟦	5	Intervention Social Distancing
🟦	6	Intervention Testing
🟪	7	Health & Environment Pop Demographics
🟪	8	Health & Environment Air Pollution
🟪	9	Health & Environment Age Distribution
🟪	10	Health & Environment Co-morbidities
🟪	11	Health & Environment Health Disparities
🟪	12	Health & Environment Hospital Beds

PVI Model Filter

☑ Limit number of results (n = 250)

Specify Counties

Selected 67/3142

Alabama ×

- Alabama, Autauga ✓
- Alabama, Baldwin ✓
- Alabama, Barbour ✓
- Alabama, Bibb ✓
- Alabama, Blount ✓
- Alabama, Bullock ✓
- Alabama, Butler ✓
- Alabama, Calhoun ✓
- Alabama, Chambers ✓
- Alabama, Cherokee ✓

ⓘ Current images filtering output ×

PVI Model (11/19/2020) basic filter returned 67/3142 counties sorted by Overall Score

Alabama, Autauga

PVI Score: **0.51** (696/3142)

HClust: 9, KMeans: 10

Model: PVI Model (11/19/2020)

🟥	1	Infection Rate Transmissible Cases	0.46
🟥	2	Infection Rate Disease Spread	0.13
🟩	3	Pop Concentration Pop Mobility	0.58
🟩	4	Pop Concentration Residential Density	0.37
🟦	5	Intervention Social Distancing	1.00
🟦	6	Intervention Testing	0.83
🟪	7	Health & Environment Pop Demographics	0.56
🟪	8	Health & Environment Air Pollution	0.52
🟪	9	Health & Environment Age Distribution	0.50
🟪	10	Health & Environment Co-morbidities	0.36
🟪	11	Health & Environment Health Disparities	0.25
🟪	12	Health & Environment Hospital Beds	0.51

Autauga
Alabama

New	3-day ave.	Declining Days?
Cases	24.3	7/14
Deaths	0.3	11/14

County Numbers

(On 11/18/2020)

Cases	2,529
Deaths	37
Deaths/Cases	1.46%
Population	55,601
Cases (per 100k)	4548.5
Deaths (per 100k)	66.5

Timelines — Cases | Deaths | PVI | PVI Rank

Daily Change — Cases | Deaths

Predictions — Cases | Deaths

Figure S1.

A screenshot of the current Dashboard. The text provides detailed descriptions of the panels and menu options. This information is also available in the Quick Start Guide at https://www.niehs.nih.gov/research/programs/coronavirus/covid19pvi/index.cfm