A short note revisiting the concentration index: Does the normalization of the concentration index matter?

John E. Ataguba1,2

1Department of Community Health Sciences, Max Rady College of Medicine, Rady Faculty of Health Sciences, University of Manitoba, Winnipeg, Manitoba, Canada
2Health Economics Unit, School of Public Health and Family Medicine, Health Sciences Faculty, University of Cape Town, Observatory, South Africa

Abstract

The concentration index, including its normalization, is prominently used to assess socioeconomic inequalities in health and health care. Wagstaff’s and Erreygers’ normalizations or corrections of the standard concentration index are the most suggested approaches when analyzing binary health variables encountered in many health economics and health services research. In empirical applications of the corrected or normalized concentration indices, researchers interpret them similarly to the standard concentration index, which may be problematic as this ignores their underlying behaviors. This paper shows that the empirical bounds of the standard concentration index, including the corrected indices, depend not only on the sample size directly but also on the sampling weight. Notably, the paper highlights critical challenges for assessing and interpreting the popular Wagstaff’s and Erreygers’ corrected concentration indices with binary health variables. Specifically, it shows that it might be misleading, for example, to assess socioeconomic health inequalities using the magnitude of the “symmetric” Erreygers’ corrected concentration index in the face of progressive improvements in the binary health variable. Also, Wagstaff’s normalized concentration index may give a spurious “concentration” of the binary health variable among the rich or the poor in certain rare instances.

Keywords

binary health variable, concentration index, Erreygers’ normalization, socioeconomic health inequality, Wagstaff’s normalization

1 | INTRODUCTION

The concentration index is commonly used to assess socioeconomic inequality in health and health care (Ataguba et al., 2011; Kakwani et al., 1997; van Doorslaer & Koolman, 2004; Wagstaff, 2005; Wagstaff et al., 1991). The theoretical values of the standard concentration index range between −1 (a case where the health variable (e.g., obesity) is concentrated on the most disadvantaged individual) and +1 (a case where it is concentrated on the most advantaged individual). The standard concentration index is positive when ill-health (or health) is more prevalent among wealthier groups and negative if otherwise. Its magnitude conveys the relative degree of concentration among poorer or richer groups. For binary variables that are common in health economics research, Wagstaff’s (2005) seminal paper showed that the range of values for the standard concentration index depends on the mean (i.e., the proportion) of the variable ($\mu_h$) and with a large sample, will be between ($\mu_h - 1$) and...
(1 − μ_h) instead of −1 and +1 for the lower and upper bounds, respectively. In general, Erreygers (2009, p.506) shows that the bounds are not unique to binary health variables as researchers can construct different upper and lower bounds for any "health variable with a finite upper value or a positive lower value.” A debate on how best to “adjust” the standard concentration index continues, mainly when, for example, a binary health variable is used (Erreygers, 2009; Wagstaff, 2005) or when finitely bounded health variables are used (Erreygers, 2009; Erreygers & Van Ourti, 2011).

This paper focuses on binary health variables often encountered in health economics research. It lays out crucial issues that researchers using the standard concentration index, Wagstaff’s and Erreygers’ normalized concentration indices may not fully internalize, especially when interpreting or comparing results. It begins by showing that the standard concentration index’s empirical bounds are determined by the sample size (something that Erreygers (2009) highlights) and by the sampling structure, especially the weight variable. It also highlights some implications of Wagstaff’s and Erreygers’ normalization for interpretation and policy.

2 EXAMINING THE LOWER AND UPPER BOUNDS FOR THE WAGSTAFF’S AND ERREYGERS’ CORRECTED CONCENTRATION INDICES

Although this is trivial in many cases, especially when the sample size is significantly large, the empirical bounds of the standard concentration index are not the same as the theoretical bounds. It is known that ignoring the sampling structure, especially the sample weights, the standard concentration index’s empirical bounds are −(n − 1)/n and (n − 1)/n, where n is the sample size (Erreygers, 2009). Also, ties in the measure of socioeconomic status could affect the estimate of the concentration index (Chen & Roy, 2009). This piece notes that the empirical bounds of the concentration index are a function of both the fractional rank (often used in empirical estimation; Kakwani et al., 1997) and the sample size.

Let us write the standard concentration index (CI) as (Kakwani et al., 1997):

\[
CI = \frac{2}{n\mu_h} \sum_{i=1}^{n} h_i R_i - 1
\]

(1)

where n is the sample size, h_i is the value of individual i’s health variable (e.g., an indicator of ill-health) with μ_h as its mean and R_i is individual i’s fractional rank in the distribution of standard of living or socioeconomic status. Assume that h is a non-negative variable and h_i ≠ 0 ∀i with individuals arranged from poorest (y_1) to richest (y_n) by the values of the socioeconomic variable, y_1,y_2,⋯,y_p,⋯,y_n. Theoretically, the standard concentration index takes on its highest value (i.e., +1) when the wealthiest individual (with y_n) reports a non-negative and non-zero value of h_i (i.e., h_n > 0) and the remaining n − 1 individuals report zero values for h (where, \( \sum_{i=1}^{n-1} h_i = 0 \)). Similarly, the standard concentration index takes on its lowest value (i.e., −1) when the poorest individual (with y_1) has a positive non-zero value for h_i (i.e., h_1 > 0) and \( \sum_{i=2}^{n} h_i = 0 \).

Assuming that the poorest j individuals have positive non-zero values for h but the wealthiest n − i individuals have zero values for h, then the mean of the health variable, in this case, can be written as \( \mu_h = \frac{\sum_{i=1}^{j} h_i}{n} = \frac{\sum_{i=1}^{j} h_i}{n} \).

The fractional rank (R_i) in Equation (1) can be written as (O’Donnell et al., 2008):

\[
R_i = \sum_{j=0}^{i-1} w_j + 0.5w_i
\]

(2)

where w_j is the relative weight of individual j and w_0 = 0.

Using the fractional rank, Equation (1) can be expanded as:

\[
CI = \frac{2}{\sum_{i=1}^{n} h_i} \left( h_1 \left( \frac{1}{n} - \delta_1 \right) + 2 h_2 \left( \frac{1}{n} - \delta_2 \right) + \cdots + (n-1) h_{n-1} \left( \frac{1}{n} - \delta_{n-1} \right) + h_n (1 - \delta_n) \right) - 1
\]

(3)

which simplifies to

\[
CI = \frac{2}{n} \left( \sum_{i=1}^{n} i \cdot h_i - n \sum_{i=1}^{n} h_i \delta_i \right) - 1
\]

(4)

where \( \delta_i = \frac{1}{2} w_i \) with \( \delta_i > 0 \ ∀i \).
So, for any health variable, including binary variables, using Equation (4), the empirical bounds \([\text{lower bound, upper bound}]\) of the standard concentration index become:

\[
\left[ \left( \frac{2}{n} - w_1 \right), 1 - w_n \right]
\]

where \(w_1\) and \(w_n\) represent the relative weights for individuals with \(y_1\) and \(y_n\), respectively. Note that \(\frac{2}{n} \to 0\) and \(w_i \to 0\) as \(n \to \infty\). Therefore, with large samples, the lower and upper bounds of the standard concentration index will approach \(-1\) and \(+1\), respectively.

Using the standard concentration index in Equation (4), the Wagstaff’s (2005) \((W_c)\) and the Erreygers’ (2009) \((E_c)\) corrected concentration indices can be written as follows:

\[
W_c = \frac{CI}{1 - \mu_h}
\]

\[
E_c = 4\mu_h \cdot CI
\]

Applying the Wagstaff’s and Erreygers’ corrections on the lower and upper bounds of the standard concentration index shown in Equation (5) will yield the lower and upper bounds of the \(E_c\) and \(W_c\) illustrated in Figure 1 as the sample size increases but assuming equal weighting.

As expected, the upper and lower bounds of \(E_c\) remain sensitive to the sample size, especially for smaller samples. However, the lower and upper bounds of \(W_c\) remain agnostic to sample size. In fact, this is expected based on the normalization scheme implemented in the Wagstaff correction.

### 3 | SHOULD WAGSTAFF’S AND ERRERYGERS’ NORMALIZED CONCENTRATION INDICES BE INTERPRETED AS THE STANDARD CONCENTRATION INDEX?

In many empirical applications of the concentration index, researchers choose between Erreygers’ or Wagstaff’s corrected concentration index. Recently, the user-friendly \(-\text{conindex}\) user-written Stata command (O’Donnell et al., 2016) has facilitated the computation and comparison of both indices. Apart from these popular normalized indices \((E_c, W_c)\), other indices proposed but not discussed in this paper include the symmetric “concentration” index (Erreygers et al., 2012) and the generalized concentration index (Clarke et al., 2002; Wagstaff et al., 1991).

Let us examine these normalized indices (i.e., \(E_c, W_c\)) in turn for possible challenges for interpretation and policy.

#### 3.1 | The Wagstaff’s corrected concentration index

In arriving at the normalization, Wagstaff (2005) shows how the mean of a binary health variable may affect the standard concentration index.

![Figure 1](https://wileyonlinelibrary.com/doi/10.1002/hec.3106) The upper and lower bounds of \(E_c\) and \(W_c\) illustrated [Colour figure can be viewed at wileyonlinelibrary.com]
Consider \( h_i \) a binary health variable where \( h_i = h_1, h_2, \ldots, h_j, h_{j+1}, \ldots, h_n \) are corresponding values of the health variable for individuals sorted by a measure of socioeconomic status (SES) \( y_1 < y_2 < \cdots < y_j < y_{j+1} < \cdots < y_n \). The standard concentration index simplifies to \( \mu_h - 1 \) when \( \sum_{i=1}^j h_i = j \) and \( \sum_{i=j+1}^n h_i = 0 \). Similarly, the standard concentration index simplifies to \( 1 - \mu_h \) when \( \sum_{i=1}^j h_i = 0 \) and \( \sum_{i=j+1}^n h_i = n - j \) (summarized based on Wagstaff, 2005).

Therefore, for a binary health variable, \( h_i \), when \( \sum_{i=1}^j h_i = j \) and \( \sum_{i=j+1}^n h_i = 0 \), the Wagstaff’s corrected concentration index \( (W_c) \) will always simplify as:

\[
W_c = \frac{\mu_h - 1}{1 - \mu_h} = -1
\]  

Similarly, when \( \sum_{i=1}^j h_i = 0 \) and \( \sum_{i=j+1}^n h_i = n - j \), \( W_c = 1 \). Wagstaff (2009) later noted that it is “reasonable” to expect that \( |W_c| = 1 \) in these cases, in response to the challenge raised by Erreygers (2009, p.508).

### 3.2 The Erreygers’ corrected concentration index

Now, let us turn to the Erreygers’ normalized index. In many empirical applications, researchers interpret \( E_c \) based on the traditional underpinning of the standard concentration index. For example, an index value estimated at 0.85 may be interpreted to mean a higher concentration of a health variable among wealthier groups compared to an index value estimated at 0.12. Unfortunately, \( E_c \), by design was not meant to be interpreted in that manner.

Now, suppose that \( n \) is sufficiently large with individuals sorted by a measure of SES such that \( y_1 < y_2 < \cdots < y_j < y_{j+1} < \cdots < y_n \), and a binary health variable, \( h_i \) exists such that \( \sum_{i=1}^j h_i = j \) and \( \sum_{i=j+1}^n h_i = 0 \). As \( j \) increases from 1 to \( n - 1 \), the empirical values of \( E_c \) decreases from zero (i.e., when \( j = 1 \)), attaining its lowest value (\( E_c = -1 \)) at the 50th SES percentile, then increases beyond this point, reaching zero at about the 100th SES percentile (i.e., when \( j = n - 1 \); see Figure 2). Similarly, although not shown, as \( j \) decreases from \( n - 1 \) to 1, when \( \sum_{i=1}^j h_i = 0 \) and \( \sum_{i=j+1}^n h_i = n - j \), the values of \( E_c \) will increase steadily from zero when \( j = n - 1 \), reaching its highest value (\( E_c = 1 \)) at the 50% percentile and declining afterward to zero when \( j = 1 \). This behavior is by design to satisfy the transfer property. On this, Erreygers (2009, p.511) writes that “the most extreme pro-poor inequality is obtained when both the rich and the poor constitute exactly half of the population and every member of the poor half has the maximum health level \( a_h \) and every member of the rich half the minimum health level \( a_h \).” In our case with a binary health variable, the lowest \( (a_h) \) and highest \( (b_h) \) values of the health variable are 0 and 1, respectively.

![Figure 2](https://wileyonlineibrary.com)  
**Figure 2** Illustrating the behavior of \( W_c \) and \( E_c \) for an arbitrary binary health variable [Colour figure can be viewed at wileyonlinelibrary.com]
What do these illustrations for Wagstaff’s and Errerygers’ normalizations mean for interpreting socioeconomic inequalities in health and for policy? Traditionally, the magnitude of the standard concentration index conveys the relative concentration of a variable in relation to the SES distribution. So, for the standard concentration index, when $CI = -0.4$ for a specific health variable, for example, this variable is relatively concentrated among poorer than wealthier groups. If the value should change to $CI = -0.6$ it means that the health variable’s concentration among poorer groups has increased, ceteris paribus. Similarly, the concentration index increasing from $CI = 0.2$ to $CI = 0.5$ implies that the variable of interest is increasingly being concentrated among wealthier groups. As shown in Figure 2, unlike the case of the standard $CI$, the “symmetric” nature of $E_c$ means that we cannot interpret its values in a similar manner as with the standard $CI$. For a binary health variable, say full immunization for children in a country, and as shown in Figure 2, the value of $E_c$ is the same if $h_l = 1$ for the bottom $x\%$ or the bottom $(100 - x)\%$ of the population. For instance, $E_c = -0.64$ when $h_l = 1$ for the bottom 20% or the bottom 80% of the population. Applying the interpretation of the standard concentration index, this result would indicate that the level of socioeconomic inequality in full immunization coverage remains the same in both cases, irrespective of whether children in the bottom 20% or the bottom 80% of the population are fully immunized. This *equivalent* socioeconomic inequality may not necessarily be the case for policy to reduce socioeconomic inequality in full immunization coverage. Also, the “worst” socioeconomic inequality estimate ($E_c = -1$) corresponds to the case where only children in the bottom 50% of the population are fully immunized. As also shown in Figure 2, a progressive policy to gradually increase the number of fully immunized children beginning with the most impoverished child “worsens” socioeconomic inequalities in full immunization coverage until the bottom 50% of children are fully immunized based on $E_c$ but continues to “improve” socioeconomic inequalities based on the standard concentration index. Clearly, one should expect that a progressive policy targeting more impoverished children through full immunization, in this case, and as shown in Figure 2, should “improve” rather than “worsen” socioeconomic inequalities.

On the other hand, $W_c$ remains the same (i.e., $W_c = -1$) whether only the bottom $x\%$ or $(x + y)\%$ of children are fully immunized, where $x, y > 0$ and $x + y = 100$. The case is similar (i.e., $W_c = 1$) if only children in the top $z\%$ of the population are fully immunized. This means that using $W_c$, socioeconomic inequality in full immunization coverage for children in this hypothetical country remains high (i.e., $W_c = -1$ or $W_c = 1$) if all children in the bottom $x\%$ or the top $z\%$ of the population are fully immunized (see Figure 1). Also, a case where only the bottom 30% of children are not fully immunized but a progressive policy is implemented to increase immunization coverage that leaves the bottom 10% not fully immunized will not “improve” socioeconomic inequalities using $W_c$. While these extreme cases for $W_c$ may rarely occur in many applications, they show that a progressive policy to increase immunization coverage among the poor would have no impact on inequality using $W_c$ when this will appear as inequality “improving” using the standard concentration index (see Figure 2), for instance.

Then, what is the appropriate index or normalization for empirical health economics analysis when using a binary health variable? Unfortunately, the answer to this question is not straightforward. However, it is essential to note that the need for normalizing the standard concentration index when using a binary health variable arose because the bounds of the standard concentration index depend on the mean (or proportion in this case) of the variable (Wagstaff, 2005). Errerygers (2009) shows that this “issue” with the bounds is not unique to binary health variables. As noted previously, it could be the case for many variables encountered in health economics research if a positive lower bound or a finite upper bound exists. So, if normalization is essential, should the standard concentration index not be normalized for most applications and not just for binary health variables? In fact, this poses a significant challenge, including when decomposing the normalized concentration index for a binary variable, for example, using the method introduced by Wagstaff et al. (2003). Thus, as shown in this paper, and for binary health variables, because $E_c$ and $W_c$ cannot always be interpreted in the same way as the standard concentration index, the intended use of the index should guide what index or normalization scheme researchers implement.

### 5 Conclusion

The concentration index remains popular for assessing socioeconomic inequalities in health (Wagstaff et al., 1991). Many applications of this index in the health economics literature use a binary health variable. With binary health variables, the normalization of the standard concentration index was initially proposed to, among other things, ensure that the value of the index lies in the $-1$ and $+1$ range. This paper begins by showing that the empirical bounds of the concentration index, including the Errerygers’ corrected index, depended on the sample size and the relative weight variable (which is also related to the population). Importantly, it demonstrates that Wagstaff’s and Errerygers’ normalization for a binary health variable may produce
results that may be “counter-intuitive” for policy. \( W_c \), for instance, is perpetually fixed at −1 or +1 if only the bottom \( x \)% of the population, respectively, records a value of one for the variable of interest. \( E_c \), on the other hand, provides the same value if only individuals in the bottom \( x \)% or the bottom \((100 − x)\)% of the population have the value of one for the health variable. By this finding, progressive achievements, when moving from the poor to the rich, will initially “worsen” socioeconomic inequality using \( E_c \) with improvements only occurring when the value of the health variable is one for more than 50% of the population. Thus, the estimated values for \( E_c \) cannot be interpreted in the same manner as the values of the standard concentration index. The “symmetric” nature of \( E_c \) will make it challenging to compare socioeconomic inequalities between any two distributions, irrespective of the mean of the binary health variable. Unfortunately, researchers implementing \( E_c \) and \( W_c \) generally ignore these nuances and interpret these indices in much the same way as the standard concentration index. While there is still room to continuously improve indices for assessing socioeconomic inequalities in health, the purpose for empirical assessment should guide the choice of an appropriate index.

ACKNOWLEDGEMENT
The research emanated from the Global Network for Health Equity (GNHE), funded by the IDRC (Grant number 106439).

CONFLICT OF INTEREST
The author declares that there are no conflicts of interest.

ETHICAL STATEMENT
The study did not use any existing datasets; there are no ethical issues.

DATA AVAILABILITY STATEMENT
Data sharing does not apply to this article as only hypothetical data were generated and used for the study. The codes to generate the hypothetical dataset are available from the author upon request.

ORCID
John E. Ataguba https://orcid.org/0000-0002-7746-3826

REFERENCES
Ataguba, J. E., Akazili, J., & McIntyre, D. (2011). Socioeconomic-related health inequality in South Africa: Evidence from general household surveys. *International Journal for Equity in Health, 10*, 48.

Chen, Z., & Roy, K. (2009). Calculating concentration index with repetitive values of indicators of economic welfare. *Journal of Health Economics, 28*, 169–175.

Clarke, P. M., Gerdtham, U.-G., Johannesson, M., Bingeors, K., & Smith, L. (2002). On the measurement of relative and absolute income-related health inequality. *Social Science & Medicine, 55*, 1923–1928.

Erreygers, G. (2009). Correcting the concentration index. *Journal of Health Economics, 28*, 504–515.

Erreygers, G., Clarke, P., & Van Ourti, T. (2012). “Mirror, mirror, on the wall, who in this land is fairest of all?”—distributional sensitivity in the measurement of socioeconomic inequality of health. *Journal of Health Economics, 31*, 257–270.

Erreygers, G., & Van Ourti, T. (2011). Measuring socioeconomic inequality in health, health care and health financing by means of rank-dependent indices: A recipe for good practice. *Journal of Health Economics, 30*, 685–694.

Kakwani, N., Wagstaff, A., & van Doorslaer, E. (1997). Socioeconomic inequalities in health: Measurement, computation, and statistical inference. *Journal of Econometrics, 77*, 87–103.

O’Donnell, O., O’Neill, S., Van Ourti, T., & Walsh, B. (2016). Conindex: Estimation of concentration indices. *The Stata Journal, 16*, 112–138.

O’Donnell, O., Van Doorslaer, E., Wagstaff, A., & Lindelow, M. (2008). *Analyzing health equity using household survey data: A guide to techniques and their implementation*. World Bank.

van Doorslaer, E., & Koolman, X. (2004). Explaining the differences in income-related health inequalities across European countries. *Health Economics, 13*, 609–628.

Wagstaff, A. (2005). The bounds of the concentration index when the variable of interest is binary, with an application to immunization inequality. *Health Economics, 14*, 429–432.

Wagstaff, A. (2009). Correcting the concentration index: A comment. *Journal of Health Economics, 28*, 516–520.

Wagstaff, A., Paci, P., & van Doorslaer, E. (1991). On the measurement of inequalities in health. *Social Science & Medicine, 33*, 545–557.

Wagstaff, A., van Doorslaer, E., & Watanabe, N. (2003). On decomposing the causes of health sector inequalities with an application to malnutrition inequalities in Vietnam. *Journal of Econometrics, 112*, 207–223.
How to cite this article: Ataguba, J. E. (2022). A short note revisiting the concentration index: Does the normalization of the concentration index matter? *Health Economics, 31*(7), 1506–1512. [https://doi.org/10.1002/hec.4515](https://doi.org/10.1002/hec.4515)