Copula-based estimation of health inequality measures*

Taoufik Bouezmarni† Mohamed Doukali‡

Université de Sherbrooke University of Liverpool

Abderrahim Taamouti§

University of Liverpool

November 8, 2024

*The authors would like to thank very much Bianca De Stavola, Serena Ng, an Associate Editor, three anonymous reviewers, and the participants at the 2022 CISEM International Symposium on Statistics and Econometrics, 2022 Africa Meeting of the Econometric Society, 7th African International Conference (AIC) on Statistic in Marrakech, 2023, University of Malaga, and INSEA for their very useful comments.

†Département de Mathématiques, Université de Sherbrooke, CIREQ, CREAS. Email addresses: Taoufik.Bouezmarni@USherbrooke.ca

‡ *Corresponding author.* Department of Economics, University of Liverpool Management School. Address: Chatham St, Liverpool L69 7ZH. E-mail: m.doukali@liverpool.ac.uk

§Department of Economics, University of Liverpool Management School. Address: Chatham St, Liverpool L69 7ZH. E-mail: Abderrahim.Taamouti@liverpool.ac.uk

Copula-based estimation of health inequality measures

ABSTRACT

This paper aims to utilize bivariate copulas for deriving estimators of the health concentration curve and Gini coefficient for health distribution. We highlight the importance of expressing health inequality measures in terms of bivariate copulas, which we in turn use to build copula-based semi and nonparametric estimators of the above measures. Subsequently, we investigate the asymptotic properties of these estimators, establishing their consistency and asymptotic normality. We provide formulas for their variances, facilitating the construction of confidence intervals and tests for the health concentration curve and Gini health coefficient with respect to a given socioeconomic variable. Through a Monte-Carlo simulation exercise, we demonstrate the superior performance of the semiparametric estimator over the smoothed nonparametric estimator, with the latter outperforming the empirical estimator in terms of Mean Squared Error. Additionally, an extensive empirical study applies our estimators, revealing that inequalities in U.S. states' socioeconomic variables, such as income/poverty and race, contribute to observed disparities in COVID-19 infections and deaths in the U.S.

Keywords: Health concentration curve, Gini health coefficient, inequality, copula, semi and nonparametric estimators, COVID-19 infections and deaths.

Journal of Economic Literature classification: C13, C14, I14.

1 Introduction

Measures of health inequalities, such as the Health Concentration Curve (hereafter CH) and the Gini coefficient with respect to a given socioeconomic variable, hold significant relevance for both healthcare professionals and policymakers. These metrics play a crucial role in identifying socioeconomic classes that face heightened risks of contracting and succumbing to specific diseases. Consequently, they serve as invaluable tools for formulating social and economic interventions aimed at safeguarding vulnerable populations. In this paper, our goal is to develop estimators for health inequality measures that leverage information regarding the dependence between health variables (e.g., COVID-19 infection, overweightness, cancer, etc.) and socioeconomic factors (e.g., income level, education level, race, etc.). This approach is designed to enhance the accuracy of estimating these measures.

The health concentration curve, with respect to a given socioeconomic variable, graphically represents the cumulative percentage of the health variable against the cumulative percentage of the population ranked according to that socioeconomic variable; see Yitzhaki, S. and Olkin (1991) and Yitzhaki and Schechtman (2013) among others. This ranking ranges from the poorest to the richest individuals. The curve illustrates the effect of socioeconomic variable on the concentration of health variable. Formally, let H represent a non-negative health variable of interest, and Y denote a non-negative socioeconomic variable, such as income. Consider the joint density function, f, for the vector (H, Y). The marginal densities of H and Y are denoted as f_H and f_Y, respectively. Additionally, $f_{H|Y}$ represents the conditional density of H given Y, while F_H and F_Y represent the marginal distributions of H and Y, respectively. For any p in the interval $(0, 1)$, the health concentration curve with respect to the socioeconomic variable Y is defined as:

$$CH(p, Y) = \frac{\int_0^p \mathbb{E}\left[H \mid Y = F_Y^{-1}(u)\right] du}{\int_0^1 \mathbb{E}\left[H \mid Y = F_Y^{-1}(u)\right] du}, \tag{1}$$

where, for $u \in (0, 1)$, $F_Y^{-1}(u) = inf\{t : u \leq F_Y(t)\}$ is the quantile function of Y. The values of $F_Y^{-1}(u)$ at $u = 0, 1$ can be set to arbitrary finite real numbers. For example, if Y is the income and H is the rate of COVID-19 mortality, $CH(p, Y)$ is the percentage of the cumulative rate of COVID-19 mortality rate of the $(100 * p)\,\%$ of the poorest population.

In this paper, we initially demonstrate that the measure of health inequality in (1) can be represented in terms of the bivariate copula function of the vector (H, Y). By exploring the information regarding the dependence between H and Y through the copula-based representation of the health concentration curve with respect to Y, we aim to develop more accurate estimators that lower Mean Square Error. As we illustrate in Section (2), when two health variables, say H_1 and H_2, exhibit similar marginal distributions, we find that the presence of health inequality depends on the nature of the dependence between Y and H_i, for $i = 1, 2$. This dependence can be captured by bivariate copula functions. Copulas are widely employed to model the dependence between variables of interest. Thanks to Sklar (1959), the copula function is directly associated with the joint distribution function. In simpler terms, the joint distribution function can be expressed in terms of marginal distributions that give information on each variable of interest, and the copula that captures the dependence between all the variables under study.

The utility of copulas extends beyond this paper, with a growing body of literature introducing novel estimation methods that leverage the advantages of copulas to improve the estimation of mean, quantile, and distributional regression functions; see Sungur (2005), Crane and Hoek (2008), Noh et al. (2013), Noh et al. (2015), Rémillard et al. (2017), Klein and Kneib (2016), Hans (2023) among others. While mean and quantile regression copulas focus on specific characteristics of the conditional distributions, distributional copulas offer a more comprehensive framework by modeling the entire conditional distribution of a response variable given a set of predictors. This distinction is critical, as it impacts the interpretation of dependence structures and the choice of statistical methods used in empirical applications. For instance, Kneib (2013) argues for moving beyond mean regression to capture more complex relationships, while Dorn et al. (2024) illustrate the importance of using distributional copula regression for understanding intersectional poverty by examining leisure time and income simultaneously. Moreover, Klein et al. (2016) demonstrate the use of structured additive conditional copula regression models, providing a Bayesian framework that integrates both mean and distributional aspects. Thus, incorporating literature on distributional copula models not only enhances methodological rigor but also broadens the analytical perspective in understanding dependencies across dimensions of many socio-economic factors such as well-being (Decancq, 2014; Klein et al., 2019). In this paper, we present estimators for the health concentration curve and Gini coefficient based on semi-parametric and nonparametric copulas.

Thereafter, we utilize the copula-based re-formulation of CH in (1) to construct semi- and nonparametric estimators for health concentration curves and the Gini coefficient, with respect to Y, of health distribution [hereafter referred to as the Gini health coefficient]. These estimators are specifically designed to quantify health inequalities. To derive the semiparametric estimator, a parametric bivariate copula is employed to model the dependence between health and socioeconomic variables. The parameter of the bivariate copula function is estimated using the maximum pseudo-likelihood estimation method, wherein the cumulative distribution of the health variable is substituted with its empirical estimator. Once the maximum likelihood estimator of the bivariate copula's parameter is acquired, it is utilized to calculate the semiparametric estimator of the health concentration curve with respect to Y. For the nonparametric estimator, we substitute the bivariate copula function with its smoothed nonparametric estimator, specifically a Bernstein copula; see Bouezmarni et al. (2010), Chen and Huang (2007), and Bouezmarni et al. (2012). The Bernstein estimator proves advantageous in avoiding potential misspecification issues encountered by the semiparametric estimator. Additionally, it remains unaffected by the boundary bias problem because, through smoothing with beta densities, the Bernstein copula does not assign weight outside its support. Furthermore, we employ the aforementioned estimators of health concentration curves with respect to Y to derive both semiparametric and nonparametric estimators of the Gini health coefficient. It is worth mentioning that a limitation of the Bernstein estimator used in this paper is its reliance on a bandwidth parameter. Selecting an optimal bandwidth parameter is not straightforward. However, we address this issue by adopting the rule of thumb for practical bandwidth parameters in our simulations and applications. Cross-validation can be employed to determine a practical value for the bandwidth parameter.

Furthermore, we investigate the asymptotic properties of the aforementioned estimators. We establish their consistency and asymptotic normality. We also provide expressions for their variances, facilitating the construction of confidence intervals and the development of tests for both the health concentration curve and the Gini health coefficient with respect to Y. A Monte Carlo simulation exercise, conducted using various data-generating processes and sample sizes, reveals that the semiparametric estimator surpasses the smoothed nonparametric estimator. Moreover, the latter performs better than the empirical estimator in terms of Integrated Mean Squared Error.

Additionally, we conduct a comprehensive empirical analysis to investigate disparities in COVID-19 infections and deaths both across and within U.S. states using county-level data. Our study uses a dataset covering 2,777 counties from 40 states, integrating health data on confirmed COVID-19 cases and deaths with key socioeconomic indicators, including income, race, poverty rate, and urban/rural status. The data are sourced from the U.S. Centers for Disease Control and Prevention (CDC) for COVID-19 statistics, and from the U.S. Census Bureau for demographic and socioeconomic information. To assess these disparities, we employ semi- and nonparametric estimators for the Health Concentration and Gini health coefficients. The results reveal substantial inequalities in COVID-19 outcomes, with infection and mortality rates varying significantly based on socioeconomic conditions, such as poverty and race, particularly in the most disadvantaged areas. This empirical approach offers valuable insights into how pre-existing socioeconomic inequalities have exacerbated the pandemic's effects.

The rest of the paper is organized as follows. In Section 2, we present some examples to illustrate the importance of expressing health inequality measures in terms of bivariate copulas. In Section 3, we introduce the notations, re-formulate the health concentration curve with respect to a socioeconomic variable as a function of bivariate copula, and derive its expression for specific copula functions. In Section 4, we formulate semiparametric and nonparametric estimators of the health concentration curve with respect to a socioeconomic variable using both parametric and nonparametric bivariate copulas. These estimators are subsequently employed to derive estimates for the Gini health coefficient. In Section 5, we explore the asymptotic properties of both the semi- and nonparametric estimators. In Section 6, we conduct Monte Carlo simulations to evaluate the performance of both the semi- and nonparametric estimators, comparing them with the empirical estimator of the health concentration curve with respect to a socioeconomic variable. Section 7 presents an extensive empirical study where our estimators are applied to analyze health inequality in U.S. COVID-19 data. Section 8 provides the conclusion. The proofs, simulation details, and empirical results are available in a separate companion appendix.

2 Motivation

This section aims to motivate the importance of expressing health inequality measures in terms of copula. This helps illustrate the roles played by the marginal distribution of the health variable, H, and the dependence structure between this variable and the socioeconomic variable, Y, in estimating health inequality measures. Through examples, we demonstrate that relying on mean regressions, as some papers do, to study health inequality may lead to misleading conclusions. For instance, when Y and H exhibit linear dependence, health inequality may or may not be present, contingent on the shape of the marginal distribution of H. Moreover, when two health variables exhibit similar marginal distributions, the existence of health inequality depends on the nature of the dependence between the socioeconomic variable and the health variable. We illustrate these points through the following examples.

Example 1. *Suppose that the health variable H is generated according to the following process:*

$$H_i = 3 + 2Y_i + \varepsilon_i, \text{ for } i = 1...n,$$

*where n is the sample size, ε_i is an error term that follows a normal distribution $N(0, 0.2)$, and the socioeconomic variable Y follows either: (i) a Weibull distribution $Weibull(3, 1)$ [hereafter **Case 1**] or (ii) a mixture of Weibull distributions $s_t * Weibull(2, 1) + (1 - s_t) * Weibull(15, 1)$, with $Pr(s_t = 0) = Pr(s_t = 1) = 1/2$ [hereafter **Case 2**]. The first and the second values in the above notations of the Weibull distribution, say $Weibull(\alpha, \beta)$, represent the values of the shape α and scale β parameters, respectively.*

Using the data in Example 1 [see Figure 1 of the separate companion appendix], we calculated the health concentration curves with respect to Y for **Case 1** and **Case 2** using formula (1), and the results are reported in Figure 2 of the appendix. Panel (a) of Figure 1 and curve (a) of Figure 2 show that when Y and H are linearly dependent, health inequality might not exist if the marginal distribution of H is unimodal and the data are well concentrated around the mean. However, under the same linear dependence between Y and H, we find that when the marginal distribution of H is bimodal as shown in Panel (b) of Figure 1, then there is a health inequality as illustrated by the curve (b) of Figure 2.

Example 1 illustrates that the use of linear regressions may not provide informative insights into the presence or absence of health inequalities. For instance, an estimator of the health concentration curve with respect to Y, relying on a parametric estimator of the marginal distribution of H, could yield misleading results if the assumed parametric distribution is misspecified [e.g., when the true distribution of H is a mixture of Weibull distributions, as seen in **Case 2**, but the assumed distribution is a Weibull, as in **Case 1**], as depicted in Figures 1 and 2.

Example 2. *Suppose that the health variable H and the socioeconomic variable Y*

*follow the Weibull distributions $Weibull(1, 2)$ and $Weibull(5, 1)$, respectively, and consider a Guassian bivariate copula for the dependence between Y and H with two different correlation coefficients: (i) $\rho = 0$ [hereafter **Case 1**] and (ii) $\rho = 0.9$ [hereafter **Case 2**].*

In line with Example 1, we employ the data generated in Example 2 to compute health concentration curves for both Case 1 and Case 2, with the results detailed in Figures 3 and 4 found in the companion appendix. As expected, the nature of the dependence between Y and H proves instrumental in indicating the presence or absence of health inequality. On the one hand, as illustrated in Panel (a) of Figure 3 and curve (a) of Figure 4, the absence of dependence between Y and H corresponds to the absence of health inequality. On the other hand, for the same marginal distribution of H featured in Panel (a) of Figure 3, Panel (b) of Figure 3, and curve (b) of Figure 4 reveal that health inequality emerges when Y and H exhibit dependence (non-linearly dependent).

Example 2 illustrates the importance of having accurate estimators for both the marginal distribution of H and the dependence between Y and H in order to obtain a reliable estimator of the health concentration curve with respect to Y. This is why the focus of this paper is on enhancing the estimation of the health concentration curve with respect to Y by providing robust estimators of the bivariate copula function that captures the dependence between Y and H, in addition to utilizing a nonparametric estimator of the marginal distribution of H.

3 Copula-based health concentration curve

The health concentration curve plots the cumulative percentage of the health variable against the cumulative percentage of the population, ranked by socioeconomic covariates such as living standards, beginning with the poorest and ending with the richest. The curve illustrates the effect of a socioeconomic variable on the concentration of a health variable such as COVID-19 infections and deaths.

Formally, the health concentration curve is given by formula (1). From this formula and as we discussed in Section (2), we note that the dependence between H and Y plays a crucial role in the calculation of health concentration curve with respect to Y $CH(p,Y)$. For instance, if H and Y are independent, then $CH(p,Y)=p$, i.e., we are in the presence of perfect equality of health variable H across the socioeconomic variable Y. However, this does not mean a perfect equality of the health concentration. As shown in the following proposition, the concentration health curve with respect to Y can be derived through the dependence structure of the vector (H,Y) using bivariate copula function [see the proof of Proposition 1 in the appendix of this paper].

Proposition 1 *Let C and c be the bivariate copula function and the copula density function of the vector (H,Y), respectively. The health concentration curve with respect to Y $CH(p,Y)$ in Equation (1) can be rewritten as a function of bivariate copula:*

$$CH(p,Y)=\frac{\int_0^1 F_H^{-1}(u)C_u(u,p)du}{\mathbb{E}(H)},\ \ \textit{for } p\in(0,1)\,, \tag{2}$$

where $C_u(u,p)$ is the partial derivative of the bivariate copula function of the vector (H,Y) :

$$C_u(u,v)=\frac{\partial C(u,v)}{\partial u}=\int_0^v c(u,z)dz.$$

The result in (2) is a copula-based representation of CH. The formula (2) offers a breakdown of the health concentration curve (CH) into two components: the marginal information about H and the dependence between Y and H through the bivariate copula. It's important to note that Y is only utilized for the copula, and any increasing transformation of Y will not affect the copula or the CH curve.We can now use this representation to derive expressions of the health concentration curve with respect to Y for some specific copula functions. We consider independence and lower and upper Fréchet–Hoeffding dependence structures. As shown below, the latter dependence structures can help illustrate the link between the concentration curve and the Lorenz curve of the health variable H with respect to Y.

Example 1 (Independence): *If H and Y are independent, then $C_u(u,v)=v$. Therefore,*

$CH(p,Y) = p$, and in this case we obtain a perfect equality, i.e. the socioeconomic variable has no effect on the concentration of the health variable.

Example 2 (Lower Fréchet–Hoeffding): Suppose that H and Y are countermonotonic variables, i.e., the dependence between the two variables can be modelled using Fréchet–Hoeffding lower bound copula $W(u,v) = \max\{u+v-1, 0\}$. In this case, $W_u(u,p) = \mathbb{I}(u > 1-p)$, where $\mathbb{I}(.)$ is an indicator function that takes value one if $u > 1-p$ and zero otherwise. Hence,

$$CH(p,Y) = \frac{\int_0^1 F_H^{-1}(u) W_u(u,p)du}{\mathbb{E}(H)} = \frac{\int_{1-p}^1 F_H^{-1}(u)du}{\mathbb{E}(H)} = 1 - L_H(1-p),$$

where L_H is the Lorenz curve of H. For example, if H represents overweightness, then $CH(p,Y)$ is the percentage of the total overweightness of the $(100 * p)\,\%$ of the richest population.

Example 3 (Upper Fréchet–Hoeffding): Suppose that H and Y are comonotone variables, i.e., the dependence between the two variables can be modelled using Fréchet–Hoeffding upper bound copula $M(u,v) = \min\{u,v\}$. In this case, $M_u(u,p) = \mathbb{I}(u < p)$, where $\mathbb{I}(.)$ is an indicator function that takes value one if $u < p$ and zero otherwise. Hence,

$$CH(p,Y) = \frac{\int_0^1 F_H^{-1}(u) M_u(u,p)du}{\mathbb{E}(H)} = \frac{\int_0^p F_H^{-1}(u)du}{\mathbb{E}(H)} = L_H(p),$$

For example, if H represents the payments people made for health care and if we assume a perfect positive correlation between Y and H, then $CH(p,Y)$ is the percentage of the total payments spent by the $(100 * p)\,\%$ of the poorest population.

4 Copula-based estimation

The copula-based representation in Equation (2) will be used to derive semiparametric and nonparametric estimators of health concentration curve and Gini health coefficient with respect to a socioeconomic variable. For the semiparametric estimator, we use a parametric bivariate copula function to model the dependence between health and socioeconomic variables. The bivariate copula's parameter is estimated using the maximum pseudo-likelihood estimator after replacing the cumulative distribution of health variable by its empirical analogue. For the nonparametric estimator, we replace the bivariate copula function by a Bernstein copula estimator. Furthermore, we use the above estimators to derive semiparametric and nonparametric estimators of Gini health coefficient.

Let us first set some notations. We denote by $\{(H_i, Y_i), i = 1, \ldots, n\}$ an independent and identically distributed sample of n copies of the vector (H, Y). From the representation in (2), the health concentration curve with respect to Y $CH(p, Y)$ can be estimated using different approaches that correspond to different ways of estimating the expectation and distribution of H, say $\mathbb{E}(H)$ and F_H, but also to different ways of estimating bivariate copula function $C_u(u, p)$. In the following, $\mathbb{E}(H)$ will be estimated using its empirical analogue (sample average of H), and F_H will be estimated using the rescaled empirical distribution:

$$\widehat{F}_H(h) = \frac{1}{n+1} \sum_{i=1}^{n} \mathbb{I}(H_i \leq h), \tag{3}$$

where $\mathbb{I}(.)$ is an indicator function that takes the value one if $H_i \leq h$ and zero otherwise. Note that alternative estimators of F_H - such as a parametric estimator or a smoothed nonparametric estimator like Kernel, Bernstein, etc.- can be considered.

In the next two sub-sections, we consider two different estimators of bivariate copula $C_u(u, p)$. In Sub-section 4.1, we introduce a semiparametric estimator of $CH(p, Y)$ in which only realizations of Y in $C_u(u, p)$ will be used for estimating the bivariate copula function, i.e., the information on the order of Y is required for modelling the dependence structure of the vector (H, Y). In Sub-section 4.2, we derive a nonparametric estimator of $CH(p, Y)$ using a nonparametric estimator of bivariate copula $C_u(u, p)$; hereafter Bernstein estimator.

4.1 Semiparametric estimation

We consider that the bivariate copula C of (H, Y) belongs to a parametric family $\{C(.,.,\theta), \theta \in \Theta\}$, with an unknown vector of parameters θ that is in the set Θ, which is a compact subset of $\mathbb{R}^q$. Denote by θ_0 the true value of θ. There exist several estimators of θ_0 and a most popular one is given by the following maximum pseudo likelihood estimator, called Canonical Maximum Likelihood method:

$$\hat{\theta}_n = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log c\left(\widehat{F}_H(H_i), \widehat{F}_Y(Y_i), \theta\right), \tag{4}$$

where c is the bivariate copula density of C, $\widehat{F}_H(h)$ is defined in Equation (3) and $\widehat{F}_Y(y) = (n+1)^{-1} \sum_{i=1}^{n} \mathbb{I}(Y_i \leq y)$. Note that $\widehat{\theta}_n$ is the estimator proposed by Shih and Louis (1995) and Genest, Ghoudi and Rivest (1995), and its asymptotic representation can straightforwardly be obtained from the proof of Theorem 1 in Tsukuhara (2005).

Using the maximum likelihood estimator of θ_0 in (4) and the nonparametric estimator of F_H in (3), a semiparametric estimator of $CH(p, Y)$ is defined as follows:

$$\widehat{CH}(p, Y) = \frac{\sum_{i=1}^{n} H_i\, C_u(\widehat{F}_H(H_i), p, \widehat{\theta}_n)}{\sum_{i=1}^{n} H_i}, \tag{5}$$

where $C_u(u, v, \theta) = \frac{\partial C(u,v,\theta)}{\partial u}$.

As an alternative to the canonical maximum likelihood approach, we can employ the inference functions for margins method; see Joe and Xu (1996). This method assumes a parametric model for the marginal distributions. However, to circumvent the need for specifying the marginal distributions, we opt for semi-parametric estimation. It's important to note that the selection of parametric distributions plays a critical role in determining the CH curve, as demonstrated in Example 1.

We next use simulations based on examples 1-3 of Section 3 to empirically illustrate the semiparametric estimator in (5). We also compare this estimator with the following empirical estimator of $CH(p, Y)$ [see Wagstaff (2002)],

$$\widehat{CH}_n(p, Y) = \frac{\sum_{i=1}^{n} H_i \; \mathbb{I}(Y_i \leq \widehat{F}_Y^{-1}(p))}{\sum_{i=1}^{n} H_i}, \tag{6}$$

where $\mathbb{I}(.)$ is an indicator function that takes the value one if $Y_i \leq \widehat{F}_Y^{-1}(p)$ and zero otherwise. The simulation results are reported in Figure 5 of the separate companion appendix, which we obtain after generating $n = 100$ observations of (H_i, Y_i) from a Gaussian bivariate copula with correlation coefficients $\rho = 0$ (Example 1), $\rho = -0.99$ (Example 2) and $\rho = 0.99$ (Example 3), and using as a marginal distribution of H and Y an exponential distribution with a parameter $\lambda = 1$. To avoid numerical problems, we excluded the cases $\rho = \pm 1$. Thereafter, we report the results for both known and unknown bivariate copulas. The unknown copula case corresponds to the situation where the copula under consideration (Gaussian bivariate copula in this case) is considered to be unknown, and we use the R package `VineCopula` [see **Nagler et al. (2023)**] - available online for free - to select a copula function that fits better our generated data [see curves of the unknown copula in Figure 5]. Furthermore, Figure 5 illustrates the estimators of $1 - L_H(1 - p)$ and $L_H(p)$, where L_H is the Lorenz curve of H; see subfigures (b) and (c) of Figure 5.

Subfigure (a) of Figure 5 shows that the copula-based estimator of $CH(p, Y)$ is much closer to the 45-degree line (true curve of $CH(p, Y)$ under independence) than the empirical estimator, and that the former is much smoother than the latter. This result holds for both known and unknown copulas, which might suggest a high performance in favour of the semiparametric estimator. We also see that the estimators of $CH(p, Y)$ using known and unknown copulas are very similar, which might indicate that the R package `VineCopula` performs well in terms of copula selection. Moreover, we find that the Lorenz curve's estimator is quite distant from the 45-degree line, which means that perfect equality using $CH(p, Y)$ does not imply perfect equality using L_H. Subfigures (b) and (c) illustrate the semiparametric and empirical estimators of $CH(p, Y)$ and $1 - L_H(1 - p)$ (or L_H) when the dependence between H and Y is given by lower and upper Fréchet–Hoeffding, respectively. From these, we see that the estimators of $CH(p, Y)$ and $1 - L_H(1-p)$ are quite similar, but again the semiparametric estimator is smoother than the empirical estimator of $CH(p, Y)$.

4.2 Nonparametric estimation

The semiparametric estimation of CH assumes that the bivariate copula C of (H, Y) belongs to a parametric family of bivariate copulas. In practice, however, the bivariate copula family is unknown and consequently the semiparametric estimator in (5) might be biased if the chosen parametric bivariate copula is misspecified. To overcome this issue, we propose a nonparametric estimation of CH based on the following nonparametric estimator of bivariate copula. Formally, we consider the following Bernstein estimator of $C(u, v)$:

$$C_{m,n}(u, v) = \sum_{k_0=0}^{m} \sum_{k_1=0}^{m} C_n \left(\frac{k_0}{m}, \frac{k_1}{m} \right) P_{m,k_0}(u) P_{m,k_1}(v), \tag{7}$$

where m is an integer that plays the role of bandwidth, C_n is the empirical bivariate copula, and

$$P_{m,k}(z) = \begin{pmatrix} m \\ k \end{pmatrix} z^k (1-z)^{m-k},$$

is the binomial distribution function with the parameters (m, k). For independent and identically distributed (*i.i.d.*) data, Sancetta and Satchell (2004) introduced a Bernstein polynomial estimator of the copula and established the asymptotic normality of the Bernstein density copula. Under some regularity conditions, Sancetta and Satchell (2004) show that any copula can be approximated by a Bernstein copula. Asymptotic properties of the Bernstein density copula for α-mixing data are studied in Bouezmarni et al. (2010). Furthermore, Janssen et al. (2012) have shown that the Bernstein copula outperforms the classical empirical copula originally proposed by Deheuvels (1979).

Moreover, it is worth noting that an estimator of the partial derivative $C_u(u, v)$ can be derived using the Bernstein bivariate copula, but not the empirical bivariate copula. Using the Bernstein bivariate copula in (7), a nonparametric estimator of the first-order partial derivative of the bivariate copula function, $C_u(u,p)$, can be obtained as follows:

$$\widetilde{C}_u(u,p) = m \sum_{k_0=0}^{m-1} \sum_{k_1=0}^{m} \left[C_n\left(\frac{k_0+1}{m}, \frac{k_1}{m}\right) - C_n\left(\frac{k_0}{m}, \frac{k_1}{m}\right)\right] P_{m-1,k_0}(u) P_{m,k_1}(p). \tag{8}$$

The above estimator was introduced and its asymptotic properties were derived in Janssen et al. (2016). From (2) and (8), a nonparametric estimator of $CH(p,Y)$ can be defined as follows:

$$\widehat{CH}_{m,n}(p,Y) = \frac{\sum_{i=1}^{n} H_i \, \widetilde{C}_u(\widehat{F}_H(H_i), p)}{\sum_{i=1}^{n} H_i}. \tag{9}$$

We now use simulations and some examples of dependence between H and Y to illustrate the nonparametric estimator in (9). We generated $n = 100$ observations of (H_i, Y_i) using a Gaussian bivariate copula with correlation coefficients: (a) $\rho = 0.6$, (b) $\rho = -0.6$, (c) dependence between H and Y is quadratic, and (d) $\rho = 0$ (independence). The simulation results reported in Figure 6 of the separate companion appendix show that the Bernstein copula-based estimator outperforms the empirical estimator as the former is much closer to the 45-degree line under independence and it is much smoother under other cases of dependence.

4.3 Gini health coefficient estimation

We now use the previously developed estimators of CH to propose estimators of Gini health coefficient with respect to a socioeconomic variable. This index is a measure of dispersion intended to quantify the health inequality within a given socioeconomic group. Formally, the Gini health coefficient with respect to socioeconomic variable Y is defined as follows:

$$G(Y) = 2\int_0^1 (p - CH(p,Y))\, dp = 2\frac{Cov(H, F_Y(Y))}{\mathbb{E}(H)}. \tag{10}$$

Under independence between H and Y, the covariance term $Cov(H, F_Y(Y))$ is equal to zero, and therefore the Gini health coefficient with respect to Y $G(Y)$ is equal to zero. The index takes values between -1 and 1. On the one hand, a negative value of $G(Y)$ indicates a pro-poor health, and in this case the health concentration curve with respect to Y will be located above the 45-degree line; see Subfigure (b) of Figure 5 of the separate companion appendix. On the other hand, a positive value of $G(Y)$ indicates a pro-rich health, and the health concentration curve with respect to Y in this case will be below the 45-degree line; see Subfigure (c) of Figure 5 of the separate companion appendix. Furthermore, $G(Y)$ is equal to zero when the concentration curve with respect to Y is close to the 45-degree line, see Subfigure (a) of Figure 5 of the separate companion appendix. Observe, however, that the Gini health coefficient with respect to Y $G(Y)$ - which is equal to the integral of the difference between the 45-degree line and the curve $CH(p, Y)$ - can be equal to zero even when the health concentration curve with respect to Y does not coincide with the 45-degree line. Thus, $CH(p, Y)$ curve should represent a better way of visualizing and detecting health inequalities.

Note that if H and Y are comonotone random variables - i.e., the dependence between the two variables can be modelled using the upper Fréchet–Hoeffding -, then $CH(p, Y) = L_H(p)$ and $G(Y)$ represents a Gini health coefficient that is constructed from Lorenz curve. In this case, we have $F_H(H) = F_Y(Y)$ and $G(Y) = \frac{2}{\mathbb{E}(H)} Cov(H, F_Y(Y)) = \frac{2}{\mathbb{E}(H)} Cov(H, F_H(H))$. Formally,

$$G(Y) = 2 \int_0^1 (p - CH(p, Y))\, dp = 2 \int_0^1 (p - L_H(p))\, dp.$$

Now, if H and Y are countermonotonic random variables - i.e., the dependence between the two variables can be modelled using the lower Fréchet–Hoeffding -, then $CH(p, Y) = 1 - L_H(1 - p)$ and $G(Y)$ will be equal to minus Gini health coefficient that is calculated based on Lorenz curve L_H. Formally,

$$G(Y) = 2 \int_0^1 (p - CH(p, Y))\, dp = 2 \int_0^1 (p - 1 + L_H(1 - p))\, dp = -2 \int_0^1 (p - L_H(p))\, dp.$$

We can show that $G(Y)$ is equal to minus Gini health coefficient that is based on Lorenz curve by observing that $F_H(H) = 1 - F_Y(Y)$ and $Cov(H, F_Y(Y)) = -Cov(H, F_H(H))$.

Thus, using the expression of $G(Y)$ in Equation (10) and the estimators of $CH(p, Y)$ in Equations (5) and (9), we propose the following semiparametric

$$\widehat{G}(Y) = 2\int_0^1 \left(p - \widehat{CH}(p, Y)\right)\, dp \tag{11}$$

and nonparametric

$$\widetilde{G}(Y) = 2\int_0^1 \left(p - \widehat{CH}_{m,n}(p, Y)\right)\, dp \tag{12}$$

estimators of Gini health coefficient, respectively.

We next establish the asymptotic properties of the semiparametric and nonparametric estimators of the health concentration curve with respect to Y $CH(p, Y)$. These properties can be used to establish the asymptotic properties of the estimators of the Gini health coefficient.

5 Asymptotic properties of the estimators of CH

We investigate the consistency and asymptotic normality of the previously developed estimators of health concentration curve with respect to Y $CH(p, Y)$. The following subsections (5.1) and (5.2) state the asymptotic i.i.d. representations of the semiparametric and nonparametric estimators $\widehat{CH}(p, Y)$ and $\widehat{CH}_{m,n}(p, Y)$ in (5) and (9), respectively.

5.1 Asymptotic properties of semiparametric estimator of CH

In this subsection, we assume that the bivariate copula function C belongs to a known parametric family of copulas $\mathcal{C} = \{C(.; \theta), : \theta \in \Theta\}$, where Θ is a compact subset of $\mathbb{R}^q$. Let $\hat{\theta}_n$ be an estimator of the true parameter θ_0 that satisfies the following condition:

Assumption A: *The estimator $\hat{\theta}_n$ of θ_0 has the following asymptotic representation:*

$$\hat{\theta}_n - \theta_0 = n^{-1} \sum_{i=1}^{n} \xi_i + o_p(n^{-1/2}), \tag{13}$$

where $\xi_i = \xi(F_H(H_i), F_Y(Y_i); \theta_0)$ is a q-dimensional random vector such that $E(\xi_i) = 0$ and $E(||\xi_i||^2) < \infty$, with $||.||$ represents the Euclidean norm.

Note that the estimators of θ_0 that have been proposed in Shih and Louis (1995) and Genest, Ghoudi and Rivest (1995) satisfy Assumption **A**. Furthermore, Tsukuhara (2005) provides an asymptotic representation of $\hat{\theta}_n$ and establishes its asymptotic properties.

Now, before we state the asymptotic representation of the semiparametric estimator of CH, we need to introduce the following notations:

- $C_{uu} = \dfrac{\partial^2 C}{\partial u^2}$ and $C_{u,\theta} = \left(\frac{\partial C_u}{\partial \theta_1}, \ldots, \frac{\partial C_u}{\partial \theta_q}\right)$.
- $r_\theta(p, \theta) = \left(\frac{\partial r(p,\theta)}{\partial \theta_1}, \ldots, \frac{\partial r(p,\theta)}{\partial \theta_q}\right)^\top$, where $r(p, \theta) = \mathbb{E}\left[HC_u(F_H(H), p, \theta)\right]$.

Using the above notations, we consider the following additional assumptions that we need in order to establish the result of Theorem 1.

Assumption B: We assume that:

B1: $E(H^2) < \infty$.

B2: $C_{u,u}(u, v, \theta)$ and $C_{u,\theta}(u, v, \theta)$ are continuous on $(0,1) \times [0,1] \times \Theta$. In addition, there is a constant c_1 such that

$$|C_{u,u}(u, v, \theta)| \leq \frac{c_1}{u(1-u)}, \ \forall (u, v) \in (0,1) \times [0,1].$$

Assuming a finite second moment of H is a classical assumption. The conditions outlined in Assumption B2 are not restrictive and are met by numerous common copula families, including Gaussian, Student, Clayton, Gumbel copulas, and others, as detailed in Omelka et al. (2009). These assumptions entail technical conditions that must be fulfilled. Both Assumptions B1 and B2 are essential to establish the law of large numbers and the asymptotic i.i.d. representation of a V-statistic.

The following theorem states the asymptotic i.i.d. representation of the semiparametric estimator $\widehat{CH}(p, Y)$ in (5) [see the proof of Theorem 1 in the Appendix of this paper].

Theorem 1 *Under Assumptions **A**, **B1-B4**, we have*

$$\widehat{CH}(p, Y) - CH(p, Y) = n^{-1} \sum_{i=1}^{n} \zeta_i + o_p(n^{-1/2}),$$

where,

$$\zeta_i = \mathbb{E}(H)^{-1} \left[H_i C_u(F_H(H_i), p, \theta_0) + \xi_i^\top r_\theta(p, \theta_0) + \eta(H_i, p, \theta_0) - H_i\, CH(p, Y) \right],$$

with $\eta(H_i, p, \theta_0) = \mathbb{E}_H \left[H \left(\mathbb{I}(H_i \leq H) - F_H(H) \right) C_{uu}(F_H(H), p, \theta) \right], r_\theta(p, \theta_0) = \mathbb{E} \left[H\, C_{u,\theta}(F_H(H), p, \theta_0) \right],$ ξ_i *is defined in Assumption **A**,* $\mathbb{E}_H$ *represents expectation with respect to* H*, and* $\mathbb{I}(.)$ *is an indicator function that takes value one if* $H_i \leq H$ *and zero otherwise.*

Theorem 1 can be used to establish the asymptotic normality of the semiparametric estimator $\widehat{CH}(p, Y)$ with zero mean and asymptotic variance:

$$Var(\widehat{CH}(p, Y)) = \mathbb{E}\left(\zeta_i^2\right) - \left(\mathbb{E}(\zeta_i)\right)^2,$$

where the expression of ζ_i is defined in Theorem 1. This variance is unknown, but it can be estimated by replacing the unknown quantities in the expressions of $\mathbb{E}(\zeta_i)$ and $\mathbb{E}\left(\zeta_i^2\right)$ by their empirical analogues. For testing and constructing confidence intervals around $\widehat{CH}(p, Y)$, and to mitigate the impact of the accuracy of the estimation of the unknown quantities involved in ζ_i on the sensitivity of the plug-in estimator of the variance, we recommend the use of bootstrap.

5.2 Asymptotic properties of nonparametric estimator of CH

We now establish the asymptotic i.i.d representation of the nonparametric estimator of CH. Let us first define some new terms. For independent random vectors $(U_1, V_1), \ldots, (U_n, V_n)$, with joint distribution function C, we define

$$W_{i,m}(u,v) = m \sum_{k_0=0}^{m-1} \sum_{k_1=0}^{m} \left[A_i(k_0,k_1) - A_i(k_0,m)C_u(u,v)\right] P_{m-1,k_0}(u) P_{m,k_1}(v),$$

where

$$A_i(k_0,k_1) = \mathbb{I}\left(\frac{k_0}{m} < U_i \leq \frac{k_0+1}{m}, V_i \leq \frac{k_1}{m}\right) - P\left(\frac{k_0}{m} < U_i \leq \frac{k_0+1}{m}, V_i \leq \frac{k_1}{m}\right).$$

The following theorem states the asymptotic i.i.d. representation of the nonparametric estimator $\widehat{CH}_{m,n}(p,Y)$ in (9) [see the proof of Theorem 2 in the Appendix of this paper].

Theorem 2 *Suppose that the second derivatives* $C_{u,u} = \frac{\partial^2 C(u,v)}{\partial^2 u}$ *and* $C_{u,v} = \frac{\partial^2 C(u,v)}{\partial u \partial v}$ *are Lipshitz continuous on* $[0,1]^2$*. If* $(m/n)\log\left(m^{1/2}n\right) \to 0$ *as* $n \to \infty$*, we have*

$$\widehat{CH}_{m,n}(p,Y) - CH(p,Y) = n^{-1} \sum_{i=1}^{n} \nu_i(p) + o_p(n^{-1/2}m^{1/4}),$$

where

$$\nu_i(p) = \mathbb{E}(H)^{-1}\left[\mathbb{E}_H\left[HW_{i,m}(F_H(H),p)\right] - H_i\, CH(p,Y)\right].$$

Theorem 2 can be used to establish the asymptotic normality of the nonparametric estimator of CH with mean zero and - by adapting the proofs in Janssen et al. (2016) - asymptotic variance of order $O(n^{-1}m^{1/2})$. This asymptotic variance, however, is unknown as it depends on C_u. In practice we suggest to use bootstrap for constructing confidence intervals around $\widehat{CH}_{m,n}(p,Y)$.

6 Monte Carlo simulations

In this section, we run Monte Carlo simulations to assess the performance of the estimators proposed previously. In particular, we calculate the Integrated Mean Square Errors (IMSE) of semiparametric and nonparametric estimators of health concentration curves with respect to socioeconomic variables, which we compare with the IMSE of the empirical estimator.

In our simulations, several data-generating processes (DGPs) are considered to model the dependence structure between health and socioeconomic variables. We use different bivariate copulas to generate data under different degrees of dependence measured by different values of Kendall's tau coefficient τ. The bivariate copulas under consideration are Gaussian copula, Student copula, Clayton copula, and Gumbel copula. The values of Kendall's tau coefficient under consideration are: (i) $\tau = -0.4$, 0.001, 0.01, 0.1, 0.2, 0.5, 0.7 for Gaussian and Student copulas and (ii) $\tau = 0.001$, 0.01, 0.1, 0.2, 0.5, 0.7 for Clatyton and Gumbel copulas. To generate the data of health and socioeconomic variables, we also use an exponential distribution with the parameter $\lambda = 1$ and a Weibull distribution with the scale and shape parameters $\lambda = 2$ and $k = 10$, respectively.

The performances of the estimators $\widehat{CH}(p, Y)$, $\widehat{CH}_{m,n}(p, Y)$ and $\widehat{CH}_n(p, Y)$ are assessed based on the IMSE that we calculate for different samples sizes: $n = 50$, 100, 200, and 500. The calculation of IMSE is obtained using $B = 1,000$ replications. Formally, the IMSE is defined as:

$$\text{IMSE} \approx \frac{1}{I}\sum_{i=1}^{I}\left(\frac{1}{B}\sum_{j=1}^{B}\left(\widehat{CH}_j(p_i, Y) - CH(p_i, Y)\right)^2\right),$$

for $p_i = 0.01, 0.02, \ldots, 0.99$, and where $\widehat{CH}_j(p_i, Y)$, for $j = 1, \ldots, B$, is the estimator of the concentration health with respect to Y $CH(p_i, Y)$ that corresponds to the j-th replication. The integral $\int(.)$ of the MSE is approximated by replacing it with the sum $\sum^I(.)$ for $I = 99$.

For the semiparametric estimator $\widehat{CH}(p, Y)$, we assume that the parametric bivariate copula is unknown and we use the **BiCopSelect** function implemented in **VineCopula R** package to select the appropriate bivariate copula from a set of bivariate copula families.

This approach is expected to alleviate the negative effect that copula misspecification might have on the semiparametric estimator. We recall that copulas can be selected according to an Akaike or Bayesian Information Criteria [AIC and BIC, respectively]; see Akaike (1973), Schwarz (1978), and Manner (2007). To do so, all available bivariate copulas are first fitted using maximum pseudo-likelihood estimation. Then the information criteria of all fitted bivariate copulas are computed, and the bivariate copula that has the minimum AIC or BIC is selected. Formally, the AIC of a bivariate parametric copula density $c(F_H(H), F_Y(Y), \theta)$ is defined as

$$AIC = -2\sum_{i}^{n} ln[c(\hat{F}_H(H_i), \hat{F}_Y(Y_i), \hat{\theta}_n) + 2q,$$

where q represents the number of parameters: e.g.; $q = 1$ for a copula that depends on one parameter, and $q = 2$ for a copula with two parameters. Similarly, the BIC of bivariate copula density $c(F_H(H), F_Y(Y), \theta)$ is given by

$$BIC = -2\sum_{i}^{n} ln[c(\hat{F}_H(H_i), \hat{F}_Y(Y_i), \hat{\theta}_n) + ln(n)q.$$

Furthermore, the bandwidth parameter m that we use to calculate $\widehat{CH}_{m,n}(p, Y)$ is selected according to the rule of thumb $m = \left[a \times n^{2.5/5}\right]$, where [.] is the integer part of $a \times n^{2.5/5}$. To assess the sensitivity of the estimation results with respect to m, we consider various values of $m \in \{1, 2, 3, 4, 5\}$, which is a common practice in nonparametric estimation when no optimal bandwidth is available. In the simulations, we find that the optimal value of a that works for all DGPs under consideration is equal to 4. Thus, to save space we only report results for $a = 4$; see Table 1. The other values of a also provide reasonable results and the latter are available upon request.

Table 1 shows that the IMSEs of all estimators are decreasing with the sample size. Interestingly, we find that the semiparametric estimator dominates both the empirical and the nonparametric estimators as it has the smallest IMSE under different bivariate copulas, degrees of dependence and sample sizes, except when the dependence between H and Y is weak or when τ takes values between 0.001 and 0.1. Thereafter, we find that the second best estimator is the nonparametric estimator that does better than the empirical estimator, except when the degree of dependence between H and Y takes values between $\tau = 0.5$ and $\tau = 0.7$ for Gaussian and Gumbel copulas.

The above results are obtained when the marginal distributions of H and Y are given by the exponential distribution with the parameter $\lambda = 1$. We have also got additional results after replacing the exponential distribution by the Weibull distribution with the parameters $\lambda = 2$ and $k = 10$. To save space, the latter results are not reported but they are available upon request. Using Weibull as a marginal distribution of H and Y, we find similar results to those discussed previously. The best estimator in terms of IMSE is the semiparametric estimator $\widehat{CH}(p, Y)$, followed by the Bernstein estimator $\widehat{CH}_{m,n}(p.Y)$ and then the empirical estimator $\widehat{CH}_n(p, Y)$.

7 Inequality in COVID-19's infections and deaths

Since January 2020, the COVID-19 pandemic has resulted in millions of infections and deaths, and caused economic distress worldwide. Recognizing the significance of measures such as lockdowns, testing, and face masks in reducing the transmission of COVID-19, concerns have emerged regarding the connection between pre-existing social and economic inequalities and disparities in COVID-19 infections and deaths, particularly affecting the most deprived classes. In the United States, the spread of COVID-19 across states has highlighted that not all Americans face an equal risk of infection and mortality from the virus. Furthermore, the World Health Organization (WHO) in the European Region has emphasized that the exposure risk and severity of health, social, and economic impacts of COVID-19 are not evenly distributed among European countries.

In this section, we apply our semi- and nonparametric estimators for the Health Concentration and Gini health coefficient to assess and analyze disparities in COVID-19 infections and deaths both across and within U.S. states. The subsequent subsection provides a description of the U.S. data employed in our empirical analysis.

7.1 Data

We begin by describing our data and presenting key descriptive statistics. Micro-level data on COVID-19 infections and deaths, combined with socioeconomic factors—such as median income, total population, percentage of the population below the poverty line, and percentages of Black, White, Asian, and Hispanic American residents, as well as the proportion of urban residents and women in the population—at the unit-record level (specifically, at the county level in the U.S.). To build our dataset, we merged county-level counts of COVID-19 cases and deaths with socioeconomic characteristics, including average income, racial demographics, income inequality, and poverty.

Our dataset includes data from 2,777 counties across 40 U.S. states. Ten states were excluded due to insufficient data; each of these states has fewer than 10 counties, making them unsuitable for estimating health concentration curves and the Gini health coefficient based on our socioeconomic variables. The excluded states are New Hampshire, Connecticut, Hawaii, Kansas, Kentucky, Louisiana, Maine, Maryland, Rhode Island, and the District of Columbia.

The county-level variables in our dataset can be categorized into three main groups:

- Health variables encompassing confirmed COVID-19 infections and deaths.
- Socioeconomic covariates that include median income, total population, percentage of the population below the poverty line, and the percentages of residents who are Black Americans, White Americans, Asian Americans, and Hispanic Americans. Additionally, the analysis includes the proportion of the population living in urban areas and the percentage of women in the population.

- Inequality measures, namely the Gini (economic) coefficient and Gini health coefficient, which were estimated using our semi- and nonparametric estimators.

For data regarding confirmed COVID-19 infections and deaths, we rely on the U.S. Centers for Disease Control and Prevention (CDC), utilizing the most recent numbers available as of writing this paper (June 10, 2021)[1]. Information on county population, population density, demographics, and poverty rates is sourced from the U.S. Census Bureau. Median income and the poverty rate are estimated from survey data, supplemented by small-area estimation methods [refer to Rao and Molina (2015)].

Descriptive statistics detailing COVID-19 infection and death rates, along with socioeconomic variables for the 2,777 counties, are presented in Table 2 of the separate companion appendix. In this table, the term "share of African Americans" pertains to the proportion of the population identifying as Black only, while "share of White" denotes the proportion identifying as White. The table reveals that the average infection and death rates per county are 9.8% and 1.9%, respectively, with standard deviations of 2.99% and 1%, respectively. Notably, Crowley County in Colorado recorded the highest infection rate at 37%, and Chattahoochee County in Georgia had the second-highest rate at 36.8%.

The overall average poverty[2] rate is 14.3%, ranging from 2.3% in Sterling County (Texas) to 54.7% in Todd County (South Dakota). The Gini (economic) coefficient is also estimated using small-area methods, exhibiting significant variation across counties, from the lowest value of 0.302 in Skagway County (Alaska) to the highest value of 0.609 in Harding County (New Mexico).

Moreover, Table 3 of the separate companion appendix reports the correlation matrix

[1] The data are available through USA Facts : https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/. An alternative source is the New York Times data site for COVID-19 obtainable from their Github repository. The NYT site, however, records infections and deaths according to the county in which they occurred, while CDC does so according to the person's place of residence.

[2] The poverty rate is a measure of how many people in the country live in economic hardship. Poverty is measured by comparing a person's or family's income to a set poverty threshold or minimum amount of income needed to cover basic needs. The poverty threshold varies by family size, number of children, and the age of the householder. For example, in 2023, the poverty threshold for a US family of four was $30,000 according to the U.S. Census Bureau which is the government agency in charge of measuring poverty.

of socioeconomic variables. Significantly, there is a positive correlation between the share of African Americans and the poverty rate, and a negative correlation between the share of White Americans and the poverty rate, both statistically significant. Subsequently, we present the results of the estimation of health concentration curves and the Gini health coefficient with respect to our socioeconomic variables.

7.2 Estimation of health concentration curve

As previously mentioned, the health concentration curve with respect to a socioeconomic variable enables the generation of plots depicting the cumulative percentage of the health variable - specifically, COVID-19's infection and death rates - against the cumulative percentage of the population ranked by the socioeconomic variable. This variable includes metrics such as the poverty rate, population density, the share of African Americans, the share of White Americans, and others. These plots offer a visual representation of health inequalities by observing the position of the health concentration curve with respect to the line of equality (45-degree line) in a two-dimensional space.

In scenarios where a higher socioeconomic variable corresponds to greater individual wealth, if the health concentration curve lies above the 45-degree line, health inequality is deemed pro-rich, indicating that the affluent possess better health than the less affluent. In such cases, the associated health concentration index is negative. Conversely, if the health concentration curve lies below the 45-degree line, health inequality is considered pro-poor, suggesting that the less affluent have better health than the affluent, and the associated health concentration index is positive. Should the health concentration curve align with the 45-degree line, it indicates the absence of socioeconomic health inequality, and the associated health concentration index is necessarily zero.

We now utilize the data detailed in the previous subsection to compute semiparametric and nonparametric estimators of the health concentration curve for COVID-19's infection and death rates concerning the aforementioned socioeconomic variables. For each state, we consider the following socioeconomic variables: income, poverty, and race (proportion of Black Americans, Whites, Asians, and Hispanics), urban area, percentage of women. While the results encompass all 40 U.S. states, we present findings for average, low, and high-income states, average, low, and high-poverty rate states, and average, low, and high-share of White states to enhance clarity. The outcomes are outlined in Figures 7 to 15 of the separate companion appendix. [3]

When employing income as a socioeconomic variable, Figures 7 and 8 reveal a distinct pro-rich inequality for COVID-19's death rate in both low and high-income states, except for West Virginia. Intriguingly, for states with average income [refer to Figure 9], the health concentration curve with respect to the poverty rate tends to align with the 45-degree line. This suggests that individuals in these states had an equal chance of succumbing to COVID-19, irrespective of their poverty level. Similar results emerge when replacing income with the poverty rate, considering that a low poverty rate is expected to correlate highly with high income, and vice versa.

Figures 10 and 11 confirm the pro-rich inequality for COVID-19's death rate in both low and high poverty rate states. The health concentration curve concerning the poverty rate lies below the 45-degree line when the poverty rate is higher, indicating greater inequality in counties with increased poverty. The result in Figure 12 echoes that obtained in Figure 9: there is no inequality in COVID-19's death rate in states with an average poverty rate, as the health concentration curve aligns with the 45-degree line.

Moreover, Figures 13 and 14 indicate that the inequality in COVID-19's death rate is generally in favor of counties with a low share of white people, except in state of California. However, the opposite happens in states with an average share of white people, where we see that the inequality in COVID-19's death rate is in favor of counties with high share of white people.

In summary, socioeconomic variables such as income, poverty rate, and ethnicity sig-

[3] Additional results related to other socioeconomic variables, such as those for Hispanic and Asian populations, are available in Figures 19 through 24 in the separate companion appendix.

nificantly influence COVID-19's death rate, disproportionately affecting the most deprived classes. In the subsequent subsections, we delve deeper into the analysis by estimating the Gini health coefficient and developing econometric models that establish connections between COVID-19's cases and deaths, Gini health coefficient, and other crucial socioeconomic factors.

7.3 Estimation of Gini health coefficient

We now use the available data to compute both semiparametric and nonparametric estimators of the Gini health coefficient for COVID-19's death rates, as expressed in Equations (11) and (12) respectively. Our emphasis on death rates stems from the belief that this measure more accurately reflects health inequalities across states characterized by diverse levels of socioeconomic variables such as income, poverty, and race.

Once the infection occurs, one might expect that COVID-19 deaths will be higher in the deprived states/areas due to the poor living conditions and underfunded local services such as health services that are essential for saving lives. In estimating the Gini health coefficient, states with fewer than 10 counties were excluded from the analysis. The affected states include New Hampshire, Connecticut, Hawaii, Kansas, Kentucky, Louisiana, Maine, Maryland, and Rhode Island. The outcomes for the remaining 40 U.S. states are presented in Figures 16 to 18 of the separate companion appendix.

Firstly, in Figure 16, we present the outcomes of estimating the Gini health coefficient for each U.S. state, with the health variable H representing COVID-19's death rates and the socioeconomic variable Y denoting median income. Given that, in this context, the lowest socioeconomic position corresponds to counties with lower median income, the observed negative Gini health coefficients across the majority of U.S. states signify a pro-rich health inequality. In simpler terms, this result implies that individuals in counties with higher median income experience lower COVID-19 mortality compared to those residing in counties with lower median income.

Secondly, Figure 17 showcases the results of semiparametric and nonparametric estimations of the Gini health coefficient, with the socioeconomic variable Y being the poverty rate. Unlike the previous scenario, where the lowest socioeconomic position for the poverty rate represented affluent counties, the positive Gini health coefficients in Figure 17 affirm a pro-rich health inequality. This implies that counties with a high poverty rate exhibit elevated COVID-19 mortality compared to counties with a low poverty rate.

Finally, in Figure 18, we present results where the socioeconomic variable is the proportion of African Americans. In this case, the lowest socioeconomic position corresponds to counties with the lowest proportion of African Americans. The outcome depicted in Figure 18 indicates a pro-non-African American health inequality across U.S. states, with counties having a high proportion of African Americans experiencing higher COVID-19 mortality compared to counties with a low proportion of African Americans.

This study provides valuable insights for policymakers aiming to mitigate health inequalities, thereby enhancing the overall chances of survival from COVID-19. The findings suggest that authorities in U.S. states should prioritize the allocation of increased health resources to counties characterized by higher deprivation and a substantial proportion of African Americans. This strategic approach is recommended to effectively address and reduce health disparities in COVID-19 mortality across various counties.

7.4 Further analysis

In this subsection, we conduct further analysis utilizing state-level data to explore the factors influencing COVID-19 infections and deaths. We consider the following mean regression model

$$Y_i = \alpha + \beta' X_i + \epsilon_i,$$

where Y_i denotes one of the following dependent variables: i) state-level COVID-19 infections, ii) state-level COVID-19 deaths. The covariates X_i incorporated in the regressions encompass state population size, poverty rate, median income, Gini economic coefficient, Gini health coefficient (calculated using the estimators developed in the previous sections), the share of the population living in urban areas, the percentage of women, share of Black Americans, share of Hispanic Americans, share of Asian Americans, share of Black women, poor women, poor Hispanics, and poor Asians. The results are presented in Tables 4-5 of the separate companion appendix.

Tables 4-5 reveal that Gini coefficient has a positive impact on COVID-19 infections and deaths when we control for all the independent variables. These effects are statistically significant. However, the effect of Gini health coefficient is statistically insignificant when we include all control variables, possibly attributed to high correlations with other covariates. Additionally, poverty rate and population size demonstrate significant and positive impacts on both COVID-19 infections and deaths. The impact of a state's share Black women and poor women is statistically significant, while the impact of a state's share of poor Asians is negative and statistically significant. The results for COVID-19 infections and deaths exhibit substantial similarity. The multiple regression framework used in our analysis enables researchers and policymakers to identify variables that may influence health outcomes, such as COVID-19 infections and deaths. The estimated parameters represent correlations between COVID-19 outcomes and specific variables—such as the proportion of Black American residents—while controlling for other variables in the model. However, we note that the mean regression model may be affected by multicollinearity, meaning that some independent variables may be highly correlated, potentially reducing precision and causing shifts in the direction of coefficients.

The above results once again confirm that increased economic and health inequalities correspond to higher rates of COVID-19 infections and deaths, disproportionately impacting the most-deprived classes. Policymakers are strongly encouraged to implement social and economic measures aimed at reducing COVID-19 infections and deaths among the most vulnerable populations. In the short term, initiatives such as providing free masks, sanitizers, and targeted vaccination campaigns should be implemented to mitigate infection and mortality rates in deprived states. In the long term, governments should implement policies addressing economic and social inequalities in society, ultimately contributing to a reduction in health disparities.

8 Conclusion

We developed semi- and nonparametric estimators for the health concentration curve and Gini health coefficient with respect to a given socioeconomic variable to assess inequalities in COVID-19 infections and deaths, identifying the social classes most susceptible to the virus. Health concentration measure was initially expressed in terms of a bivariate copula function, forming the basis for our estimators. In the semiparametric approach, a parametric bivariate copula modeled the dependence between health and socioeconomic variables, with parameters estimated through maximum pseudo-likelihood estimation using the empirical distribution of the health variable. For the nonparametric estimator, we employed the Bernstein bivariate copula estimator. Additionally, health concentration curve estimators were used to derive semi- and nonparametric estimators for the Gini health coefficient. We established the consistency and asymptotic normality of the health concentration curve estimators.

Monte Carlo simulations, encompassing various data-generating processes and sample sizes, revealed superior performance of the semiparametric estimator over the Bernstein-copula-based estimator, which, in turn, outperformed the empirical estimator in terms of Integrated Mean Squared Error. An extensive empirical study highlighted the importance of health concentration curves and Gini health coefficient estimators in analyzing COVID-19 inequality in the U.S. The results demonstrated that socioeconomic variables, such as poverty, race, and income at the state level, can help explain the observed disparities in COVID-19 infections and deaths.

The methodology proposed in this paper is applicable only for a univariate Y, presenting a limitation in estimating health concentration curves using just one socio-economic variable. This emerges as focusing on a single variable oversimplifies the intricate interplay of diverse health determinants. This restriction arises from the technical issues associated with the use of multivariate quantiles, making interpretation and estimation challenging. While considering several socio-economic variables simultaneously might provide a more comprehensive understanding, the complexities involved in deciphering the combined impact of multiple factors hinder the practicality of such analyses. An alternative approach involves constructing an index that combines diverse socio-economic variables, yet unraveling the nuanced interpretation of this composite index proves to be a difficult task. These issues will be subject to further investigation in future research.

References

[1] Akaike, H. (1998). "Information theory and an extension of the maximum likelihood principle," In Selected papers of hirotugu akaike (pp. 199-213). New York, NY: Springer New York, pp. 199-213.

[2] Allison, R. A. and Foster, J. E. (2004). "Measuring health inequality using qualitative data," *Journal of Health Economics,* vol. 23 (3), pp. 505-524.

[3] Bouezmarni, T., Rombouts, J.V. and Taamouti, A. (2010). "Asymptotic properties of the Bernstein density copula estimator for -mixing data," *Journal of Multivariate Analysis*, vol. 101, pp. 1-10.

[4] Bouezmarni, T., Rombouts, J.V. and Taamouti, A. (2012). "Nonparametric copula-based test for conditional independence with applications to Granger causality," *Journal of Business & Economic Statistics,* vol. 30 (2), pp. 275-287.

[5] Chen, S. X. and Huang, T.M. (2007). "Nonparametric estimation of copula functions for dependence modelling," *Canadian Journal of Statistics,* vol. 35 (2), pp. 265-282.

[6] Crane, G. J. and Hoek, J. V. D. (2008). "Conditional expectation formulae for copulas," *Australian & New Zealand Journal of Statistics,* vol. 50 (1), pp. 53-67.

[7] Decancq, K. (2014). "Copula-based measurement of dependence between dimensions of well-being," *Oxford Economic Papers*, vol. 66(3), pp. 681-701.

[8] Deheuvels, P. (1979). "La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d'indépendance," *Bulletins de l'Académie Royale de Belgique*, vol. 65(1), pp. 274-292.

[9] Dorn, F., Radice, R., Marra, G. and Kneib, T. (2024). "A bivariate relative poverty line for leisure time and income poverty: Detecting intersectional differences using distributional copulas," R*eview of Income and Wealth*, vol. 70(2), pp. 395-419.

[10] Erreygers, G. and Van Ourti, T. (2011). "Measuring socioeconomic inequality in health, health care and health nancing by means of rank-dependent indices: a recipe for good practice," *Journal of Health Economics*, vol. 30(4), pp. 685-694.

[11] Genest, C., Ghoudi, K. and Rivest, L. P. (1995). "A semiparametric estimation procedure of dependence parameters in multivariate families of distributions," *Biometrika,* vol. 82(3), pp. 543-552.

[12] Hans, N., Klein, N., Faschingbauer, F., Schneider, M., and Mayr, A. (2023). "Boosting distributional copula regression," *Biometrics*, vol. 79(3), pp. 2298-2310.

[13] Janssen, P., Swanepoel, J. and Veraverbeke, N. (2012). "Large sample behavior of the Bernstein copula estimator," *Journal of Statistical Planing and Inference*. vol. 142, pp. 1189-1197.

[14] Janssen, P., Swanepoel, J. and Veraverbeke, N. (2016). "Bernstein estimation for a copula derivative with application to conditional distribution and regression functionals," TEST vol. 25 (2), pp. 351-374.

[15] Joe, H. and Xu, J. (1996). "The estimation method of inference functions for margins for multivariate models," Technical Report No. 166, Department of Statistics, University of British Columbia, Vancouver.

[16] Klein, N. and Kneib, T. (2016). "Simultaneous inference in structured additive conditional copula regression models: a unifying Bayesian approach," *Statistics and Computing*, vol. 26(4), pp. 841-860.

[17] Klein, N., Kneib, T., Marra, G., Radice, R., Rokicki, S. and McGovern, M. E. (2019). "Mixed binary-continuous copula regression models with application to adverse birth outcomes," *Statistics in Medicine*, vol. 38(3), pp. 413-436.

[18] Kneib, T. (2013). "Beyond mean regression," *Statistical Modelling*, vol. 13(4), pp. 275-303.

[19] Krieger, N., Chen, J. T., Waterman, P. D., Rehkopf, D. H. and Subramanian, S. V. (2005). "Painting a truer picture of US socioeconomic and racial/ethnic health inequalities: the Public Health Disparities Geocoding Project," *American Journal of Public Health*, vol. 95(2), pp. 312-323.

[20] Krieger, N., Chen, J. T., Waterman, P. D., Rehkopf, D. H. and Subramanian, S. V. (2003). "Race/ethnicity, gender, and monitoring socioeconomic gradients in health: a comparison of area-based socioeconomic measures-the public health disparities geocoding project," *American Journal of Public Health*, vol. 93(10), pp. 1655-1671.

[21] Krieger N., Chen, J.T, Waterman, P.D, Soobader, M-J, Subramanian, S.V. and Carson, R. (2003a). "Monitoring socioeconomic inequalities in sexually transmitted infections,

tuberculosis, and violence: Geocoding and choice of area-based socioeconomic measures-The public health disparities geocoding project (US)," *Public Health Reports*, vol. 118, pp. 240-260.

[22] Krieger N., Chen, J.T, Waterman, P.D, Soobader, M-J, Subramanian, S.V. and Carson, R. (2003b). "Choosing area-based socioeconomic measures to monitor social inequalities in low birthweight and childhood lead poisoning –The public health disparities geocoding project (US)," J*ournal of Epidemiology & Community Health*, vol. 57(3), pp. 186-199.

[23] Krieger N., Chen J.T, Waterman P.D, Soobader M-J, Subramanian, S.V. and Carson R. (2002). "Geocoding and monitoring US socioeconomic inequalities in mortality and cancer incidence: Does choice of area-based measure and geographic level matter?-The public health disparities geocoding project," *American Journal of Epidemiology*, vol. 156(5), pp. 471-82.

[24] Manner, H. (2007). "Estimation and model selection of copulas with an application to exchange rates," METEOR, Maastricht research school of Economics of Technology and Organizations.

[25] Nagler, T., Schepsmeier, U., Stoeber, J., Brechmann, E.C., Graeler, B., Erhardt, T., Almeida, C., Min, A., Czado, C., Hofmann, M., Killiches, M., Joe, H., and Vatter, T. (2023). "VineCopula: Statistical inference of Vine copulas," R Package, version 2.5.0, URL https://github.com/tnagler/VineCopula

[26] Noh, H., Ghouch, A. E. and Bouezmarni, T. (2013). "Copula-based regression estimation and inference," *Journal of the American Statistical Association,* vol. 108 (502), pp. 676-688.

[27] Noh, H., Ghouch, A. E. and Van Keilegom, I. (2015). "Semiparametric conditional quantile estimation through copula-based multivariate models," *Journal of Business & Economic Statistics,* vol. 33 (2), pp. 167-178.

[28] Omelka, M., Gijbels, I. and Veraverbeke, N. (2009). "Improved kernel estimation of copulas: Weak convergence and goodness-of-fit testing," *The Annals of Statistics*, vol. 37, pp 3023-3058.

[29] Rao, J. N. and Molina, I. (2015). Small area estimation. John Wiley & Sons.

[30] Rémillard, B., Nasri, B. and Bouezmarni, T. (2017). "On copula-based conditional quantile estimators," *Statistics & Probability Letters,* vol. 128, pp. 14-20.

[31] Shih, J. H. and Louis, T. A. (1995). "Inferences on the association parameter in copula models for bivariate survival data," *Biometrics*, pp. 1384-1399.

[32] Sancetta, A. and Satchell, S. (2004). "The Bernstein copula and its applications to modeling and approximations of multivariate distributions," *Econometric Theory,* vol. 20, pp. 535-562.

[33] Schwarz, G. (1978). "Estimating the dimension of a model," *The Annals of Statistics*, pp. 461-464.

[34] Sklar, A. (1959). "Fonctions de repartitionan dimensions et leurs marges," *Publications de l'Institut de Statistique de l'Universite de Paris,* vol. 8, pp. 229-231.

[35] Sungur, E. A. (2005). "Some observations on copula regression functions," *Communications in Statistics- Theory and Methods,* vol. 34 (9-10), pp. 1967-1978.

[36] Tsukahara, H. (2005). "Semiparametric estimation in copula models," *Canadian Journal of Statistics*, vol. 33(3), pp. 357-375.

[37] Wagstaff , A. (2002). "Inequality aversion, health inequalities and health achievement," *Journal of Health Economics,* vol. 21, pp. 627-641.

[38] Wu, X., Nethery, R. C., Sabath, M. B., Braun, D. and Dominici, F. (2020). "Exposure to air pollution and COVID-19 mortality in the United States," medRxiv, 2020.04.05.20054502.

[39] Yitzhaki, S. and Olkin, I. (1991). "Concentration indices and concentration curves," Lecture Notes-Monograph Series, pp.380-392.

[40] Yitzhaki, S. and Schechtman, E. (2013). "The Gini methodology: a primer on a statistical methodology," New York: Springer, pp.11-31.

[41] Zheng, B. (2011). "A new approach to measure socioeconomic inequality in health," *The Journal of Economic Inequality,* vol. 9 (4), pp. 555-577.

Copula-based estimation of health inequality measures:

Online Appendix*

Taoufik Bouezmarni[†] Mohamed Doukali[‡]

Université de Sherbrooke University of Liverpool

Abderrahim Taamouti[§]

University of Liverpool

November 8, 2024

*The authors would like to thank very much Bianca De Stavola, Serena Ng, an Associate Editor, three anonymous reviewers, and the participants at the 2022 CISEM International Symposium on Statistics and Econometrics, 2022 Africa Meeting of the Econometric Society, 7th African International Conference (AIC) on Statistic in Marrakech, 2023, University of Malaga, and INSEA for their very useful comments.

[†]Département de Mathématiques, Université de Sherbrooke, CIREQ, CREAS. Email addresses: Taoufik.Bouezmarni@USherbrooke.ca

[‡] *Corresponding author.* Department of Economics, University of Liverpool Management School. Address: Chatham St, Liverpool L69 7ZH. E-mail: m.doukali@liverpool.ac.uk

[§]Department of Economics, University of Liverpool Management School. Address: Chatham St, Liverpool L69 7ZH. E-mail: Abderrahim.Taamouti@liverpool.ac.uk

Copula-based estimation of health inequality measures: Online Appendix

ABSTRACT

This online appendix contains the proofs of the theoretical results and all the simulations and empirical results stated in the main paper entitled "Copula-based estimation of health inequality measures".

Keywords: Health concentration curve, Gini health coefficient, inequality, copula, semi and nonparametric estimators, COVID-19 infections and deaths.

Journal of Economic Literature classification: C13, C14, I14.

Appendix: Proofs

This appendix contains the proofs of the main results in the text.

Poof of Proposition 1: Observe that

$$\begin{aligned}
CH(p,Y) &= \frac{\int_0^p \mathbb{E}\left[H \mid Y = F^{-1}(u)\right] du}{\int_0^1 \mathbb{E}\left[H \mid Y = F_Y^{-1}(u)\right] du} \\
&= \frac{\int_0^p \int_0^{+\infty} h f_H(h) c(F_H(h), u) dh du}{\int_0^1 \int_0^{+\infty} h f_H(h) c(F_H(h), u) dh du} \\
&= \frac{\int_0^{+\infty} h f_H(h) C_u(F_H(h), p) dh}{\int_0^{+\infty} h f_H(h) C_u(F_H(h), 1) dh} \\
&= \frac{\int_0^1 F_H^{-1}(u) C_u(u,p) du}{\mathbb{E}(H)},
\end{aligned}$$

where the last equality is due to the fact that $\int_0^{+\infty} h f_H(h) C_u(F_H(h), 1) dh = \int_0^{+\infty} h f_H(h) dh = \mathbb{E}(H)$ since $C_u(u,1) = 1$.

Poof of Theorem 1: We start with the following decomposition

$$\widehat{CH}(p,Y) - CH(p,Y) = \frac{\hat{m}(p) - \bar{H} CH(p,Y)}{\mathbb{E}(H)} + \frac{(\mathbb{E}(H) - \bar{H})(\hat{m}(p) - \bar{H}\, CH(p,Y))}{\mathbb{E}(H)\bar{H}}, \quad (1)$$

where $\bar{H}$ is the empirical mean of H and $\hat{m}(p) = n^{-1}\sum_{i=1}^n H_i C_u(\widehat{F}_H(H_i), p, \hat{\theta}_n)$. We study the first term in (1) since the second term is negligible with respect to the first term. Using Taylor expansion of C_u around $(F_H(H_i), p, \theta_0)$, we obtain

$$\hat{m}(p) = n^{-1} \sum_{i=1}^{n} H_i\, C_u(F_H(H_i), p, \theta_0) + I_{n,1} + I_{n,2}, \quad (2)$$

where

$$I_{n,1} = n^{-1} \textstyle\sum_{i=1}^n H_i\, C_{uu}(\tilde{F}_H(H_i), p, \tilde{\theta})(\widehat{F}_H(H_i) - F_H(H_i))$$

$$I_{n,2} = n^{-1} \textstyle\sum_{i=1}^n H_i\, (\hat{\theta}_n - \theta_0)^\top C_{u,\theta}(\tilde{F}_H(H_i), p, \tilde{\theta}),$$

with

$$\tilde{F}_H(H_i) = F_H(H_i) + \alpha(\widehat{F}_H(H_i) - F_H(H_i)),\ \tilde{\theta} = \theta_0 + \alpha(\hat{\theta}_n - \theta_0),\ \alpha \in (0,1).$$

Under Assumptions B1 and B2, we have:

$$I_{n,1} = n^{-1} \sum_{i=1}^{n} H_i \, C_{uu}(F_H(H_i), p, \theta_0)(\widehat{F}_H(H_i) - F_H(H_i)) + o_p(n^{-1/2}),$$

and, from Assumption B4, we obtain

$$\begin{aligned} I_{n,2} &= n^{-1} \sum_{i=1}^{n} H_i \, (\hat{\theta}_n - \theta_0)^\top C_{u,\theta}(F_H(H_i), p, \theta_0) + o_p(n^{-1/2}) \\ &= (\hat{\theta}_n - \theta_0)^\top n^{-1} \sum_{i=1}^{n} H_i \, C_{u,\theta}(F_H(H_i), p, \theta_0) + o_p(n^{-1/2}) \\ &= (\hat{\theta}_n - \theta_0)^\top \mathbb{E} \left[H \, C_{u,\theta}(F_H(H), p, \theta_0) \right] + o_p(n^{-1/2}) \\ &= \xi_i^\top r_\theta(p, \theta_0) + o_p(n^{-1/2}), \end{aligned} \tag{3}$$

where, $r_\theta(p, \theta_0) = \mathbb{E} \left[H \, C_{u,\theta}(F_H(H), p, \theta_0) \right]$.

Now, observe that $I_{1,n}$ is a V-statistic with the kernel $h_1(u, v, \theta_0) = \frac{1}{2} \left[h_2(u, v, \theta_0) + h_2(v, u, \theta_0) \right]$, where

$$h_2(u, v, \theta_0) = u C_{uu}(F_H(u), p, \theta_0)(\mathbb{I}(v \leq u) - F_H(u)).$$

It can be shown that $\mathbb{E}(h_1(H_i, H_j, \theta_0)) = 0$, and from Assumption B3, we have $\mathbb{E}(h_1^2(H_i, H_j, \theta_0)) < \infty$.

Hence, we obtain

$$I_{n,1} = n^{-1} \sum_{i=1}^{n} \eta(H_i, p, \theta_0) + o_p(n^{-1/2}), \tag{4}$$

where

$$\eta(H_i, p, \theta_0) = E_H \left[H \, (\mathbb{I}(H_i \leq H) - F_H(H)) C_{uu}(F_H(H), p, \theta_0) \right].$$

From (2), (3), and (4), we conclude the proof of Theorem 1.

Poof of Theorem 2: First, we recall the expression of the nonparametric estimator of

$CH(p,Y)$:

$$\widehat{CH}_{m,n}(p,Y) = \frac{\frac{1}{n}\sum_{i=1}^{n} H_i\, \widetilde{C}_u(\widehat{F}_H(H_i),p)}{\frac{1}{n}\sum_{i=1}^{n} H_i} = \frac{\tilde{m}(p)}{\bar{H}}.$$

Note that from Janssen et al. (2016), and using the conditions of the theorem, we have

$$\widetilde{C}_u(\widehat{F}_H(h),p) = \frac{1}{n}\sum_{i=1}^{n} W_{i,m}(F_H(h),p) + o_p(n^{-1/2}m^{1/4}).$$

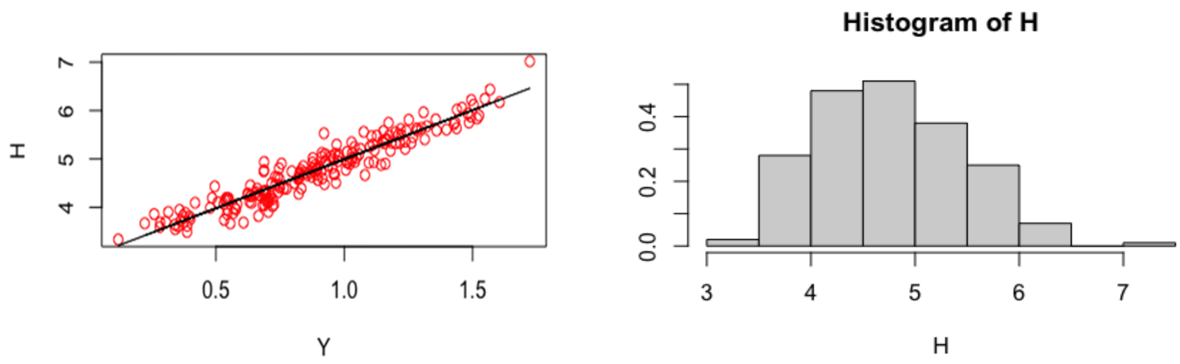
Using the above result, the numerator of $\widehat{CH}_{m,n}(p,Y)$, can be written as:

$$\begin{aligned}
\frac{1}{n}\sum_{j=1}^{n} H_j \widetilde{C}_u(\widehat{F}_H(H_j),p) &= \frac{1}{n}\sum_{j=1}^{n}\left[H_j \frac{1}{n}\sum_{i=1}^{n} W_{i,m}(F_H(H_j),p)\right] + o_p(n^{-1/2}m^{1/4}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{n}\sum_{j=1}^{n} H_j W_{i,m}(F_H(H_j),p)\right] + o_p(n^{-1/2}m^{1/4}) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_H\left[H W_{i,m}(F_H(H),p)\right] + o_p(n^{-1/2}m^{1/4}). \qquad (5)
\end{aligned}$$

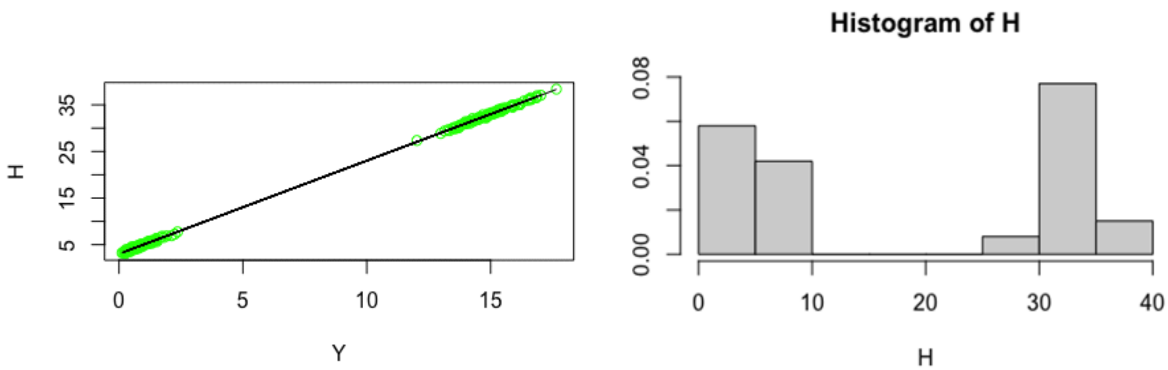
The proof of Theorem 2 follows from Equation (5) and the following decomposition:

$$\widehat{CH}_{m,n}(p,Y) - CH(p,Y) = \frac{\tilde{m}(p) - \bar{H}CH(p,Y)}{\mathbb{E}(H)} + \frac{(\mathbb{E}(H)-\bar{H})(\tilde{m}(p) - \bar{H}\, CH(p,Y))}{\mathbb{E}(H)\bar{H}}.$$

Appendix: Simulation and empirical results



Panel (a)



Panel (b)

Figure 1: This figure illustrates different marginal distributions of H and linear dependence structure between H and Y.

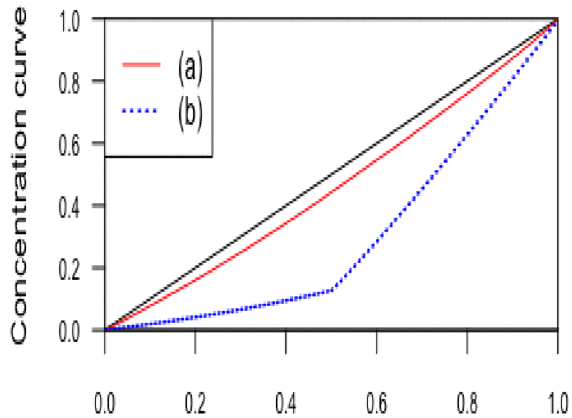
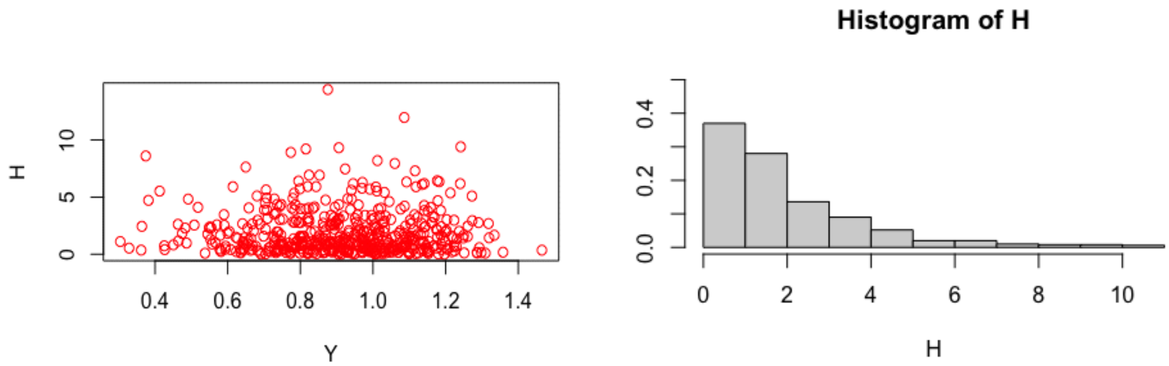
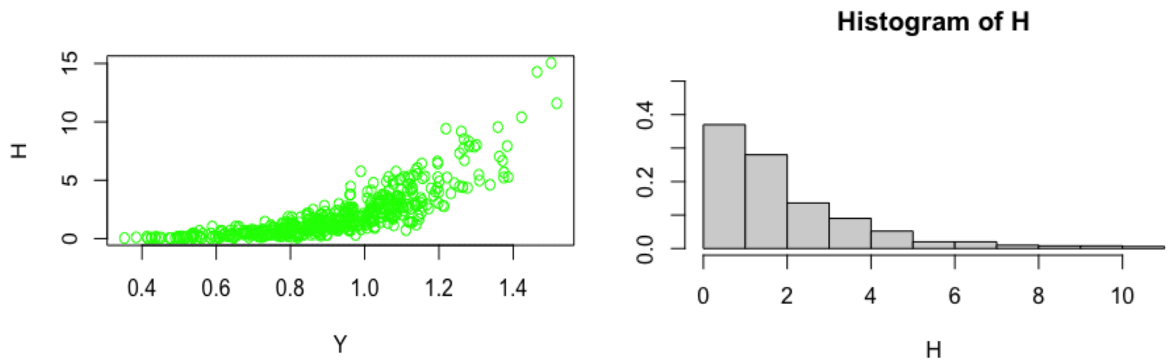


Figure 2: This figure illustrates health concentration curves with respect to Y for the marginal distributions of H and linear dependence between H and Y reported in Figure 1; see panels (a) and (b) of Figure 1.



Panel (a)



Panel (b)

Figure 3: This figure illustrates similar marginal distributions of H and different structures for the dependence between H and Y.

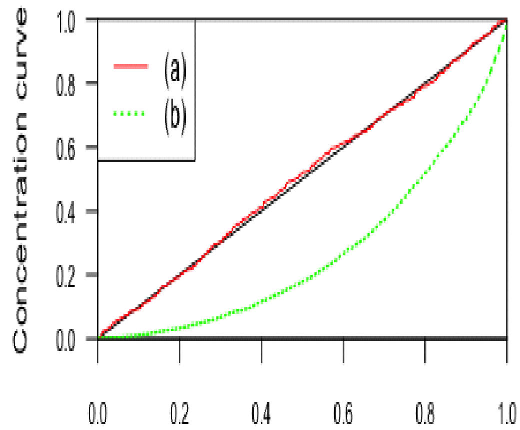
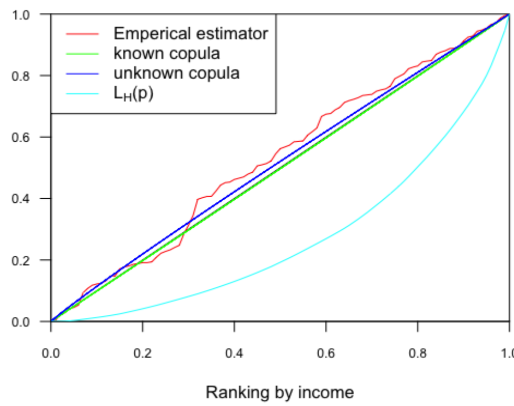
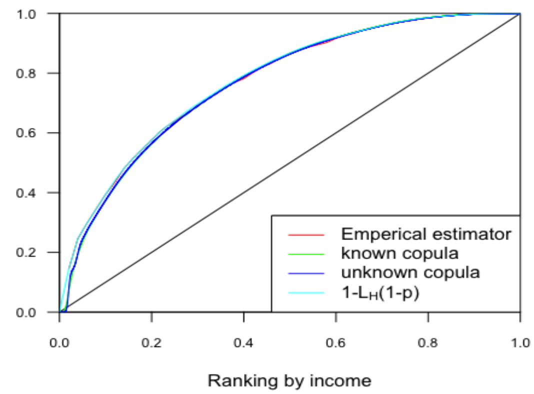


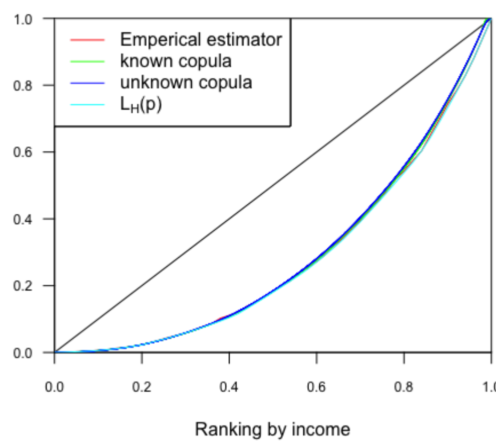
Figure 4: This figure illustrates health concentration curves with respect to Y for the marginal distributions of H and structures of dependence between H and Y reported in Figure 3; see panels (a) and (b) of Figure 3.



(a) H and Y are independent

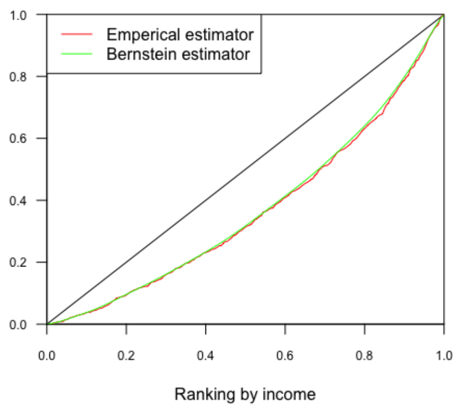


(b) Correlation between H and Y equal to -0.99

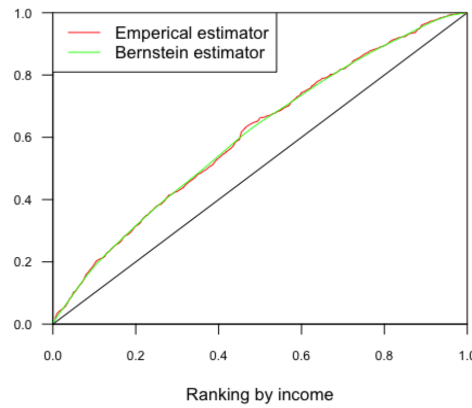


(c) Correlation between H and Y equal to 0.99

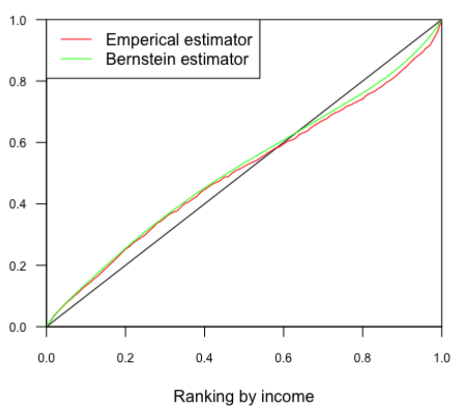
Figure 5: *This figure illustrates the copula-based semiparametric estimator and the empirical estimator of the health concentration curve with respect to Y $CH(p, Y)$ for different structures of dependence between H and Y.*



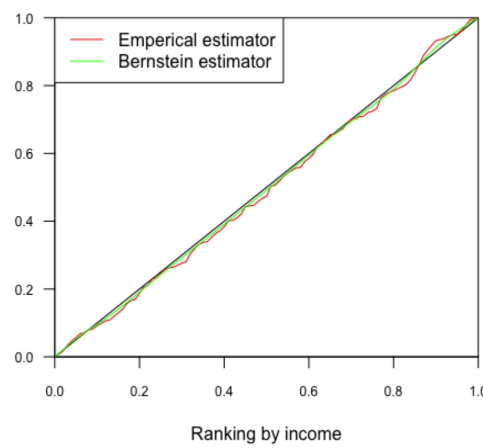
(a) Correlation between H and Y is positive



(b) Correlation between H and Y is negative



(c) Correlation between H and Y changes sign



(d) H and Y are independent

Figure 6: *This figure illustrates the Bernstein copula-based estimator and the empirical estimator of the health concentration curve with respect to Y $CH(p, Y)$ for different degrees of dependence between H and Y.*

Table 1: Integrated Mean Squared Error of Semiparametric ($\widehat{CH}$), Empirical ($\widehat{CH}_n$) and Nonparametric ($\widehat{CH}_{m,n}$) estimators of CH. Simulation results based on 1000 replicates of $10^3 \times$ IMSE of the concentration health estimators.

Copula	τ	$n = 50$			$n = 100$			$n = 200$			$n = 500$		
		$\widehat{CH}$	$\widehat{CH}_n$	$\widehat{CH}_{m,n}$	$\widehat{CH}$	$\widehat{CH}_n$	$\widehat{CH}_{m,n}$	$\widehat{CH}$	$\widehat{CH}_n$	$\widehat{CH}_{m,n}$	$\widehat{CH}$	$\widehat{CH}_n$	$\widehat{CH}_{m,n}$
Gaussian	-0.4	1.52405	2.37677	2.14225	0.81371	1.15701	1.06885	0.38623	0.58638	0.53311	0.12977	0.22836	0.22454
	0.001	2.10629	3.44384	1.88307	1.07944	1.68181	1.00998	0.50564	0.85069	0.49509	0.19713	0.32445	0.19456
	0.01	2.13191	3.38774	1.85736	1.06366	1.73533	0.96506	0.51634	0.82562	0.51147	0.20159	0.32712	0.20469
	0.1	1.79796	3.20303	1.85608	0.87136	1.62803	1.04206	0.41574	0.81374	0.51939	0.17935	0.31983	0.22626
	0.2	1.50939	3.10968	1.86665	0.79499	1.50698	0.97972	0.46025	0.77342	0.51299	0.18221	0.30089	0.25418
	0.5	1.45913	1.98083	2.04345	0.84333	1.06459	1.17494	0.36143	0.51315	0.54520	0.10408	0.19161	0.16909
	0.7	1.08239	1.22984	1.44666	0.64150	0.64480	0.85689	0.22934	0.31696	0.27078	0.08308	0.12031	0.11829
Student	-0.4	1.50998	2.41594	2.09419	0.91477	1.32553	1.14913	0.45419	0.64521	0.57807	0.15306	0.25293	0.23254
	0.001	2.69858	3.84649	2.17324	1.34693	1.89724	1.13122	0.66351	0.98423	0.62257	0.22396	0.40470	0.24812
	0.01	2.79426	3.90271	2.18947	1.37388	1.85705	1.17449	0.68979	1.00119	0.62375	0.21079	0.38427	0.25512
	0.1	2.34374	3.67801	2.05408	1.15942	1.84015	1.14661	0.52027	0.95107	0.60335	0.20992	0.38355	0.26273
	0.2	2.00519	3.63507	2.12335	1.04043	1.70332	1.15646	0.53195	0.86679	0.61781	0.20103	0.34301	0.28498
	0.5	1.41260	2.04994	2.02757	0.74987	1.02980	1.01611	0.36943	0.48307	0.44746	0.15646	0.2214	0.24387
	0.7	1.05205	1.27782	1.44762	0.57896	0.62840	0.75690	0.34523	0.36431	0.48286	0.12641	0.14386	0.21322
Clayton	0.001	2.09081	3.42189	1.87387	1.00368	1.62497	0.91140	0.53735	0.83230	0.50405	0.21267	0.34863	0.20759
	0.01	2.04309	3.39695	1.87465	1.06874	1.62283	1.01943	0.51462	0.84717	0.51277	0.21135	0.34748	0.21627
	0.1	1.95105	3.44810	1.91312	0.91821	1.61878	1.02497	0.38758	0.82406	0.50239	0.13228	0.32879	0.20838
	0.2	1.55803	3.16271	1.76030	0.76749	1.56814	1.00488	0.34245	0.79809	0.49569	0.13989	0.31672	0.21340
	0.5	1.05728	2.33026	1.46983	0.46329	1.13629	0.88551	0.27416	0.55912	0.45783	0.09539	0.24638	0.20293
	0.7	0.84169	1.63300	1.34834	0.42886	0.76142	0.66166	0.20152	0.40266	0.35363	0.07494	0.16145	0.15101
Gumbel	0.001	2.23206	3.52253	1.96451	1.10604	1.72069	1.01641	0.51837	0.83419	0.50979	0.22490	0.36063	0.21245
	0.01	2.17886	3.53497	1.87014	1.02629	1.63997	0.98007	0.54707	0.83069	0.49453	0.21463	0.34587	0.20453
	0.1	2.08629	3.47387	1.99008	1.11354	1.71672	1.15681	0.54978	0.89121	0.60417	0.21383	0.35531	0.27247
	0.2	2.06448	3.18206	2.31921	1.01851	1.53778	1.27509	0.56014	0.81521	0.70188	0.20091	0.30514	0.39143
	0.5	1.35059	1.81019	1.91188	0.77963	0.94400	1.04940	0.37581	0.46746	0.50481	0.13160	0.18465	0.17968
	0.7	1.08024	1.15756	1.47898	0.62124	0.64797	0.85754	0.31388	0.32589	0.40235	0.09029	0.11647	0.11790

Table 2: Summary statistics

	Mean	**Std Dev.**	**Min**	**Max**
Cases	0.098	0.0299	0.0029	0.375
Deaths	0.019	0.010	0.000	0.102
Population	106542	341079	66	10081570
Population density	214.77	778.96	0.000	17179
Poverty rate	0.143	0.058	0.023	0.547
Median income	27776	5850	8641	70390
Gini Index	0.444	0.035	0.302	0.609
Share of African Americans	0.089	0.144	0	0.872
Share of White Americans	0.756	0.203	0.006	1
Share of Asian Americans	0.013	0.028	0	0.417

Note: This table provides the descriptive statistics of COVID-19's infection and death rates and socioeconomic variables for 2777 counties of 45 U.S. states. Data on COVID-19's cases and deaths rates are obtained from U.S. Centers for Disease Control and Protection (CC) and covers the period until June 10, 2021. Demographic variables are drawn primarily from the US Census and CDC.

Table 3: Correlation coefficients between socio-economic variables

	Density	**Median income**	**Population**	**Poverty rate**
Median income	0.296			
Population	0.366	0.227		
Poverty rate	-0.052	-0.742	-0.076	
Share of African Americans	0.128	-0.263	0.057	0.441
Share of White Americans	-0.185	0.194	-0.224	-0.475

Note: The table shows the correlation coefficients matrix of socioeconomic variables we use in our empirical analysis. These variables are described in Table 2.

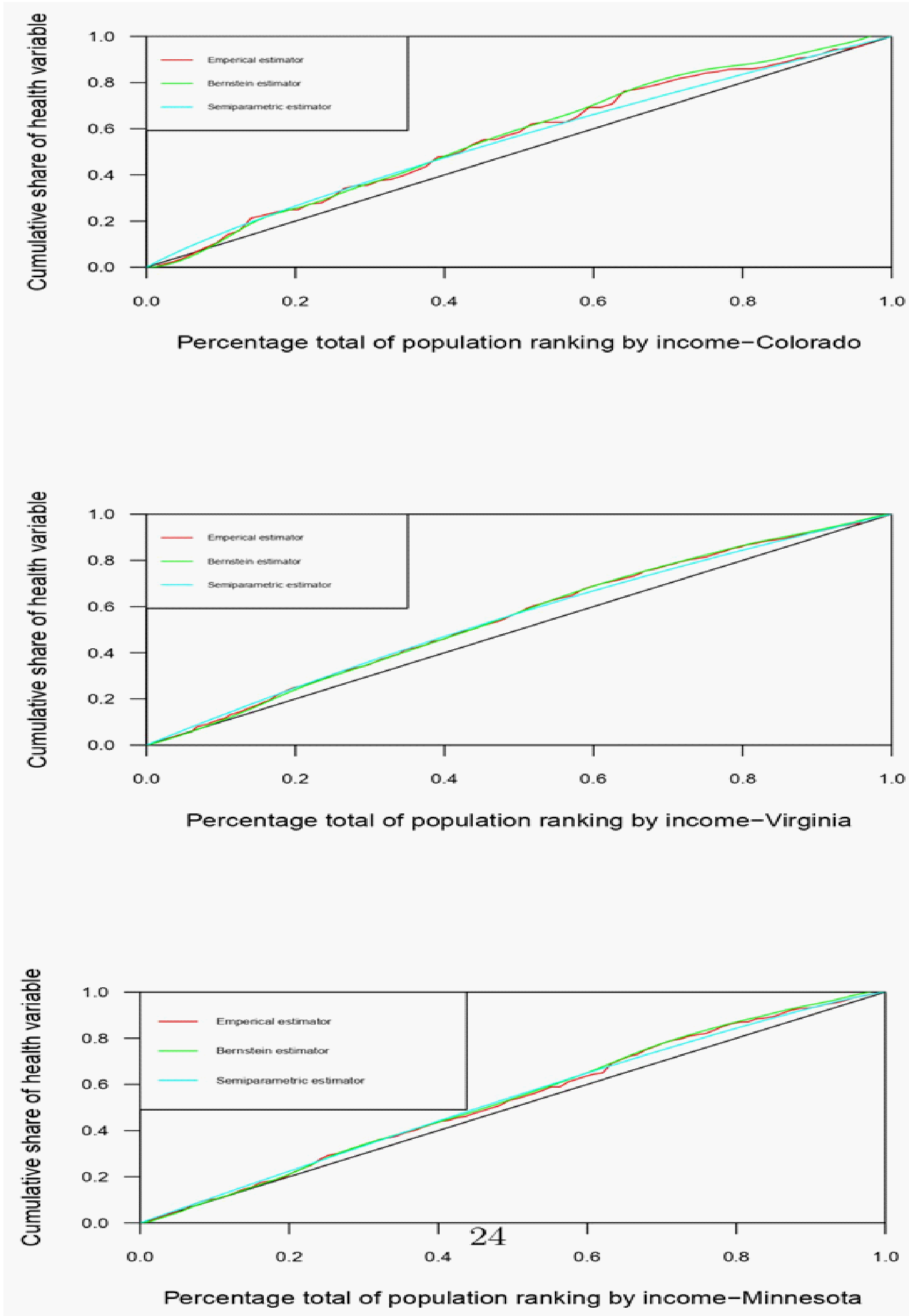


Figure 7: The empirical estimator, the semiparametric and the nonparametric estimator of the health concentration curve with respect to socioeconomic variables for COVID-19 deaths. High income states.

Table 4: Regressions for COVID1-19 deaths

Dependent variable: Log(COVID-19's deaths)			
	(1)	**(2)**	**(3)**
Constant	27.73*** (2.60)	7.59 (5.94)	−0.83 (5.7)
Gini index	24.83*** (3.40)	9.00** (4.30)	6.06*** (3.83)
Gini health index	2.15 (2.04)	2.24 (1.89)	−0.35 (1.81)
Population		0.10*** (0.01)	0.15*** (0.02)
Share of African Americans		−0.03 (1.48)	−4.33** (2.12)
Share of Asian Americans		−7.63*** (2.07)	4.56*** (1.61)
Share of Hispanic Americans		−4.31*** (1.50)	−10.5*** (5.37)
Share of women		0.0004 (0.005)	0.04 (0.002)
Urban area (%)		0.04*** (0.01)	0.05*** (0.01)
Median income		0.0008 (0.0006)	0.0015** (0.0005)
Poverty rate		19.18** (11.2)	5.24*** (1.51)
Share of Black women			0.13* (0.06)
Share of women in poverty			0.049** (0.23)
Share of Hispanic individuals in poverty			3.37 (3.12)
Share of Asian individuals in poverty			−5.57*** (1.7)

Note: This table reports the estimation results of regressing state-level COVID-19's death [log of COVID-19's deaths] on state's population size, full set of socioeconomic variables under consideration, and several interaction terms as described in the main text. Standard errors in parentheses. *p<0.1; **p<0.05; ***p<0.01.

Table 5: Regressions for COVID1-19 infections

Dependent variable: Log(COVID-19's cases)			
	(1)	**(2)**	**(3)**
Constant	27.19*** (2.57)	11.31** (4.98)	5.28 (5.12)
Gini index	18.75*** (3.36)	4.55 (3.61)	2.33*** (3.44)
Gini health index	2.37 (2.01)	3.08** (1.58)	1.27 (1.62)
Population		0.10*** (0.01)	0.14*** (0.02)
Share of African Americans		−0.16 (1.24)	−3.94** (1.90)
Share of Asian Americans		−8.20*** (1.73)	25.3*** (12.5)
Share of Hispanic Americans		−3.85*** (1.26)	−6.22 (4.18)
Share of women		0.0003 (0.004)	0.03 (0.002)
Urban area (%)		0.03*** (0.009)	0.04*** (0.009)
Median income		0.00002 (0.0009)	0.0007 (0.0005)
Poverty rate		10.03 (9.46)	34.59** (13.6)
Share of Black women			0.12** (0.06)
Share of women in poverty			0.37** (0.12)
Share of Hispanic individuals in poverty			11.9 (2.79)
Share of Asian individuals in poverty			−3.51*** (1.52)

Note: This table reports the estimation results of regressing state-level COVID-19's infections [log of COVID-19's cases] on state's population size, full set of socioeconomic variables under consideration, and several interaction terms as described in the main text. Standard errors in parentheses. *p<0.1; **p<0.05; ***p<0.01.

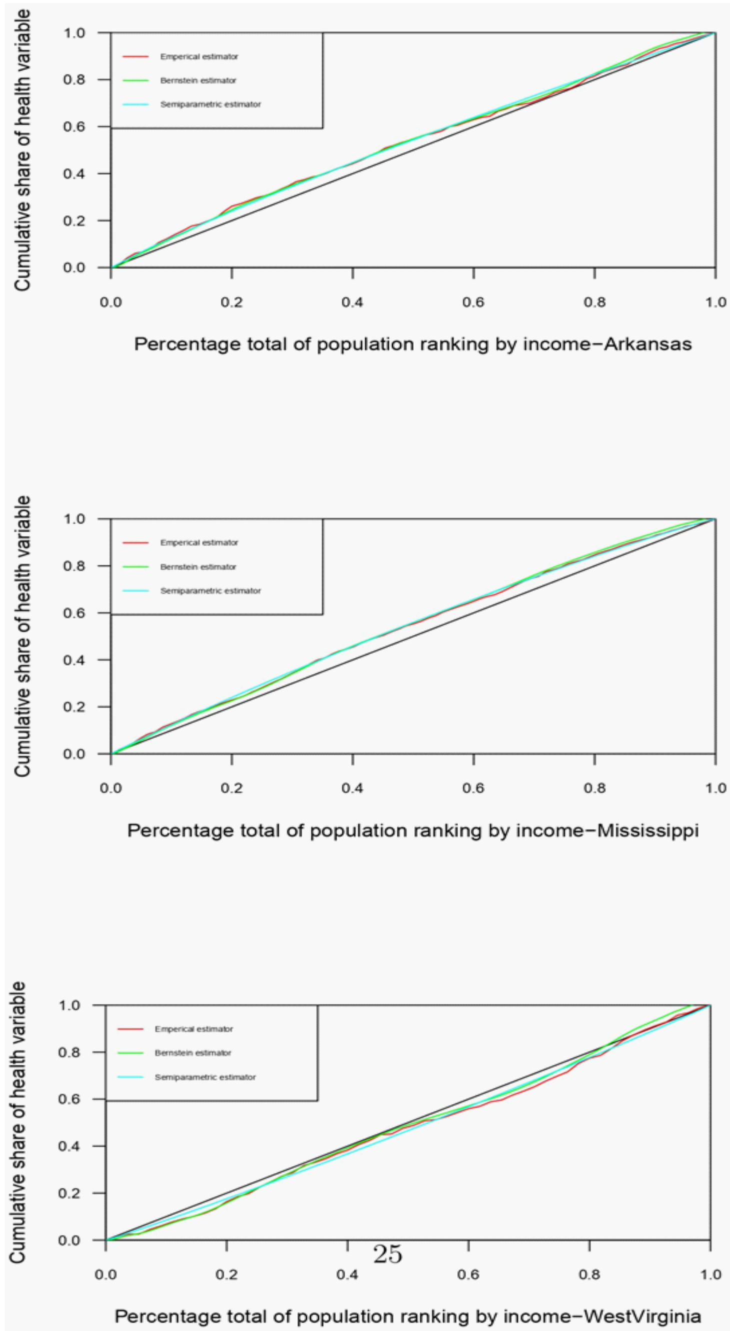


Figure 8: The empirical estimator, the semiparametric and the nonparametric estimator of the health concentration curve with respect of socioeconomic variable for COVID-19 deaths. Low income states.

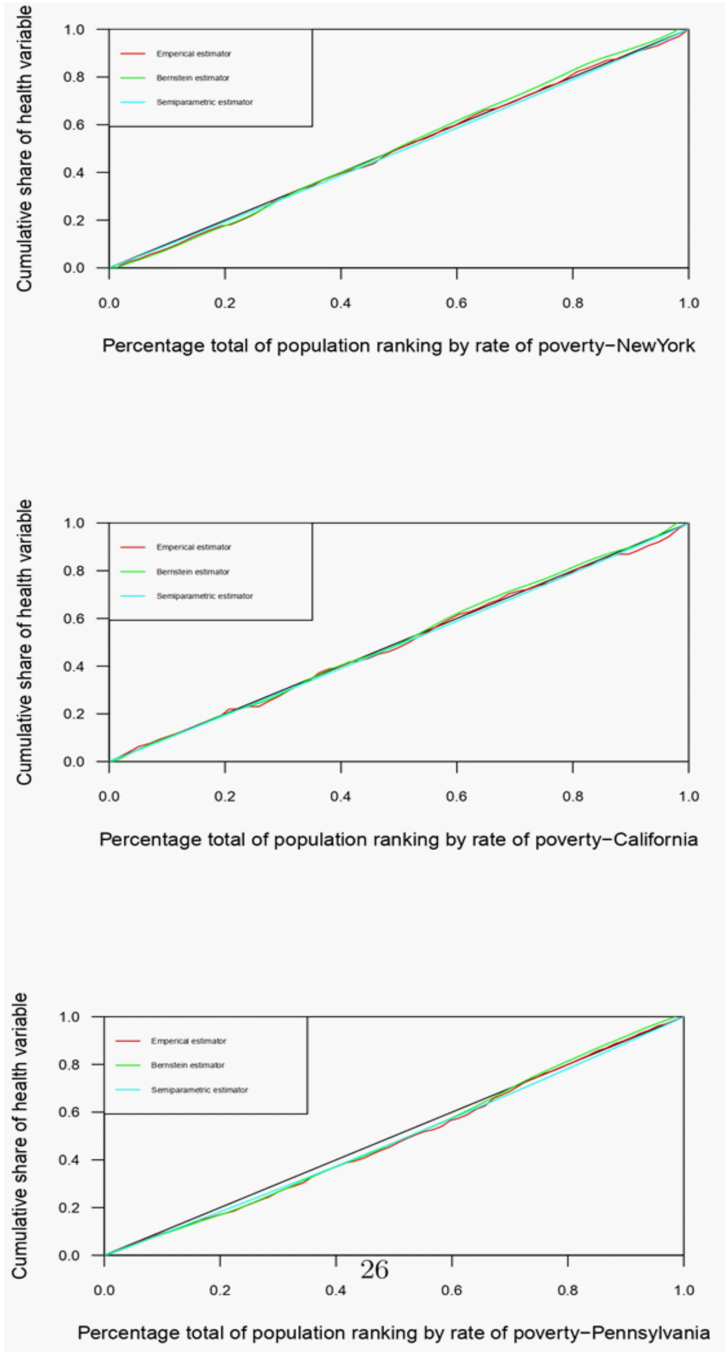


Figure 9: The empirical estimator, the semiparametric and the nonparametric estimator of the health concentration curve with respect of socioeconomic variable for COVID-19 deaths. Average income states.

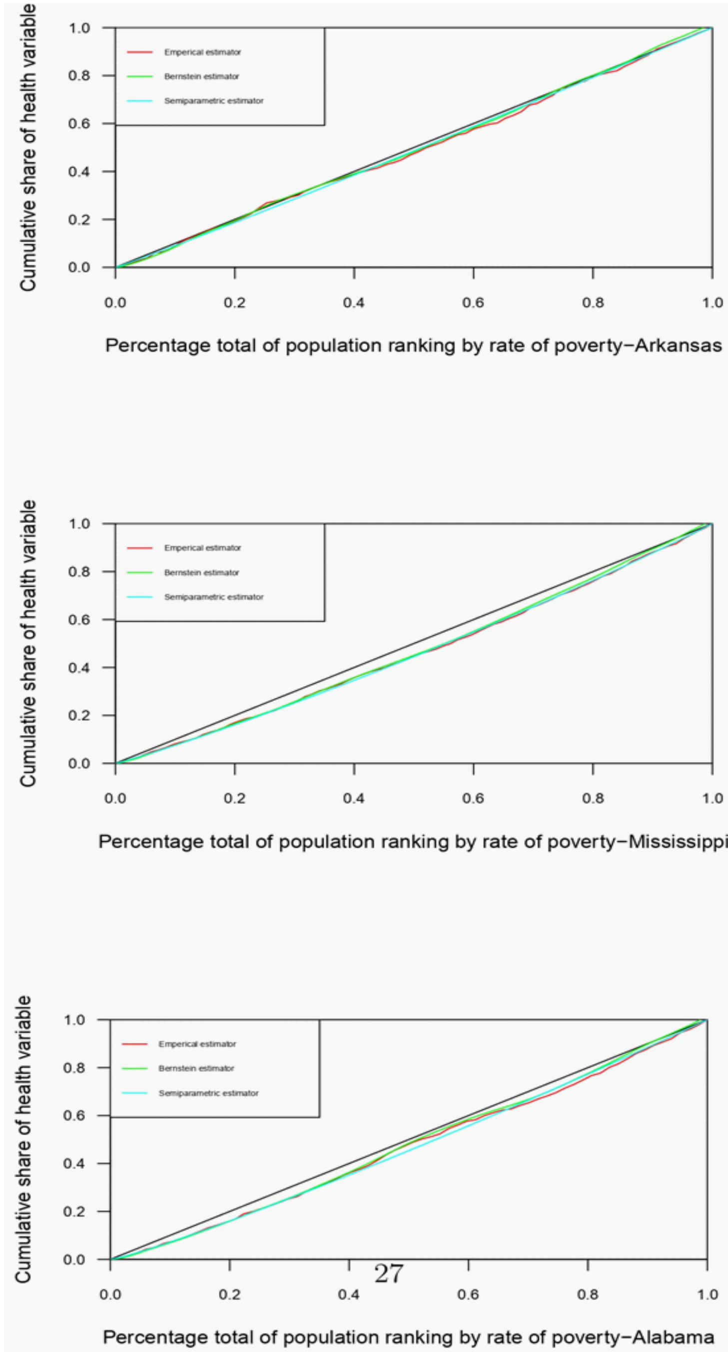


Figure 10: The empirical estimator, the semiparametric and the nonparametric estimator of the health concentration curve with respect of socioeconomic variable for COVID-19 deaths. High poverty rate states.

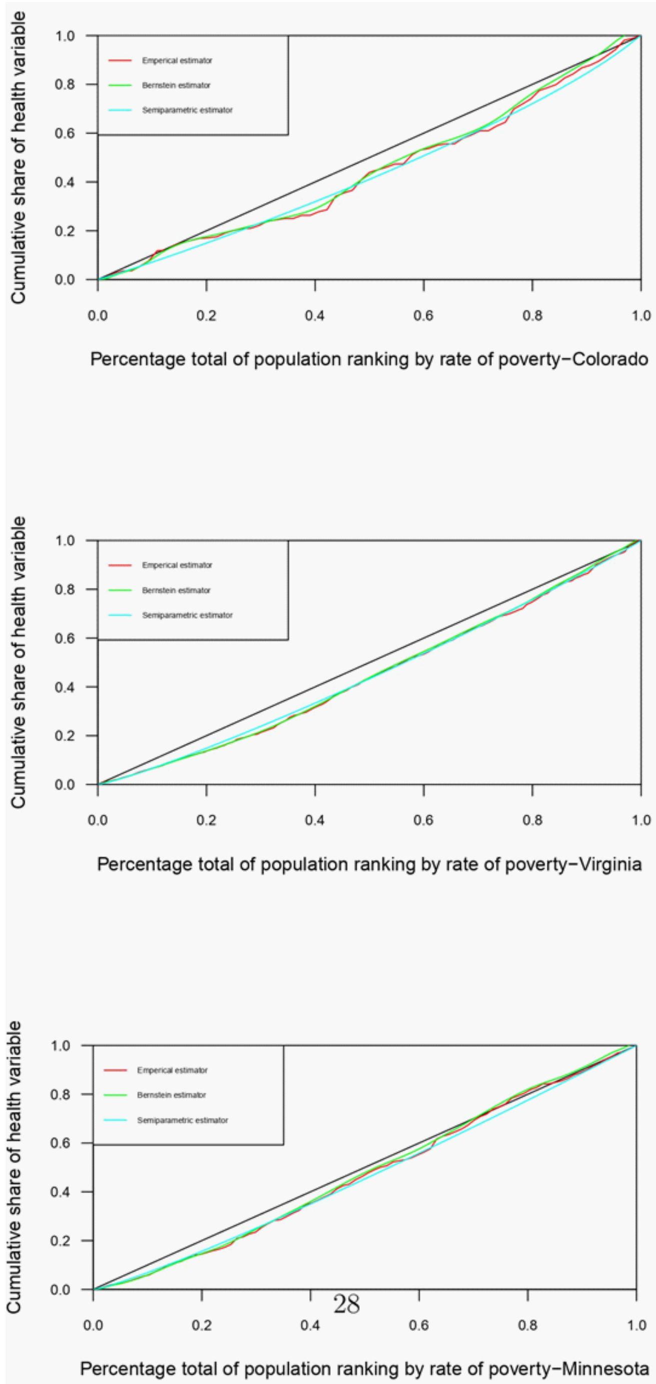


Figure 11: The empirical estimator, the semiparametric and the nonparametric estimator of the health concentration curve with respect of socioeconomic variable for COVID-19 deaths. Low poverty rate states.

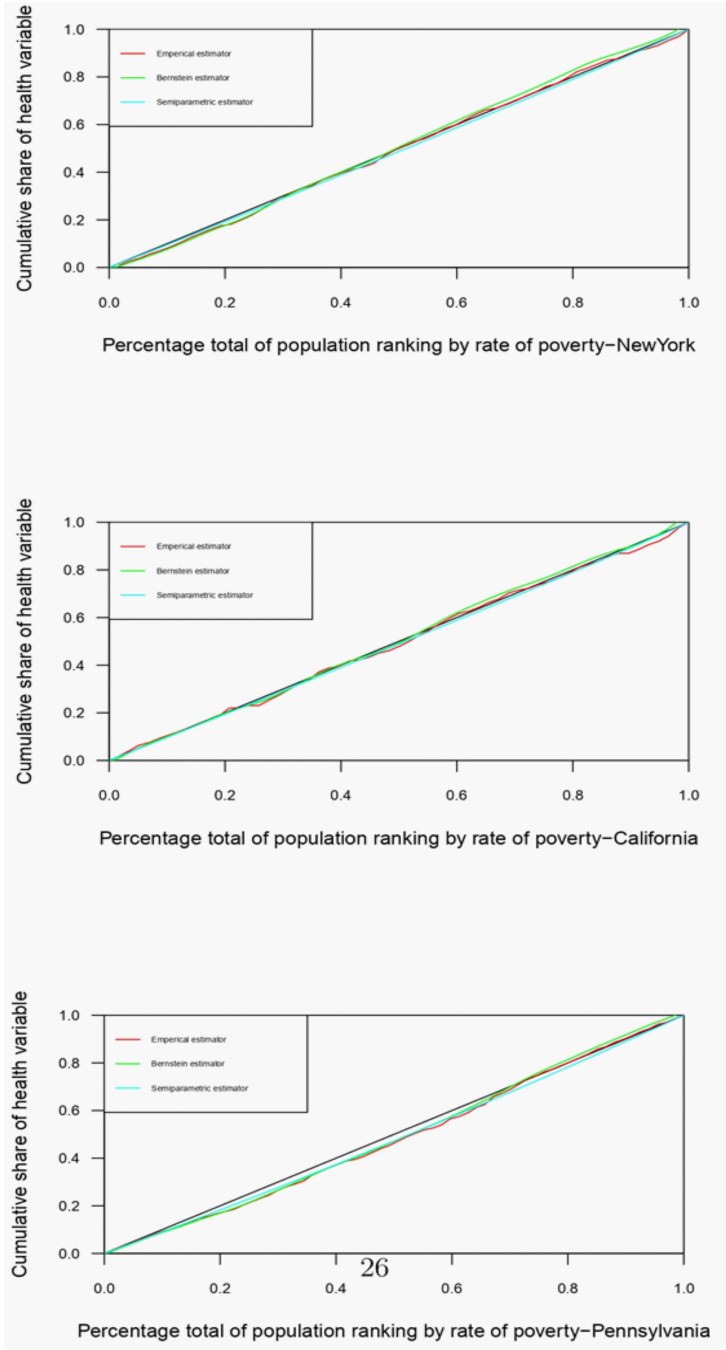


Figure 12: The empirical estimator, the semiparametric and the nonparametric estimator of the health concentration curve with respect of socioeconomic variable for COVID-19 deaths. Average poverty rate states.

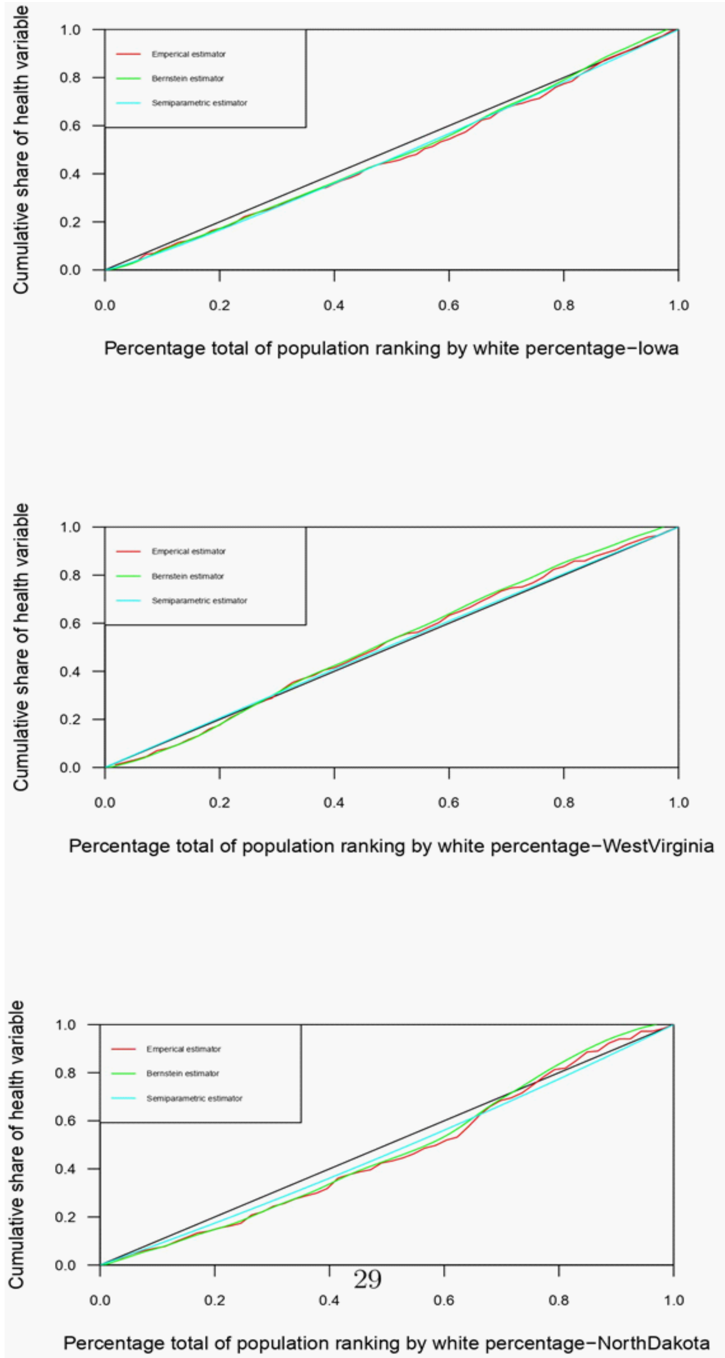


Figure 13: The empirical estimator, the semiparametric and the nonparametric estimator of the health concentration curve with respect of socioeconomic variable for COVID-19 deaths. High rate of white people states.

Figure 14: The empirical estimator, the semiparametric and the nonparametric estimator of the health concentration curve with respect of socioeconomic variable for COVID-19 deaths. Low rate of white people states.

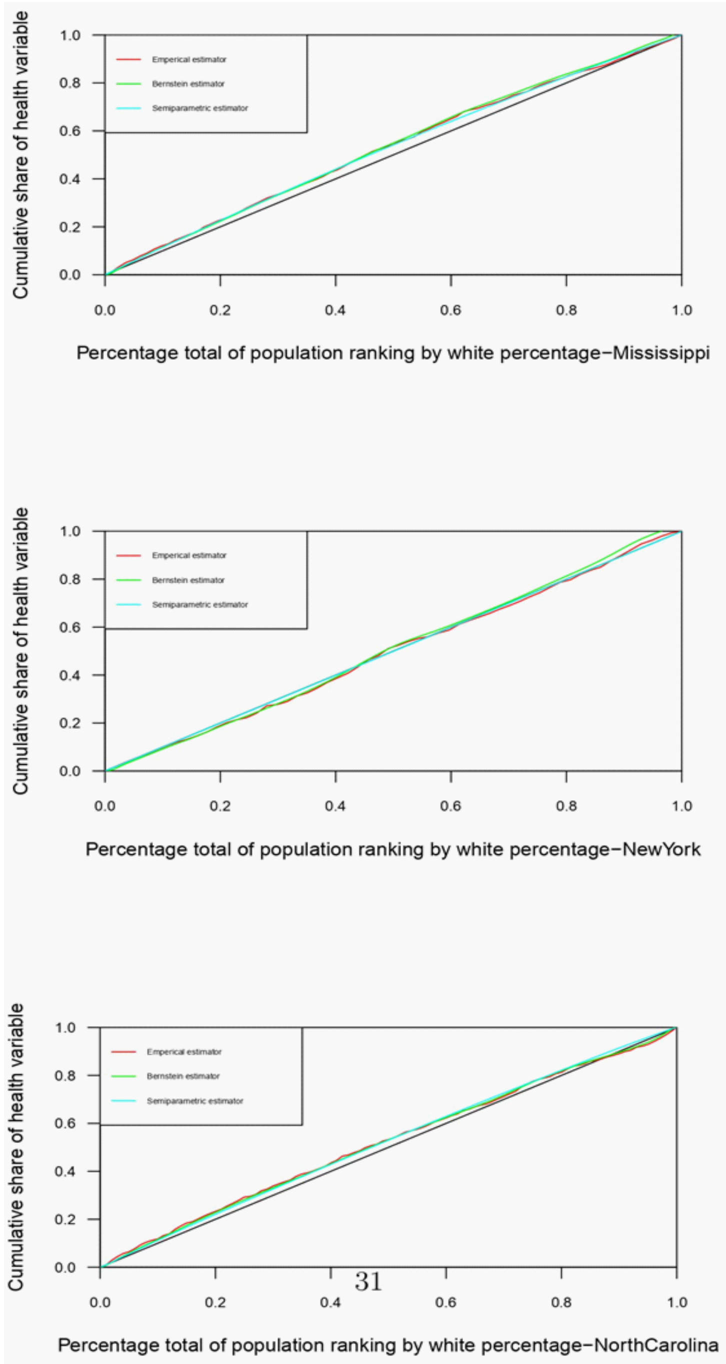
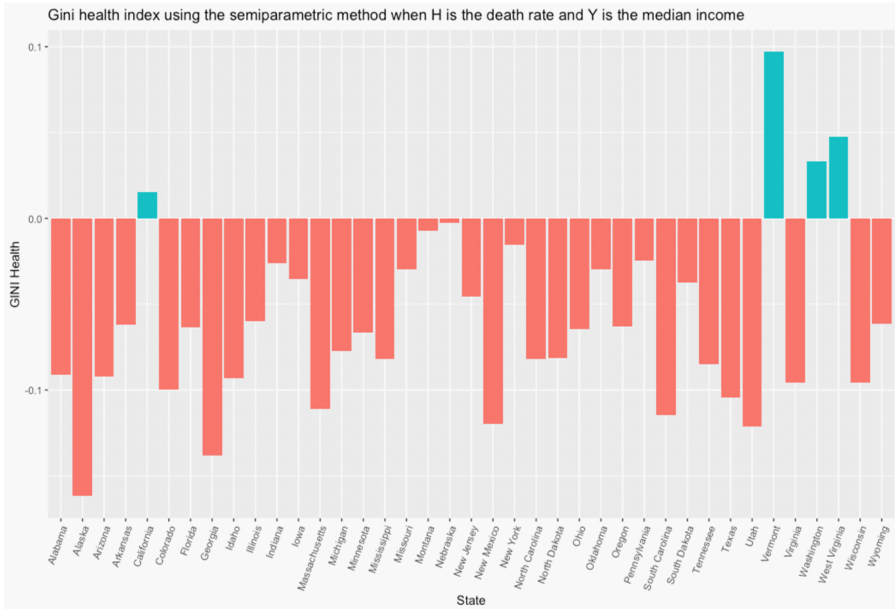
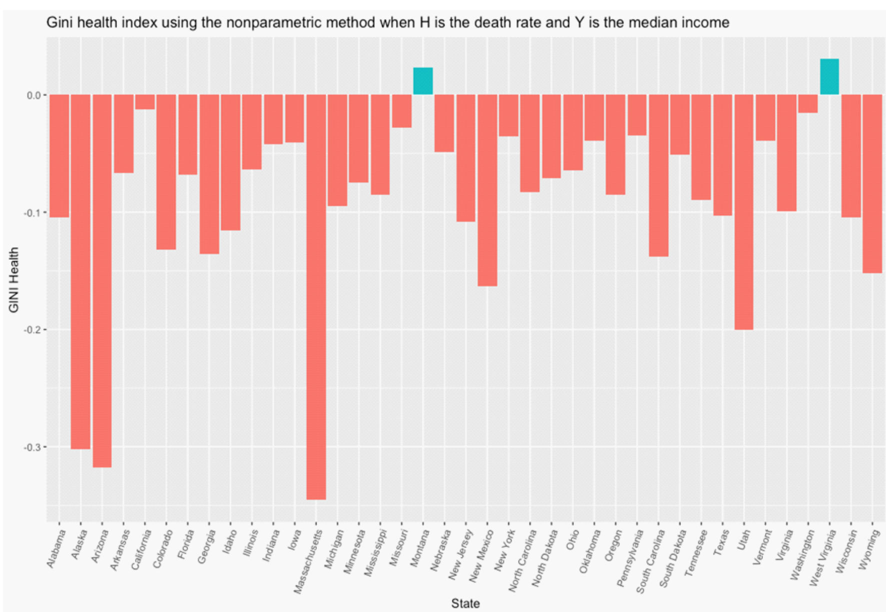


Figure 15: The empirical estimator, the semiparametric and the nonparametric estimator of the health concentration curve with respect of socioeconomic variable for COVID-19 deaths. Average rate of white people states.

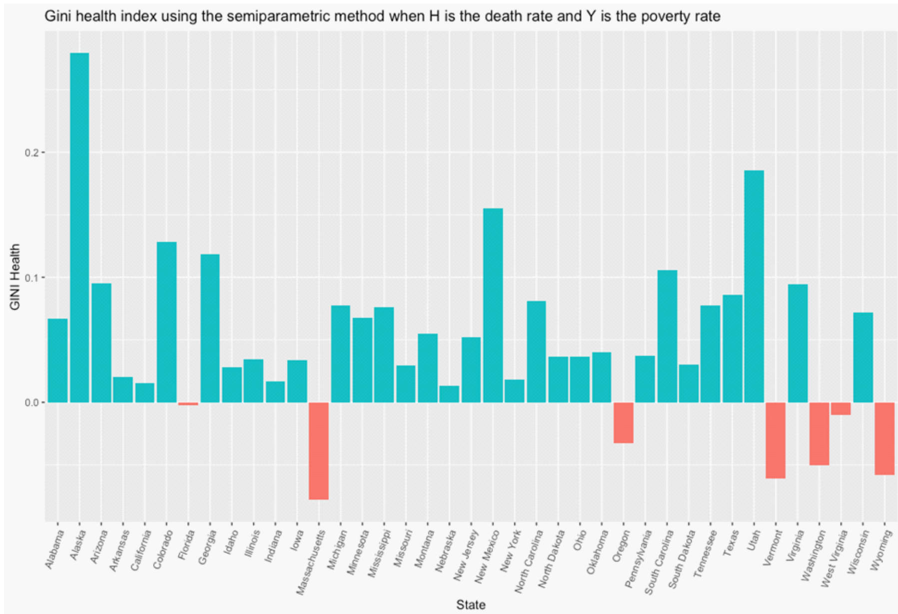


(a) Semi-parametric Estimator

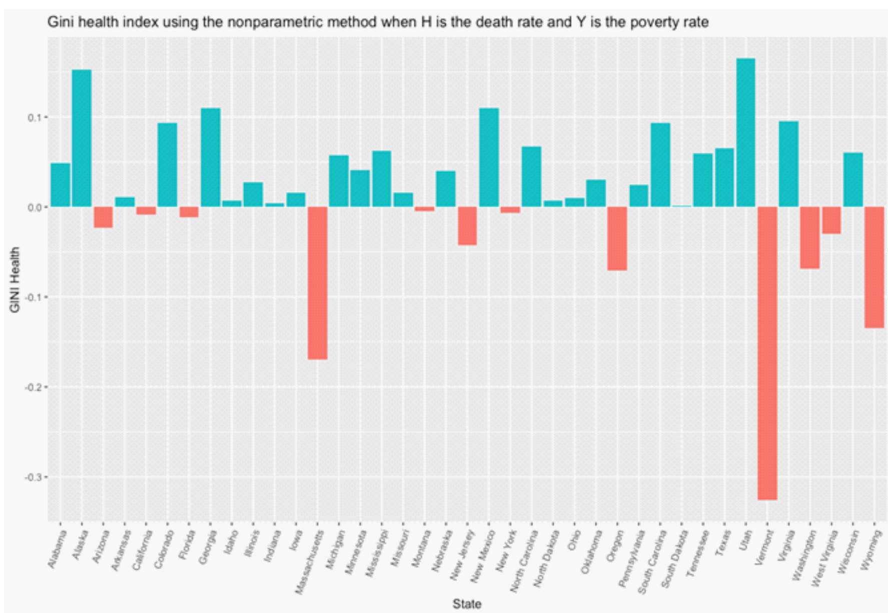


(b) Non-parametric Estimator

Figure 16: *The figure reports the semi/non-parametric estimators of Gini health index for 40 U.S. states when the health variable H represents COVID-19's death rates and the socioeconomic variable Y is the median income.*

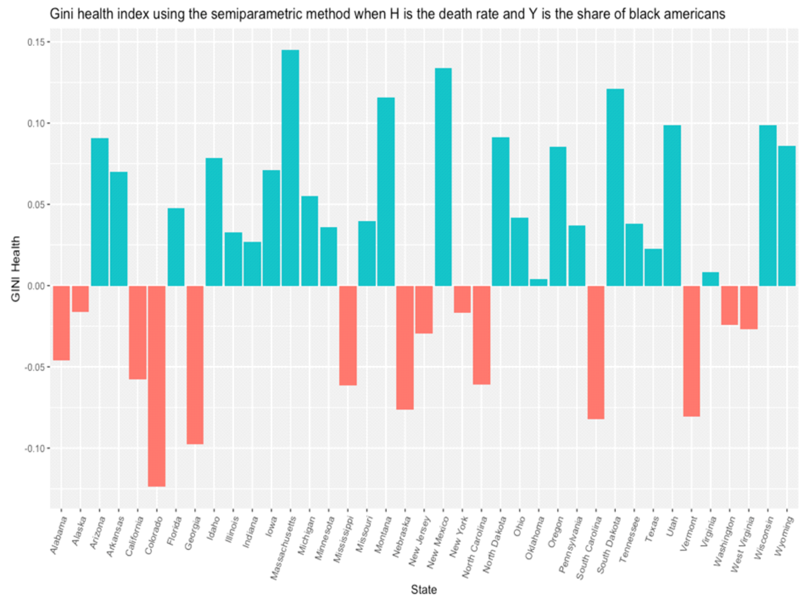


(a) Semi-parametric Estimator

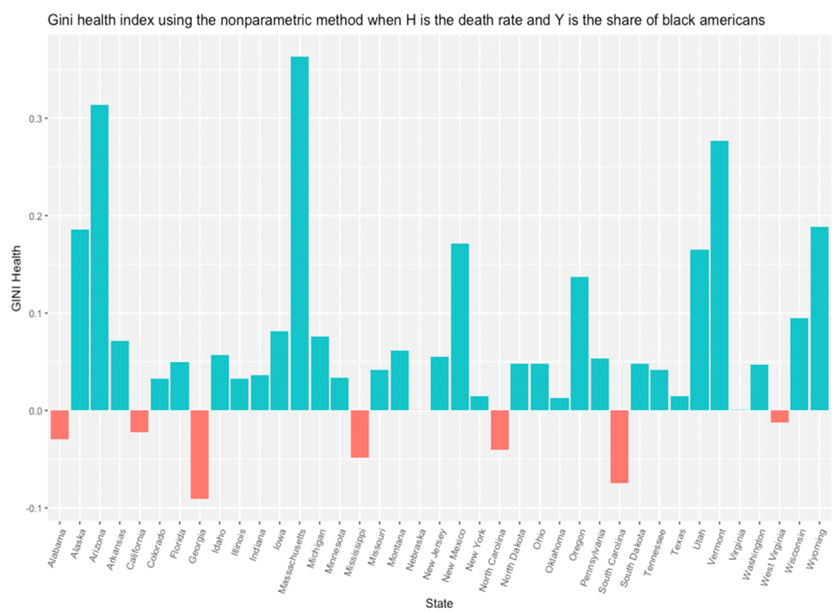


(b) Non-parametric Estimator

Figure 17: *The figure reports the semi/non-parametric estimators of Gini health index for 40 U.S. states when the health variable H represents COVID-19's death rates and the socioeconomic variable Y is the poverty rate.*



(a) Semi-parametric Estimator



(b) Non-parametric Estimator

Figure 18: *The figure reports the semi/non-parametric estimators of Gini health index for 40 U.S. states when the health variable H represents COVID-19's death rates and the socioeconomic variable Y is the share of African Americans.*

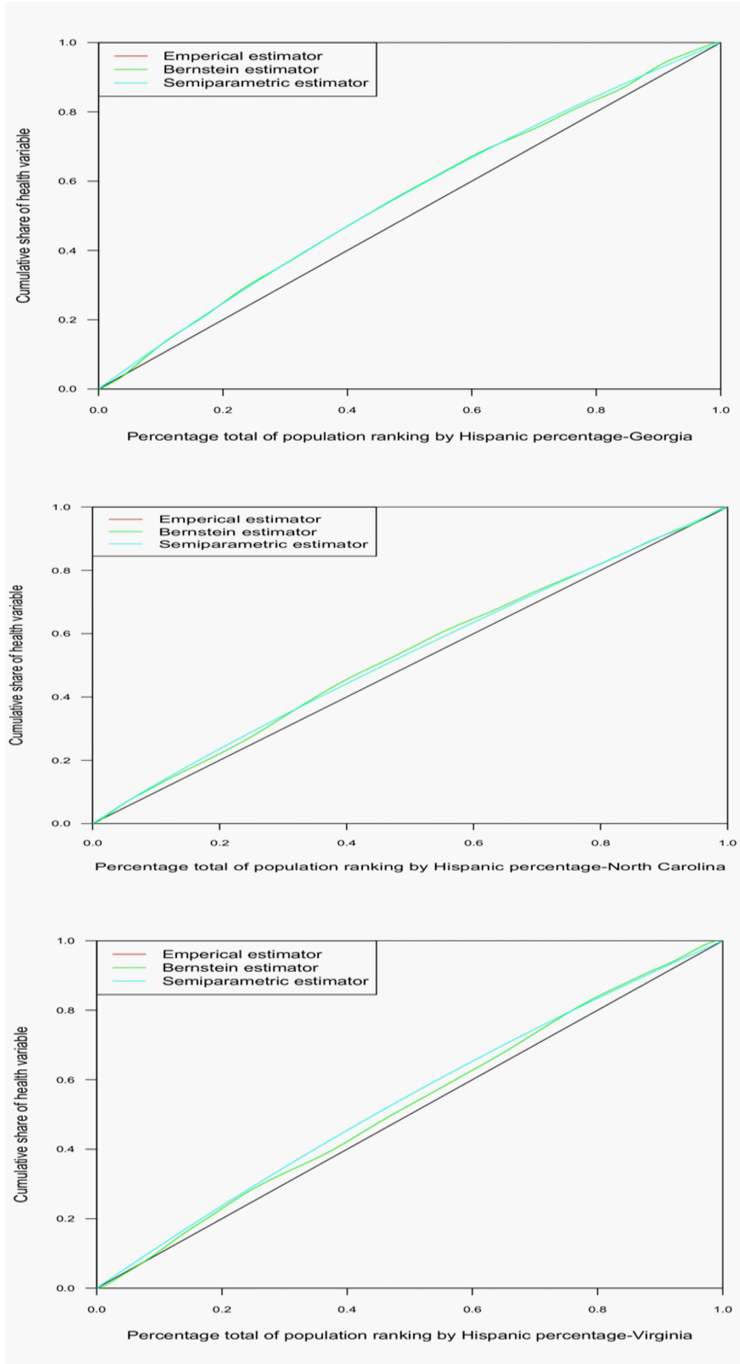


Figure 19: The empirical estimator, the semiparametric and the nonparametric estimator of the health concentration curve with respect of socioeconomic variable for COVID-19 deaths. Average rate of Hispanic population states.

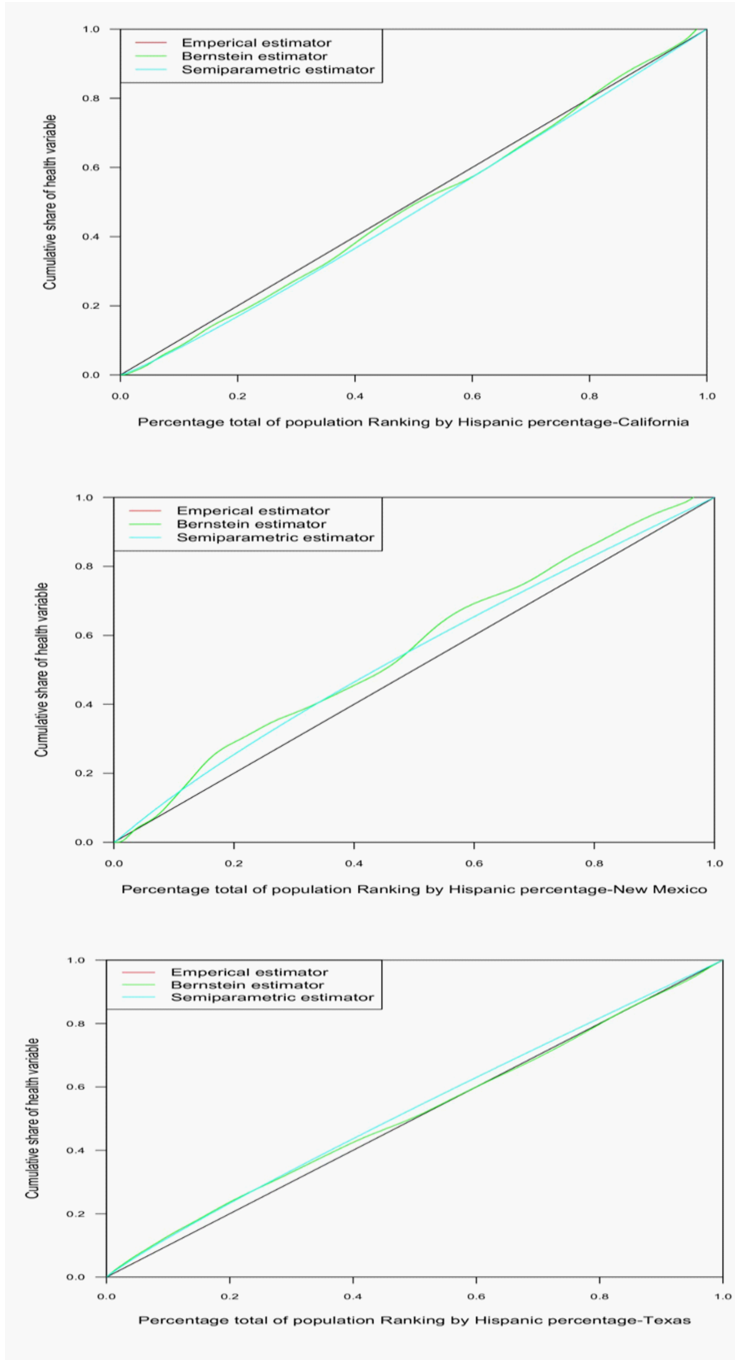


Figure 20: The empirical estimator, the semiparametric and the nonparametric estimator of the health concentration curve with respect of socioeconomic variable for COVID-19 deaths. High rate of Hispanic population states).

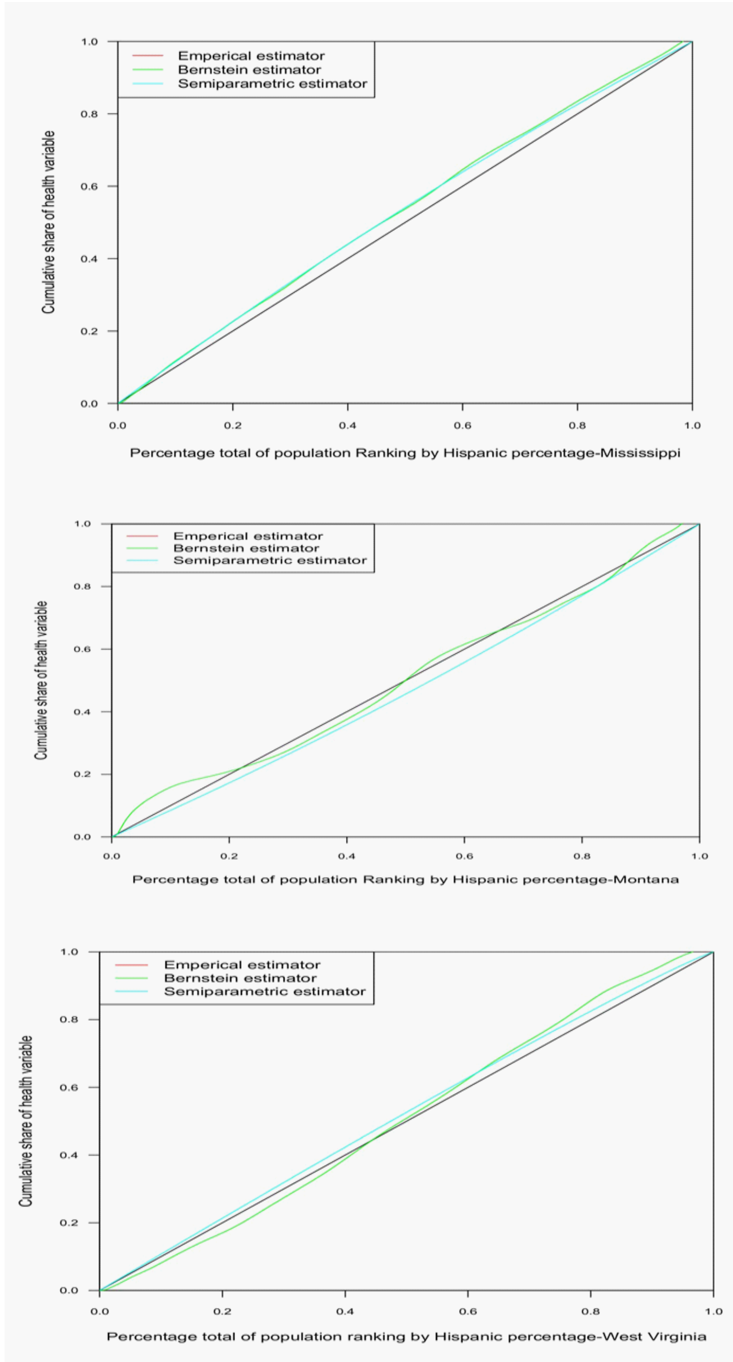


Figure 21: The empirical estimator, the semiparametric and the nonparametric estimator of the health concentration curve with respect of socioeconomic variable for COVID-19 deaths. Low rate of Hispanic population states.

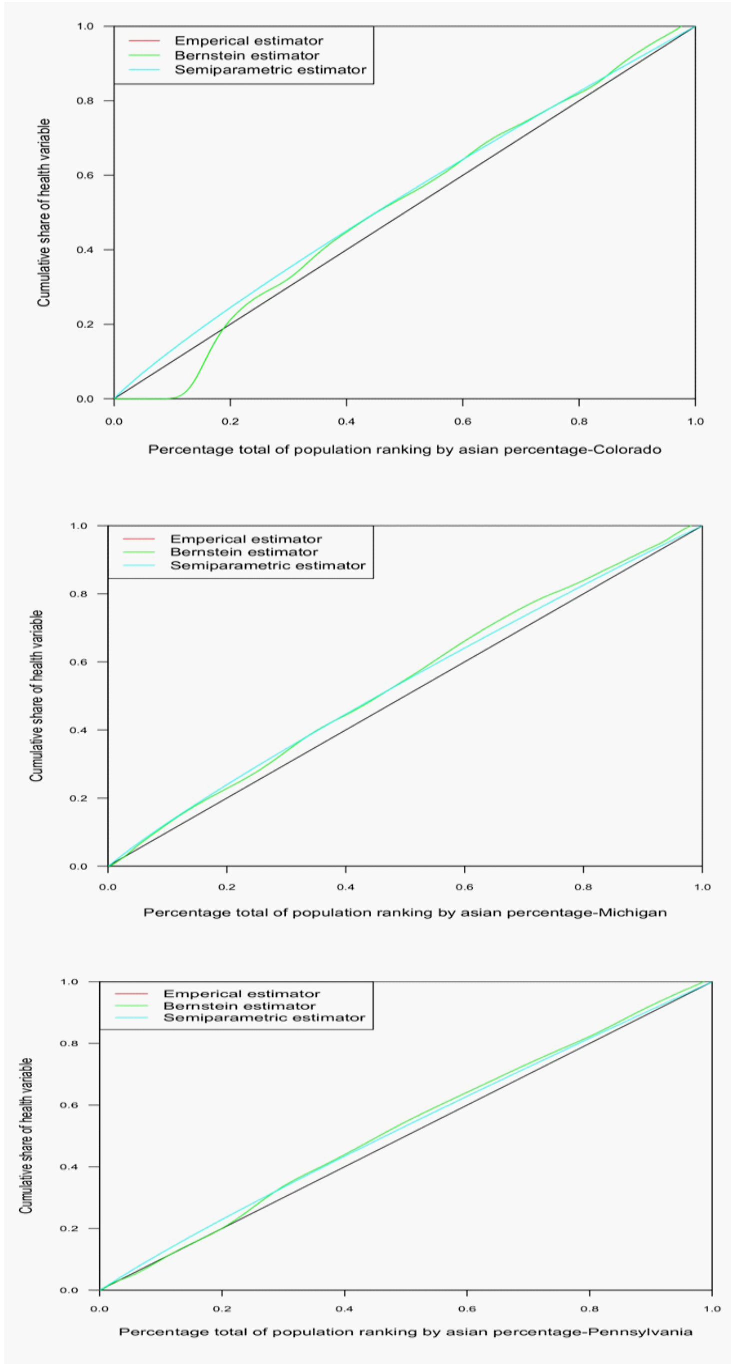


Figure 22: The empirical estimator, the semiparametric and the nonparametric estimator of the health concentration curve with respect of socioeconomic variable for COVID-19 deaths. Average rate of Asian population states.

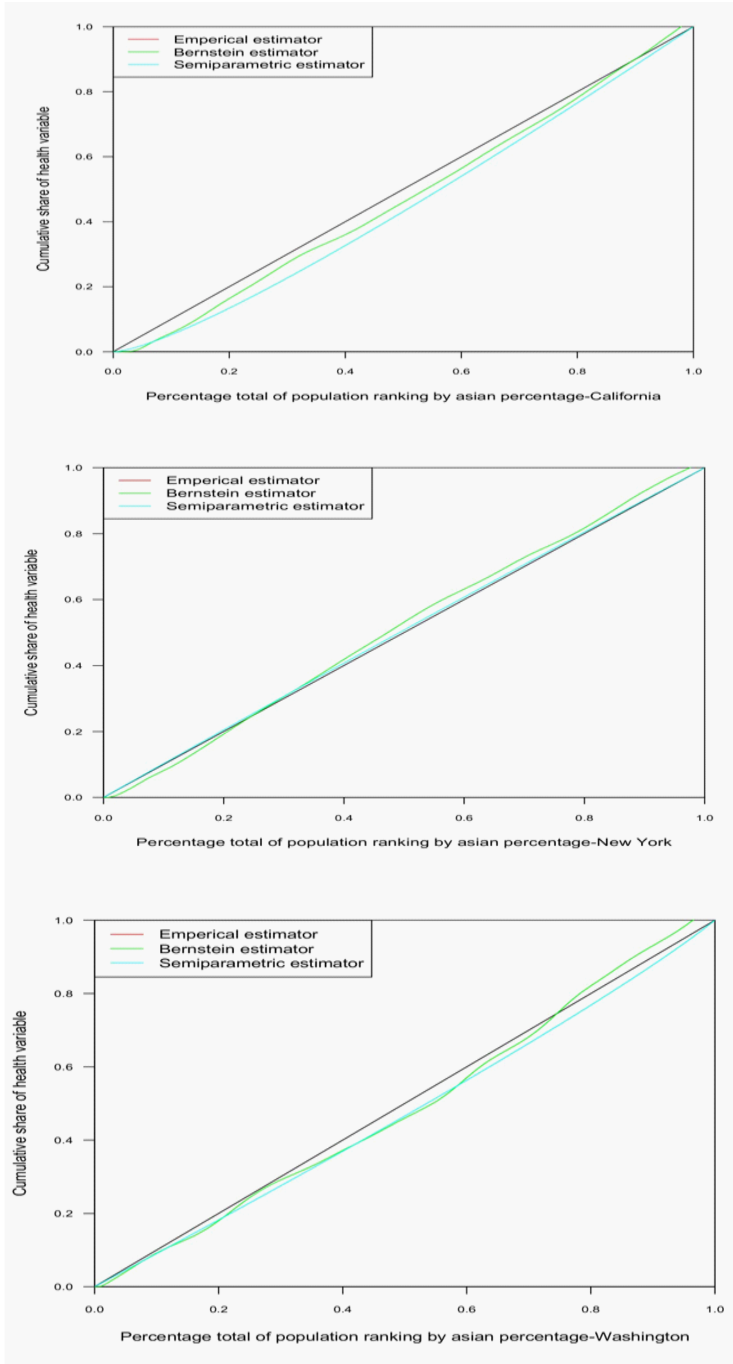


Figure 23: The empirical estimator, the semiparametric and the nonparametric estimator of the health concentration curve with respect of socioeconomic variable for COVID-19 deaths. High rate of Asian population states.

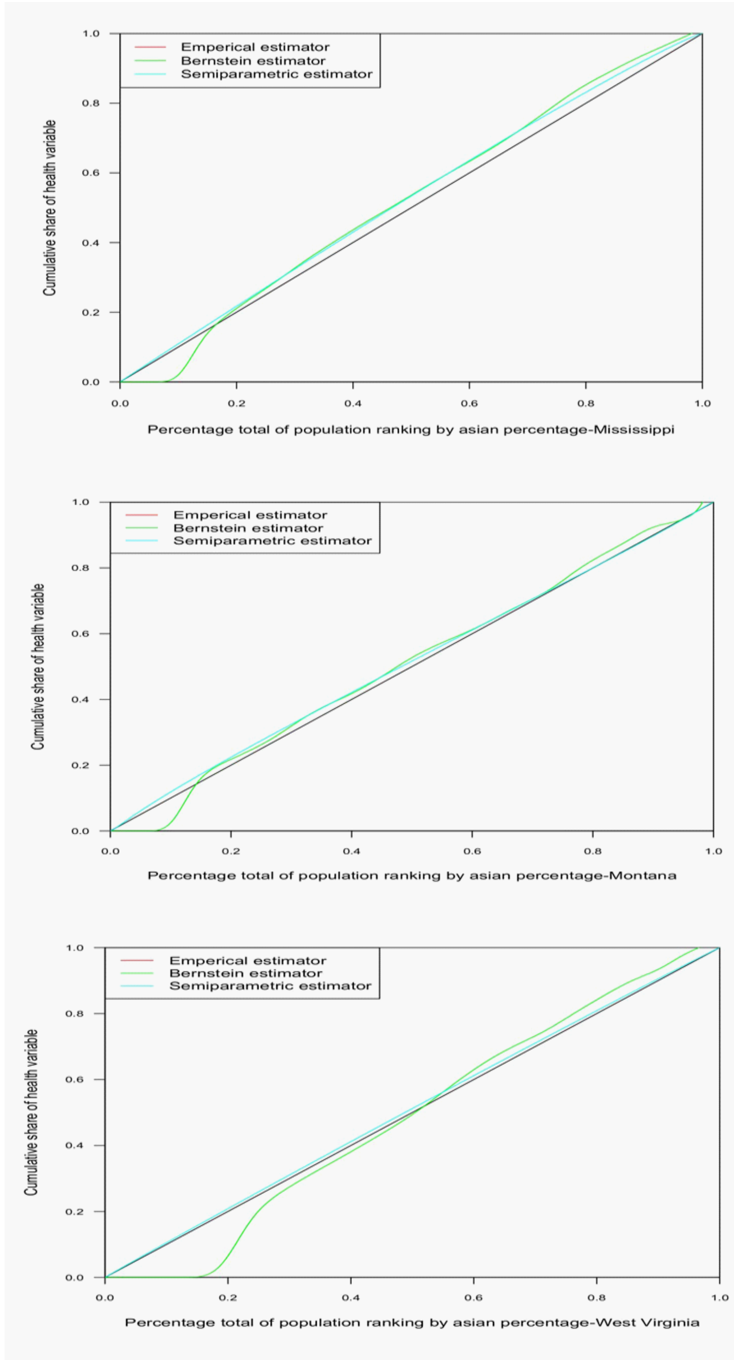


Figure 24: The empirical estimator, the semiparametric and the nonparametric estimator of the health concentration curve with respect of socioeconomic variable for COVID-19 deaths. Low rate of Asian population states.