Lorenz Curves and Treatment-Covariate Interactions in Clinical Trials

Marco Bonetti[1*], **Elena Colicino**[2], **and Pietro Muliere**[3]

[1] Bocconi University and Carlo F. Dondena Centre for Research on Social Dynamics and Public Policies, Milan, Italy

[2] Harvard School of Public Health, Boston, MA, U.S.A.

[3] Bocconi University, Milan, Italy

[*] Corresponding Author: marco.bonetti@unibocconi.it

ABSTRACT

A common objective in comparative two-treatment randomized clinical trials is the study of the possible heterogeneity of the treatment effect across subgroups of patients, with the objective of identifying patients who benefit the most (or the least) from a new treatment. Here we describe the connection that exists between an exploratory approach to such problem (STEPP, or the Subpopulation Treatment Effect Pattern Plot approach) and the Lorenz curve, and in particular the generalized Lorenz curve. We exploit such connection to construct a test for the absence of interaction between a continuous covariate and the difference in the mean of a continuous outcome between the two treatment groups. We also review some recent developments in the study of concentration for right censored survival data, which are also closed related to the Lorenz curve.

1. Introduction

Consider the general setting of comparing a new treatment against standard therapy in a two-arm randomized clinical trial. It is often of interest to study the heterogeneity of the treatment effect across groups of patients, to try and identify subgroups of patients who may benefit the most (or the least) from the new therapy, so that treatment can be tailored to the individual patient. Here we focus on the case in which subgroups are defined with respect to a one-dimensional covariate X – say, a biomarker or a baseline risk index.

The presence of an interaction effect is defined as the situation in which the effect of the new treatment on the outcome measure Y as compared to the standard treatment varies as X varies.

One may test for the presence of an interaction effect within a regression model. For example, one could use a fractional polynomial (FP) model both overall and within each treatment group (see Royston and Sauerbrei, 2008). In that case, the first step is to construct a multivariable adjustment model which may contain binary covariates and fractional polynomial transformations of any continuous covariates. The second step involves the fractional polynomial modelling within the adjustment model. This approach is more flexible than the standard polynomial regression and it is useful when one wishes to preserve the continuous nature of the covariate in the regression model, but suspects that some or all of the relationships may be non linear. Potential shortcomings of the fractional polynomial approach are its limited power to detect nonlinear interactions, and the possible sensitivity to extreme values at either end of the distribution of X. Owing to insufficient sample size, variables with a modest or weak effect may not be selected or by default, linear effects may be chosen instead of more realistic non linear functions.

In clinical applications, on the other hand, the patient population is often categorized into two or more disjoint groups according to the value of X based on cut-points placed over its support, and the interaction is studied within a model that includes main effects as well as treatment-covariate interaction terms (Altman et al., 1994). Such approach suffers from its dependence on the choice of the cut-points and from the loss of power due to categorization.

An alternative approach to the investigation of interactions, that sits between these two extremes, was first introduced in Bonetti and Gelber (2000) within the survival analysis setting, and it is called the Subpopulation Treatment Effect Pattern Plot (STEPP) approach. The approach is reviewed in Section 2. It is based on dividing the observations into subpopulations defined with respect to a collection of cutoff values of the covariate X, and on estimating the treatment effect within each subpopulation. To increase the number of patients that contribute to each point estimate (and hence the precision of the individual estimates) subpopulations are allowed to overlap. Plots showing the estimated treatment effects in the subpopulations can be used to investigate the possible interaction between X and treatment. STEPP is essentially a smoothing-by-binning approach, and it is appealing because it clearly defines the groups of

patients on which customary treatment effect measures are calculated using standard methods.

Different implementations of STEPP exist that depend on the nature of the outcome of interest. For example, in survival data clinical trials interest may be in the estimated difference between the survival probabilities at some fixed time point t^* between the two treatment groups, across the subpopulations. Other possibilities include the Cox proportional hazards model, or the cumulative incidence function within the competing risks setting (Bonetti and Gelber, 2004; Lazar et al., 2010).

Here we note the connections that exist between the STEPP construction and ideas related to the Lorenz curve, and in particular to its generalized version based on the regression function. Indeed, the Lorenz curve has been introduced to measure the level of inequality in a population, typically with respect to income or wealth. In Section 3 we recall this concept, and the notion of generalized Lorenz curve (Lorenz, 1905; Kakwani, 1977). In particular, we focus on the generalized Lorenz curve of the regression function as a tool to measure the dependence between two variables X (real) and Y (positive). By comparing two generalized Lorenz curves of the regression function $E(Y|X = x)$ between two groups in a population, one can study the presence of an interaction, i.e. whether the presence of a differential effect of the continuous predictor X on the outcome variable Y varies between the two groups. We relate this concept to the STEPP construction, show some results, and use them to construct a test for treatment-covariate interactions. In addition, we highlight other recent developments in subgroup identification in clinical studies based directly on concentration ideas, also closely related to the Lorenz curve. We close with some discussion in Section 4.

2. The Subpopulation Treatment Effect Pattern Plot (STEPP)

STEPP extends the idea of looking at treatment effect within subgroups of patients. In particular, the approach studies patterns of treatment effect across overlapping subpopulations of patients. Consider n patients in a two-arm (randomized) clinical trial, with a continuous baseline covariate $X \in [x_{min}, x_{max}] \subset \Re^1$ observed on all patients, and define K overlapping subpopulations of patients with respect to their values of X. Subpopulations $(\mathcal{P}_j, j = 1, ..., K)$ can be constructed according to a *sliding window* pattern: assign a patient i to subpopulation $\mathcal{P}_j$ when $x_i \in [l_j, u_j]$, where the two nondecreasing sets of numbers

$\{l_j\}$ and $\{u_j\}$ are such that $\{l_j \in [x_{min}, x_{max}], u_j = \inf\{u \geq l_j | P_n(l_j \leq X \leq u) \geq p)\}$ for some fixed $p \in (0, 1)$, with P_n the empirical distribution of X in the data. Subpopulations in the sliding window construction have an overlapping part and a part that differs from the neighboring supopulations. The patient with index i having covariate value x_i belongs to subpopulation $\mathcal{P}_j$ if $l_j \leq x_i \leq u_j$.

As another possibility, the *tail oriented* construction produces subpopulations by removing patients from the whole patient population, starting with the ones with the highest (or lowest) values of X. With increasing distance from the center, more and more patients with high covariate values (or low covariate values) are dropped. The cutoffs are defined as $\{l_j \in [X_{min}, X_{max}], u_j = X_{max}\}$ or $\{l_j = X_{min}, u_j \in [X_{min}, X_{max}]\}$. Note that, unlike the sliding window approach, the tail oriented construction includes among the subpopulations also the subpopulation that coincides with the set of all patients in the study.

Figure 2.1 illustrates the two constructions. The horizontal axis indicates the various subpopulations within which treatment effects are estimated, and shows on the vertical axis the range of covariate values used to define the cohort of patients included in each subpopulation.
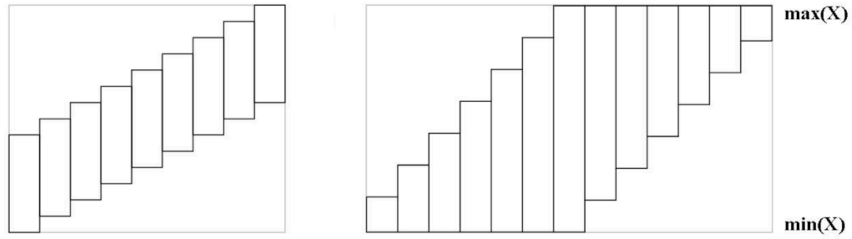

Figure 2.1: Sliding window (left) and Tail Oriented (right) construction of the subpopulations in STEPP.

For the jth subpopulation, an estimate $\widehat{\theta}_j$ of treatment effect is then computed. The plot of the treatment effect estimates $\widehat{\theta}_j$ within the subpopulations vs. the median value of X within each subpopulation is called a STEPP plot, and its examination may suggest the presence of patterns in the treatment effects as the covariate of interest varies.

The possible implementations of STEPP include the case in which treatment effect is defined as $\widehat{\theta}_j = \widehat{S}_{A,j}(t^*) - \widehat{S}_{B,j}(t^*)$, with $\widehat{S}(t)_{G,j}$ the Kaplan-Meier estimator of survival at time t within treatment group G inside subpopulation

$\mathcal{P}_j$, and with t^* a suitably chosen time point (Bonetti and Gelber, 2004). Figure 2.2 shows the plot of the four-year disease free survival (DFS) estimates in two of the four arms of a breast cancer clinical trial across the sliding window subpopulations (panel A), and the associated STEPP plot of their difference (panel B). The covariate of interest here is a well-known predictor of breast cancer prognosis, the tumor proliferation fraction (Ki-67), which is associated with the degree of effectiveness of chemotherapy. The prognostic and predictive value of Ki-67 LI were evaluated in the BIG 1-98 study, an international, double-blind phase III clinical trial of 8,010 postmenopausal women with early stage invasive breast cancer, who were randomly assigned to one of four adjuvant endocrine therapy arms: letrozole, tamoxifen, or sequences of these agents (letrozole to tamoxifen, tamoxifen to letrozole). We refer to Lazar et al. (2010) for a detailed description of that STEPP analysis.
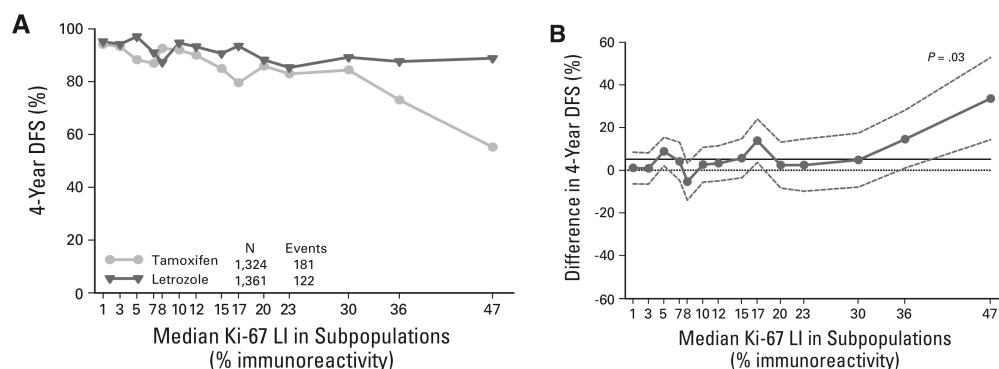

Figure 2.2: STEPP plots for 4-year disease-free survival (DFS) vs. Ki-67 labeling index (LI) in letrozole vs. tamoxifen. Each subpopulation contains approximately 150 patients, with approximately 50 overlapping patients. The left panel (A) shows the estimated 4-year DFS within each subpopulation for the two treatment groups; the right panel (B) shows the difference in the 4-year DFS estimates, with the 95% marginal confidence intervals (Reprinted with permission from Figure 1 in Lazar et al., 2010).

Another implementation of STEPP is based on the hazards ratio obtained from fitting a Cox PH model (Bonetti and Gelber, 2000). This implementation presents some difficulties in its interpretation due to the fact that the same patient typically belongs to more than one subpopulation, and that it is therefore not clear what probability model its survival time follows. Nevertheless, the desire to capture the difference in two survival distributions beyond the probability of

surviving beyond up to a specific time point makes this hazards ratio implementation useful.

Other implementations include the case in which treatment effect is defined to be the difference in a quantile of the survival function (say, the median survival time) as described in Bonetti et al. (2009b). The survival medians in treatment group G ($\in \{1, 2\}$) within the K subpopulations are obtained by inversion of $(\widehat{S}_{G,1}(\cdot), \dots, \widehat{S}_{G,K}(\cdot))$, the vector of the Kaplan-Meier estimators of the survival functions $S_{G,j}(\cdot)$, $j = 1, \dots, K$.

Lastly, treatment effect can be defined within the context of competing risks as the difference in the estimated cumulative incidence function (of, say, the cause of death of interest c) at some fixed time point t^* within each subpopulation: $F_1(t^*) = P(T \leq t^*, \text{cause} = c)$, or as the treatment effect parameter in a Fine and Gray semiparametric proportional hazards model for the competing risk of interest (Fine and Gray, 1999; Lazar et al., 2010). Recently, the graphical part of STEPP has been extended to the case of two continuous covariates (Pogue-Geile et al., 2013).

Apart from the graphical examination of the STEPP plot, statistical tests of significance can be performed to test for the homogeneity of the treatment effects across the subpopulations. To test the null hypothesis of no interaction between the covariate of interest (i.e. across subpopulations) and treatment effect one can use one of several possible test statistics. For example, the test statistic

$$T = \max \left\{ \frac{|\widehat{\theta}_j - \widehat{\theta}_{ALL}|}{\left[\widehat{\text{var}}(\widehat{\theta}_j - \widehat{\theta}_{ALL})\right]^{1/2}}, \, j = 1, \dots, K \right\}$$

can be used, where $\widehat{\theta}_{ALL}$ is the measure of treatment effect computed on all patients in the study. The distribution of T can be estimated by sampling repeatedly from the estimated asymptotic distribution of $(\widehat{\theta}_1, \dots, \widehat{\theta}_K, \widehat{\theta}_{ALL})$, and a p-value can thus be produced.

Lastly, a simultaneous confidence band around the plot of the estimated treatment effects can be obtained by solving numerically for γ the equation

$$P \left(\bigcap_{j=1}^{K} \left\{ \theta_j \in \left[\widehat{\theta}_j \pm \gamma(1.96)[\widehat{\text{var}}(\widehat{\theta}_j)]^{1/2} \right] \right\} \right) = 1 - \alpha$$

for a sample of random variables generated from the estimated asymptotic distribution of the estimates. The parameter γ represents the widening of the

---

marginal confidence intervals that is necessary to produce the desired simultaneous coverage of the band. Note that Bonferroni adjustment can also be used, as well as marginal confidence intervals for each subpopulation. Alternatively to large sample theory, in general one can also perform more accurate testing by using permutation-based inference (Bonetti et al., 2009b).

Several examples of clinical applications of the various implementations of STEPP can be found in Lazar et al. (2010). The R *stepp* package for both the asymptotic and the permutation-based procedures is available on CRAN.

3. Lorenz-Curve, Generalized Lorenz Curve of the Regression Function, and Concentration

3.1. The Lorenz Curve

The Lorenz curve (LC) has been used since 1905 to describe concentration and inequality in distributions of resources. Its main importance is in economics, when dealing with income and wealth: it is still considered as a simple method for visualizing distributions of income or wealth with respect to their inherent inequality or concentration, through the graph of the cumulative proportion of total income or wealth owned, against the cumulative proportion of the population owning it.

For a nonnegative random variable $Y \sim f(y)$ with finite and strictly positive expected value $\mu = E(Y)$, the Lorenz curve is defined as the function $(F, C)(y)$ given by

$$(F, C)(y) = (F(y), C(y)) = \left(\int_0^y f(u)du, \frac{1}{\mu} \int_0^y uf(u)du \right), \ y \geq 0$$

(Lorenz, 1905). Note that this can also be written as

$$L_Y(p) = \frac{1}{E(Y)} \int_0^p F_Y^{-1}(z)dz, \ p \in [0, 1], \quad (3.1)$$

where the quantile function F_Y^{-1} is defined as the pseudoinverse of the cumulative distribution function (cdf), F_Y,

$$F_Y^{-1}(p) = \inf\{y | F_Y(y) \geq p\}, \ 0 \leq p \leq 1 \quad (3.2)$$

as in Pietra (1915).

Note that any distribution supported on the non-negative half line with a finite and positive first moment admits a Lorenz curve. It is a direct consequence of

the definition that the Lorenz curve $L(p)$ is continuous on $[0,1]$, with $L(0) = 0$ and $L(1) = 1$, increasing, and convex.

The Lorenz curve is often represented through its dual. The Dual Lorenz curve (DLC) of a non negative positive random variable Y with finite expectation μ is the graph of

$$ L_Y^*(p) = \frac{1}{\mu} \int_{1-p}^{1} F_Y^{-1}(z) dz $$

where $p \in [0,1]$. The dual Lorenz curve can be denoted as $L_Y^*(p) = 1 - L_Y(1-p)$, so it is the centrally symmetric curve to the LC, and it is therefore a concave, increasing and continuous function on $[0,1]$, with $L^*(0) = 0$ and $L^*(1) = 1$.

The empirical counterparts of these curves can be expressed in terms of order statistics. In particular, considering n points, let $y_{(i)n}$ denote the i-th smallest value of the random variable Y, with $i = 1, ..., n$. Then the empirical Lorenz curve is defined as:

$$ L\left(\frac{k}{n}\right) = \frac{\sum_{i=1}^{k} y_{(i)n}}{\sum_{i=1}^{n} y_{(i)n}} \tag{3.3} $$

with $k = 1, \ldots, n$ and with $L(0) = 0$. (The empirical dual Lorenz curve is defined similarly). Essentially, the Lorenz curve describes the amount of total wealth (say) owned by the $p \times 100\%$ poorest individuals, for $p \in [0,1]$ (see also Gastwirth, 1971).

3.2. The Generalized Lorenz Curve of the Regression Function

A generalization of the Lorenz curve has been proposed in Kakwani (1977). Let $g(x)$ be a continuos function of x such that its first derivative exists, and $g(x) \geq 0 \ \forall x$. If $0 < E(g(X)) < +\infty$, then the generalized Lorenz curve (GLC) is defined as

$$ L_{g(X)}(p) = \frac{1}{E(g(X))} \int_0^{x_p} g(t) dF_X(t) = \frac{1}{E(g(X))} \int_0^{p} g(F_X^{-1}(z)) dz \tag{3.4} $$

where $p \in [0,1]$ and $x_p = F_X^{-1}(p)$. It turns out that the GLC can be used to study orderings of the dependence between two variables: Blitz and Brittain (1964) introduced a monotone dependence structure based on the generalized

Lorenz curve of the regression function $g(x) = E(Y|X = x)$:

$$L_{E(Y|X)}(p) = \frac{1}{E(E(Y|X))} \int_0^{x_p} E(Y|X = t)dF_X(t)$$

$$= \frac{1}{E(Y)} \int_0^p E(Y|X = F_X^{-1}(z))dz, \qquad (3.5)$$

also with $p \in [0, 1]$ and $x_p = F_X^{-1}(p)$ (Blitz and Brittain, 1964). As a consequence, bivariate distributions can be ordered according to monotone dependence. The symbol $\doteq$ is used below to denote the equivalence of random variables Z_1 and Z_2: $Z_1 \doteq Z_2$ iff the probability of the event $\{Z_1 = Z_2\}$ is equal to one. In particular, the following properties hold:

1. $L_{E(Y|X)}(p)$ passes through the points $(0, 0)$ and $(1, 1)$.

2. $L_{E(Y|X)}(p)$ is increasing if and only if $g(x) > 0$ for all x.

3. $L_{E(Y|X)}(p)$ is concave if and only if $g(x)$ is nondecreasing for all x.

4. $L_Y(p) \leq L_{E(Y|X)}(p) \leq L_Y^*(p)$, where $L_Y(p)$ is the Lorenz curve and $L_Y^*(p)$ is its dual.

5. $L_{E(Y|X)}(p) = p$ for all $p \in (0, 1)$ if and only if $E(Y|X) \doteq E(Y)$.

6. $L_{E(Y|X)}(p) = L_Y(p)$ for all $p \in (0, 1)$ if and only if $g(x)$ is increasing and such that $E(Y|X) \doteq Y$.

7. $L_{E(Y|X)}(p) = L*_Y(p)$ for all $p \in (0, 1)$ if and only if $g(x)$ is decreasing and such that $E(Y|X) \doteq Y$.

8. If $g(x)$ is nondecreasing for all x then $L_{E(Y|X)}(p) \leq p$. The inequalities are reversed if $g(x)$ is nonincreasing. In both cases, cov(X, Y)=0 and $L_{E(Y|X)}(p) = p$ hold if and only if $E(Y|X) \doteq E(Y)$.

9. $L_{E(Y|X)}(p)$ is invariant under F_X-increasing transformation of X [a function $h : \mathbb{R} \to \mathbb{R}$ is F_X-increasing if for all $s, t \in \mathbb{R}$: $F_X(s) < F_X(t)$ implies $h(s) < h(t)$].

10. Let $(X, Y), (X', Y') \in \pi(F_X, F_Y)$, where $\pi(F_X, F_Y)$ is the class of bivariate nonnegative random variables (X, Y) with continuous marginal distribution functions F_X and F_Y. Then $L_{E(Y|X)}$ will be above (below) $L_{E(Y'|X')}$ if the elasticity of $E(Y|X = x)$ is less (greater) than the elasticity of $E(Y'|X' = x)$.

For the proofs of these properties we refer to Muliere and Petrone (1992).

Example 1 Consider the bivariate random variable (X, Y) with joint probability density function (pdf) $f_{X,Y}(x, y) = 1_{[0,1]}(x)1_{[x,1+x]}(y)$. Through standard calculations one obtains the Lorenz curve as

$$L_Y(p) = \begin{cases} \frac{(2p)^{3/2}}{3} & \text{if } p \in (0, 0.5) \\ 2p - 1 + \frac{(2(1-p))^{3/2}}{3} & \text{if } p \in [0.5, 1) \end{cases};$$

the dual Lorenz curve as

$$DL_Y(p) = \begin{cases} 1 - \frac{(2(1-p))^{3/2}}{3} & \text{if } p \in (0, 0.5) \\ 2p - \frac{(2p)^{3/2}}{3} & \text{if } p \in [0.5, 1) \end{cases};$$

and the generalized Lorenz curve of the regression function $E(Y|X) = x + \frac{1}{2}$ as $L_{E(Y|X)}(p) = \frac{1}{2}(p^2 + p)$.

These functions are represented in Figure 3.1 (left panel). As noted above, the population GLC curve is contained between the LC and the DLC.
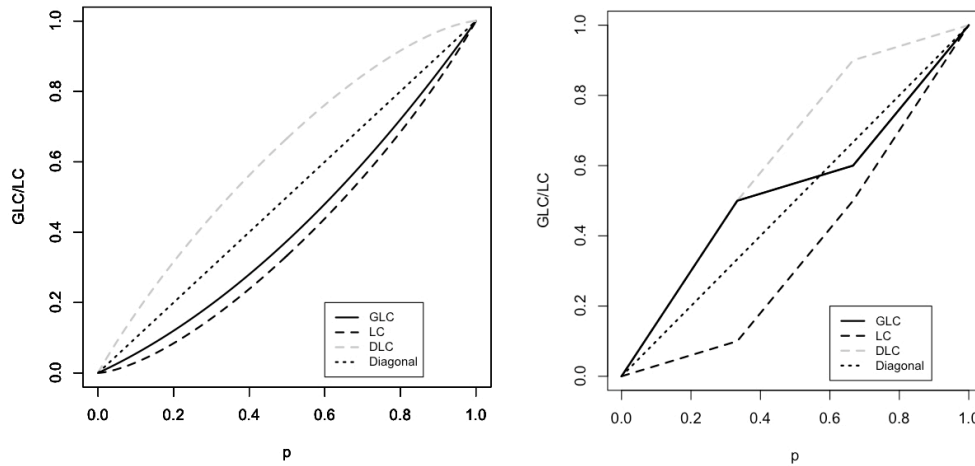

Figure 3.1: Generalized Lorenz curve (GLC) of $E(Y|X = x)$, Lorenz curve (LC), and Dual Lorenz curve (DLC) of Y for Example 1 (left panel). Generalized Lorenz curve (GLC) of the regression function, Lorenz curve (LC), ad Dual Lorenz curve (DLC) of Y corresponding to the sample $\{X = \{3, 7, 2\}, Y = \{1, 4, 5\}\}$ (right panel).

By assuming the absolute continuity of (X, Y), one can prove the following equivalent expression of the generalized Lorenz curve of the regression function:

$$L_{E(Y|X)}(p) = \frac{p}{E(Y)} E(Y|X \le x_p) \qquad (3.6)$$

(see, e.g., Muliere and Petrone, 1992).

Now, let us consider the observed sample $(\mathbf{x}, \mathbf{y}) = ((x_1, y_1), ..., (x_n, y_n))$ extracted by a bivariate population described by the absolutely continuous random vector (X, Y). The empirical version of the generalized Lorenz curve can be obtained by using an estimator $\widehat{g}(x)$ of the regression function $g(x) = E(Y|X = x)$. Or, one may choose to use $\widehat{g}(x) = y(x)$ for the observed values (x_i, y_i), thus constructing the empirical GLC from (3.6) as

$$L_{E(Y|X)}(p) = \frac{p}{\frac{\sum_{i=1}^n y_i}{n}} \frac{\sum_{i=1}^n y_i 1(x_i \le x_p)}{\sum_{i=1}^n 1(x_i \le x_p)} = \frac{\sum_{i=1}^n y_i 1(x_i \le x_p)}{\sum_{i=1}^n y_i}. \qquad (3.7)$$

An example of empirical LC and GLC is shown in the right panel of Figure 3.1. The figure shows the Lorenz curve and the dual Lorenz curve corresponding to the sample $\{X = \{3, 7, 2\}, Y = \{1, 4, 5\}\}$, as well as the estimated generalized Lorenz curve of $E(Y|X = x)$.

3.3. Asymptotic Behavior of the Empirical GLC of the Regression Function, and a Test for Interaction

The following result describes the asymptotic behavior of the (scaled) empirical GLC of the regression function as constructed in (3.7) above.

Theorem 3.1. *Let (X, Y) be an absolutely continuous random vector with support $(\mathbb{R}^+)^2$, and with finite first and second marginal moments. Let $F(x, y)$ and $\widehat{F}(x, y)$ be the joint cdf and the joint empirical cdf from an i.i.d. sample extracted from F, respectively. For $p \in [0, 1]$ define the map $\phi(F(x, y), p)$ and its empirical counterpart $\phi(\widehat{F}(x, y), p)$ as:*

$$
\begin{aligned}
\phi(F(x, y), p) &= pE(Y|X \le F_X^{-1}(p)) = \int_{(\Re^+)^2} y\, 1(x \le F_X^{-1}(p))) dF(x, y) \\
&= \int_{(\Re^+)^2} y\, 1(x \le F_X^{-1}(p))) dF(x, y)
\end{aligned}
$$

and

$$\phi(\widehat{F}(x,y),p) = p\widehat{E}(Y|X \le \widehat{F}_X^{-1}(p)) = \int_{(\Re^+)^2} y\, 1(x \le \widehat{F}_X^{-1}(p))d\widehat{F}(x,y)$$

$$= \frac{1}{n}\sum_{i=1}^{n} Y_i\, 1(X_i \le \widehat{F}_X^{-1}(p)),$$

where $\widehat{F}(x,y)$ is the ecdf of (X,Y) and $\widehat{F}_X^{-1}$ is the empirical quantile function of X.

Then, $\sqrt{n}(\phi(\widehat{F}(x,y),p) - \phi(F(x,y),p))$ is asymptotically equivalent to

$$\int_{(\Re^+)^2} v\, 1(u \le F_X^{-1}(p))d(\mathbb{G}_{F(u,v)}), \tag{3.8}$$

where $\mathbb{G}_{F(u,v)}$ is a zero mean Gaussian process with known covariance function.

We refer to Appendix A for a sketch of the proof of Theorem 3.1. While this result describes the large sample behavior of $\widehat{E}(Y|X \le \widehat{F}_X^{-1}(p))$, the fact that its expression involves an integration with respect to a process makes it hard to use in practice. Indeed, we recommend the use of permutation-based inference instead.

Note that above we have defined the empirical generalized Lorenz curve of the regression function as:

$$L_{E(Y|X)}(p) = \frac{\sum_{i=1}^{n} y_i 1(x_i \le x_p)}{\sum_{i=1}^{n} y_i}.$$

The GLC of the regression function is therefore related to an increasing sequence of subpopulations defined by the values $x_p = F_X^{-1}(p)$ of the predictor X, across all values of $p \in [0,1]$. In particular, for a fixed value p the GLC of the regression function is the ratio between the conditional expected value of Y given that $X \le F_X^{-1}(p)$ and the overall expected value of Y. This clearly resembles the tail oriented version of STEPP described in Section 2 when one defines as treatment effect the (estimated) differences in conditional means of exactly the form $E(Y|X \le F_X^{-1}(p))$. For the subpopulations containing all patients that have values of X smaller than the empirical percentiles x_p, these are therefore estimated nonparametrically.

This parallel suggests alternative tests for interaction, motivated by the comparison of the level of dependence between X and the outcome Y in the two

treatment groups, when such dependence is expressed by the GLC of the regression function. In particular, one may study the presence of an interaction effect through the difference between the estimated GLC of the regression function for the two treatment groups $\left(\widehat{L}^{(1)}_{E(Y|X)}(p) \text{ and } \widehat{L}^{(2)}_{E(Y|X)}(p)\right)$ across all values of $p \in [0, 1]$. For example, one may consider the following two test statistics:

$$T_1 \;=\; \int_0^1 \left|\widehat{L}^{(1)}_{E(Y|X)}(p) - \widehat{L}^{(2)}_{E(Y|X)}(p)\right| dp$$

and

$$T_2 \;=\; \int_0^1 \left(\widehat{L}^{(1)}_{E(Y|X)}(p) - \widehat{L}^{(2)}_{E(Y|X)}(p)\right)^2 dp.$$

Inference for the proposed test statistics could be based on Theorem 3.1. However, as noted above, it is more practical to make use of permutation testing. Permutation tests are conditional statistical procedures, where the conditioning is with respect to the permutation set of the observed data which plays the role of reference set for the inference. Under the null hypothesis and assuming exchangeability, the conditional probability distribution of a generic element that belongs to the set of the permutations of the observed data is independent of the distribution P of the data. This allows permutation inference to be invariant with respect to P within H_0. Due to this invariance property, permutation tests are therefore distribution free and non parametric. To perform a permutation test one follows the customary four steps: (i) define the null hypothesis, the alternative hypothesis, and the assumptions; (ii) choose the appropriate test statistic and calculate its value using the observed data; (iii) calculate the value of the test statistic for all permutations of the data; and (iv) reject or not reject the null hypothesis using the estimated distribution of the test statistic under permutation. When the number of all permutations is very large one typically uses the conditional Monte Carlo approach that chooses a sample of permutations, thus producing an estimate of the exact p-value (Good, 1994). In our problem one may develop a permutation test under the null hypothesis that $F_{X,Y}$ is equal in the two treatment-defined populations by permuting the pairs of values $(X_i Y_i)$ across the two patient groups.

3.4. Concentration

An alternative way of studying the differential effect of a new treatment on a subgroup of patients exists that is based directly on the Lorenz curve con-

structed on the two treatment groups. Disregarding covariates, let us go back to the Lorenz curve of a positive random variable Y. The area between the 45 degree line and the Lorenz curve is a measure of the *concentration* of the distribution of Y.

Let the random variable of interest be $Y \geq 0$ with cdf F_Y, survival function $S_Y(y) = 1 - F_Y(y)$, density function $f_Y(y)$, finite expected value $\mu > 0$, and variance $Var(Y)$. The coefficient of mean difference is defined as

$$\delta = \int_{(\Re^+)^2} |y_1 - y_2| \, dF_Y(y_1) dF_Y(y_2).$$

The Gini coefficient of concentration for F_Y is $G = \delta/(2\mu)$, and it can be estimated from an i.i.d. sample $Y_1, \ldots, Y_n$ drawn from the population by $\widehat{G}_H = D/(2\overline{Y}_n)$, with $D = (n(n-1))^{-1} \sum_{j \neq k} |Y_j - Y_k|$ and $\overline{Y}_n$ the sample mean (Gini, 1912). Asymptotic normality of the empirical Gini index is well known (Hoeffding, 1948). Importantly, the Gini index G is equal to twice the concentration area, or the area between the 45 degree line and the Lorenz curve. Thus G is consistent with orderings of distributions that are induced by the Lorenz curves. It is therefore reasonable to compare distribution with respect to their concentration.

In particular, testing for the equality of two population Gini indices may capture differences in the two distributions that may not be revealed by other measures. Rejection of the null hypothesis of equality of the two outcome distributions of the two patient groups from the point of view of their concentration may suggest that *subgroups of patients exist for whom treatment has a strong positive (or detrimental) effect*. Equivalently, differences in concentration between groups may suggest the presence of a *differential* treatment effect on some patient groups.

Following these ideas, the use of the Gini index has recently been extended to right-censored survival data arising from clinical studies to detect differences in concentration between the survival time distributions of two treatment groups of patients (Bonetti et al., 2009a). An alternative expression for G is

$$G = 1 - \frac{\int_{\Re^+} S_Y^2(u) du}{\int_{\Re^+} S_Y(u) du}.$$

One may consider the following *restricted* version of G:

$$G_t = 1 - \frac{\int_0^t S_Y^2(u)du}{\int_0^t S_Y(u)du},$$

and estimate it by replacing S_Y by its Kaplan-Meier estimator $\widehat{S}_Y$. Under some regularity conditions, the restricted Gini statistic has an asymptotic normal asymptotic distribution, and a plug-in estimator of its asymptotic variance is available (Bonetti et al., 2009a). A test is then constructed that compares the estimated (restricted) Gini concentration coefficient between the two treatment groups. The Gini test complements other tests in the sense that it captures differences in survival distributions not typically detected by other tests (e.g. Wilcoxon, Log-rank, or Gray-Tsiatis), and a large simulation study suggests that the Gini index should be considered together with other existing tests to detect differences in survival distributions. Here, too, permutation distribution inference is recommended, but only for the case of small and unbalanced groups (Gigliarano and Bonetti, 2013). Additional results in that reference apply specifically to cure rate models. An R function for the implementation of the restricted Gini test for right censored survival data is available from the Authors.

4. Conclusions

As we have seen above, STEPP is an exploratory tool which is easy to interpret, and it provides an opportunity to detect interactions beyond those that may be apparent based on regression models. STEPP may suggest the presence of a pattern in treatment effects, and it falls somewhere between the simple subgroup analyses performed on disjoint patient subgroups and full modeling. When implementing the methods, it is important to assess the robustness of the results of the analysis to the choice of the parameters that define the subpopulations.

As we have discussed here, interesting connections exist between STEPP and Lorenz curves. In particular, starting from the definition of the generalized Lorenz curve of the regression function, we have suggested another way to explore the presence of the interaction effect between treatment and a continuous predictor X. Using test statistics based on the difference of two generalized Lorenz curve of the regression function, one may construct increasing overlapping subpopulations of patients (as in the tail-oriented version of STEPP) and

compare subpopulation-specific conditional means of the outcome Y.

A connection also exists between the GLC of the regression function and the problem of identifying the optimal cutoff to dichotomize a population into two groups with respect to the (continuous) covariate X. Such problem can be described as the problem of finding the value $p \in (0,1)$, if it exists, such that it maximizes the absolute value of some distance $\Delta(p)$ between the two conditional distributions of $Y|X \leq x_p$ and $Y|X > x_p$, where we again indicate with x_p the percentile $F_X^{-1}(p)$. A particular case is that in which such distance between the two conditional distributions is defined to be the difference between the two conditional expectations, i.e.

$$p^* = \mathrm{argmax}_{p \in (0,1)} |\Delta(p)| = \mathrm{argmax}_{p \in (0,1)} |E(Y|X > x_p) - E(Y|X \leq x_p)|$$

(see also the related maximally selected chi square approach in Miller and Siegmund, 1982). The problem of finding p^* can equivalently be expressed in terms of the GLC of the regression function. In particular, it is easy to show that $\Delta(p) = \frac{1-p}{p} E(Y) \left[p - L_{E(Y|X)}(p) \right]$.

Here we have seen that the Lorenz curve can also be exploited directly to detect a differential treatment effect in different subgroups of the patient population by studying differences in concentration, both in the uncensored case and in the right censored case. Whatever the approach that one chooses to exploring heterogeneities in treatment effects, positive results should be confirmed by analysing results from other datasets investigating similar treatment comparisons. These (and other) approaches may also help provide further justification to subsequent subgroup analyses in a study, i.e. for the study of treatment effect within subgroups of patients. Such practice is still quite controversial but we believe that it is an ethical necessity, especially when studying survival in life threatening diseases.

Appendix A - Outline of Proof of Theorem 1

The asymptotic study of the empirical generalized Lorenz curve of the regression function (with $\widehat{E}(Y|X = x) = y(x)$) is based on the concepts of Hadamard differentiability, the functional delta method, and the central limit theorem for empirical measures (see, e.g., van der Vaart, 1998).

Let $h_n(x, p) = [1(x \leq \widehat{F}_X^{-1}(p)) - 1(x \leq F_X^{-1}(p))]$, and write

$$\int_{(\Re^+)^2} h_n(x, p)\, y\, d\widehat{F}_{X,Y}(x, y) =$$

$$= \int_{(\Re^+)^2} h_n(x, p)\, y\, dF_{X,Y}(x, y) +$$

$$+ \int_{(\Re^+)^2} h_n(x, p)\, y\, d\left[\widehat{F}_{X,Y}(x, y) - F_{X,Y}(x, y)\right]$$

$$= A_n(p) + B_n(p).$$

Application of the dominated convergence theorem yields $A_n(p) \xrightarrow{\text{P}} 0$ uniformly in p, as does the term $B_n(p)$ since $|B_n(p)| \leq |\widehat{E}(Y) - E(Y)|$. Now, consider the map $\phi : (F(x, y), p) \mapsto \int 1(u \leq x_p)v dF(u, v)$, where $x_p = F_X^{-1}(p)$. The map ϕ is Hadamard differentiable at $F(x, y)$ if there is a continuos linear map, the Hadamard derivative, $\phi\prime_{F(x,y)}$ such that:

$$\phi'_{F(x,y)}(h) = \lim_{t \to 0} \frac{1}{t} \left[\phi\left(F(x, y) + th_t, p\right) - \phi(F(x, y), p)\right],$$

where the sequence $h_t \to h$ uniformly in $D[0, 1]$, h is continuous, and $F(x, y) + th_t$ is contained in $D[0, 1]$ for all t. One can find

$$\phi'_{F(x,y)}(h) = \lim_{t \to 0} \frac{1}{t} \left[\int_{(\Re^+)^2} y\, 1(x \leq x_p) d(F(x, y) + th_t) - \right.$$

$$\left. - \int_{(\Re^+)^2} y\, 1(x \leq x_p) dF(x, y)\right]$$

$$= \int_{(\Re^+)^2} y\, 1(x \leq x_p) dh.$$

The class of lower rectangles is P-Donsker for any law of $(X_1, Y_1), ..., (X_n, Y_n)$, and the indicator function is square integrable (van der Vaart and Wellner, 1996; Das Gupta, 2008). By the central limit theorem for empirical measures, the sequence $\sqrt{n}(\widehat{F}(x, y) - F(x, y))$, converges weakly to is a zero-mean Gaussian process $\mathbb{G}_{F(x,y)}$ with covariance function given by $Cov(\mathbb{G}_{F(x_i,y_i)}, \mathbb{G}_{F(x_j,y_j)}) = F((x_i \wedge x_j), (y_i \wedge y_j)) - F(x_i, y_i)F(x_j, y_j)$, where $\wedge$ denotes the minimum and $i, j = 1, .., n$ (van der Vaart and Wellner, 1996). Lastly, by the functional delta method, one concludes that $\sqrt{n}(\phi(\widehat{F}(x, y), p) - \phi F((x, y), p)$ is asymptotically equivalent to $\int y\, 1(x \leq x_p) d(\mathbb{G}_{F(x,y)})$.

Acknowledgements

References

1. Altman, D. G., Lausen, B., Sauerbrei, W. and Schumacher, M. (1994). Dangers of using "optimal" cutpoints in the evaluation of prognostic factors, *Journal of National Cancer Institute* 86: 829–835. DOI:10.1093/jnci/86.11.829

2. Blitz, R. C. and Brittain, J. A. (1964). An extension of the Lorenz Diagram to the Correlation to Two Variables, *Metron* XXIII137–143.

3. Bonetti, M. and Gelber, R. D. (2000). A graphical method to assess treatment-covariate interactions using the Cox model on subsets of the data. *Statistics in Medicine* 19: 2595–2609. DOI:10.1002/1097-0258(20001015)19:19<2595::AID-SIM562>3.0.CO;2-M

4. Bonetti, M. and Gelber, R. D. (2004). Patterns of treatment effects in subsets of patients in clinical trials. *Biostatistics* 5:465–481. DOI:10.1093/biostatistics/kxh002

5. Bonetti, M., Gigliarano, C. and Muliere, P. (2009). The Gini Concentration Test for Survival Data. *Lifetime Data Analysis* 15:493–518. DOI:10.1007/s10985-009-9125-5

6. Bonetti, M., Zahrieh, D., Cole, B. F. and Gelber, R. D. (2009). A small sample study of the STEPP approach to assessing treatment-covariate interactions in survival data. *Statistics in Medicine* 28:1255–68. DOI:10.1002/sim.3524

7. Das Gupta, A. (2008). *Asymptotic theory of statistics and probability*, Springer: New York. DOI:10.1007/978-0-387-75971-5

8. Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *J. Amer. Statist. Assoc.* 94:496–509. DOI:10.2307/2670170

9. Gastwirth, J. L. (1971). A general definition of the Lorenz curve. *Econometrica* 39:1037–1039. DOI:10.2307/1909675

10. Gigliarano, C. and Bonetti, M. (2013). The Gini Test for Survival Data in Presence of Small and Unbalanced Groups, *Epidemiology, Biostatistics and Public Health*. DOI:10.2427/8762

11. Gini, C. (1912). Variabilità e Mutabilità. Contributo Allo Studio Delle Distribuzioni e Relazioni Statistiche. *Studi Economico-Giuridici dell'Università di Cagliari* III.

12. Good, P. (1994). *Permutation tests: a practical guide to resampling methods for testing hypothesis.* Springer Verlag: New York.

13. Hoeffding, W. (1948). A Class of Statistics With Asymptotically Normal Distribution. *The Annals of Mathematical Statistics* 19:293–325. DOI:10.1214/aoms/1177730196

14. Kakwani, N. (1977). Applications of Lorenz curves in economic analysis. *Econometrica* 45:719–727. DOI:10.2307/1911684

15. Lazar, A. A., Cole, B. F., Bonetti, M. and Gelber, R. D. (2010). Evaluation of Treatment–Effect Heterogeneity Using Biomarkers Measured on a Continuous Scale: Subpopulation Treatment-Effect Pattern Plot (STEPP). *Journal of Clinical Oncology (Statistics in Oncology series)* 28: 4539–4544. DOI:10.1200/JCO.2009.27.9182

16. Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *Quarterly Publications of the American Statistical Association* 9:209–219. DOI:10.2307/2276207

17. Miller, R. and Siegmund, D. (1982). Maximally selected chi square statistics. *Biometrics* 38:1011–1016. DOI:10.2307/2529881

18. Muliere, P. and Petrone, S. (1992). Generalized Lorenz curve and monotone dependence orderings. *Metron - International Journal of Statistics* L(3-4):19–38.

19. Pietra, G. (1915). Delle relazioni tra gli indici di variabili. *Atti del reale istituto veneto di scienze, lettere ed arti*.

20. Pogue-Geile, K.L., Kim, C., Jeong, J.-H. et al. (2013). Predicting degree of benefit from adjuvant trastuzumab in NSABP Trial B-31. *J Natl. Cancer Inst.* 105(23):1782–1788. DOI:10.1093/jnci/djt321

21. Royston, P. and Sauerbrei, W. (2008). *Multivariable Model Building: A Pragmatic Approach to Regression Analysis based on Fractional Polynomials for Modelling Continuous Variables.* Wiley Series in Probability and Statistics: New York. DOI:10.1002/9780470770771

22. Van der Vaart, A. (1998). *Asymptotic statistics.* Cambridge University press: Cambridge. DOI:10.1017/CBO9780511802256

23. Van der Vaart, A. and Wellner, J. (1996). *Weak convergence and empirical processes*, Springer-Verlag: New York. DOI:10.1007/978-1-4757-2545-2