Variance Estimation and Confidence Intervals for 11 Commonly Used Health Disparity Measures

abstract

There is increased interest in eliminating health disparities in the United States and worldwide. Broadly defined, health disparities refer to preventable inequalities in health status, such as cancer to ethnicity, socioeconomic status, gender, education, environment, and geographic locations. To make informed health policy decisions, it is essential to precisely measure the magnitude of disparities and assess trends over time. The Health Disparities Calculator (HD*Calc) is free statistical software that calculates 11 commonly used measures of health disparities and provides corresponding 95% CIs for the 11 measures using either an analytic method or a Monte Carlo simulation–based method; however, the derivation of SEs and coverage properties of the CIs have not been formally evaluated. We used simulation studies to assess the coverage properties of these CIs. We have also conducted bias analyses for measures implemented in HD*Calc using age-adjusted cancer incidence rates from national, state, and county level SEER data. The results of these analyses indicate that HD*Calc should be used with caution to construct CIs for some health disparity measures when the proportion of zero event counts is greater than 25%.

Clin Cancer Inform. © 2018 by American Society of Clinical Oncology

INTRODUCTION

In recent years, there has been increasing interest in eliminating health disparities not only expressed by US government agencies,¹⁻⁴ but also by worldwide institutions, such as the WHO and the World Bank. Broadly, health disparities refer to preventable inequalities in health status, such as cancer prevalence or incidence and measures of health, access to care, and quality of health care receipts, across social groups defined by ethnicity, socioeconomic position, age, gender identity, education, environment, and geographic locations. It is essential for health-related policymakers or members of economic government bodies to be provided with precise estimates of the magnitude of health disparities and correct assessments of trends over time. With proper evaluations and accurate monitoring of the effects of current policies, both health-related policymakers and biomedical researchers can help improve current interventions so as to lead to the elimination of health disparities, as well as to eventually “improve the health of all groups to the highest level of health.”⁴

To measure health disparities between well-defined groups across different social dimensions,

researchers have used a number of quantitative measures. Currently, a range of health disparities measures have been proposed to provide a complete description of the magnitude of health disparities.^{5,6} The measurement of health disparities is complicated because of a number of issues, including whether to measure disparities on the relative or absolute scale and whether to weigh social groups by their population size. For example, in the context of a recent decline in cervical cancer incidence rates, absolute disparities—for example, the range difference—have decreased, whereas relative disparities—for example, the range ratio—have increased.⁷ Such discrepancies between measures of health disparities are a consequence of the value judgments that are unavoidable when trying to quantify a complex concept, such as inequality.⁸⁻¹⁰ In many cases, there is no unambiguous way to measure health disparities, and a suite of measures may be needed to provide a complete picture.

The Health Disparity Calculator (HD*Calc)⁷ is statistical software that calculates 11 commonly used health disparities measures, together with their corresponding 95% CIs, using both analytical and Monte Carlo simulation (MCS)–based methods. The 11 measures include both relative

Jaeil Ahn
Sam Harper
Mandi Yu
Eric J. Feuer
Benmei Liu
George Luta

Author affiliations and support information (if applicable) appear at the end of this article.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Corresponding author:

Jaeil Ahn, PhD, Department of Biostatistics, Georgetown University, 3900 Reservoir Rd NW, Basic Science 255, Washington, DC 20057; e-mail: ja1030@georgetown.edu.

and absolute measures of disparities. Absolute measures include range difference (RD), absolute concentration index (ACI), slope index of inequality (SII), and between-group variance (BGV). Relative measures include range ratio (RR), index of disparity (IDisp), Theil's index (T), mean log deviation (MLD), Kunst-Mackenbach relative index (KMI), relative concentration index (RCI), and relative index of inequality (RII). In addition to RD and RR, the software allows for the calculation of the rate difference and rate ratio for any two prespecified groups. In practical situations, HD*Calc can be used to calculate the 11 health disparities measure not only for cancer-related data, but also for other population-based health outcomes.

Despite the convenience of HD*Calc, details on SE estimation or coverage performance of CIs produced by HD*Calc has not been formally evaluated. In this article, we provide a detailed review of the 11 measures of health disparities and the derivation of their SEs. We then evaluate the coverage properties of CIs and provide advice on when the CIs can be reliably used in practice on the basis of the results of bias analyses using SEER data.

This article is organized as follows. In Background and Notation, we briefly introduce the notation and describe variance estimation methods, including analytic and MCS methods. In The 11 Health Disparities Measures, we review the 11 health disparities measures implemented in HD*Calc and describe their characteristics. Through simulation studies, we evaluate the coverage performance of the CIs constructed using both analytic and MCS methods in Simulation Studies. In Conclusion, we investigate when the two approaches are recommended to be used in practice through comprehensive bias analyses using SEER data. We conclude with a brief remark in the Appendix. Mathematical details on the derivations of SEs using the analytical methods, including the Taylor series expansion approximation method, the direct derivation method, the quadratic form method, and the weighted least squares estimation method, are presented in the Appendix.

BACKGROUND AND NOTATION

For health statistics, age-adjusted rates are commonly used to address the confounding effect of age on the associations of exposures of interest with health outcomes. Let A_k denote the observed

rate in the k -th age group. Rate A_k is expressed as X_k/n_k , where X_k is the number of events and n_k is the number of person-years or the population size corresponding to the k -th age group. Let Y denote the age-adjusted estimate obtained as $Y = \sum_k w_k A_k$ when w_k is the weight for the k -th age group from the reference population. Then, the variance of Y is given as $Var(Y) = \sum_k w_k^2 Var(A_k)$ by the independence of rates across the age groups. The number of events, X_k , is frequently assumed to follow a Poisson distribution with mean $n_k \lambda_k$ —that is, $X_k \sim Poi(n_k \lambda_k)$ with the true rate λ_k . Then,

$$\begin{aligned} Var(Y) &= \sum_k w_k^2 Var(A_k) = \sum_k w_k^2 Var(X_k/n_k) \\ &= \sum_k w_k^2/n_k^2 Var(X_k) = \sum_k (w_k^2/n_k) \lambda_k. \end{aligned}$$

When it comes to health status comparisons, we are interested in comparing J mutually exclusive social groups. Let Y_j , $j = 1, \dots, J$ be the age-adjusted rate that corresponds to the j -th social group, where $\mathbf{Y} = (Y_1, \dots, Y_J)^T$ and $\mathbf{y}$ is a realization of the vector $\mathbf{Y}$. We denote by μ_j and σ_j^2 the unknown mean and variance of Y_j . To estimate μ_j and σ_j^2 , we consider their unbiased estimators, then plug them in while deriving the formulas for SE. Specifically, we use the unbiased estimator, Y_j , as $\hat{\mu}_j$ and $\sum_k w_k^2/n_k^2 X_k$ as $\hat{\sigma}_j^2$. The overall or population-averaged μ is expressed as $\sum_j p_j \mu_j$, where p_j is the relative size of group j —that is, $\sum_j p_j = 1$. Similarly, we can use $\sum_j p_j Y_j$ as $\hat{\mu}$, the unbiased estimator of μ . For relative measures of health disparities, we often use a relative ratio, $R_j = Y_j/\hat{\mu}$, that represents the ratio of the health status of the j -th population to the average health status.

For groups defined using an ordinal scale of social status, such as household income or years of education, the population of each social group corresponds to a part of the ordered cumulative distribution, which is scaled overall from 0 to 1. We introduce the socioeconomic rank, Z , where $Z = 0$ at the hypothetical bottom of the distribution, and $Z = 1$ at the hypothetical top of the distribution. In HD*Calc, Z_j is the relative rank of the j -th social group, which is defined as

$$Z_j = \sum_{k=1}^j p_k - 0.5p_j$$

where Z_j indicates the cumulative share of the populations up to the midpoint of the social group j . For example, when we use social groups

that correspond to the deciles of the household income distribution, for the bottom group, $Z_1 = 0.05$, and for the top group, $Z_{10} = 0.95$. We also define the weighted average rank, $\bar{Z} = \sum_j p_j Z_j$. We assume the independence of the Z s (additional discussion can be also found in Kakwani¹¹).

To estimate the variance of estimators of health disparity measures in HD*Calc, both analytic methods, including the Taylor series expansion approximation method, the direct derivation method, the quadratic form approximation method, and the weighted least squares method, as well as an MCS method, are used. HD*Calc allows users to select either type of method to estimate variance and construct CIs. We give a mathematical overview of all these methods in the following section.

Variance Estimation Using the Taylor Series Approximation Method

When the analytic approach is chosen, HD*Calc uses a Taylor series approximation method for variance estimation for the majority of the 11 health disparities measures. If θ is the true health disparities measure estimated by $\hat{\theta} = F(\mathbf{Y})$, we want to estimate $\text{Var}(\hat{\theta})$ and construct a CI for θ . To do so, we expand F using a first-order Taylor series around μ , where $\mathbf{Y}$ has the mean, μ , and variance, σ^2 :

$$\text{Var}(\hat{\theta}) \approx \text{Var} \left\{ \sum_{j=1}^J \left(\frac{\partial F}{\partial y_j} \right)_{y=\mu} (Y_j - \mu_j) \right\},$$

where the second-order derivative or higher-order terms are considered to be negligible. If we assume the independence of each component of $\mathbf{Y}$, because they correspond to different social groups, the above variance will be approximated by

$$\text{Var}(\hat{\theta}) \approx \sum_{j=1}^J \left(\frac{\partial F}{\partial y_j} \right)_{y=\mu}^2 \sigma_j^2. \quad (1)$$

Using this mathematical tool, we describe how the variances of RR, RII, KMI, IDisp, T, MLD, RCI, and ACI are obtained in the Appendix. In practice, to obtain $\widehat{\text{Var}}(\hat{\theta})$, we replace the unknown parameters, $\mu_j \sigma_j^2$, and μ , with their unbiased estimators introduced at the beginning of this section.

Variance Estimation Using the MCS-Based Method

For the MCS approach, we randomly generate new age-adjusted rates, y_j^{new} , for each social group, j , according to a truncated normal distribution instead of the Poisson distribution because of its continuous nature. When doing so, the observed mean and variance of each age-adjusted rate are used for sampling. We draw age-adjusted rates, y_j^{new} , for $j = 1, \dots, J$ at each iteration; a negative simulated value would be rejected and another value would be redrawn. We repeat the procedure B times to obtain B simulated samples. We then calculate the health disparities measure for each simulated sample, say, $\hat{\theta}_1, \dots, \hat{\theta}_B$. Estimated SE is the sample standard deviation

$$(B-1)^{-1} \sum_{b=1}^B (\hat{\theta}_b - \bar{\theta})^2,$$

where $\bar{\theta} = B^{-1} \sum_b \hat{\theta}_b$.

As a default setting for HD*Calc, $B = 1,000$ iterations are used to calculate the SE.

THE 11 HEALTH DISPARITIES MEASURES

According to the documentation of the HD*Calc Web site,⁷ we group the 11 measures as absolute measures, including RD, ACI, SII, and BGV; and relative measures, including RR, IDisp, T, MLD, KMI, RCI, and RII. The associated analytic variance estimators are provided in the Appendix.

Absolute Measures of Health Disparities

RD. RD is defined as $RD = Y_{\max} - Y_{\min}$, where $Y_{\max}$ and $Y_{\min}$ are the maximum and minimum age-adjusted rates among the J social groups, respectively. RD is the largest value, among all $\binom{J}{2}$ possible pairwise comparisons, of the absolute differences between two age-adjusted rates for two social groups. RD compares the social group with the worst health status with the social group with the best health status, which represents the largest gap possible. An RD value away from zero indicates disparity. HD*Calc also allows users to choose two specific social groups, j and j' , so that users can obtain the rate difference, $Y_j - Y_{j'}$, as opposed to the RD.

ACI. In contrast to the RCI (see RCI), ACI measures whether any particular social group

received concentrated health or illness on the absolute scale. It is applicable to ordered social groups. The absolute version of the concentration index is RCI multiplied by the mean rate:

$$ACI = \hat{\mu}RCI = 2 \sum_{j=1}^J p_j Y_j Z_j - \hat{\mu},$$

where Z_j is the relative rank of the social group j , and $\hat{\mu}$ is the average rate in the population (see Background and Notation for details).

SII. SII captures the difference in the average health status between a person in the highest social group and a person in the lowest social group. SII is derived from a simple linear regression model—that is,

$$E(Y_j|Z_j) = \beta_0 + \beta_1 Z_j.$$

Here, β_0 is the health status of a hypothetical person at the bottom of the social group hierarchy ($Z = 0$) and $\hat{\beta}_1$, or SII, is the estimated difference in the average health status between the hypothetical top group and the hypothetical bottom group, assuming linearity. In this sense, the estimated coefficient, $\hat{\beta}_1$, is comparable to RD. The larger the SII, the more disparity exists.

BGV. Measures that are based on variance are often used to summarize deviations from a population mean. BGV calculates the weighted sum of squares of the differences between the group-average rates and the population average. BGV is defined as

$$BGV = \sum_{j=1}^J p_j (Y_j - \hat{\mu})^2 = \sum_{j=1}^J p_j Y_j^2 - \hat{\mu}^2.$$

We can interpret that a larger BGV indicates more severe disparities between social groups. BGV can serve as a useful indicator of the absolute disparities as it is sensitive to the magnitude of larger deviations from the population average.

Relative Measures of Health Disparities

RR. RR is calculated by dividing $Y_{\max}$ by $Y_{\min}$ so that the rate of the social group with the worst health status and that of the social group with the best health status are compared. We write it as $RR = Y_{\max}/Y_{\min}$. The further that RR is from unity, the more disparities exist. Similar to RD, RR is the largest among the $\binom{J}{2}$ possible ratios of two age-adjusted rates for two social groups.

HD*Calc also allows users to fix two groups, j and j' , in advance so that users can calculate the relative ratio as well.

IDisp. IDisp is calculated by first summing the absolute differences between social group rates and a reference rate, then dividing it by the reference rate.¹² It is expressed as

$$ID_{isp} = \frac{1}{J-1} \left(\sum_{j=1, j \neq \text{ref}}^J \frac{|Y_j - Y_{\text{ref}}|}{Y_{\text{ref}}} \right) \times 100, \quad (2)$$

where Y_{ref} is the rate in the reference population. If the social group with the best health status—that is, the social group with the smallest rate—is used as the reference group, then the absolute value is not needed in the formula. Hence, we can restate IDisp as

$$ID_{isp} = \left[\frac{1}{(J-1)} \sum_{j=1, j \neq \text{ref}}^J \frac{Y_j}{Y_{\text{ref}}} - 1 \right] \times 100, \quad (3)$$

We can interpret that a larger value indicates more severe disparities relative to the social group with the best health status.

T. The T index measures general disproportionality.¹³ It is a summary of the relative distance that the population is away from the egalitarian state when everyone has the same health status. T is defined as

$$T = \sum_{j=1}^J p_j R_j \ln(R_j).$$

Because of its definition, T is more sensitive to health differences further away from the average as a result of the multiplication of the logarithm. For example, a larger absolute value—either positive or negative—indicates a state that is further away from the egalitarian state. T has the advantage that it can be used for both ordered social groups (eg, education and/or household income) and unordered social groups (eg, gender and/or race).

MLD. Similar to T, MLD is a population-weighted measure of relative disparity, and it is sensitive to differences from the average rate:

$$\begin{aligned} MLD &= \sum_{j=1}^J p_j \ln\left(\frac{1}{R_j}\right) = - \sum_{j=1}^J p_j \ln\left(\frac{Y_j}{\hat{\mu}}\right) \\ &= \ln(\hat{\mu}) - \sum_{j=1}^J p_j \ln(Y_j). \end{aligned}$$

This measure is also useful for both ordered and unordered social groups.

RII. Whereas SII is a measure of absolute disparities between the two extreme situations

under a linearity assumption, a relative disparities measure, RII, can be obtained by dividing SII by the mean rate.¹⁴ We can write it as $RII = SII/\hat{\mu} = \hat{\beta}_1/\hat{\mu}$. It can be interpreted as a relative SII with regard to the population average rate.

KMI. Kunst and Mackenbach modified the RII by dividing the estimated health of a hypothetical person at the bottom of the social group distribution, when $Z = 0$, by the estimated health of a hypothetical person at the top of the social group distribution ($Z = 1$). Thus, KMI is defined as

$$KMI = \frac{\hat{\beta}_0}{\hat{\beta}_0 + \hat{\beta}_1} = \frac{\hat{\beta}_0}{\hat{\beta}_0 + SII}.$$

KMI compares the health status of the least advantaged and most advantaged social groups. For example, if $KMI = 2$ for lung cancer incidence rates, then the lung cancer age-adjusted incidence rate of the most disadvantaged social group is twice as high as that of the most advantaged social group.

RCI. RCI measures the extent to which health or illness is concentrated among particular social groups. RCI was originally intended for social groups that have an inherent ranking, such as income or education groups. Compared with KMI, which compares the two extreme groups, RCI summarizes the health distribution across all social groups. The formula for RCI was proposed by Kakwani¹¹ and is given by

$$RCI = \frac{\sum_{j=1}^J p_j Y_j Z_j}{\hat{\mu}} - 1 = \frac{\sum_{j=1}^J p_j (2Z_j - 1) Y_j}{\hat{\mu}}.$$

RCI reflects the socioeconomic gradient of inequalities in health. In other words, a downward health tendency, such that health worsens with social group rank, results in a positive RCI, whereas an upward health tendency results in a negative RCI.

SIMULATION STUDIES

We performed a series of simulation studies to evaluate the coverage performance of the CIs implemented in HD*Calc. We focus on coverage probabilities (CPs) computed as the frequency of 95% CIs that cover the true value of the health disparities measure. To be realistic, the 11 true values of the measures of health disparities were based on real data.

To generate simulated data sets, we used publicly available cancer data sets from the National

Cancer Institute SEER program from 2008 to 2012.¹⁵ Because different cancer types were involved, cancer-related data sets had different rates of events (Table 1). We have selected the specific social groups, cancer types, and regions of interest to get a range for the number of social groups, cancer rates, and proportions of zeros. SEER data sets are uniquely formatted with 19 standard 5-year age groups ($k = 1, \dots, 19$) and a varying number of ordered or unordered social groups (eg, $J = 10$ for ordered social groups on the basis of deciles of household income, and $J = 6$ for unordered social groups on the basis of race).

Given the age group and social group combination specific λ_{jk} and n_{jk} obtained from SEER data, we iteratively generate the age-adjusted rates, y_j , as follows. At the n -th iteration, we sample $x_{jk}^{(n)}$ from $Poi(n_{jk}\lambda_{jk})$ for $j = 1, \dots, J$ and $k = 1, \dots, 19$. Given a set of $x_{jk}^{(n)}$ for all k , we compute the age-adjusted rate, $y_j^{(n)} = \sum_k w_k a_{jk}^{(n)}$, where $a_{jk}^{(n)}$ is $x_{jk}^{(n)}/n_j$ and w_k is fixed. When the observed rates, λ_{jk} , are zero for some k , the Poisson sampling leads to zero values for $x_{jk}^{(n)}$ as a result of assumed zero true mean (and variance). To avoid issues related to the estimation of the variability of age-adjusted rates, we add $1/n_{jk}$ to the mean of Poisson distribution, $n_{jk}\lambda_{jk}$, across all age groups. We repeat this procedure $n = 10,000$ times. On the basis of these 10,000 estimates for each of the 11 measures of health disparities, we compute the coverage probabilities. For measures for ordered groups, such as RII, RCI, ACI, and KMI, we use HD*Calc only for the decile household income groups. We obtain the estimated CPs from both the analytic and MCS methods.

Table 2 lists the results when nationwide cervical cancer, breast cancer, and lung cancer incidence data are used. Cervical and breast cancer data are based on unordered racial and ethnic groups ($J = 2, 6$), respectively, whereas lung cancer data are based on ordered social groups, where the 10 groups are formed using the deciles of county-level percent of families below the poverty level derived from the 5-year average of the American Community Survey 2008 to 2012. Regardless of the number of groups, cancer type, and whether the groups are ordered or not, in general both the analytic and MCS methods produce CPs that are close to the nominal

Table 1. Partial SEER Data for Female Brain, Breast, Cervical, and Lung Cancer From 12 States, Excluding Alaska; Nationwide; and King County (Washington state)

Cancer by State*	Group	Min AR†	Max AR*	Min Var*	Max Var*	Average Count	Proportion Zeros	Population Size
California	6.0	29.91	58.15	1.161	10.359	47.69	0.07	93,802,864
Connecticut	6.0	0.0	58.68	0.0	7.851	5.97	0.46	9,167,187
Georgia	6.0	0.0	57.23	0.0	7.198	12.74	0.28	24,826,892
Hawaii	6.0	0.0	63.76	0.0	11.547	1.24	0.59	3,394,788
Iowa	6.0	12.96	74.40	2.841	52.613	5.51	0.66	7,693,357
Female brain								
Kentucky	6.0	0.0	64.28	0.0	22.099	7.67	0.54	11,028,460
Louisiana	6.0	14.09	55.41	2.724	14.089	6.11	0.44	11,570,363
Michigan	6.0	33.74	64.92	2.817	49.025	6.02	0.40	10,002,307
New Jersey	6.0	0.0	59.95	0.0	4.402	13.71	0.24	22,552,663
New Mexico	6.0	10.39	52.31	4.066	14.469	2.52	0.49	5,195,854
Utah	6.0	6.58	59.42	3.374	14.599	3.78	0.58	6,881,752
Washington	6.0	15.40	66.84	2.717	13.736	7.92	0.35	11,548,466
Breast								
Nationwide	6.0	906.45	1,340.42	3.016	23.915	3,120.40	0.17	217,963,069
Cervical								
Nationwide	2.0	73.19	98.72	0.649	1.675	450.18	0.13	217,963,069
Lung								
Nationwide	10.0	478.93	755.57	2.526	5.923	1,347.72	0.09	430,457,544
Breast								
King County	6.0	982.413	1,664.028	20.100	225.407	78.41	0.32	4,865,310
Cervical								
King County	2.0	60.30	104.81	3.540	24.400	8.58	0.34	4,865,310

NOTE. Included are the number of social groups, minimum (Min) and maximum (Max) of age-adjusted rates (AR), minimum and maximum variance (Var) of AR, average event count per social group, proportion of zero event counts ($x_{jk} = 0$), and population size across the $J \times K$ age group by social group combinations. Full SEER data Table is shown in Appendix [Table A1](#).

*Some states among 13 states per each cancer are not shown because of the table size limit.

†Values are multiplied by 10^6 .

level of 0.95, showing less than 1% bias, where bias is defined as the estimated CP minus 95%; however, compared with the MCS method, the analytic method for IDisp yields conservative coverage estimates, with more than 98% estimated CPs when breast and lung cancer data are used. Similarly, [Table 3](#) lists the results when Georgia cervical cancer, breast cancer, and lung cancer incidence data are used.

Simulation Results: Example 1

We used three cancer data sets containing nationwide cervical cancer, breast cancer, and lung cancer incidence rates. Social groups for these data sets are two ethnic groups (Spanish Hispanic Latino and non-Spanish Hispanic Latino [reference]), six racial/ethnic groups

(Hispanic [reference], white, black, Asian, Asian Pacific Islander, and unknown), and 10 groups on the basis of deciles of household income in 2008 (the lowest decile group is the reference). For each data set, age is categorized into 19 standard age groups using 5-year intervals ($K = 19$). Rates are per 100,000 persons and are age adjusted to the year 2000 nationwide standard population (single ages to 84 to Census P25-1130).

Simulation Results: Example 2

Similar to the previous example, the corresponding three cancer data sets from the Georgia Cancer Registry are used. It is important to note that the population size of Georgia is relatively small compared with the nationwide population.

Table 2. Coverage Probabilities for 11 Health Disparity Measures, Using Analytic and MCS Methods With Nationwide SEER Data

Measure	Cervical Cancer (J = 2)		Breast Cancer (J = 6)		Lung Cancer (J = 10)	
	Analytic Methods*	MCS Method	Analytic Methods*	MCS Method	Analytic Methods*	MCS Method
RD†	0.9481	0.9555	0.9499	0.9501	0.9552	0.9533
BGV	0.9463	0.9553	0.9491	0.9459	0.9507	0.9451
ACI‡	NA	NA	NA	NA	0.9516	0.9590
SII‡	NA	NA	NA	NA	0.9516	0.9590
RR†	0.9457	0.9529	0.9508	0.9465	0.9552	0.9527
IDisp†	0.9465	0.9532	0.9987	0.9484	0.9839	0.9451
MLD	0.9450	0.9529	0.9506	0.9479	0.9521	0.9498
T	0.9448	0.9529	0.9515	0.9470	0.9514	0.9499
RCI‡	NA	NA	NA	NA	0.9520	0.9588
RII‡	NA	NA	NA	NA	0.9520	0.9588
KMI‡	NA	NA	NA	NA	0.9524	0.9588

Abbreviations: ACI, absolute concentration index; BGV, between-group variance; IDisp, index of disparity; KMI, Kunst-Mackenbach index; MCS, Monte Carlo simulation; MLD, mean log deviation; RCI, relative concentration index; RD, range difference; RII, relative index of inequality; RR, range ratio; SII, slope index of inequality; T, Theil index.

*Direct derivation for RD, quadratic form for BGV, weighted least squares estimation for SII, and Taylor series approximation for the other measures.

†SE calculation assumes the independence of the statistical components.

‡Health disparity measures are defined only for ordered social groups.

Bias Analyses

Our simulation studies indicate that HD*Calc yields accurate calculations of SEs with less than 1% absolute bias of CPs for all health disparities measures with the exception of IDisp when nationwide SEER data sets are used; however, we start to see increased bias for the majority of health disparities measures for data for Georgia—for

example, 4.92% bias for BGV for cervical cancer and two ethnic groups. A natural question arises about what characteristics of the data set affect the empirical coverage of the 95% CIs, and if simple rules can be identified with regard to when HD*Calc should not be used to construct 95% CIs for the 11 measures of health disparities.

To answer these question, we have investigated biases of CPs for all 11 health disparities

Table 3. Coverage Probabilities for 11 Health Disparity Measures, Using Analytic and MCS Methods With Georgia SEER Data

Measure	Cervical Cancer (J = 2)		Breast Cancer (J = 6)		Lung Cancer (J = 10)	
	Analytic Methods*	MCS Method	Analytic Methods*	MCS Method	Analytic Methods*	MCS Method
RD†	0.9825	0.9808	0.9345	0.9239	0.9495	0.9489
BGV	0.9992	0.9802	0.9281	0.9390	0.9546	0.9154
ACI‡	NA	NA	NA	NA	0.9491	0.9476
SII‡	NA	NA	NA	NA	0.9491	0.9476
RR†	0.9669	0.9803	0.9348	0.9241	0.9496	0.9486
IDisp†	0.9900	0.9807	0.9797	0.9293	0.9839	0.9421
MLD	0.9602	0.9807	0.9380	0.9275	0.9504	0.9106
T	0.9591	0.9809	0.9285	0.9340	0.9502	0.9086
RCI‡	NA	NA	NA	NA	0.9497	0.9472
RII‡	NA	NA	NA	NA	0.9497	0.9472
KMI‡	NA	NA	NA	NA	0.9504	0.9472

Abbreviations: ACI, absolute concentration index; BGV, between-group variance; IDisp, index of disparity; KMI, Kunst-Mackenbach index; MCS, Monte Carlo simulation; MLD, mean log deviation; RCI, relative concentration index; RD, range difference; RII, relative index of inequality; RR, range ratio; SII, slope index of inequality; T, Theil index.

*Direct derivation for RD, the quadratic form for BGV, weighted least squares estimation for SII, and Taylor series approximation for the other measures.

†SE calculation assumes the independence of the statistical components.

‡Health disparity measures are defined only for ordered social groups.

measures by performing additional simulation studies using SEER data sets from nationwide, 13 states, and King county from Washington state for four cancer types, including cervical, breast, lung, and female brain cancer incidence rates and different numbers of social groups. We have considered the following seven factors: the number of social groups, minimum and maximum of age-adjusted rates (AR), minimum and maximum variance of AR, average event count per social group, and proportion of zero event counts ($x_{jk} = 0$) across the $J \times K$ age group by social group combinations. These characteristics are for the data sets from which we simulated samples and do not necessarily correspond to the characteristics of the observed data. For example, the proportion of zero event counts is an expected proportion of zero counts under the assumed Poisson distributions, and it may be different from the observed proportion of zero counts across samples. Appendix Table A1 presents these characteristics for selected data sets used in the simulations. HD*Calc cannot generate results when there are missing data for any age group by social group combination or when the population size for a specific age group by social group combination (n_{jk}) is zero. Hence, HD*Calc users need to address such situations before using HD*Calc (eg, by excluding age or social groups or collapsing scarce groups). For situations that do not involve these problems, we can calculate health disparities measures and construct the corresponding 95% CIs

On the basis of the results of our simulations, we estimate the biases of CPs for the situation of a 95% nominal level. For each measure, we fit two multivariable linear regression models, one for the analytic methods and another for the MCS method, where bias is regressed on the seven factors listed previously. Figures 1A and 1C show the distributions of the P values that correspond to the seven predictors on the x-axis for the analytic method (top) and MCS method (bottom), respectively. The proportion of zeros across the 19 age groups by J social group combinations is consistently associated with bias across all 11 measures, with $P < .05$ for both methods. In contrast, P values for the other predictors do not indicate statistically significant results ($P < .05$), with small variation among the 11 measures. For example, BGV seems to be negatively associated with the number of social groups and maximum of age-adjusted rates but positively associated

with the maximum variance of AR. When unadjusted simple linear regression was fitted for each outcome, the proportion of zeros is the sole significant factor after false discovery rate adjustment as shown in Appendix Table A2. We further examine the scatter plots of biases of the 11 measures against the proportion of zeros for the analytic and MCS methods (Figs 1B and 1D). For both methods, we observe larger biases (absolute bias $\geq 10\%$) when the proportion of zeros is greater than 25%. In particular, KMI, T, MLD, and BGV may be sensitive to the proportion of zeros when analytic methods are used, whereas MLD, T, IDisp, BGV, and RD may be sensitive when the MCS method is used. We also performed a similar analysis using an indicator of good coverage as our binary outcome variable in which we define good coverage as having estimated CP that ranges from 94% to 96%. We used a logistic regression model with the same predictors as those used previously. The proportion of zeros across the 19 age groups by J social groups combinations was again the strongest predictor for most measures (data not shown).

On the basis of the results of the bias analyses, we conclude that the 95% CIs generated by HD*Calc have empirical coverage close to the nominal level for situations when data are not sparse, specifically, when the proportion of zero event counts across the age group by social group combinations is 25% or less. We also observed that analytic methods and the MCS method perform comparably without noticeable differences.

CONCLUSION

We have presented 11 health disparities measures implemented in HD*Calc, together with the derivation of their SEs. Our simulation studies indicate that the 95% CIs that are based on the analytic methods and MCS method perform well except for sparse data sets. HD*Calc users must be aware that for situations when the proportion of zero event counts is greater than 25%, the 95% CIs for some of the 11 measures of health disparities should be used with caution. These measures include KMI, T, MLD, and BGV when analytic methods are used, and MLD, T, IDisp, BGV, and RD when the MCS method is used.

It is important to note that issues related to small sample sizes or rare diseases are analogous to the sparse data problem in the multiway contingency table analysis. Hence, HD*Calc users can use approaches proposed for contingency

Fig 1. (A and C) Plots for the P values from the multivariable linear regression models that regressed the bias of the coverage probabilities (CPs) of the 11 health disparity measures on the seven predictors: the number of social groups (N group), minimum and maximum of age-adjusted rates (Min AR and Max AR), minimum and maximum variance of age-adjusted rates (Min Var and Max Var), proportion of zero event counts across the age group by social group combinations (Prop Zero), and average event count per social group (Avg A) using the (A) analytic and (C) Monte Carlo simulation (MCS)-based methods. (B and D) Scatter plots of bias of CPs from the 11 measures against the proportion of zero event counts using the (B) analytic and (D) MCS methods. ACI, absolute concentration index; BGV, between-group variance; IDisp, index of disparity; KMI, Kunst-Mackenbach index; MLD, mean log deviation; RCI, relative concentration index; RD, range difference; RII, relative index of inequality; RR, range ratio; SII, slope index of inequality; T, Theil index.

tables in the presence of avoidable sparse data. For example, one can consider in advance either to collapse adjacent age groups or to redefine and expand social groups, time frames, and regions to make sure that each age group has a sufficiently large number of events. In contrast, for HD*Calc developers and statisticians,

future research work should focus on applying other methods for sparse data, such as using a zero-inflated Poisson distribution to construct valid CIs.

Another limitation is that the analytic SE estimation methods implemented in HD*Calc for RD

and RR are valid only when the two social groups to be compared are fixed a priori. Otherwise, the SE formulas for RD and RR are not correct because they do not account for the dependence between Y_{max} and Y_{min} as indicated in the Appendix. Similarly, the analytic SE estimation method for IDisp is valid only when $Y_j = Y_{min}$ is fixed a priori, so that the first-order derivatives for the Taylor series approximation method exist and are correct; however, in general, as

$$\frac{\partial}{\partial y_j} \left(\frac{y_j}{\min(y_1, \dots, y_j)} \right) \neq \frac{1}{\min(y_1, \dots, y_j)},$$

the SE formula is not correct. As for RD and RR, for IDisp HD*Calc posits that the social group

with the observed minimum rate is the social group with the true minimum rate—and also the social group with the observed maximum rate is the social group with the true maximum rate. Because of these statistical issues regarding SE derivations for RR, RD, and IDisp, we recommend that HD*Calc users use the MCS method for these specific measures. Additional research should address these issues for RD, RR, and IDisp so as to expand the scope of application for future versions of HD*Calc.

DOI: <https://doi.org/10.1200/CCI.18.00031>
Published online on ascopubs.org/journal/cci on
September 10, 2018.

AUTHOR CONTRIBUTIONS

Conception and design: All authors

Financial support: Eric J. Feuer

Collection and assembly of data: Sam Harper, Benmei Liu, George Luta

Data analysis and interpretation: Jaeil Ahn, Sam Harper, Mandi Yu, Benmei Liu, George Luta

Manuscript writing: All authors

Final approval of manuscript: All authors

Accountable for all aspects of the work: All authors

Jaeil Ahn

No relationship to disclose

Sam Harper

No relationship to disclose

Mandi Yu

No relationship to disclose

Eric J. Feuer

No relationship to disclose

Benmei Liu

No relationship to disclose

George Luta

No relationship to disclose

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/jco/site/ifc.

ACKNOWLEDGMENT

We thank William Davis, MD, for the initial development of the variance estimation formulas presented in this paper and the reviewers for their insightful comments to improve this manuscript.

Affiliations

Jaeil Ahn and **George Luta**, Georgetown University, Washington, DC; **Sam Harper**, McGill University, Montreal, Quebec, Canada; **Mandi Yu**, **Eric J. Feuer**, and **Benmei Liu**, National Cancer Institute, Bethesda, MD; and **George Luta**, Aarhus University Hospital, Aarhus N, Denmark.

Support

Supported in part by the National Institutes of Health, National Cancer Institute Grant No. HHSN261201500408P.

REFERENCES

1. Center for Disease Control and Prevention: CDC health disparities & inequalities report (CHDIR). <https://www.cdc.gov/minorityhealth/chdireport.html/>
2. National Information Center on Health Services Research and Health Care Technology (NICHSR): Health disparities. <https://www.nlm.nih.gov/hsrinfo/disparities.html>
3. US Department of Health and Human Services: Healthy People 2010. <http://www.healthypeople.gov/2010/>

4. U.S. Department of Health and Human Services: Healthy People 2020. <https://www.healthypeople.gov/>
5. Mackenbach JP, Kunst AE: Measuring the magnitude of socio-economic inequalities in health: An overview of available measures illustrated with two examples from Europe. *Soc Sci Med* 44:757-771, 1997
6. Harper S, Lynch J; National Cancer Institute: NCI Cancer Surveillance Monograph Series, Number 6: Methods for Measuring Cancer Disparities: A Review Using Data Relevant to Healthy People 2010 Cancer-Related Objectives. https://surveillance.cancer.gov/publications/factsheets/SEER_Health_Disparities-1.pdf
7. Breen N, Scott S, Percy-Laurry A, et al: Health disparities calculator: A methodologically rigorous tool for analyzing inequalities in population health. *Am J Public Health* 104:1589-1591, 2014
8. Sen A: *Inequality Reexamined*. Cambridge, MA, Harvard University Press, 1992
9. Keppel K, Percy J, Wagener D: *Trends in Racial and Ethnic-Specific Rates for the Health Status Indicators: United States, 1990-98*. Hyattsville, MD, National Center for Health Statistics, 2002
10. Harper S, King NB, Meersman SC, et al: Implicit value judgments in the measurement of health inequalities. *Milbank Q* 88:4-29, 2010
11. Kakwani N: *Income Inequality and Poverty: Methods of Estimation and Policy Applications*. New York, NY, Oxford University Press, 1980
12. Percy JN, Keppel KG: A summary measure of health disparity. *Public Health Rep* 117:273-280, 2002
13. Theil H: *Economics and Information Theory*. Amsterdam, the Netherlands, North-Holland, 1967
14. Pamuk ER: Social class inequality in mortality from 1921 to 1972 in England and Wales. *Popul Stud (Camb)* 39:17-31, 1985
15. National Cancer Institute: Surveillance, Epidemiology, and End Results program. <https://seer.cancer.gov/>

Range difference (RD): RD is defined as the difference between Y_{max} and Y_{min} (see RD). Instead of the SE calculation for RD, which should account for the correlation between Y_{max} and Y_{min} , HD*Calc provides the SE calculation for the rate difference (RD*) for which the two age-adjusted rates for two prespecified social groups, say Y_j and $Y_{j'}$, are independent of each other. Then the variance of RD* is:

$$Var(RD^*) = Var(Y_j - Y_{j'}) = Var(Y_j) + Var(Y_{j'}) = \sigma_j^2 + \sigma_{j'}^2.$$

So $\widehat{Var}(RD^*) = \hat{\sigma}_j^2 + \hat{\sigma}_{j'}^2$, where unbiased estimators for $\hat{\sigma}_j^2$ and $\hat{\sigma}_{j'}^2$ are given in Background and Notation.

Absolute concentration index (ACI): With its simpler representation, the application of the Taylor series expansion method gives:

$$\frac{\partial ACI}{\partial y_j} = p_j(2Z_j - 1),$$

Then

$$\widehat{Var}(ACI) \approx \sum_{j=1}^J p_j^2 (2Z_j - 1)^2 \hat{\sigma}_j^2,$$

where $\hat{\sigma}_j^2$ are given in Background and Notation.

Slope index of inequality (SII): First, the estimation of β_1 is simply carried out using the weighted least squares solution with the group j specific weight being p_j . Estimation of β_1 can be done as follows:

$$\begin{aligned} \hat{\beta}_1 &= \underset{\beta_1}{\operatorname{argmin}} \sum_{j=1}^J p_j (y_j - \beta_0 - \beta_1 Z_j)^2 \\ &= \frac{\sum_{j=1}^J p_j Z_j (y_j - \hat{\mu})}{\sum_{j=1}^J p_j Z_j^2 - \bar{Z}^2}. \end{aligned}$$

Then,

$$Var(SII) = Var(\hat{\beta}_1) = \frac{\sum_{j=1}^J p_j^2 Z_j^2 \sigma_j^2 + \bar{Z}^2 \sum_{j=1}^J p_j^2 \sigma_j^2 - 2\bar{Z} \sum_{j=1}^J p_j^2 Z_j \sigma_j^2}{(\sum_{j=1}^J p_j Z_j^2 - \bar{Z}^2)^2}.$$

Then, to obtain $Var(SII)$, we replace σ_j^2 with its unbiased estimator $\hat{\sigma}_j^2$, as in Background and Notation.

Between-group variance (BGV): If y is assumed to follow a multivariable normal distribution with mean $\mu = (\mu_1, \dots, \mu_J)^T$ and variance-covariance matrix, Σ , it is known that the variance of a quadratic form of $y = (y_1, \dots, y_J)^T$, on the basis of a symmetric matrix, A , can be expressed as

$$Var(y^T A y) = 2tr(A \Sigma A \Sigma) + 4\mu^T A \Sigma A \mu, \quad (4),$$

where tr denotes the trace of a matrix. Using Equation 4, we can express BGV as

$$BGV = y^T A y$$

$$A = (a_{ij})_{i,j=1}^{J,J}, \text{ where } a_{ij} = (\delta_{ij} p_j - p_i p_j).$$

Here, δ_{ij} is the Kronecker delta function—that is, $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise. If we assume $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_J^2)$ we can compute the quadratic solution from Equation 4. For simplification, let $B = A \Sigma = (b_{ij})_{i,j=1}^{J,J}$, where $b_{ij} = (\delta_{ij} p_j \sigma_j^2 - p_i p_j \sigma_i^2 - p_j p_j \sigma_j^2)$. The first trace term of Equation 4 is computed as

$$\begin{aligned} tr(A \Sigma A \Sigma) &= \sum_{i=1}^J \sum_{j=1}^J b_{ij} b_{ji} \\ &= \sum_{i=1}^J \sum_{j=1}^J (\delta_{ij} p_j \sigma_j^2 - p_i p_j \sigma_i^2) (\delta_{ji} p_i \sigma_i^2 - p_i p_j \sigma_j^2) \\ &= \sum_{i=1}^J p_i^2 \sigma_i^4 + \left(\sum_{i=1}^J p_i^2 \sigma_i^2 \right) \left(\sum_{j=1}^J p_j^2 \sigma_j^2 \right) - 2 \left(\sum_{j=1}^J p_j^3 \sigma_j^4 \right) \end{aligned}$$

$$= \sum_{i=1}^J p_i^2 (1-p_i)^2 \sigma_i^4 + \left(\sum_{i=1}^J p_i^2 \sigma_i^2 \right)^2 - \left(\sum_{i=1}^J p_i^4 \sigma_i^4 \right)$$

$$= \sum_{i=1}^J p_i^2 (1-p_i)^2 \sigma_i^4 + S_2^2 - S_4,$$

where $S_4 = \sum_{i=1}^J p_i^4 \sigma_i^4$ and $S_2 = \sum_{i=1}^J p_i^2 \sigma_i^2$.

By trivial matrix multiplication, it can be shown that

$$\mathbf{C} = \mathbf{A}\Sigma\mathbf{A} = (c_{ij}) = (\delta_{ij} p_i^2 \sigma_i^2 + p_i p_j [S_2 - p_j \sigma_j^2 - p_i \sigma_i^2]).$$

It follows that

$$\boldsymbol{\mu}^T \mathbf{A}\Sigma\mathbf{A}\boldsymbol{\mu} = \sum_{j=1}^J (p_j^2 \sigma_j^2 \mu_j^2 - 2\mu_j p_j p_j^2 \sigma_j^2 + \mu_j^2 p_j^2 \sigma_j^2)$$

$$= \sum_{j=1}^J (\mu_j - \mu)^2 p_j^2 \sigma_j^2.$$

If we replace μ_j , μ with their unbiased estimators (See Background and Notation), we obtain:

$$\widehat{Var}(BGV) \approx 4 \sum_{j=1}^J p_j^2 \hat{\sigma}_j^2 \left(y_j - \sum_{j=1}^J p_j y_j \right)^2$$

$$+ 2 \left[\hat{S}_2^2 - \hat{S}_4 + \sum_{j=1}^J p_j^2 (1-p_j)^2 \hat{\sigma}_j^4 \right],$$

where $\hat{S}_2$ and $\hat{S}_4$ are defined using $\hat{\sigma}_j^2$ instead of σ_j^2 .

Range ratio (RR): Similar to RD, Y_{max} and Y_{min} used for the calculation of RR are not independent. HD*Calc provides the SE calculation for the rate ratio (RR*) for which the two age-adjusted rates for the two prespecified social groups, say Y_j and $Y_{j'}$, are independent of each other. We apply the Taylor series approximation method as introduced in Variance Estimation Using the Taylor Series Approximation Method with $F(y) = y/y_{j'}$.

We have $\frac{\partial F}{\partial y_j} = 1/y_{j'}$ and then $\frac{\partial F}{\partial y_{j'}} = -y_j/(y_{j'})^2$ and

$$Var(RR^*) = Var(Y_j/Y_{j'}) = \left(\frac{\partial F}{\partial y_j} \right)^2 \sigma_j^2 + \left(\frac{\partial F}{\partial y_{j'}} \right)^2 \sigma_{j'}^2$$

$$= \left(\frac{1}{\mu_{j'}} \right)^2 \sigma_j^2 + \left(\frac{\mu_j}{\mu_{j'}^2} \right)^2 \sigma_{j'}^2$$

$$= \frac{\sigma_j^2}{\mu_{j'}^2} + \frac{\mu_j^2 \sigma_{j'}^2}{\mu_{j'}^4}.$$

For $Var(RR^*)$, we replace μ_j , $\mu_{j'}$, σ_j^2 , and $\sigma_{j'}^2$ with their unbiased estimates (Background and Notation).

Index of disparity (IDisp): For simplicity, we assume that y_j is y_{ref} , which corresponds to the healthiest group (ie, y_j is the minimum among all y_j , $j = 1, \dots, J$). Then $|y_j - y_{ref}| = y_j - y_{ref}$. In what follows, we ignore the fact that y_{ref} is the minimum function of $y_1, \dots, y_J$.

We can restate Equation 2 as

$$IDisp = \left[\frac{1}{J-1} \left(\sum_{j=1}^{J-1} y_j/y_J \right) - 1 \right] \times 100.$$

Applying the Taylor series expansion, we have

$$\frac{\partial IDisp}{\partial y_j} = \left[\frac{1}{J-1} \frac{1}{y_j} - 1 \right] \times 100, \text{ for } j < J$$

$$\frac{\partial IDisp}{\partial y_J} = \left[\frac{1}{J-1} \sum_{j=1}^{J-1} -y_j/y_J^2 \right] \times 100 = (IDisp + 100) \times \frac{-1}{y_J}, \text{ for } j = J.$$

Thus,

$$\widehat{Var}(IDisp) = \frac{1}{y_J^2} \left(\frac{1}{(J-1)^2} \sum_{j=1}^{J-1} \hat{\sigma}_j^2 + (IDisp + 100)^2 \hat{\sigma}_J^2 \right).$$

In the current version of HD*Calc, y_{min} is fixed a priori to be y_{ref} , so that the above derivation holds. If y_{ref} is set at a value other than y_{min} , the above derivation is not valid.

Theil's index (T): Applying the Taylor series expansion method, we have

$$\frac{\partial T}{\partial y_j} = \frac{\partial}{\partial y_j}(p_j R_j \ln(R_j)) + \sum_{k \neq j} p_k \frac{\partial}{\partial y_j}(R_k \ln(R_k)).$$

Then, the first term above becomes

$$\frac{\partial}{\partial y_j}(p_j R_j \ln(R_j)) = p_j \left(\frac{\partial R_j}{\partial y_j} \ln(R_j) + \frac{R_j \partial \ln(R_j)}{\partial y_j} \right),$$

where

$$\frac{\partial R_j}{\partial y_j} = \frac{\hat{\mu} - p_j y}{\hat{\mu}^2} = \hat{\mu}^{-1} (1 + p_j R_j)$$

and

$$\frac{\partial (R_j \ln(R_j))}{\partial y_j} = \ln(R_j) \frac{\partial R_j}{\partial y_j} + \frac{\partial R_j}{\partial y_j} = (\ln(R_j) + 1) \hat{\mu}^{-1} (1 + p_j R_j).$$

Similarly, we can show that

$$\frac{\partial}{\partial y_j}(p_k R_k \ln(R_k)) = -\hat{\mu}^{-1} p_j p_k R_k (1 + \ln(R_k)), \quad \text{for } k \neq j$$

Then

$$\frac{\partial T}{\partial y_j} = p_j \hat{\mu}^{-1} \left\{ (1 + \ln(R_j)) - \sum_{k \neq j} p_k R_k (1 + \ln(R_k)) \right\}.$$

We then conclude that

$$\widehat{Var}(T) \approx \sum_{j=1}^J \left(\frac{\partial T}{\partial y_j} \right)^2 \hat{\sigma}_j^2$$

where the unbiased estimator of μ , $\sum p_j y_j$, is used in $\frac{\partial T}{\partial y_j}$.

Mean log deviation (MLD): We employ the Taylor series expansion method and calculate the partial derivatives

$$\frac{\partial MLD}{\partial y_j} = \left(\frac{p_j}{\hat{\mu}} - \frac{p_j}{\hat{\mu}_j} \right).$$

Thus,

$$\widehat{Var}(MLD) \approx \sum_j p_j^2 \hat{\sigma}_j^2 \left(\frac{1}{\hat{\mu}} - \frac{1}{\hat{\mu}_j} \right)^2.$$

Relative index of inequality (RII): Given the estimated SII, $\hat{\beta}_I$, RII can be written as

$$RII = \frac{1}{\sum_{j=1}^J p_j (Z_j - \bar{Z})^2} \left(\frac{\sum_j p_j Z_j Y_j}{\sum_j p_j Y_j} - \sum_j p_j Z_j \right).$$

The variance of RII is computed as

$$Var(RII) = \frac{1}{\left\{ \sum_{j=1}^J p_j (Z_j - \bar{Z})^2 \right\}^2} Var \left(\frac{\sum_j p_j Z_j Y_j}{\sum_j p_j Y_j} \right),$$

where we apply the Taylor series expansion method for the second term,

$$Var \left(\frac{\sum_j p_j Z_j Y_j}{\sum_j p_j Y_j} \right), \text{ and evaluate it at } y_j = \mu_j$$

$$\frac{1}{\partial y_j} \left(\frac{\sum_j p_j Z_j y_j}{\sum_j p_j y_j} \right) = \left\{ \frac{p_j Z_j \hat{\mu} - p_j \sum_j p_j Z_j y_j}{(\sum_j p_j y_j)^2} \right\}$$

We can get an approximate $Var(RII)$ using Equation 1 as

$$\widehat{Var}(RII) \approx \frac{\sum_j (p_j Z_j \hat{\mu} - p_j \sum_j p_j Z_j y_j)^2 \hat{\sigma}_j^2}{\left\{ \hat{\mu}^2 \sum_{j=1}^J p_j (Z_j - \bar{Z})^2 \right\}^2},$$

where $\hat{\sigma}_j^2, \hat{\mu}$ are the unbiased estimators of σ_j^2 and μ , respectively. The alternative asymptotic estimation of the SEs of RII that is using individual-level data are described (Kakwani N, et al: J Econom 77:87-103, 1997).

Kunst-Mackenbach relative index (KMI): When $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{Z}$ is plugged in, KMI can be rewritten as

$$KMI = 1 - \frac{\hat{\beta}_1}{\hat{\mu} - \hat{\beta}_1(\bar{Z} - 1)}.$$

Using the Taylor series expansion method and the result regarding $\frac{\partial \hat{\beta}_1}{\partial y_j}$ from the calculation for SII, we get

$$\begin{aligned} \frac{\partial KMI}{\partial y_j} &= \frac{\left\{ \hat{\mu} - \hat{\beta}_1(\bar{Z} - 1) \right\} \frac{\partial \hat{\beta}_1}{\partial y_j} - \hat{\beta}_1 \frac{\partial}{\partial y_j} \left\{ \hat{\mu} - \hat{\beta}_1(\bar{Z} - 1) \right\}}{\left\{ \hat{\mu} - \hat{\beta}_1(\bar{Z} - 1) \right\}^2} \\ &= \frac{\hat{\beta}_1 \frac{\partial \mu}{\partial y_j} - \hat{\mu} \frac{\partial \hat{\beta}_1}{\partial y_j}}{(\hat{\beta}_0 + \hat{\beta}_1)^2} \\ &= \frac{p_j \hat{\beta}_1}{(\hat{\beta}_0 + \hat{\beta}_1)^2} - \frac{p_j (Z_j - \bar{Z}) \hat{\mu}}{(\hat{\beta}_0 + \hat{\beta}_1)^2 \left\{ \sum_{j=1}^J p_j (Z_j - \bar{Z})^2 \right\}^2}. \end{aligned}$$

We plug in $\frac{\partial KMI}{\partial y_j}$ in Equation 1 to complete the computation of $Var(KMI)$;

$$\widehat{Var}(KMI) \approx \sum_{j=1}^J \left[\frac{p_j \hat{\beta}_1}{(\hat{\beta}_0 + \hat{\beta}_1)^2} - \frac{p_j (Z_j - \bar{Z}) \hat{\mu}}{(\hat{\beta}_0 + \hat{\beta}_1)^2 \left\{ \sum_{j=1}^J p_j (Z_j - \bar{Z})^2 \right\}^2} \right]^2 \hat{\sigma}_j^2.$$

As an alternative approach, researchers used Fieller's method to calculate the variance of KMI because $\hat{\beta}_1$ and $\bar{Z}$ are uncorrelated in the least squares estimation (Hayes LJ, et al: J Epidemiol Community Health 56:762-765, 2002). They showed that the two methods are empirically identical through numerical studies.

Relative concentration index (RCI): If we apply the Taylor expansion method,

$$\frac{\partial RCI}{\partial y_j} = 2 p_j \frac{Z_j \hat{\mu} - p_j \sum_k p_k y_k Z_k}{\hat{\mu}^2},$$

the estimated variance of RCI is computed as

$$\widehat{Var}(RCI) \approx \sum_{j=1}^J 4 \frac{p_j^2}{\hat{\mu}^4} (Z_j \hat{\mu} - p_j \sum_{k=1}^J p_k y_k Z_k)^2 \hat{\sigma}_j^2.$$

Table A1. Complete SEER Data Characteristics for Cervical, Lung, Breast, and Female Brain Cancer From Nationwide; 12 States, Excluding Alaska; and King County (Washington state)

Cancer by State*	Group	Min AR†	Max AR†	Min Var*	Max Var*	Average Count	Proportion Zeros	Population Size
Cervical								
California	2.0	69.79	98.3	1.06	1.96	191.92	0.16	93,802,864
Connecticut	2.0	55.43	123.69	2.634	12.186	15.82	0.24	9,167,187
Georgia	2.0	79.63	109.98	1.860	10.77	53.05	0.26	24,826,892
Hawaii	2.0	77.52	78.36	5.112	20.138	7.00	0.39	3,394,788
Iowa	2.0	68.83	124.21	3.174	27.018	14.00	0.37	7,693,357
Kentucky	2.0	87.50	99.50	2.868	25.245	26.21	0.37	11,028,460
Louisiana	2.0	92.93	94.72	2.942	16.359	28.39	0.26	11,570,363
Michigan	2.0	73.84	132.45	2.753	22.605	20.89	0.32	10,002,307
New Jersey	2.0	75.21	106.82	1.939	5.844	51.55	0.24	22,552,663
New Mexico	2.0	70.03	82.18	5.052	6.274	10.32	0.26	5,195,854
Utah	2.0	49.27	79.85	3.071	13.135	8.21	0.32	6,881,752
Washington	2.0	67.98	100.04	2.494	14.606	22.11	0.26	11,548,466
Lung								
California	10.0	443.56	593.29	2.937	11.806	478.1	0.15	186,567,726
Connecticut	10.0	NA	NA	NA	NA	77.56	0.74	17,867,309
Georgia	10.0	564.29	802.48	9.421	26.973	157.02	0.26	48,564,764
Hawaii	10.0	NA	NA	NA	NA	22.42	0.77	6,809,648
Iowa	10.0	NA	NA	NA	NA	69.97	0.43	15,239,059
Kentucky	10.0	810.21	1028.21	10.875	115.448	113.38	0.34	21,701,249
Louisiana	10.0	NA	NA	NA	NA	84.5	0.4	22,649,957
Michigan	10.0	NA	NA	NA	NA	96.32	0.73	19,356,328
New Jersey	10.0	NA	NA	NA	NA	172.53	0.51	43,973,787
New Mexico	10.0	NA	NA	NA	NA	23.78	0.49	10,273,905
Utah	10.0	NA	NA	NA	NA	17.77	0.53	13,830,529

(Continued on following page)

Table A1. Complete SEER Data Characteristics for Cervical, Lung, Breast, and Female Brain Cancer From Nationwide; 12 States, Excluding Alaska; and King County (Washington state) (Continued)

Cancer by State*	Group	Min AR†	Max AR†	Min Var*	Max Var*	Average Count	Proportion Zeros	Population Size
Washington	10.0	NA	NA	NA	NA	85.32	0.68	23,020,659
Breast								
California	5.0	890.44	1,404.53	5.310	41.249	1264.51	0.23	93,802,864
Connecticut	5.0	761.19	1,410.66	12.975	197.502	159.72	0.33	9,167,187
Georgia	5.0	518.15	1,258.01	8.717	107.557	336.46	0.31	24,826,892
Hawaii	5.0	1,253.48	2,114.93	21.631	656.001	54.87	0.36	3,394,788
Iowa	5.0	663.25	1,243.27	12.316	240.512	119.64	0.38	7,693,357
Kentucky	5.0	136.58	1,332.25	10.325	75.159	165.15	0.43	11,028,460
Louisiana	5.0	289.39	1300.00	12.065	67.000	164.68	0.32	11,570,363
Michigan	5.0	677.59	1,303.37	12.807	126.636	159.28	0.33	10,002,307
New Jersey	5.0	489.23	1,406.67	9.063	103.667	364.27	0.31	22,552,663
New Mexico	5.0	534.88	1,256.02	20.933	114.814	69.58	0.29	5,195,854
Utah	5.0	490.42	1,159.40	15.252	181.984	68.97	0.33	6,881,752
Washington	5.0	952.24	1,820.16	12.001	122.998	189.88	0.29	11,548,466
Female brain								
California	5.0	29.91	58.15	1.161	10.359	47.69	0.07	93,802,864
Connecticut	5.0	0	58.68	0	7.851	5.97	0.46	9,167,187
Georgia	5.0	0	57.23	0	7.198	12.74	0.28	24,826,892
Hawaii	5.0	0	63.76	0	11.547	1.24	0.59	3,394,788
Iowa	5.0	12.96	74.4	2.841	52.613	5.51	0.66	7,693,357
Kentucky	5.0	0	64.28	0	22.099	7.67	0.54	11,028,460
Louisiana	5.0	14.09	55.41	2.724	14.089	6.11	0.44	11,570,363
Michigan	5.0	33.74	64.92	2.817	49.025	6.02	0.40	10,002,307
New Jersey	5.0	0	59.95	0	4.402	13.71	0.24	22,552,663
New Mexico	5.0	10.39	52.31	4.066	14.469	2.52	0.49	5,195,854

(Continued on following page)

Table A1. Complete SEER Data Characteristics for Cervical, Lung, Breast, and Female Brain Cancer From Nationwide; 12 States, Excluding Alaska; and King County (Washington state) (Continued)

Cancer by State*	Group	Min AR†	Max AR†	Min Var*	Max Var*	Average Count	Proportion Zeros	Population Size
Utah	5.0	6.58	59.42	3.374	14.599	3.78	0.58	6,881,752
Washington	5.0	15.40	66.84	2.717	13.736	7.92	0.35	11,548,466
Breast								
Nationwide	5.0	906.45	1,340.42	3.016	23.915	3,120.4	0.17	217,963,069
Cervical								
Nationwide	2.0	73.19	98.72	0.649	1.675	450.18	0.13	217,963,069
Lung								
Nationwide	10.0	478.93	755.57	2.526	5.923	1,347.72	0.09	430,457,544
Breast								
King county	5.0	982.413	1,664.028	20.100	225.407	78.41	0.32	4,865,310
Cervical								
King county	2.0	60.30	104.81	3.540	24.400	8.58	0.34	4,865,310

NOTE. Included are the number of social groups, minimum (Min) and maximum (Max) of age-adjusted rates (AR), minimum and maximum variance (Var) of AR, average event count per social group, proportion of zero event counts ($x_k = 0$), and population size across the $J \times K$ age group by social group combinations.

*Some states among 13 states per each cancer are not shown because of the table size limit.

†Values are multiplied by 10^6 .

Table A2. False-Discovery Rate Adjusted *P* Values From the Unadjusted Simple Linear Regression Models That Regressed the Bias of the Coverage Probabilities of the 11 Healthy Disparity Measures on Each of the Seven Predictors, Using the Analytic and MCS Methods

Method and Measure	No. of Social Groups	Min AR	Max AR	Min VA	Max VA	Prop_Zero	Avg_A
RD	0.585	0.260	0.260	0.530	0.530	0.012	0.530
BGV	0.695	0.203	0.203	0.402	0.402	0.021	0.532
ACI	0.617	0.242	0.242	0.524	0.524	0.014	0.525
SII	0.617	0.242	0.242	0.524	0.524	0.014	0.525
Analytics methods							
RR	0.565	0.284	0.284	0.543	0.543	0.016	0.553
IDisp	0.803	0.179	0.179	0.451	0.455	0.02	0.455
MLD	0.803	0.163	0.163	0.387	0.387	0.022	0.487
T	0.811	0.144	0.144	0.371	0.371	0.017	0.469
RCI	0.619	0.241	0.241	0.524	0.524	0.013	0.524
RII	0.619	0.241	0.241	0.524	0.524	0.013	0.524
KMI	0.748	0.058	0.058	0.227	0.304	0.015	0.334
RD	0.791	0.199	0.199	0.476	0.476	0.009	0.480
BGV	0.728	0.284	0.284	0.554	0.639	0.006	0.554
ACI	0.554	0.284	0.284	0.547	0.547	0.014	0.547
SII	0.558	0.280	0.280	0.546	0.546	0.014	0.546
RR	0.773	0.215	0.215	0.492	0.492	0.010	0.492
MCS							
IDisp	0.608	0.230	0.230	0.468	0.468	0.011	0.474
MLD	0.686	0.317	0.317	0.620	0.686	0.005	0.620
T	0.709	0.304	0.304	0.584	0.681	0.006	0.584
RCI	0.559	0.279	0.279	0.532	0.532	0.014	0.532
RII	0.559	0.279	0.279	0.532	0.532	0.014	0.532
KMI	0.505	0.325	0.325	0.505	0.505	0.016	0.505

Abbreviations: ACI, absolute concentration index; Avg_A, average event count per social group; BGV, between-group variance; IDisp, index of disparity; KMI, Kunst-Mackenbach index; Max/Min AR, minimum and maximum of age-adjusted rates; Max/Min VA, minimum and maximum variance of age-adjusted rates; MCS, Monte Carlo simulation; MLD, mean log deviation; Prop_Zero, proportion of zero event counts across the age group by social group combinations; RCI, relative concentration index; RD, range difference; RII, relative index of inequality; RR, range ratio; SII, slope index of inequality; T, Theil index.