Algorithms to Reduce Biases in Disease Rate Estimates Caused by Data Suppression

Fariba Afrin Irany, *Member, IAENG,* Sundos Al Subhi, Rubenia Borge Flores, Chetan Tiwari

***Abstract*—Data custodians are required to safeguard personal health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA). Nevertheless, the practice of data suppression, which is frequently employed to handle small-count records, can induce biases that result in underestimating disease loads in rural regions. This work provides a formal description and analysis of three methods that are used to estimate suppressed data values. The algorithms are applied to synthetic heart disease mortality data at the county level. These algorithms vary depending on their utilization of demographic adjustments, area illness risk, and level of implementation complexity. Although population-based estimation of suppressed data is the most straightforward to do, it may be more precise for diseases that have a natural spatial element. Estimates can be enhanced by integrating risk, whether it is specified at a broader geographical level (such as the state level) or at a more localized level (such as a group of nearby counties). Gaining insight into these trade-offs can aid in predicting concealed data using illness attributes. This can also aid in reducing biases caused by data suppression.**

***Index Terms*—Algorithms, Data estimation, Data generation, Data suppression.**

I. Introduction

As stated in Act [1], Nelson [2], the Health Insurance Portability and Accountability Act of 1996 (HIPAA) mandates that custodians of public health data, like CDC WONDER, must establish measures to safeguard and restrict the disclosure of an individual's identifiable health information. This is achieved by employing diverse methodologies, such as data consolidation as suggested by Croner [3], Tiwari et al. [4], masking Allshouse et al. [5], Kwan et al. [6], Leitner and Curtis [7], suppression Tiwari et al. [8], or a combination of these three. Various federal and state public health authorities, such as the CDC, employ a blend of aggregation and suppression techniques to safeguard the confidentiality of individuals. The process of consolidating data into larger geographic units, such as counties, is typically the initial measure taken to protect the confidentiality of health data at the individual level. Nevertheless, in regions with limited populations, just combining the data into a less detailed format may not offer sufficient safeguards against the possible exposure of personal health information at an individual level. Data suppression is frequently a subsequent measure taken to guarantee privacy safeguarding. Data custodians commonly employ suppression rules that are based on two primary criteria: (1) the suppression of all cells in a data table that have 10 or fewer observations, and (2) the classification of data as "unreliable" when any data cell contains less than 20 observations, as stated by the Centers for Disease Control and Prevention (2022). Although aggregation and suppression are crucial for safeguarding the potential identification of health information for people, they can introduce biases that affect the accurate representation of disease loads in a certain geographic area. Prior studies have revealed that rural regions with lower populations are more prone to experiencing data suppression in contrast to metropolitan areas with greater populations. [8] stated that if end users utilize suppressed data without making proper demographic modifications, it will continuously result in underestimating disease burdens in rural areas throughout the United States. Utilizing secondary data such as area disease risk and demographic structures might help estimate the suppressed data values, thus reducing the possible influence of spatial biases. Nevertheless, the geographical delineation of regional risk can impact the computation of suppressed data values estimations. We present three different algorithms that employ distinct regional risk definitions to estimate suppressed data values and analyze the tradeoffs between implementation complexity and accuracy level for each of the algorithm.

Manuscript received October 2, 2023; revised January 31, 2023. This work was supported in part by the Texas Department of State Health Services.

Fariba Afrin Irany is a Ph.D. student at the University of North Texas, Denton, Texas 76207, USA (Corresponding Author, phone: 940-465-9618; email: faribaafrinirany@my.unt.edu).

Sundos Al Subhi is a PhD candidate of the Georgia State University, Atlanta, Georgia 30302, USA. (e-mail: salsubhi1@student.gsu.edu).

Rubenia Borge Flores is a Ph.D. student at the University of North Texas, Denton, TX 76207, USA. (e-mail: rubeniaborgeflores@my.unt.edu).

Chetan Tiwari is an Associate Professor of Georgia State University, Atlanta, Georgia 30302, USA. (e-mail: ctiwari@gsu.edu).

II. Literature Review

In recent decades, researchers have uncovered biases in several illness areas and suggested techniques to alleviate their impact. Within this section, I will critically examine several scholarly articles pertaining to biases and the various approaches employed to mitigate their effects. Now, I provide the results of my review for a couple of related papers below. In their study, Lipsitch et al. [9] initially discuss the inherent biases that hinder the accurate interpretation of CFR, a metric used to gauge the severity of a disease outbreak. They subsequently suggest methods to mitigate these biases and conclude by examining the circumstances in which risk factors for mortality may be influenced by these biases. [10] examine the biases related to estimate procedures for DRS, a metric used to manage confounding in non-experimental investigations. They argue that typical estimation tactics are unable to effectively reduce these biases and provide methods to address this issue. Chiang et al. [11] report the findings of their investigation into the impact of rater bias and the methodology used to assess the severity of a disease. Additionally, they address the significance of diminishing biases, noting that estimated biases amplify the potency of the hypothesis. Conner et al. [12] provide a method to detect

the presence of bias in prevalence estimates obtained from harvest samples, which offer insights into disease trends. Accorsi et al. offered a way to address biases that occurred in different types of observational research on COVID-19. Their findings were published in a paper titled "Detecting and Correcting for Bias in Observational Studies of COVID-19" [13]. Huang et al. [14] introduce a novel approach called Biased Sentinel Hospital-based Area Disease Estimation (B-SHADE) to mitigate geographical biases in search engine data. The study conducted by Bower et al. [15] aimed to examine the biases commonly found in electronic health record (EHR)-based studies on cardiovascular disease (CVD) risk. The researchers offered various approaches to mitigate the influence of these biases. The study conducted by Alonzo et al. [16] examines the impact of a test without a definitive standard on bias, and identifies a correlation between the screening test and an inappropriate reference test. Wood et al. [17] propose a technique to reduce bias in sample and multiscale entropy in fMRI data. They argue that this technique can be applied to other methodologies used to investigate abnormal brain activity in various brain disorders. The authors Angelopouloset al. [18] contend that the estimation of time- and severity-dependent reporting of cases is subject to biases. To address these biases, they first examine biases associated with CFR estimation. Subsequently, they propose a corrected estimator that involves testing the contacts of infected individuals regardless of symptoms. In order to address the biases related to the examination of disease progression, Mitchell et al.'s research [19] examines statistical techniques and their application, while Pack et al.'s work [20] highlights the biases associated with randomized control trials (RCTs) where the treatment for obstructive sleep apnea (OSA) did not reduce the occurrence of cardiovascular failure. The researchers focused solely on estimating prejudice rather than developing techniques to mitigate biases. For instance, the study conducted by Hall et al. [21] examines the presence of racial bias among healthcare workers and its resulting consequences. Shan et al. [22], Croskerry [23], Sajeev et al. [24] assess the cognitive strategy biases. Dawson et al. [25] discuss the concept of hindsight bias, which refers to the challenges associated with accurately estimating the probability of clinicopathologic conference impediment. Jensen et al. [26] discuss the presence of bias in estimations of illness prevalence. Rudolph et al. [27] highlight the significance of measuring bias in respondent-driven samples. Baines et al. [28] discuss how mixed mode administration can help eliminate bias. [29] conduct a systematic analysis to compare bias and accuracy in the context of renal disease, specifically focusing on bias reduction techniques for electronic health record data. Czeisler et al. [30] discuss survival bias in relation to mental health surveys conducted during the COVID-19 pandemic. Lachish and Kris [31] identify the origins of bias in a study on disease ecology. Our research is the first to offer techniques for mitigating the biases in illness rate estimation caused by data suppression in CDC Wonder, which is a public health data repository. This publication builds upon the research conducted by Tiwari et al. [8] by introducing methodologies to mitigate the impact of data suppression. The proposed method is essential to demonstrate that the existing strategy employed by the CDC Wonder to hide data fails to safeguard persons' privacy.

III. MATERIALS AND METHODS

Initially, we discuss the dataset that was utilized for our research endeavors. We created synthetic data at the county level to evaluate the effectiveness of three proposed algorithms in calculating suppressed values for heart disease mortality. The number of instances (i.e., deaths) for each county (i) and age group (g) was simulated using the following procedure:

i. Define the regional risk by utilizing a spatial weights file that delineates the influence of n neighboring counties on the rate of heart disease mortality in age group g within county i. The regional influence on each county i is believed to be distinct and is determined by randomly selecting neighboring counties within a range of 0 to 150 miles from the centroid of i.

ii. The crude rate for age group g in county i is calculated by averaging the rates for age group g in all nearby counties, as obtained in step (i) above. Excluded from the analysis are data points that had missing rates, either due to suppression or missing data. If rates for all surrounding counties (n) are unavailable, the crude rate for age group g in i is designated as "Not Available."

iii. The case counts for each county i and age group g are determined by multiplying the age-group-specific rate obtained from (ii) by the corresponding population. Cells lacking rate information are also designated as "Not Available." iv. One hundred synthetic datasets of heart disease mortality were generated by altering the level of geographical impact, as outlined in step (I).

For each synthetic dataset, a corresponding suppressed dataset was generated by eliminating all cells that had fewer than ten observations. Subsequently, the three methods were implemented on every dataset. Prior to delving into the techniques and their significance, we outline the crucial computations required for implementing such techniques on a synthetic dataset.

A. Necessity of Developing Methods to Estimate Suppressed Value

Information regarding mortality rates and other health indicators can frequently be obtained at various geographical levels, such as county and state levels. Typically, the number of cases or deaths recorded at the state level is unlikely to be less than 10 and is not withheld or concealed. Data obtained from smaller populations, such as at the county or census tract level, are more prone to being withheld due to the higher likelihood of suppression. The discrepancy between the total number of reported cases for a specific state and the combined count of available (unrestricted) cases at the county level within that state will indicate the exact number of instances that have been withheld or suppressed. The same reasoning is applicable to age-specific data that is being requested at the county level. In this scenario, the likelihood of data suppression is lower when data is collected for a county without age stratification, as opposed to when individual age-stratified data counts are sought for the same county. The discrepancy between the combined unsuppressed data (e.g., total reported count for a whole county) and the total of available (i.e., unsuppressed) counts for the categorized data (e.g., total counts across all unsuppressed

age groups in that county) will determine the number of cases that are withheld (S_i). The three approaches described below utilize data on population distribution, statewide or regional risk, or a mix of both to redistribute the count of missing data (S_i) among suppressed data cells. Subsequently, we will examine a specific scenario in order to gain a deeper comprehension of the necessity for employing these strategies.

If data for a certain age group in a county are withheld, the projected count for that county can be determined using the following equation:

$$\forall \mathbf{i} \in \mathbf{I} \quad \mathbf{g_s} \in \mathbf{G} \quad \mathbf{c'_{i,g_s}} = \mathbf{D_i} - \mathbf{\Sigma_{g=1}^{m-1}} \mathbf{d_i, g_{us}} \tag{1}$$

In Equation 1, the symbol denotes a county from the set of all counties I, where there is just one data cell that has been suppressed. The variable g_s denotes the age group that has withheld information among the m age groups in the set G. The variable c'_{i,g_s} denotes the estimated count for a specific county and age group when the data is intentionally withheld. The variable is denoted as g_s. D_i is the overall count (not divided into age groups) provided for the county (i), whereas $\Sigma_{g=1}^{m-1} d_i, g_{us}$ indicates the total of counts available for age groups with unsuppressed data. In such instances, the precise value for the suppressed cell is calculated by straightforwardly subtracting this value from D_i, so ensuring no margin of error. In this instance, the derived estimate (c'_{i,g_s}) is equivalent to S_i (as explained in the calculation procedure given above under "Necessary Calculations for Methods"). The counties depicted on the maps in Figure 1, 2, 3 below are represented by the class that is shaded white, without any errors. In situations when data is withheld for multiple age groups within a county, it is necessary to develop techniques for computing the suppressed values.

Presently, we put up three distinct approaches to calculate concealed data values. The computation process of the method parameter S_i is outlined in subsection "Necessary Calculations for Methods." The methods are presented below:

B. Method 1: Population-derived Estimates for Counties with More Than One Suppressed Cell

In Method 1, local population structures obtained from the US Census Bureau American Community Survey (ACS) are used to estimate missing data in suppressed cells as follows:

$$\forall \mathbf{i} \in \mathbf{I} \quad \forall \mathbf{g_s} \in \mathbf{G} \quad \mathbf{c'_{i,g_s}} = \frac{\mathbf{p_{i,g_s}}}{\mathbf{\Sigma_{g_s=1}^{m_s}} \mathbf{p_{i,g_s}}} \times S_i \tag{2}$$

While the basic steps for method 1 are presented in algorithm 1, we explain the algorithm using the equation 2 for ease of understanding. In Equation 2, the variable i represents a county from the set of all n counties (I). The variable g_s represents an age group with suppressed data. The variable m_s represents the total number of age groups with suppressed data. The variable c'_{i,g_s} represents the estimated count for county i and age group. g_s denotes the suppressed age group, and p_{i,g_s} represents the population in county i for that age group. The expression $\Sigma_{g_s=1}^{m_s} p_i, g_s$ denotes the cumulative population of all age groups that have been suppressed in county i. On the other hand, S_i indicates the overall count of suppressed cases in county i. When it is not possible to calculate S_i, procedure 1 is unable to provide an estimate for the county with suppressed data (c'_{i,g_s}). In this approach, we calculate the value of the suppressed cell by dispersing the value of S_i among the suppressed cells, using the population proportion as a weighting factor. One drawback of this strategy is that it is only applicable to diseases in which the likelihood of infection or mortality is influenced by the population structure. In other words, the weights used to redistribute S_i based on population proportions must be relevant to the specific disease being studied. This strategy tends to exaggerate mortality in younger population groups and underestimate mortality in older populations, including heart disease mortality based on our synthetic data.

Algorithm 1 Population-derived Estimates

Require: set of counties I, set of age group G, county-specific death count for an age group $d_{i,g_{us}}$, county-specific population for suppressed age group p_{i,g_s}, county-level reported death count D_i, suppressed age group set for each county G_{is}

Ensure: estimated death count for a county for an age group c'_{i,g_s}

for $i \in I$ **do**
  n $\leftarrow$ length(G)
  t $\leftarrow$ length(G_{is})
  $D_{us} \leftarrow \Sigma_{g=1}^{n} d_i, g_{us}$
  $S_i \leftarrow D_i$ - D_{us}
  $p_o \leftarrow p_{i,g_s}$
  $p_{all} \leftarrow \Sigma_{g=1}^{t} p_{i,g_s}$
  p $\leftarrow p_o \div p_{all}$
  $c'_{i,g_s} \leftarrow p \times S_i$
end for
  return c'_{i,g_s}

C. Method 2: Population and Statewide Risk Derived Estimates for Counties with More Than One Suppressed Cell

Method 2 utilizes local population structures derived from the American Community Survey (ACS) and regional or statewide risk to estimate missing data in suppressed cells in the following manner:

$$\forall \mathbf{i} \in \mathbf{I} \quad \forall \mathbf{g_s} \in \mathbf{G_s} \quad \mathbf{c'_{i,g_s}} = \frac{\mathbf{d_{j,g_s}}}{\mathbf{p_{j,g_s}}} \times p_{i,g_s} \tag{3}$$

We explain the algorithm 2 using the equation 3. The notations i, I, g_s, G_s, c'_{i,g_s}, and p_{i,g_s} in equation 3 are explained in Method 1 as mentioned before. In addition, the variable 'j' denotes the specific state in which county 'i' is situated among all states represented by 'J'. The expression $\frac{d_{j,g_s}}{p_{j,g_s}}$ reflects the calculated statewide risk by dividing the total count of individuals in age group g_s in the state (d_{j,g_s}) by the corresponding population (p_{j,g_s}). If the state-level counts for any age-group g_s data are missing, procedure 2 is unable to calculate the statewide risk estimate. Consequently, it is not possible to compute an estimate for the suppressed cell. This approach is suitable when the regional likelihood of a disease is comparable to the risk throughout the entire state. Please be aware that the estimations generated by this method, namely c'_{i,g_s}, do not take into account the documented count of suppressed cases (S_i). The discrepancy

Algorithm 2 Population and Statewide Risk Derived Estimates

Require: set of counties I, set of age group G, set of states J, state-specific death count for suppressed age group d_{j,g_s}, state-specific population for suppressed age group p_{j,g_s}, county-specific population for suppressed age group p_{i,g_s}.

Ensure: estimated death count for a county for an age group p_{j,g_s}

for $j \in J$ **do**
  n $\leftarrow$ length(G)
  c $\leftarrow$ length(I)
  $d_{j,g_s} \leftarrow \Sigma_{i=1}^{c} d_j, g_s$
  $p_{j,g_s} \leftarrow \Sigma_{i=1}^{c} p_j, g_s$
  $s_r \leftarrow d_{j,g_s} \div p_{j,g_s}$
  $c'_{i,g_s} \leftarrow s_r \times p_{i,g_s}$
end for
  return c'_{i,g_s}

between the total estimated counts for the county i ($\Sigma c'_{i,g_s}$) and S_i is resolved by randomly adding or subtracting counts until S_i is met.

D. Method 3: Population and Local Risk-Derived Estimates

Method 3 incorporates the local population patterns derived from the American Community Survey (ACS) and the user-defined local risk to estimate missing data in suppressed cells.

Algorithm 3 Population and Local Risk-Derived Estimates

Require: set of counties I, county-specific death count for an age group for neighboring county $d_{i,g_s,k}$, county-specific population for suppressed age group neighboring county $p_{i,g_s,k}$, county-specific population for suppressed age group p_{i,g_s}, neighboring county set N for each county.

Ensure: estimated death count for a county for an age group c'_{i,g_s}

for $i \in I$ **do**
  x $\leftarrow$ length(N)
  $D \leftarrow \Sigma_{k=1}^{x} d_i, g_s, k$
  $P \leftarrow \Sigma_{k=1}^{x} p_i, g_s, k$
  $L \leftarrow D \div P$
  $c'_{i,g_s} \leftarrow$ L $\times\ p_{i,g_s}$
end for
  return c'_{i,g_s}

$$\forall_{\mathbf{i}} \in \mathbf{I} \quad \forall_{\mathbf{g_s}} \in \mathbf{G_s} \quad \mathbf{c'_{i,g_s}} = \frac{\mathbf{\Sigma_{k=1}^{x} d_{i,g_s,k}}}{\mathbf{\Sigma_{k=1}^{x} p_{i,g_s,k}}} \times p_{i,g_s} \qquad (4)$$

While the basic steps of method 3 are presented in algorithm 3, we use equation 4 to explain algorithm 3 for clarity. The notations i, I, g_s, G_s, c'_{i,g_s}, and p_{i,g_s} are defined in Method 1 as stated in Equation 4. Furthermore, the variable k symbolizes each individual county within a collection of x counties denoted as K. The set K is specified for each county (i) in the research region, specifically as the set of counties next to county i. This cluster of contiguous counties is utilized to compute the regional susceptibility to disease. The expression $\frac{\Sigma_{k=1}^{x} d_{i,g_s,k}}{\Sigma_{k=1}^{x} p_{i,g_s,k}}$ indicates the local risk. The expression $\Sigma_{k=1}^{x} d_{i,g_s,k}$ denotes the total number of deaths in all counties inside the set K that are linked to county i. Similarly, $\Sigma_{k=1}^{x} p_{i,g_s,k}$ reflects the total population of all counties in K that are affiliated with county i. Similar to technique 2, any discrepancy between the estimated counts for county i ($\Sigma c'_{i,g_s}$) and S_i is rectified by randomly increasing or decreasing counts until S_i is reached. If there is a lack of data for all counties in K, the user has the option to exclude the county with missing data or reconsider the definition of the local region. In cases where there are few inhabitants and numerous data gaps, researchers may need to resort to the aforementioned approach 2. Similar to technique 2, any disparity between the projected number of fatalities and the officially reported number for the entire state (S_i) is rectified by introducing random additions or subtractions of deaths until S_i is fulfilled. Furthermore, it should be noted that the local risk for a different county will be calculated based on its own distinct group of neighboring counties.

TABLE I
SUMMARY STATISTICS OF SPATIAL INFLUENCE

Neighboring County Statistics of Counties	**Value**
mean	35.63
std	37.81
min	1.0
25%	6.0
50%	21.0
75%	55.0
max	192.0

TABLE II
SUMMARY STATISTICS FOR METHOD 1

Method 1	**Mean Error**	**Standard Deviation**
mean	1.55	1.78
std	.033	.033
min	1.45	1.71
25%	1.52	1.76
50%	1.54	1.78
75%	1.57	1.81
max	1.63	1.85

TABLE III
SUMMARY STATISTICS FOR METHOD 2

Method 2	**Mean Error**	**Standard Deviation**
mean	1.33	1.62
std	.030	.046
min	1.26	1.53
25%	1.30	1.59
50%	1.33	1.62
75%	1.35	1.65
max	1.39	1.77

Availability of datasets and code: materials.

IV. RESULTS AND DISCUSSION

Figure 1,2,3,4 summarize the error in estimated counts obtained from methods 1, 2, and 3, respectively. For each

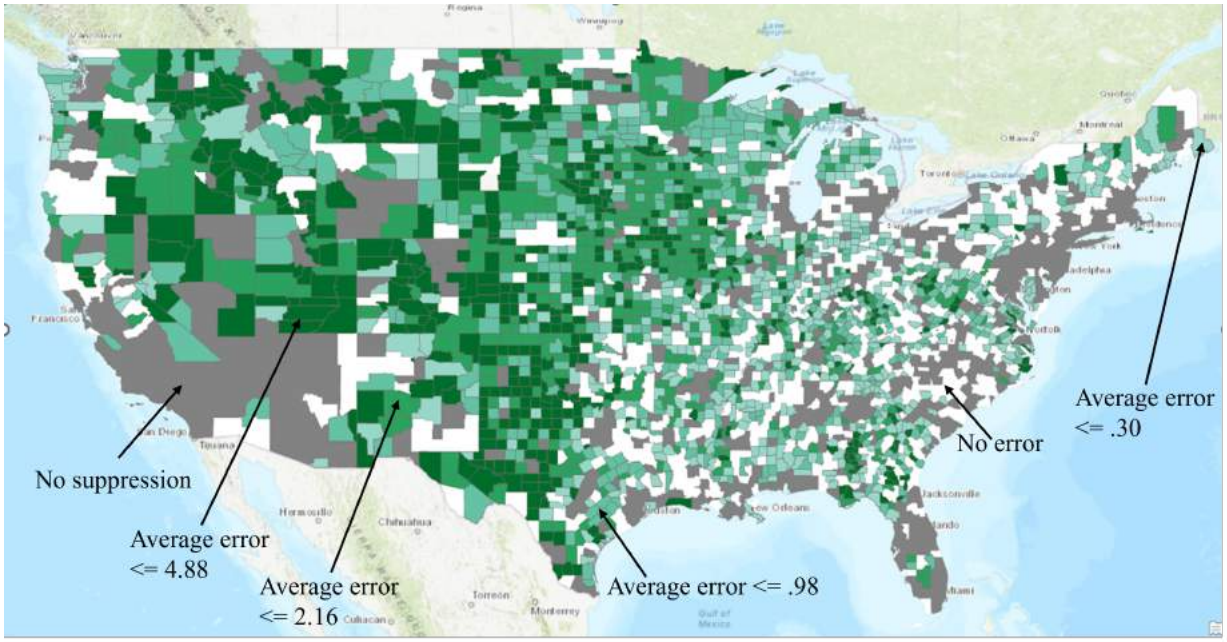

Fig. 1. Methods 1 for Estimating Suppressed Data Values

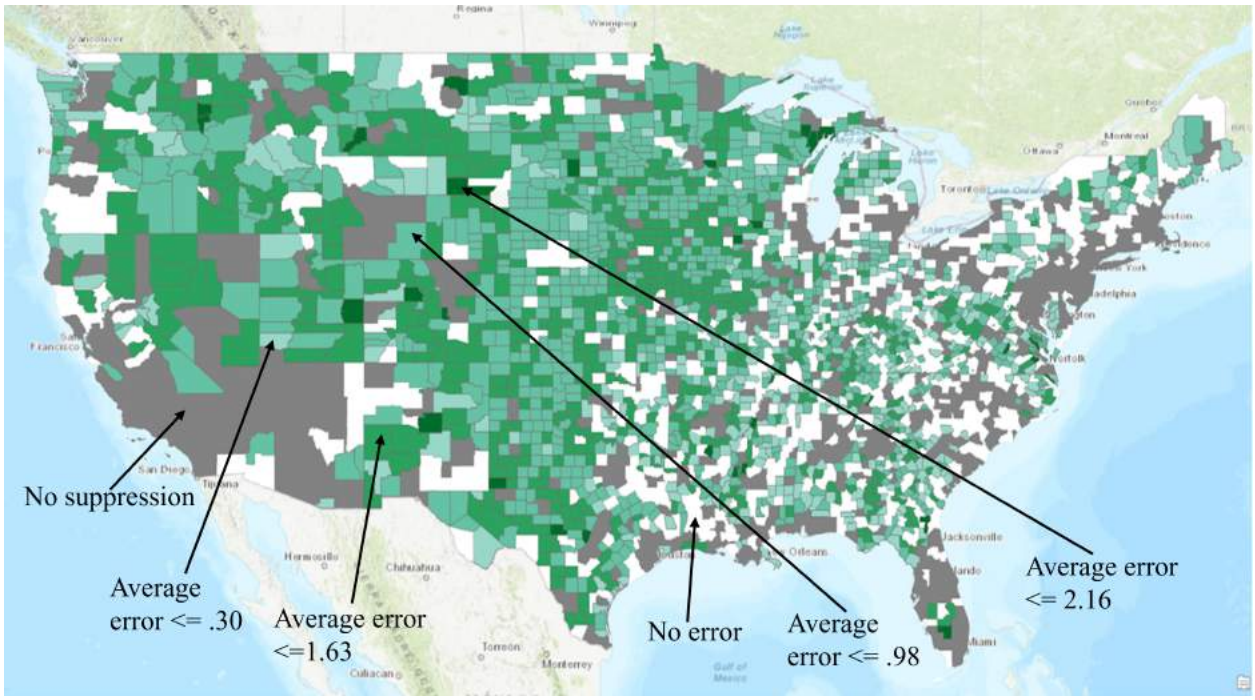

Fig. 2. Methods 2 for Estimating Suppressed Data Values

TABLE IV
SUMMARY STATISTICS FOR METHOD 3

Method 3	Mean Error	Standard Deviation
mean	1.21	1.44
std	.026	.030
min	1.13	1.37
25%	1.19	1.42
50%	1.21	1.44
75%	1.23	1.46
max	1.27	1.53

county, the error is defined as the average difference between the true value (from the synthetic dataset) and the estimated value (using methods described above) for all suppressed cells. The colors on the maps in Figure 1,2,3 show the average error for each county across all 100 simulations of the synthetic dataset. The error term was grouped into five classes (Class 1 through Class 5) using the quintiles classification method. Lighter colors on the map indicate a lower average error. A visual examination of the map associated with method 1 (Figure 1) shows a greater intensity of darker colors, indicating greater average error than methods 2 and 3 (Figure 2 and Figure 3, respectively). Although method 1 only requires information on population proportions and is

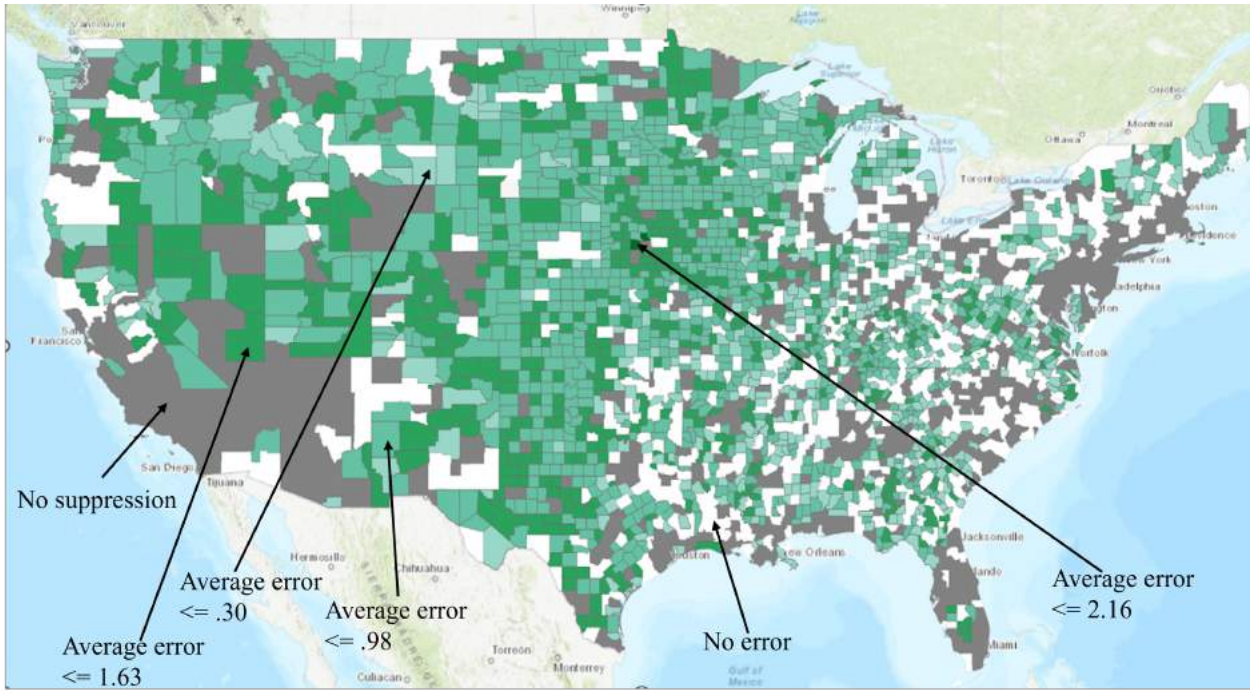

Fig. 3. Methods 3 for Estimating Suppressed Data Values

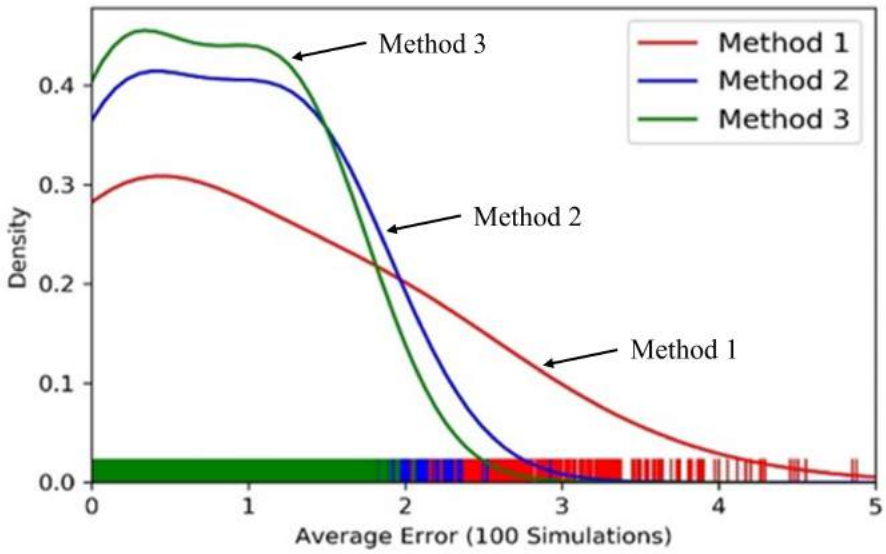

Fig. 4. Average Error Across Three Different Methods

TABLE V
PERFORMANCE OF DIFFERENT METHODS

Map class	**Method 1**	**Method 2**	**Method 3**	**% Difference between Method 1 and Method 2**	**% Difference between Method 2 and Method 3**
No Suppression	608	608	608	0	0
No error	619	619	619	0	0
1($\leq$ .30)	385	276	301	-28.31	9.05
2($\leq$ .98)	381	519	580	36.22	11.75
3($\leq$ 1.63)	383	852	933	122.45	9.50
4($\leq$ 2.16)	383	248	99	-35.24	-60.08
5($\leq$ 4.88)	381	18	0	-95.27	-100

TABLE VI
PERFORMANCE OF DIFFERENT METHODS USING URBAN-RURAL CLASSIFICATION SCHEME

County Code	Total Counties	Suppressed Cells	No Suppression	% of Counties in Highest Error Classes (i.e., 4 and 5) for Method 1	% of Counties in Highest Error Classes (i.e., 4 and 5) for Method 2	% of Counties in Highest Error Classes (i.e., 4 and 5) for Method 3
Large Central Metro	68	0(0%)	68	0	0	0
Large Fringe Metro	368	192(52.17%)	176	25	20	4
Medium Metro	372	191(51.34%)	181	34	20	8
Small Metro	358	241(67.31%)	117	34	15	6
Micropolitan	641	596(92.97%)	45	24	10	3
Noncore	1333	1312(98.42%)	21	47	13	5

relatively easy to implement, it is only appropriate for outcomes where the risk of an adverse outcome is proportional to population size (e.g., fatal traffic accidents). In the case study presented in this paper – heart disease mortality – this method tends to overestimate missing values in the younger age groups, given their relatively large population proportions and lower risk for heart disease mortality. Methods 2 and 3 are an improvement over method 1 as they use a combination of population size and risk when estimating missing values. In the case of method 2, the risk is defined as the statewide risk of heart disease mortality. This value is multiplied by the local population to produce an estimate of the missing value. Although method 2 requires more information than method 1, it explicitly includes a measure of risk while considering local population structures. However, this method assumes little to no variation in risk within the local areas for which the estimate is being computed. This method is appropriate where risk is spread over large geographic areas. Method 3 is an improvement over method 2, including a measure of local risk. In this case, we defined local risk for every county as the average rate of heart disease mortality of its surrounding counties. This method is appropriate for cases where one expects to see local variations in risk. The improvement in the estimate obtained using method 3 compared to methods 1 and 2 is unsurprising, given known local variations in heart disease mortality risk. While our synthetic dataset was constructed using some regional definition of risk, we chose to modify the spatial influence randomly across all 100 simulations of the dataset. Table I displays the data pertaining to the created file used to alter the spatial influence. The simulation provides data on the neighboring county number for all counties. The properties mean, std, min, 25%, 50%, 75%, max represent statistical measures of the number of neighboring counties. The mean represents the average number of neighboring counties, std represents the standard deviation which measures the variability of the neighbor number, min and max represent the smallest and largest number of neighboring counties. The values 25%, 50%, 75% represent the quartiles, indicating that 25%, 50%, 75% of the neighbor numbers fall within these values.

Tables II, III, and IV display the error data from 100 simulations for method 1, method 2, and method 3, respectively. The properties listed in these tables have the following significance:

mean:The mean error throughout each table corresponds to the average value of the average error, while the standard deviation shows the average variability across 100 simulations; **std**: The mean error across each table corresponds to the standard deviation of the average error. The standard deviation number shows the standard deviation over 100 simulations; **min,max**: the mean error throughout each table represents the lowest and highest values of the average error, respectively. The minimum and maximum values of the standard deviation are the lowest and highest standard deviations across 100 simulations; **25%,50%,75%**: The values of 25%, 50%, and 75% for the mean error across each table indicate that 25%, 50%, and 75% of the average errors are within these respective values. The values of 25%, 50%, and 75% of the standard deviation indicate that 25%, 50% and 75% of the standard deviation values fall within this range throughout 100 simulations.

The graphs in Figure 4 show the error distribution obtained by the three methods. The extended tail of method 1 indicates a greater number of counties with significantly larger error values than methods 2 and 3. Table V presents the performance of different methods concerning each other. Method 2 reduces the number of counties in the highest error class (class 5) by slightly over 95%, i.e., method 2 only has 18 counties in class 5 compared to 381 in method 1. Similarly, method 2 also contains fewer counties classified number of counties in classes 2 and 3 improved by $\sim$36% and $\sim$122%, respectively. Despite this assumption, we find that method 3 provides a substantial improvement over method 2, with no counties in the highest error classification (class 5) and $\sim$60% reduction in the number of counties included in the 4th highest error classification (class 4). Although the distributions for methods 2 and 3 in Figure 4 follow similar trends, the number of observations with lower error values is greater for method 3 (green line) than method 2 (blue line). We used the 2013 National Center for Health Statistics (NCHS) Urban-Rural classification scheme for counties to better understand where suppression was most likely to occur in our synthetic dataset and to evaluate the efficacy of the three algorithms in estimating counts of counties with suppressed data. Of 3140 counties analyzed in this study, 608 were not subject to any suppression ($\sim$19%), 619 contained one suppressed cell for which an exact value was determined ($\sim$20%), and 1913 counties contained more than two suppressed cells which were estimated using the three methods described above ($\sim$ 61%). Under the NCHS classification scheme, counties categorized as large central metro were not subject to data suppression (Table VI). Among

the other groups, a little over half the counties classified as large fringe metro and medium metro (Table VI) were subject to suppression (52.17% and 51.34%, respectively). The highest levels of suppression were observed in counties classified as micropolitan and non-core (92.97 and 98.42% respectively). Among the three methods used to estimate suppressed cells, method 1 performed poorly across all urban-rural classifications. Most notably, estimates obtained by method 1 showed the highest error in medium metro, non-core, and small metro, where 35% to 50% of estimated values (Table VI) fell within the highest two error classes (Class 4 and 5 in Figure 1,2,3). In comparison, estimates obtained using method 2 show substantial improvement in estimates across all urban-rural classifications, with very few counties falling within the two highest error classes (Class 4 and 5 in Figure 1, 2, 3). Finally, method 3 results in the lowest error across all urban-rural classes with no counties falling in the highest error class (Class 5 in Figure 1, 2, 3). In the case of counties exhibiting the most suppression (i.e., micropolitan and non-core), estimates obtained using methods 2 and 3 provide substantial improvement over method 1 (Table VI). To summarize, method 1 shows relatively consistent error values across the 5 error classes, while methods 2 and 3 perform similarly across error classes 1, 2 and 3. Method 3 outperforms method 2 in the number of counties contained within the highest two error classes.

V. CONCLUSION

The privacy and confidentiality of individual-level health data are safeguarded by mechanisms that employ suppression and aggregation principles. These principles involve the elimination of data and/or reduction of spatial resolution in regions with low population density. Prior studies indicate that these regulations are more prone to be applicable in rural regions and can result in a partial perspective of the geographical distribution of illnesses. In the absence of suitable modifications, the use of data suppression can result in disease rates that underestimate the actual disease burden. This research presents three methodologies for approximating the number of case counts that were excluded owing to suppression. Our analysis demonstrates that each method involves trade-offs in terms of use case, implementation complexity, and the accuracy of the estimations given. We contend that approach 1 is suitable for diseases that are predominantly influenced by demographic patterns and lack an inherent geographical component. Both techniques 2 and 3 utilize data on local population structures, but they employ distinct conceptualizations of disease risk. Method 2 operates under the assumption that risk is distributed across expansive geographic areas and relies on utilizing available estimates for entities such as states. Method 3 adopts a more refined geographical interpretation of risk. For the sake of this paper, we use the assumption that the risk of disease in a particular area is represented by calculating the average risk of the surrounding counties. Although the idea of risk is rather straightforward, we demonstrate that method three greatly enhances our capacity to estimate the values of suppressed data cells. After employing three methodologies on the synthetic heart disease dataset, it is evident that technique 3 yields the most favorable outcome. Specifically, this methodology allows for the estimation of 60% of the suppressed data with an error margin of no more than 1%. Furthermore, we would like to draw attention to the practices of data suppression and aggregation on a wider scale. Our analysis reveals that approximately 25% of all suppressed cells may have their exact values determined using secondary information available on CDC WONDER, particularly for outcomes such as heart disease mortality. Approximately 46% of the remaining suppressed data cells were approximated with a margin of error of $\pm 1\%$. Considering the capacity to analyze concealed values through straightforward GIS methods and readily accessible data, we raise doubts about the effectiveness of suppression practices and if they cause more harm than benefit. Our future plans involve utilizing local spatial autocorrelation and other GIS approaches to enhance our understanding of localized risk and refine our estimations of suppressed data values.

ACKNOWLEDGMENT

I want to express my great appreciation to Dr. Chetan Tiwari for his valuable and constructive suggestions during the planning and development of this research work. His willingness to give his time so generously has been very much appreciated. I would also like to thank all the co-authors for their continuous effort to finish the research.

REFERENCES

[1] A. Act, "Health insurance portability and accountability act of 1996," *Public law*, vol. 104, p. 191, 1996.

[2] G. S. Nelson, "Practical implications of sharing data: a primer on data privacy, anonymization, and de-identification," in *SAS global forum proceedings*, 2015, pp. 1–23.

[3] C. M. Croner, "Public health, gis, and the internet," *Annual Review of Public Health*, vol. 24, no. 1, pp. 57–82, 2003.

[4] C. Tiwari, D. Sterling, and L. Allsopp, "Linking disease outcomes to environmental risks: The effects of changing spatial scale," in *Geospatial Technology for Human Well-Being and Health*. Springer, 2022, pp. 265–280.

[5] W. B. Allshouse, M. K. Fitch, K. H. Hampton, D. C. Gesink, I. A. Doherty, P. A. Leone, M. L. Serre, and W. C. Miller, "Geomasking sensitive health data and privacy protection: an evaluation using an e911 database," *Geocarto international*, vol. 25, no. 6, pp. 443–452, 2010.

[6] M.-P. Kwan, I. Casas, and B. Schmitz, "Protection of geoprivacy and accuracy of spatial information: How effective are geographical masks?" *Cartographica: The International Journal for Geographic Information and Geovisualization*, vol. 39, no. 2, pp. 15–28, 2004.

[7] M. Leitner and A. Curtis, "A first step towards a framework for presenting the location of confidential point data on maps—results of an empirical perceptual study," *International Journal of Geographical Information Science*, vol. 20, no. 7, pp. 813–822, 2006.

[8] C. Tiwari, K. Beyer, and G. Rushton, "The impact of data suppression on local mortality rates: the case of cdc wonder," *American journal of public health*, vol. 104, no. 8, pp. 1386–1388, 2014.

[9] M. Lipsitch, C. A. Donnelly, C. Fraser, I. M. Blake, A. Cori, I. Dorigatti, N. M. Ferguson, T. Garske, H. L. Mills, S. Riley *et al.*, "Potential biases in estimating absolute and relative case-fatality risks during outbreaks," *PLoS neglected tropical diseases*, vol. 9, no. 7, p. e0003846, 2015.

[10] R. Wyss, M. Lunt, M. A. Brookhart, R. J. Glynn, and T. Stürmer, "Reducing bias amplification in the presence of unmeasured confounding through out-of-sample estimation strategies for the disease risk score," *Journal of causal inference*, vol. 2, no. 2, pp. 131–146, 2014.

[11] K.-S. Chiang, C. Bock, M. El Jarroudi, P. Delfosse, I. Lee, and H. Liu, "Effects of rater bias and assessment method on disease severity estimation with regard to hypothesis testing," *Plant Pathology*, vol. 65, no. 4, pp. 523–535, 2016.

[12] M. M. Conner, C. W. McCarty, and M. W. Miller, "Detection of bias in harvest-based estimates of chronic wasting disease prevalence in mule deer," *Journal of Wildlife Diseases*, vol. 36, no. 4, pp. 691–699, 2000.

[13] E. K. Accorsi, X. Qiu, E. Rumpler, L. Kennedy-Shaffer, R. Kahn, K. Joshi, E. Goldstein, M. J. Stensrud, R. Niehus, M. Cevik *et al.*, "How to detect and reduce potential sources of biases in studies of sars-cov-2 and covid-19," *European Journal of Epidemiology*, vol. 36, pp. 179–196, 2021.

[14] D.-C. Huang, J.-F. Wang, J.-X. Huang, D. Z. Sui, H.-Y. Zhang, M.-G. Hu, and C.-D. Xu, "Towards identifying and reducing the bias of disease information extracted from search engine data," *PLoS Computational Biology*, vol. 12, no. 6, p. e1004876, 2016.

[15] J. K. Bower, S. Patel, J. E. Rudy, and A. S. Felix, "Addressing bias in electronic health record-based surveillance of cardiovascular disease risk: finding the signal through the noise," *Current epidemiology reports*, vol. 4, pp. 346–352, 2017.

[16] T. A. Alonzo, J. T. Brinton, B. M. Ringham, and D. H. Glueck, "Bias in estimating accuracy of a binary screening test with differential disease verification," *Statistics in Medicine*, vol. 30, no. 15, pp. 1852–1864, 2011.

[17] L. Wood, M. Egger, L. L. Gluud, K. F. Schulz, P. Jüni, D. G. Altman, C. Gluud, R. M. Martin, A. J. Wood, and J. A. Sterne, "Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study," *bmj*, vol. 336, no. 7644, pp. 601–605, 2008.

[18] A. Nikolas Angelopoulos, R. Pathak, R. Varma, and M. I. Jordan, "On identifying and mitigating bias in the estimation of the covid-19 case fatality rate," *arXiv e-prints*, pp. arXiv–2003, 2020.

[19] R. E. Mitchell, A. E. Hartley, V. M. Walker, A. Gkatzionis, J. Yarmolinsky, J. A. Bell, A. H. Chong, L. Paternoster, K. Tilling, and G. D. Smith, "Strategies to investigate and mitigate collider bias in genetic and mendelian randomisation studies of disease progression," *PLoS Genetics*, vol. 19, no. 2, p. e1010596, 2023.

[20] A. I. Pack, U. J. Magalang, B. Singh, S. T. Kuna, B. T. Keenan, and G. Maislin, "Randomized clinical trials of cardiovascular disease in obstructive sleep apnea: understanding and overcoming bias," *Sleep*, vol. 44, no. 2, p. zsaa229, 2021.

[21] W. J. Hall, M. V. Chapman, K. M. Lee, Y. M. Merino, T. W. Thomas, B. K. Payne, E. Eng, S. H. Day, and T. Coyne-Beasley, "Implicit racial/ethnic bias among health care professionals and its influence on health care outcomes: a systematic review," *American journal of public health*, vol. 105, no. 12, pp. e60–e76, 2015.

[22] L. Shan, S. Wang, L. Wu, and F.-S. Tsai, "Cognitive biases of consumers' risk perception of foodborne diseases in china: examining anchoring effect," *International Journal of Environmental Research and Public Health*, vol. 16, no. 13, p. 2268, 2019.

[23] P. Croskerry, "Achieving quality in clinical decision making: cognitive strategies and detection of bias," *Academic emergency medicine*, vol. 9, no. 11, pp. 1184–1204, 2002.

[24] G. Sajeev, J. Weuve, J. W. Jackson, T. J. VanderWeele, D. A. Bennett, F. Grodstein, and D. Blacker, "Late-life cognitive activity and dementia: a systematic review and bias analysis," *Epidemiology (Cambridge, Mass.)*, vol. 27, no. 5, p. 732, 2016.

[25] N. V. Dawson, H. R. Arkes, C. Siciliano, R. Blinkhorn, M. Lakshmanan, and M. Petrelli, "Hindsight bias: an impediment to accurate probability estimation in clinicopathologic conferences," *Medical Decision Making*, vol. 8, no. 4, pp. 259–264, 1988.

[26] E. T. Jensen, S. F. Cook, J. K. Allen, J. Logie, M. A. Brookhart, M. D. Kappelman, and E. S. Dellon, "Enrollment factors and bias of disease prevalence estimates in administrative claims data," *Annals of epidemiology*, vol. 25, no. 7, pp. 519–525, 2015.

[27] A. E. Rudolph, C. M. Fuller, and C. Latkin, "The importance of measuring and accounting for potential biases in respondent-driven samples," *AIDS and Behavior*, vol. 17, pp. 2244–2252, 2013.

[28] A. D. Baines, M. R. Partin, M. Davern, and T. H. Rockwood, "Mixed-mode administration reduced bias and enhanced poststratification adjustments in a health behavior survey," *Journal of clinical epidemiology*, vol. 60, no. 12, pp. 1246–1255, 2007.

[29] E. C. McFadden, J. A. Hirst, J. Y. Verbakel, J. H. McLellan, F. R. Hobbs, R. J. Stevens, C. A. O'Callaghan, and D. S. Lasserson, "Systematic review and metaanalysis comparing the bias and accuracy of the modification of diet in renal disease and chronic kidney disease epidemiology collaboration equations in community-based populations," *Clinical chemistry*, vol. 64, no. 3, pp. 475–485, 2018.

[30] S. Lachish and K. A. Murray, "The certainty of uncertainty: potential sources of bias and imprecision in disease ecology studies," *Frontiers in veterinary science*, vol. 5, p. 90, 2018.

[31] M. É. Czeisler, J. F. Wiley, C. A. Czeisler, S. M. Rajaratnam, and M. E. Howard, "Uncovering survivorship bias in longitudinal mental health surveys during the covid-19 pandemic," *Epidemiology and psychiatric sciences*, vol. 30, p. e45, 2021.