Revisiting the Concentration Curves and Indices as Useful Tools for Assessing Relative and Attributable Risks

Yuejen Zhao1 and Andy H. Lee2
1Adjunct Senior Research Fellow, Institute of Advanced Studies, Charles Darwin University, Darwin, NT, Australia
2Professor of Biostatistics, School of Public Health, Curtin University, Australia

Abstract

Accurate assessment of the association between exposure and response is central to identifying causality in medical research. The concentration index has been commonly used to study income inequality and socioeconomic related health inequality. This study generalizes applications of the concentration index to measure the relative and attributable risks for describing exposure-response relationships in medical research. Based on cumulative distribution functions, a new measure of correlation is proposed to quantify the association between exposure and response. The connection between the new and existing measures is discussed. The method enables the semi-parametric analysis of overall association and disparity by risk factors. Both grouped and continuous data situations are considered with two applications. The first example illustrates the relationships between the concentration index, relative and attributable risks. The second example demonstrates how the concentration index can assist in evaluating the association between the radiation dose and the incidence of leukaemia. Logistic regression based decomposition is compared with the new approach. We found the concentration index analysis useful not only for examining socioeconomic determinants of health, but also for assessing quantitative relations between exposures to health risks and ill-health outcomes.

Keywords: Dose-response relationship; Health status disparities; Odds ratio; Risk factors

Introduction

Accurate quantitative assessment of the association between exposure and response is central to identifying causality in medical research [1,2]. Multiple rate ratios or odds ratios are commonly used for quantifying the exposure-response associations [3]. However, dichotomizing continuous exposure may produce biased estimates and result in a loss of statistical efficiency, while multiple inferences can lead to false positive results [4-6]. The Lorenz curve and Gini index (GI) provide an alternative to assessing the overall relationship between continuous exposure and response [7-9]. The approach utilizes an integrated quantitative and graphical framework to make more efficient use of information. In addition to performing univariate assessment of inequality for highly skewed variables (such as individual income) [10,11], the GI has been applied to examine continuous exposure-response relations [7,12,13]. On the other hand, when bivariate relations between the exposure and response are of interest, a more general form of GI, the concentration index (CI), is available [14]. Although CI is well suited for measuring socioeconomic inequality in health [15-17], there are still potentials for more general applications [9].

The present study generalizes the application of GI and CI in medical research and demonstrates their usefulness for summarizing rate ratios, odds ratios and attributable risks. A correlation measure is proposed to assess and summarize overall associations between risk factors and ill-health outcome. Two examples illustrate applications of the methodology in comparison with the regression based decomposition. Pros and cons of the approach are also considered in the medical research context. The variance of rate ratio and the derivation of continuous data are given in the appendices.

Methods

Gini and concentration indices

We first review the GI, Lorenz curve, CI, and their role in assessing exposure and response using grouped data. Consider a $p$-level exposure $X_i$ and an ill-health response $Y_i$, where $Y_i = d/n_i$ denotes the response proportions sorted in ascending order $(Y_1 \leq \ldots \leq Y_i \leq \ldots \leq Y_p)$; $d$ and $n_i$ represent respectively the number of ill-health events and population size for group $i$, with $f_i = n_i/N$ being the frequency proportion of the total population $N$. Let $E = \sum_n f_i$ be the cumulative frequency proportion and $\Gamma = \sum_i f_i Y_i$ be the mean. Under this setting, the Lorenz curve is defined by $L_n = \sum f_i Y_i / \Gamma$ plotted on the ordinate against $Fi$ along the abscissa [14,15]. It is the cumulative proportion of cases ($L_n$) compared to the cumulative proportion of the at-risk population ($F_i$), ordered by the level of risk. If $L_n = F_i$, the Lorenz curve coincides with the diagonal line, implying that $Y$ is distributed in line with $f$ so that the ill-health is evenly distributed. Otherwise, it lies beneath the diagonal line. Further the Lorenz curve deviates from the diagonal line, the greater is the degree of disparity. Let $\text{cov} \{Y, F\}$ be the covariance between $Y$ and $F$. The GI can be given by

$$GI = \frac{2}{\Gamma} \text{cov}\{Y,F\} = \sum_{i=1}^{p} (F_{i-1} - F_i) (L_{i} - L_{i-1}),$$

which represents twice the area between the diagonal line and the Lorenz curve [14]. If every group has exactly the same risk, $GI = 0$ representing perfect equality. If one group owns all the ill-health

*Corresponding author: Yuejen Zhao, Adjunct Senior Research Fellow, Institute of Advanced Studies, Charles Darwin University, Principal Health Economist, Department of Health, Darwin Plaza, Level 1, 41 Smith St, Darwin, NT 0800, Australia, Tel: 61-8-89858077; Fax: 61-8-89858075; E-mail: Yuejen.Zhao@nt.gov.au

Received March 06, 2012; Accepted July 20, 2012; Published July 25, 2012


Copyright: © 2012 Zhao Y, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
risks, GI = 1 representing perfect inequality. Typically, GI varies between 0 and 1, indicating the level of inequality in ill-health risks between groups. At the individual level, the total inequality is clearly \( 0.5 - 0.5 \sum_{i} d_{i} / N \) = 1 - \( \bar{Y} \), by ranking \( d \) from 0 (alive) to 1 (dead) for the Lorenz curve, where \( \eta = 1 \) and \( \bar{Y} \) is the mortality rate. For grouped data, the GI actually reflects the degree of inequality under the current groupings.

Let \( Y_{jk} \) represents the \( Y \) being reordered by exposure level \( X_{jk} \), where \( X_{j1} \leq \ldots \leq X_{jn} \). We have \( f_{jk} = n_{jk} / N \), where \( n_{jk} \) is the number of observations in group \( k \). The concentration curve is defined by plotting

\[
L_{(k)} = \sum_{j=1}^{n_{k}} f_{jk} Y_{j} / \bar{Y}
\]

against \( F L_{(k)} = \sum_{j=1}^{n_{k}} f_{jk} \) [17]. The CI is then given by

\[
CI = \frac{2}{\bar{Y}} \text{cov}[Y, F] = \sum_{j=1}^{n_{k}} (F j_{(k)} - \bar{F})_{j} (L j_{(k)} - \bar{L})_{j},
\]

which is twice the area between the equitarian line and the concentration curve [17]. Here the groups are ranked by \( X \) instead of \( Y \). Unlike GI, CI can be either positive or negative. If the exposure is harmful, 0 < CI ≤ GI. If it is protective, GI ≤ CI < 0. The standard errors of GI and CI can also be estimated [16]. The absolute value of the ratio

\[
CI/GI = \frac{\text{cov}[Y, F]}{\text{cov}[Y, F]}
\]

indicates the inequality explainable by the exposure [14]. In this context, the concentration curve, CI and the ratio between CI and GI are often used to analyze socioeconomic inequality of health [15-17]. Assuming a regression model \( \bar{Y} = g^{-1}(Z) \) and the residual \( e = Y - \bar{Y} \), where \( Z \) represents the predictor(s) and \( g(\cdot) \) is a generalized linear link function, the GI and CI can be decomposed into a deterministic component and a residual component: [18]

\[
\frac{2}{\bar{Y}} \text{cov}[\bar{Y}, F^*] + \frac{2}{\bar{Y}} \text{cov}[e, F^*],
\]

where \( F^* \) is either \( F \) or \( F \), for GI or CI decomposition respectively. In this paper, we examine the situation \( Z = X \).

**Correlation measure**

It is known that the above CI/GI ratio can overestimate the contribution of the exposure responsible for the health inequality [19]. A new correlation measure is proposed for assessing the exposure-response relationship. The correlation between exposure and response can be examined by changes in the frequency proportions from being sorted by \( Y \) to being sorted by \( X \). Let \[ \text{var}[\cdot] \] denote the variance of a quantity. We propose a correlation coefficient between \( F \) and \( F^* \),

\[
\rho = \frac{\text{cov}[F, F^*]}{\sqrt{\text{var}[F] \text{var}[F^*]}} = \frac{\text{cov}[F, F^*]}{\text{var}[F]} = \frac{\sum_{j=1}^{n_{k}} f_{jk} (F - \bar{F}) (F - \bar{F})}{\sum_{j=1}^{n_{k}} f_{jk} (F - \bar{F})},
\]

as an overall measure of association between the exposure and the ill-health response. Note that \( \rho \) assesses the correlation between exposure distribution and response distribution based on cumulative functions, while \( F \) and \( F^* \) are rearranged using \( f_{jk} = f_{jk}^{(i)} \), and \( \text{cov}[F, F^*] = \text{cov}[F, F^*] \). If \( 0 < p \leq 1 \), the proportional fractions ranked by \( Y \) and those ranked by \( X \) are positively correlated. Otherwise, \( 1 \leq p < 0 \) implies a negative correlation. The coefficient of determination is then

\[
\rho^2 = \frac{\text{cov}[F, F^*]}{\text{var}[F]} = \frac{\sum_{j=1}^{n_{k}} f_{jk} (F - \bar{F}) (F - \bar{F})}{\sum_{j=1}^{n_{k}} f_{jk} (F - \bar{F})},
\]

which yields the proportion of disparity in \( Y \) explained by \( X \). Let \( \rho \) and \( \text{var}[\cdot] \) represent the residual terms for \( F \) and \( F^* \) predicted by the ranked \( Y \) and those ranked by \( X \). GI and CI are algebraically related to \( \rho \). Similar to the GI and CI, \( \rho \) can be decomposed into a model component and a residual component:

\[
\rho = \frac{\text{cov}[Y, F] \text{cov}[Y, F^*] \text{cov}[e, e]}{\text{var}[Y] \text{var}[F] \text{var}[F^*]}
\]

\[
= \frac{\text{CI} \cdot \text{GI} \cdot \bar{Y} \text{cov}[Y, F^*] \text{cov}[e, e]}{4 \text{var}[Y] \text{var}[F] \text{var}[F^*]},
\]

**Risk assessment**

Some basic properties of the CI measures are studied further in terms of risk assessment in this section. The relative risk (RR) may be considered as a ratio of the excess risk estimated by a rate ratio, or a density ratio of incremental change in ill-health in response to the change in exposure [20,21]. RR is thus the slope of the tangent line of the concentration curve, evaluated at a point \( Y_{(m)} \), viz,

\[
RR_{(m)} = (L_{(m)} - \bar{L}) / (F_{(m)} - \bar{F}).
\]

Using the concentration curve, it equals to the magnitude of the risk in comparison with the expectation (i.e., the average risk) [22]. Clearly, the RR is slightly different from the usual case in epidemiology, based on the minimum level of exposure: \( RR_{(a)} = RR_{(1)} / RR_{(m)} = Y_{(a)} / Y_{(m)} \). More generally, let \( RR_{(m)} = RR_{(1)} / RR_{(m)} = Y_{(m)} / Y_{(m)} \). The variance of \( RR_{(m)} \) can be derived as

\[
\text{var}[RR_{(m)}] = \frac{Y_{(m)}(1 - Y_{(m)})}{n_{(m)} Y_{(m)}^2} + \frac{Y_{(m)}(1 - Y_{(m)})}{n_{(m)} Y_{(m)}^2}.
\]

(see the derivation in the first section of Appendix). Note that RR is monotonically increasing for the Lorenz curve by definition, but this is not necessarily the case for the concentration curve. Application of RR is more meaningful in the context of concentration curve, because it involves both exposure and response, whereas the Lorenz curve involves only the response.

In medical research, the RR is often approximated by the odds ratio (OR). Denoting the total number of ill-health events by \( D = \sum_{i=1}^{n} d_{i} \), then we have

\[
OR_{(m)} = \left( d_{(m)} / (N - D) \right) \left( D(d_{(m)} - d_{(m)}) \right).
\]

As classically defined, attributable risk (AR) is the percentage of cumulative proportion of total population developing a disease over a specified interval, caused by an exposure [23]. The AR also gives the

---

**Table 1:** Summary of the examples.

<table>
<thead>
<tr>
<th>Index</th>
<th>95% Confidence Interval</th>
<th>Logistic model decomposition (Contribution)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Model</td>
</tr>
<tr>
<td>Colorectal polyps</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GI</td>
<td></td>
<td>0.073</td>
</tr>
<tr>
<td>CI</td>
<td></td>
<td>0.065 to 0.082</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.036(49%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.037(51%)</td>
</tr>
<tr>
<td>N = 976</td>
<td></td>
<td>( \bar{Y} = 0.5 )</td>
</tr>
<tr>
<td></td>
<td>( \bar{Y} = 0.5 )</td>
<td>( r = 0.583 )</td>
</tr>
<tr>
<td>Leukaemia</td>
<td></td>
<td>0.374</td>
</tr>
<tr>
<td>GI</td>
<td></td>
<td>0.367 to 0.382</td>
</tr>
<tr>
<td>CI</td>
<td></td>
<td>0.313(84%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.062(16%)</td>
</tr>
<tr>
<td>N = 61902</td>
<td></td>
<td>( \bar{Y} = 0.614*10^3 )</td>
</tr>
<tr>
<td></td>
<td>( \bar{Y} = 0.614 *10^3 )</td>
<td>( r = 0.819 )</td>
</tr>
</tbody>
</table>
proportion of ill-health events that can be avoided if the exposure is reduced to 0.429 rather than being totally eliminated [4,24]. In the case of grouped data, the AR can be viewed as the proportion of the incidence of ill-health that will be reduced if the exposure is reduced to X(i,k) rather than complete elimination of the risk. For example, a smoking cessation policy intends to reach a nominated target level X(i,k) rather than complete elimination of the risk. For example, a smoking cessation policy intends to reach a nominated target level X(i,k) rather than achieving an unrealistic zero smoking prevalence. The AR can take on negative values, if the AR is used for studying protective factors rather than being totally eliminated [24]. When the comparison standard is exchanged (the exposure of concern is changed from being harmful to being protective), AR = 1 - 1/RR [24,25]. The CI is in fact the weighted average of AR. The derivations of RR, AR and ρ for continuous data are given in the second section of the Appendix.

### Examples

**Colorectal polyps:** To demonstrate the relationships between CI, RR and AR, let us consider the matched case-control study of the associations of vegetables, fruits, and grain intakes with colorectal polyps [24]. The results of the analysis are summarized in the upper part of Table 1. The total individual level inequality for the matched case-control design is 1 - Y = 0.5 = 0.5 and the GI is 0.073, showing that the case-control grouping reflects 15% (0.073/0.5) of the total

<table>
<thead>
<tr>
<th>Mean servings Xk</th>
<th>Cases d</th>
<th>Total n</th>
<th>Rate Y</th>
<th>Cumulative proportions</th>
<th>Concentration index CI</th>
<th>Attributable risk AR</th>
<th>Actual RR</th>
<th>Logistic mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>13</td>
<td>17</td>
<td>0.765</td>
<td>0.017</td>
<td>0.027</td>
<td>-0.0004</td>
<td>-0.5294</td>
<td>1.529</td>
</tr>
<tr>
<td>1</td>
<td>36</td>
<td>60</td>
<td>0.600</td>
<td>0.079</td>
<td>0.100</td>
<td>-0.0013</td>
<td>-0.2057</td>
<td>1.200</td>
</tr>
<tr>
<td>2</td>
<td>55</td>
<td>99</td>
<td>0.556</td>
<td>0.180</td>
<td>0.213</td>
<td>-0.0040</td>
<td>-0.1231</td>
<td>1.111</td>
</tr>
<tr>
<td>3</td>
<td>70</td>
<td>137</td>
<td>0.511</td>
<td>0.321</td>
<td>0.357</td>
<td>-0.0046</td>
<td>-0.0507</td>
<td>1.022</td>
</tr>
<tr>
<td>4</td>
<td>77</td>
<td>151</td>
<td>0.510</td>
<td>0.475</td>
<td>0.514</td>
<td>-0.0084</td>
<td>-0.0494</td>
<td>1.020</td>
</tr>
</tbody>
</table>

**Table 2:** Concentration Index and Relative Risk for the Fruit and Vegetable Intake and Colon Polyps.

<table>
<thead>
<tr>
<th>Radiation dose Xk</th>
<th>Incidence /1000 Y (2)</th>
<th>i (3)</th>
<th>k (4)</th>
<th>F (5)</th>
<th>F(y) (6)</th>
<th>L(y) (7)</th>
<th>L(y)(x) (8)</th>
<th>L_n(x)(k) (9)</th>
<th>GI (10)</th>
<th>CI (11)</th>
<th>RR_m (12)</th>
<th>OR_m (13)</th>
<th>AR (14)</th>
<th>RR (15)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;250</td>
<td>0.244</td>
<td>1</td>
<td>1</td>
<td>0.132</td>
<td>0.132</td>
<td>0.053</td>
<td>0.132</td>
<td>0.063</td>
<td>0.002</td>
<td>0.002</td>
<td>1.000</td>
<td>1.000</td>
<td>0.602</td>
<td>1.000</td>
</tr>
<tr>
<td>250-499</td>
<td>0.290</td>
<td>2</td>
<td>2</td>
<td>0.167</td>
<td>0.299</td>
<td>0.299</td>
<td>0.299</td>
<td>0.132</td>
<td>0.007</td>
<td>0.026</td>
<td>1.187</td>
<td>1.187</td>
<td>0.537</td>
<td>1.354</td>
</tr>
<tr>
<td>500-749</td>
<td>0.593</td>
<td>5</td>
<td>3</td>
<td>0.164</td>
<td>0.733</td>
<td>0.474</td>
<td>0.463</td>
<td>0.289</td>
<td>0.007</td>
<td>-0.006</td>
<td>2.425</td>
<td>2.425</td>
<td>0.192</td>
<td>1.833</td>
</tr>
<tr>
<td>750-999</td>
<td>0.343</td>
<td>3</td>
<td>4</td>
<td>0.188</td>
<td>0.487</td>
<td>0.237</td>
<td>0.651</td>
<td>0.395</td>
<td>0.019</td>
<td>0.069</td>
<td>1.405</td>
<td>1.405</td>
<td>0.410</td>
<td>2.483</td>
</tr>
<tr>
<td>1000-1249</td>
<td>0.752</td>
<td>7</td>
<td>5</td>
<td>0.172</td>
<td>0.930</td>
<td>0.711</td>
<td>0.823</td>
<td>0.605</td>
<td>0.045</td>
<td>0.15</td>
<td>3.081</td>
<td>3.079</td>
<td>0.178</td>
<td>3.363</td>
</tr>
<tr>
<td>1250-1499</td>
<td>0.588</td>
<td>4</td>
<td>6</td>
<td>0.082</td>
<td>0.570</td>
<td>0.316</td>
<td>0.905</td>
<td>0.684</td>
<td>0.038</td>
<td>0.045</td>
<td>2.409</td>
<td>2.408</td>
<td>0.225</td>
<td>4.554</td>
</tr>
<tr>
<td>1500-1749</td>
<td>1.231</td>
<td>8</td>
<td>7</td>
<td>0.039</td>
<td>0.970</td>
<td>0.789</td>
<td>0.945</td>
<td>0.763</td>
<td>0.039</td>
<td>0.006</td>
<td>5.042</td>
<td>5.037</td>
<td>0.126</td>
<td>6.166</td>
</tr>
<tr>
<td>1750-1999</td>
<td>0.645</td>
<td>6</td>
<td>8</td>
<td>0.025</td>
<td>0.758</td>
<td>0.500</td>
<td>0.970</td>
<td>0.789</td>
<td>0.074</td>
<td>0.039</td>
<td>2.641</td>
<td>2.640</td>
<td>0.179</td>
<td>8.348</td>
</tr>
<tr>
<td>2000-2249</td>
<td>2.130</td>
<td>9</td>
<td>9</td>
<td>0.015</td>
<td>0.985</td>
<td>0.842</td>
<td>0.842</td>
<td>0.842</td>
<td>0.022</td>
<td>0.097</td>
<td>8.732</td>
<td>8.716</td>
<td>0.105</td>
<td>11.301</td>
</tr>
<tr>
<td>2250-2499</td>
<td>7.859</td>
<td>12</td>
<td>10</td>
<td>0.008</td>
<td>1.000</td>
<td>1.000</td>
<td>0.993</td>
<td>0.947</td>
<td>0.000</td>
<td>0.022</td>
<td>32.404</td>
<td>32.157</td>
<td>-0.037</td>
<td>15.295</td>
</tr>
<tr>
<td>2500-2749</td>
<td>3.534</td>
<td>11</td>
<td>11</td>
<td>0.005</td>
<td>0.989</td>
<td>0.868</td>
<td>0.998</td>
<td>0.974</td>
<td>0.024</td>
<td>0.024</td>
<td>14.507</td>
<td>14.459</td>
<td>0.012</td>
<td>20.694</td>
</tr>
<tr>
<td>≥2750</td>
<td>6.623</td>
<td>11</td>
<td>12</td>
<td>0.002</td>
<td>0.992</td>
<td>0.895</td>
<td>1.000</td>
<td>1.000</td>
<td>0.007</td>
<td>0.000</td>
<td>27.273</td>
<td>27.099</td>
<td>0.000</td>
<td>28.005</td>
</tr>
<tr>
<td>Total</td>
<td>0.614</td>
<td>1</td>
<td>100</td>
<td>0.374</td>
<td>0.338</td>
<td>2.513</td>
<td>2.512</td>
<td>0.338</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Gl: Gini Index; CI: Concentration Index; RR_m: Rate Ratio rebased on minimum level of exposure; OR_m: Odds Ratio rebased on minimum level of exposure; AR: Attributable Risk; *cumulative sum with respect to i.

**Table 3:** Concentration Curve and Index for the Radiation–induced Leukaemia Data, United Kingdom, 1935-1954.
Radiation–induced leukaemia: The second example is taken from an investigation of leukaemia among patients treated with X-ray for ankylosing spondylitis at 81 British radiotherapy centres between 1935 and 1954 [26]. The study aimed to determine the relationship between the doses of radiation given and the incidence of leukaemia. Details of radiation were recorded in the mean spinal-marrow dose (roentgens). The 38 leukaemia cases included definite, probable and presumptive diagnoses. The men-years at risk (61,902 in total) were used to estimate the incidence. We reanalyze the data using the proposed concentration curve approach. The GI and CI analyses are summarized in the lower part of Table 1.

The total individual level inequality for the study design is 1 – \( Y = 1 – 0.000614 = 0.999 \) and the GI is 0.374, showing that the study grouping reflects 37% \( (0.374/0.999) \) of the total leukaemia incidence inequality (see Table 3). Re-ranking \( Y \) by \( X \) (radiation dose), CI has the value 0.338. In accord with the new coefficient of determination \( r^2 \), the radiation dose accounts for 67% of the leukaemia inequality. This result seems more plausible than the CI/GI ratio (90%) and the logistic model GI decomposition (84%). The fitted logistic model is

\[
\hat{Y} = \exp\left(0.001213X - 8.643\right) / \left[1 + \exp\left(0.001213X - 8.643\right)\right].
\]

A clear gradient is observed in the RR estimates (columns 12 and 13 of Table 3). Specifically, radiation dose over 2,750 roentgens could increase the leukaemia risk by about 27 times above that at the minimum radiation level. From the AR calculation (column 14 of Table 3), about 60% of leukaemia incidence in the spondylitic patients could be avoided, if the radiation exposure is reduced to the minimum level. If the radiation exposure level is reduced to the average level, 33.8% of the leukaemia incidence could be avoided. The logistic modelled RRs are listed in column 15 of Table 3.

Figure 1 shows that the concentration curve almost coincided with the Lorenz curve and the high radiation dose rankings explain the majority of leukaemia incidence disparity. The correlation coefficient \( \rho \) is 0.819, which means a high association between the leukaemia risk and radiation exposure; see Figure 2.

Note that when the exposure is used for the CI estimation and also used as the predictor for the logistic model based decomposition, the factor contributions for the exposure in the decomposed CI based on the logistic model are 98% and 112% respectively for the colorectal polyps and leukaemia example in Table 1. This indicates that the regression based CI decomposition can over-estimate the contribution of the exposure if the exposure variable is used as the underlying variable and explanatory variable simultaneously.

**Discussion**

Applications of GI and CI for assessing risk factors can potentially provide more insightful information about the association between exposure and response in medical research [7,17]. The approach is appealing and straightforward, which summarizes RR, OR, AR, correlation coefficient and logistic regression model in a coherent manner. It can analyze three types of variable simultaneously: the exposure or underlying variable (X), the response or ill-health outcome (Y) and the predictors or determinants (Z). This feature is particularly attractive for analyzing the determinants of socioeconomic related health inequality [27]. The method brings together the inequality and relative risk analysis in a unified framework and enables researchers to assess overall exposure-response association. Our study has demonstrated that the concentration curves and indices are closely linked with RR, AR and regression analysis. The instrumental method provides
another approach to investigate the structure of exposure and response relationship. Different levels of exposure and response are modeled to allow a more detailed examination on the interplay between exposure and response in a graphical manner. The percentile based analysis is appropriate for skewed data and free of the underlying distribution. As demonstrated by the two examples, the method provides a powerful alternative for analyzing the cause - effect relationship for ordinal or continuous variables. It also allows further decomposition by multiple factors for identification of health determinants or adjustment of confounders using multivariate models [19,28]. A new variance estimate of rate ratio was derived using Taylor expansion without data transformation. The new correlation and determination coefficients are based on a semi-parametric approach to estimate factor contributions. Comparing the contribution estimates, this new method appears more plausible and robust than the $[\text{CI/GI}]$ ratio and regression based decomposition. The logistic regression decomposition is a parametric model directly using the predictor information. If the model is chosen appropriately, the contribution estimates may be more accurate than the semi-parametric approach. There is empirical evidence that the exposure cannot be used simultaneously as the underlying variable for the $\text{CI}$ and the predictor in the regression model $Y = g^{-1}(Z)$, as previously recommended [27]. This may overestimate the contribution of the exposure variable due to double-counting.

Several limitations of the method should be noted. The concentration curve analysis is semi-parametric. The measure is relative rather than absolute. It does not use the exposure levels directly but the rankings instead. The same applies to $\text{RR}$, $\text{OR}$ and $\text{AR}$. Use of spline regression to define knots of exposure categories might be helpful to address this shortcoming [9]. This limitation can also be addressed by jointly using logistic regression based $\text{GI}$ decomposition as demonstrated in the examples. The $\text{GI}$ and $\text{CI}$ estimation by grouped data may underestimate the true association, because they overlook the within-group variation [29]. The coefficient of determination $r^2$ is a conservative measure of the exposure-response association based on the cumulative distributions. The $\text{CI}$ is best used in a monotonic exposure-disease relation. With a non-monotonic situation (e.g. quadratic function), the positive and negative contributions may cancel out in the aggregate $\text{CI}$, although the concentration curve will reflect the detailed positive and negative areas, and $\text{OR}$ and $\text{RR}$ still remain valid.

In conclusion, the concentration curve approach provides a simple and useful alternative for risk factor analysis. It has desirable properties for assessing quantitative relationships between cause and effect. This approach is valuable for the overall assessment of exposure-response relationships, as the focus of health studies shifts from the proximal causes to the distal risk factors [30].

References


Appendix

Variance of the relative risk

Consider two independent binomial random variables $d_i \sim \text{Binomial}(n_i, p)$, where $i = (1, 2)$ and $p$ indicates the probability of ill health events in $n_i$ observations.
For $Y = \delta_i/n_i$ within the domain $(0, 1)$, $r(Y) = Y/\delta_i$. The first order Taylor expansion about $Y$ is

$$r(Y) = r(p) + r'(p)(Y - p) + \text{remainder}$$

Assuming $\rho = (E[Y], E[Y])$, the second order approximation for $E[r(Y)]$ is

$$E[r(Y)] = \frac{E[Y]}{E[Y]} + \frac{E[Y]E[Y] - E[Y]}{E[Y]}$$

The variance is

$$\text{var}[r(Y)] = \frac{E[Y]}{E[Y]} + \frac{E[Y]E[Y] - E[Y]}{E[Y]}$$

Because of the independence between $Y$ and $Y_i$, we therefore have

$$\text{var}[r(Y)] = \frac{E[Y]}{E[Y]} + \frac{E[Y]E[Y] - E[Y]}{E[Y]}$$

Derivation of measures for continuous data

To lighten notations, when there is no ambiguity, we adopt the symbols similar to those for the discrete grouped data. Let $f(Y)$ be the probability density function (pdf) of the continuous and non-negative ill-health random variable $Y$. The Lorenz curve of $Y$ is obtained by plotting $L(x) = \int_0^x f(Y)dY$ on the ordinate, against the cumulative distribution function (cdf) $F(x)$ along the abscissa, where the expectation $E[Y] = \int_0^\infty Yf(Y)dY$ exists [14]. By definition, $L(x) \geq 0$ and $L(x) \leq 1$. The health inequality may be measured by $GI = 1 - 2\int_0^1 L(F(x))dF(x)$.

We further define $X$, a continuous and non-negative exposure or risk factor of $Y$ with pdf $f(X) = f(Y)$. The cdf for $X$ is

$$F(x) = \int_0^x f(Y)dY$$

Unlike $f(Y)$ in which $f(Y)$ is integrated with respect to $Y$, $F(x)$ is given by integrating $f(Y)$ with respect to $X$. The concentration curve of $X$ can be obtained by plotting

$$L(x) = \int_0^x f(Y)dY$$

against $F(x)$ [16]. This definition means that $RR$ may be measured by taking the first order differentiation of $L(F(x))$ with respect to $F(x)$ as

$$RR = L(F(x)) = \frac{dL(F(x))}{dF(x)} = \frac{Y}{E[Y]}$$

which is the tangent of the concentration curve at a given value of $x$. In this case, $OR$ is a function of $Y$,

$$OR = \frac{1 - L(F(x))}{1 - L(F(x))}$$

The AR is

$$AR = 1 - L(F(x)) - RR(1 - F(x))$$

evaluated at $F(x)$ [13]. Health inequality attributable to $X$ can be measured by $GI$ in the form of

$$GI = 1 - 2\int_0^1 L(F(x))dF(x)$$

In this paper, $F$ and $F_1$ are rearranged using $f(Y) = f(X)$. As a measure of overall association between the risk factor and ill-health, the correlation coefficient $\rho$ between $F$ and $F_1$ is defined by

$$\rho = \frac{\text{cov}[F(Y), F_1(x)]}{\sqrt{\text{var}[F(Y)]\text{var}[F_1(x)]}}$$

Since $\text{var}[F(Y)] = \text{var}[F_1(x)]$, we have

$$\rho = \frac{\text{cov}[F(Y), F_1(x)]}{\text{var}[F(Y)]}$$