Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Zhao and Lee, J Biomet Biostat 2012, S:7
DOI: 10.4172/2155-6180.S7-019

Research Article Open Access

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

Yuejen Zhao[1]* and Andy H. Lee[2]

[1]*Adjunct Senior Research Fellow, Institute of Advanced Studies, Charles Darwin University, Darwin, NT, Australia*

[2]*Professor 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 $Y_i = d_i/n_i$ denotes the response proportions sorted in ascending order ($Y_1 \le \ldots \le Y_i \le \ldots \le Y_p$); d_i 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 $F_i = \sum_{j=1}^{i} f_j$ be the cumulative frequency proportion and $\bar{Y} = \sum_{i=1}^{p} f_i Y_i$ be the mean. Under this setting, the Lorenz curve is defined by $L_i = \sum_{j=1}^{i} \frac{f_j Y_j}{\bar{Y}}$ plotted on the ordinate against F_i along the abscissa [14,15]. It is the cumulative proportion of cases (L_i) compared to the cumulative proportion of the at-risk population (F_i), ordered by the level of risk. If $L_i = F_i$, 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. The further the Lorenz curve deviates from the diagonal line, the greater is the degree of disparity. Let cov[Y,F] be the covariance between Y and F. The *GI* can be given by

$$GI = \frac{2}{\bar{Y}} cov[Y, F] = \sum_{i=1}^{p-1} (F_i L_{i+1} - F_{i+1} L_i),$$

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

Citation: Zhao Y, Lee AH (2012) Revisiting the Concentration Curves and Indices as Useful Tools for Assessing Relative and Attributable Risks. J Biomet Biostat S7-019. doi:10.4172/2155-6180.S7-019

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 $2\left(0.5-0.5\sum_i d_i \Big/ N\right)=1-\overline{Y}$, by ranking d_i from 0 (alive) to 1 (dead) for the Lorenz curve, where $n_i = 1$ and $\overline{Y}$ is the mortality rate. For grouped data, the GI actually reflects the degree of inequality under the current groupings.

Let $Y_{(k)}$ represents the Y_i being reordered by exposure level $X_{(k)}$, where $X_{(1)} \le \ldots \le X_{(k)} \le \ldots \le X_{(p)}$. We have $f_{(k)} = n_{(k)}/N$, where $n_{(k)}$ is the number of observations in group k. The concentration curve is defined by plotting

$$L_{(k)} = \sum_{j=1}^{(k)} \frac{f_j Y_j}{\overline{Y}}$$

against $F1_{(k)} = \sum_{j=1}^{(k)} f_j$ [17]. The CI is then given by

$$CI = \frac{2}{\overline{Y}} cov[Y, F1] = \sum_{k=1}^{p-1} (F1_{(k)} L_{(k+1)} - F1_{(k+1)} L_{(k)}),$$

which is twice the area between the equalitarian 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, $0 < CI \le GI$. If it is protective, $-GI \le CI < 0$. The standard errors of GI and CI can also be estimated [16]. The absolute value of the ratio

$$CI/GI = \frac{cov[Y, F1]}{cov[Y, F]}$$

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 $\hat{Y} = g^{-1}(Z)$ and the residual $e = Y - \hat{Y}$, 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]

$$\frac{2}{\overline{Y}} cov[\hat{Y}, F^*] + \frac{2}{\overline{Y}} cov[e, F^*],$$

where F^* is either F or $F1$ 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 var[·] denote the variance of a quantity. We propose a correlation coefficient between F and $F1$,

$$\rho = \frac{cov[F, F1]}{\sqrt{var[F]\, var[F1]}} = \frac{cov[F, F1]}{var[F]} = \frac{\sum_{i=1}^{p} f_i (F_i - \overline{F})(F1_{(i)} - \overline{F})}{\sum_{i=1}^{p} f_i (F_i - \overline{F})^2},$$

as an overall measure of association between the exposure and the ill-health response. Note that ρ assesses the correlation between exposure distribution and response distribution based on cumulative functions, while F and $F1$ are rearranged using $f_i = f_{(k)}$, and $\text{cov}[F, F1] = \text{cov}[F1, F]$. If $0 < \rho \le 1$, the proportional fractions ranked by Y and those ranked by X are positively correlated. Otherwise, $-1 \le \rho < 0$ implies a negative correlation. The coefficient of determination is then

$$\rho^2 = \frac{cov[F, F1]^2}{var[F]^2} = \frac{\sum_{i=1}^{p} f_i (\hat{F}_i - \overline{F})^2}{var[F]},$$

which yields the proportion of disparity in Y explained by X. Let ε and $\varepsilon 1$ represent the residual terms for F and $F1$ predicted by the ranked Y and those ranked by X. GI and CI are algebraically related to ρ. Similar to the GI and CI, ρ can be decomposed into a model component and a residual component:

$$\rho = \frac{cov[Y, F]\, cov[Y, F1]}{var[Y]\, var[F]} + \frac{cov[\varepsilon, \varepsilon 1]}{var[F]}$$
$$= \frac{CI \cdot GI \cdot \overline{Y}^2}{4\, var[Y]\, var[F]} + \frac{cov[\varepsilon, \varepsilon 1]}{var[F]}.$$

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 $Y_{(k)}$, viz,

$$RR_{(k)} = (L_{(k)} - L_{(k-1)}) / (F1_{(k)} - F1_{(k-1)}) = Y_{(k)} / \overline{Y}\ .$$

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: $RR_{(k)1} = RR_{(k)} / RR_{(1)} = Y_{(k)} / Y_{(1)}$. More generally, let $RR_{(k)m} = RR_{(k)} / RR_{(m)} = Y_{(k)} / Y_{(m)}$. The variance of $RR_{(k)m}$ can be derived as

$$var\left[RR_{(k)m}\right] = \frac{Y_{(k)}(1 - Y_{(k)})}{n_{(k)} Y_{(m)}^2} + \frac{Y_{(k)}^2 (1 - Y_{(m)})}{n_{(m)} Y_{(m)}^3}$$

(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 $D = \sum_i d_i$, then we have

$$OR_{(k)} = \left(d_{(k)}(N - D)\right) \Big/ \left(D(n_{(k)} - d_{(k)})\right).$$

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

	Index	95% Confidence Interval	Logistic model decomposition (Contribution)	
			Model	Residual
Colorectal polyps GI CI	 0.073 -0.060	 0.065 to 0.082 -0.069 to -0.051	 0.036(49%) -0.059(98%)	 0.037(51%) -0.001 (2%)
N = 976	$\overline{Y}$ = 0.5	|CI/GI|=82%	ρ = -0.583	ρ^2 = 34%
Leukaemia GI CI	 0.374 0.338	 0.367 to 0.382 0.330 to 0.346	 0.313(84%) 0.378(112%)	 0.062(16%) -0.04(-12%)
N = 61902	$\overline{Y}$ = 0.614×10^{-3}	|CI/GI|=90%	ρ = 0.819	ρ^2 = 67%

Table 1: Summary of the examples.

proportion of ill-health events that can be avoided if the exposure is eliminated. When the exposure is continuous, the $AR_{(k)}$ can be viewed as the proportion of the incidence of ill-health that will be reduced if the exposure is reduced to $X_{(k)}$, rather than being totally eliminated [4,24]. In the case of grouped data, the $AR_{(k)}$ takes the form

$$AR_{(k)} = 1 - L_{(k-1)} - \frac{(L_{(k)} - L_{(k-1)})(1 - F1_{(k-1)})}{F1_{(k)} - F1_{(k-1)}},$$

which measures the health effect of a more relevant reduction in the risk rather than complete elimination of the risk. For example, a smoking cessation policy intends to reach a nominated target level $X_{(k)}$ rather than achieving an unrealistic zero smoking prevalence. The *AR* can take on negative values, if the *AR* is used for studying protective factors and the concentration curve lies above the diagonal line. As noted by Llorca and Delgado-Rodriguez [13], when $Y = 0$, $AR = 1 - RR$. 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 $1 - \bar{Y} = 1 - 0.5 = 0.5$ and the *GI* is 0.073, showing that the case-control grouping reflects 15% (0.073/0.5) of the total

Mean servings *Xk*	Cases *d*	Total *n*	Rate *Y*	Cumulative proportions		Concentration index *CI*	Attributable risk *AR*	Actual *RR*	Logistic mode *RR*
				F(*k*)	*L*(*k*)				
0	13	17	0.765	0.017	0.027	-0.0004	-0.5294	1.529	1.179
1	36	60	0.600	0.079	0.100	-0.0013	-0.2057	1.200	1.145
2	55	99	0.556	0.180	0.213	-0.0040	-0.1239	1.111	1.112
3	70	137	0.511	0.321	0.357	-0.0046	-0.0507	1.022	1.078
4	77	151	0.510	0.475	0.514	-0.0084	-0.0494	1.020	1.044
5	59	125	0.472	0.603	0.635	0.0004	-0.0096	0.944	1.010
6	54	102	0.529	0.708	0.746	-0.0087	-0.0551	1.059	0.976
7	33	74	0.446	0.784	0.814	-0.0003	-0.0063	0.892	0.941
8	33	64	0.516	0.849	0.881	0.0002	-0.0365	1.031	0.907
9	24	46	0.522	0.897	0.930	-0.0159	-0.0383	1.043	0.874
10	10	36	0.278	0.933	0.951	-0.0061	0.0122	0.556	0.840
11	6	18	0.333	0.952	0.963	-0.0032	0.0048	0.667	0.807
12	9	21	0.429	0.973	0.982	-0.0071	-0.0044	0.857	0.774
6	54	102	0.529	0.708	0.746	-0.0087	-0.0551	1.059	0.976
7	33	74	0.446	0.784	0.814	-0.0003	-0.0063	0.892	0.941
8	33	64	0.516	0.849	0.881	0.0002	-0.0365	1.031	0.907
9	24	46	0.522	0.897	0.930	-0.0159	-0.0383	1.043	0.874
10	10	36	0.278	0.933	0.951	-0.0061	0.0122	0.556	0.840
11	6	18	0.333	0.952	0.963	-0.0032	0.0048	0.667	0.807
12	9	21	0.429	0.973	0.982	-0.0071	-0.0044	0.857	0.774
14	4	15	0.267	0.989	0.990	-0.0010	0.0042	0.533	0.711
18	5	11	0.455	1.000	1.000	0.0000	0.0000	0.909	0.591
Total	488	976	0.500			-0.0603	-0.0603	1.000	1.000

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

Radiation dose *X* (1)	Incidence /1000 *Y* (2)	*i* (3)	*k* (4)	*f* (5)	$F_i(y)^a$ (6)	$L_i(y)^a$ (7)	$F_{(k)}(x)$ (8)	$L_{(k)}(x)$ (9)	*GI* (10)	*CI* (11)	RR_m (12)	OR_m (13)	*AR* (14)	$\hat{RR}$ (15)
<250	0.244	1	1	0.132	0.132	0.053	0.132	0.053	0.002	0.002	1.000	1.000	0.602	1.000
250-499	0.290	2	2	0.167	0.299	0.132	0.299	0.132	0.007	0.026	1.187	1.187	0.537	1.354
500-749	0.593	5	3	0.164	0.733	0.474	0.463	0.289	0.007	-0.006	2.425	2.425	0.192	1.833
750-999	0.343	3	4	0.188	0.487	0.237	0.651	0.395	0.019	0.069	1.405	1.404	0.410	2.483
1000-1249	0.752	7	5	0.172	0.930	0.711	0.823	0.605	0.045	0.015	3.081	3.079	0.178	3.363
1250-1499	0.588	4	6	0.082	0.570	0.316	0.905	0.684	0.038	0.045	2.409	2.408	0.225	4.554
1500-1749	1.231	8	7	0.039	0.970	0.789	0.945	0.763	0.039	0.006	5.042	5.037	0.126	6.166
1750-1999	0.645	6	8	0.025	0.758	0.500	0.970	0.789	0.074	0.039	2.641	2.640	0.179	8.348
2000-2249	2.130	9	9	0.015	0.985	0.842	0.985	0.842	0.022	0.097	8.732	8.716	0.105	11.301
2250-2499	7.859	12	10	0.008	1.000	1.000	0.993	0.947	0.000	0.022	32.404	32.157	-0.037	15.295
2500-2749	3.534	10	11	0.005	0.989	0.868	0.998	0.974	0.024	0.024	14.507	14.459	0.012	20.694
≥2750	6.623	11	12	0.002	0.992	0.895	1.000	1.000	0.097	0.000	27.273	27.099	0.000	28.005
Total	0.614			1.000					0.374	0.338	2.513	2.512	0.338	

GI: 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. [a]cumulative sum with respect to *i*.

Table 3: Concentration Curve and Index for the Radiation–induced Leukaemia Data, United Kingdom, 1935-1954.

inequality. According to the new coefficient of determination ρ^2, about 34% of the disparity in polyp incidence is explained by the mean servings of fruits and vegetables (X). This result appears more plausible than the $|CI/GI|$ ratio (82%) and the logistic model GI decomposition assessment (49%). The fitted logistic model is

$$\hat{Y} = exp(0.362 - 0.068X) / [1 + exp(0.362 - 0.068X)].$$

The CI or total AR is negative (-0.0603), indicating the concentration curve is above the diagonal line and the fruit and vegetable intake is a protective factor. An increase in the fruit and vegetable intakes to the average (5 to 6 servings) could potentially decrease the number of colon polyps by 6%. Decrease in the fruit and vegetable intake to zero can potentially increase colon polyps by 53% ($AR_{(1)} = -0.5294$, see Table 2). In other words, the decreased levels of fruit and vegetable intake are associated with an increased risk of polyps among the matched pairs. As shown in Table 2, the RR decreases with an increased level of mean servings per day. Detailed AR and RR estimates are listed in Table 2, in comparison with the logistic model estimates.

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 $1 - \overline{Y} = 1 - 0.000614 = 0.999$ 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 ρ^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

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

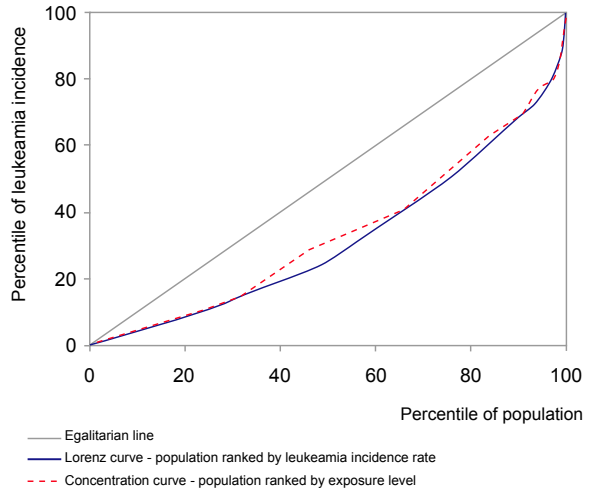
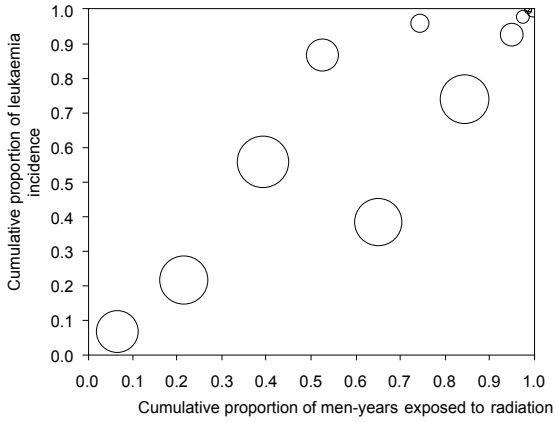


Figure 1: Lorenz and concentration curves for radiation induced leukaemia.



(Note: Bubble size indicates men-years)

Figure 2: Scatter diagram for correlation between leukaemia risk and radiation exposure.

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 ρ 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 variables 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 $|CI/GI|$ 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 *CI* and the predictor in the regression model $\hat{Y} = g^{-1}(Z)$, 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 *RR, OR* and *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 *GI* decomposition as demonstrated in the examples. The *GI* and *CI* estimation by grouped data may underestimate the true association, because they overlook the within-group variation [29]. The coefficient of determination ρ^2 is a conservative measure of the exposure-response association based on the cumulative distributions. The *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 *CI*, although the concentration curve will reflect the detailed positive and negative areas, and *OR* and *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

1. Steenland K, Deddens JA (2004) A practical guide to dose-response analyses and risk assessment in occupational epidemiology. Epidemiology 15: 63-70.
2. Checkoway H, Pearce N, Kriebel D (2004) Research Methods in Occupational Epidemiology (2ndedn), Oxford University Press, USA.
3. van Wijngaarden E (2006) A graphical method to evaluate exposure-response relationships in epidemiologic studies using standardized mortality or morbidity ratios. Dose Response 3: 465-473.
4. Greenland S (2001) Attributable fractions: bias from broad definition of exposure. Epidemiology 12: 518-520.
5. Birkett NJ (1992) Effect of nondifferential misclassification on estimates of odds ratios with multiple levels of exposure. Am J Epidemiol 136: 356-362.
6. Zhao LP, Kolonel LN (1992) Efficiency loss from categorizing quantitative exposures into qualitative exposures in case-control studies. Am J Epidemiol 136: 464-474.
7. Lee WC (1997) Characterizing exposure-disease association in human populations using the Lorenz curve and Gini index. Stat Med 16: 729-739.
8. Chatterjee N, Graubard BI, Gastwirth JL (2009) The use of the risk percentile curve in the analysis of epidemiologic data. Stat Interface 2: 123-131.
9. Greenland S (1995) Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. Epidemiology 6: 356-365.
10. Schneider MP (2004) Measuring Inequality: The Origins of the Lorenz Curve and the Gini Coefficient. La Trobe University.
11. Giorgi GM (1990) Bibliographic portrait of the Gini concentration ratio. Metron 183-221.
12. Gail MH (2009) Applying the Lorenz curve to disease risk to optimize health benefits under cost constraints. Stat Interface 2: 117-121.
13. Llorca J, Delgado-Rodríguez M (2002) Visualising exposure-disease association: the Lorenz curve and the Gini index. Med Sci Monit 8: MT193-197.
14. Kakwani NC (1980) Income Inequality and Poverty: Methods of Estimation and Policy Applications. Oxford University Press.
15. Wagstaff A, Paci P, van Doorslaer E (1991) On the measurement of inequalities in health. Soc Sci Med 33: 545-557.
16. Kakwani N, Wagstaff A, van Doorlsaer E (1997) Socioeconomic inequalities in health: measurement, computation, and statistical inference. J Econom 77: 87-103.
17. Wagstaff A, van Doorslaer E (2004) Overall versus socioeconomic health inequality: a measurement framework and two empirical illustrations. Health Econ 13: 297-301.
18. van Doorslaer E, Masseria C (2004) Income-Related Inequality in the Use of Medical Care in 21 OECD Countries. OECD Health Working Papers.
19. van Doorslaer E, Jones AM (2003) Inequalities in self-reported health: validation of a new approach to measurement. J Health Econ 22: 61-87.
20. Zhao LP, Kristal AR, White E (1996) Estimating relative risk functions in case-control studies using a nonparametric logistic regression. Am J Epidemiol 144: 598-609.
21. Pereira JC, de Barros LC (2006) Derivative ratio as a measure of effect: sex over age of occurrence of myocardial infarction in Brazil. Eur J Epidemiol 21: 263-266.
22. Goddard GH, Lewis CM (2010) Risk categorization for complex disorders according to genotype relative risk and precision in parameter estimates. Genet Epidemiol 34: 624-632.
23. Rockhill B, Newman B, Weinberg C (1998) Use and misuse of population attributable fractions. Am J Public Health 88: 15-19.
24. Rothman KJ, Greenland S, Lash TL (2008) Modern Epidemiology. Lippincott Williams & Wilkins.
25. Eide GE, Heuch I (2001) Attributable fractions: fundamental concepts and their visualization. Stat Methods Med Res 10: 159-193.
26. Upton AC (1961) The dose-response relation in radiation-induced cancer. Cancer Res 21: 717-729.
27. 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. The World Bank, Washington, DC.
28. Shorrocks AF (1982) Inequality decomposition by factor components. Econometrica 50: 193-211.
29. Deltas G (2003) The small-sample bias of the Gini coefficient: results and implications for empirical research. Rev Econ Statist 85: 226-234.
30. World Health Organization (2010) Priorities for Research on Equity and Health. World Health Organization.

Appendix

Variance of the relative risk

Consider two independent binomial random variables $d_i \sim Binomial(n_i, p_i)$,

where l = (1, 2) and p_l indicates the probability of ill health events in n_l observations. For $Y_l = d_l/n_l$ within the domain (0, 1], $r(Y) = Y_1/Y_2$. The first order Taylor expansion about Y is

$$r(Y) = r(p) + r_1'(p)(Y_1 - p_1) + r_2'(p)(Y_2 - p_2) + \text{remainder}.$$

Assuming $p = (E[Y_1], E[Y_2])$, the second order approximation for $E[r(Y)]$ is

$$E[r(Y)] \approx \frac{E[Y_1]}{E[Y_2]} + \frac{E[Y_1 - E[Y_1]]}{E[Y_2]} - \frac{E[Y_1]E[Y_2 - E[Y_2]]}{E^2[Y_2]} = \frac{E[Y_1]}{E[Y_2]} = r(p)$$

The variance is

$$var[r(Y)] = E\left[\{r(Y) - E[r(Y)]\}^2\right] \approx E\left[\{r_1'(p)(Y_1 - E[Y_1]) + r_2'(p)(Y_2 - E[Y_2])\}^2\right]$$

$$= E\left\{\left[\frac{Y_1 - E[Y_1]}{E[Y_2]} - \frac{E[Y_1](Y_2 - E[Y_2])}{E^2[Y_2]}\right]^2\right\} = \frac{var[Y_1]}{E^2[Y_2]} - \frac{2E[Y_1]cov[Y_1, Y_2]}{E^3[Y_2]} + \frac{E^2[Y_1]var[Y_2]}{E^4[Y_2]}.$$

Because of the independence between Y_1 and Y_2, we therefore have

$$var[r(Y)] \approx \frac{var[Y_1]}{E^2[Y_2]} + \frac{E^2[Y_1]var[Y_2]}{E^4[Y_2]} = \frac{p_1(1-p_1)}{n_1 p_2^2} + \frac{p_1^2(1-p_2)}{n_2 p_2^3}.$$

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

$$L(y) = \frac{1}{E[Y]}\int_0^y Yf(Y)dY$$

on the ordinate, against the cumulative distribution function (cdf) $F(y)$ along the abscissa, where the expectation $E[Y] = \int_0^\infty Yf(Y)dY$ exists [14]. By definition, $L'(y) \geq 0$ and $L''(y) \geq 0$. The health inequality may be measured by

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

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

$$F1(x) = \int_0^x f(X)dX = \int_0^x f(Y)dY.$$

Unlike $F(y)$ in which $f(Y)$ is integrated with respect to Y, $F1(x)$ is given by integrating $f(Y)$ with respect to X. The concentration curve of X can be obtained by plotting

$$L(x) = \frac{1}{E[Y]}\int_0^x Yf(Y)dY$$

against $F1(x)$ [16]. This definition means that RR may be measured by taking the first order differentiation of $L(F1(x))$ with respect to $F1(x)$ as

$$RR = L'(F1(x)) = \frac{dL(F1(x))}{dF1(x)} = \frac{Y}{E[Y]},$$

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

$$OR = \frac{1 - E[Y]}{E[Y](1/Y - 1)}.$$

The AR is

$$AR = 1 - L(F1(x)) - RR(1 - F1(x))$$

evaluated at $F1(x)$ [13]. Health inequality attributable to X can be measured by CI in the form of

$$CI = 1 - 2\int_0^1 L(F1(x))dF1(x).$$

In this paper, F and $F1$ are rearranged using $f(Y) = f(X)$. As a measure of overall association between the risk factor and ill-health, the correlation coefficient ρ between F and $F1$ is defined by

$$\rho = \frac{cov[F(y), F1(x)]}{\sqrt{var[F(y)]var[F1(x)]}}.$$

Since $var[F(y)] = var[F1(x)]$, we have

$$\rho = \frac{cov[F(y), F1(x)]}{var[F(y)]}.$$

This article was originally published in a special issue, **Medical statistics: Clinical and experimental research** handled by Editor(s). Dr. Herbert Pang, Duke University, USA.