Chemotherapeutic Impact on Pain and Global Health-Related Quality of Life in Hormone-Refractory Prostate Cancer: Dynamically Modified Outcomes Analysis (DYNAMO) of a Randomized Controlled Trial

Carol M. Moinpour[a]

Gary W. Donaldson[b]

Yoshio Nakamura[b]

[a]Southwest Oncology Group Statistical Center

Public Health Sciences Division

Fred Hutchinson Cancer Research Center

Seattle, WA

Email: cmoinpou@fhcrc.org

[b]Pain Research Center

Department of Anesthesiology

University of Utah

Salt Lake City, UT

Key Words: Causal effect, Individual differences, Mediation, Moderated intervention, Relational Outcome

Abbreviations: DYNAMO = Dynamic Modified Outcomes, M+P = mitoxantrone plus prednisone, D+E = docetaxel plus estramustine, GHRQL = global health-related quality of life, SWOG = Southwest Oncology Group, PS = performance status [0 (Fully active); 1 (Restricted in physically strenuous activity but ambulatory and able to do light work); 2 (Ambulatory and capable of self-care but unable to carry out any work activities); 3 (Capable of limited self-care, confined to bed or chair more than 50% of

waking hours); 4 (Completely disabled)], MPQ = McGill Pain Questionnaire, EORTC QLQ-C30 PR25 = European Organization for Research and Treatment of Cancer Quality of Life Questionnaire Core 30 and Prostate Cancer Module, DCE = Direct Causal Effect, ACE = Average Causal Effect, ICE=Individual Causal Effect

Text Word Count: 3893/Abstract Word Count: 199

Abstract

Purpose. This paper applies the Dynamically Modified Outcomes (DYNAMO) model to a clinical trial of two chemotherapeutic regimens on global health-related quality of life (GHRQL) in hormone-refractory prostate cancer. **Methods**. DYNAMO identifies the causal influences operating in a clinical trial and their mediation, moderation, and modulation by uncontrolled variables. Southwest Oncology Group Trial S9916 randomized assignment to mitoxantrone plus prednisone (M+P) versus docetaxel plus estramustine (D+E) treatments. In this application, we examine baseline-adjusted impacts of Worst Pain (McGill Pain Questionnaire) on GHRQL (EORTC Quality of Life Questionnaire-C30)at 10 weeks. **Results**. Average treatment levels of Pain did not differ, hence the average mediated effect of treatment on GHRQL was zero. Nonetheless, M+P reduced the impact (the relational outcome) of Pain on GHRQL by 54% relative to D+E. Individual variation in the relational outcome (modulation) was of the same magnitude as the average difference between arms. Performance status moderated the direct effects of treatment, with D+E more effective in good, but not poor, performance strata. **Conclusions**. The DYNAMO approach comprehensively accounted for treatment effects. Rather than a single average effect, there were three distinct treatment effects: one direct effect for each performance status level, and a direct effect on the relationship between pain and GHRQL.

The original publication is available at www.springerlink.com

Introduction

This paper applies the Dynamically Modified Outcomes (DYNAMO) causal analysis approach (Donaldson et al., this issue) to a chemotherapeutic trial targeting health-related quality of life (HRQL) in hormone-refractory prostate cancer. The theoretical framework for this approach explicates the direct causal effects of an intervention, and the mediators, moderators, and modulators that qualify them in the context of a clinical trial. This example extends earlier work, which discussed general latent trait analysis of variance models for clinical trials [1], the prominence of individual differences in treatment response [2], and the advantages that accrue when symptom outcomes are integrated within multivariate longitudinal analysis of general HRQL domains [3].

In this paper we apply the new DYNAMO approach to Southwest Oncology Group (SWOG) trial S9916, which compared mitoxantrone plus prednisone and docetaxel plus estramustine for men with hormone-refractory prostate cancer (HRPC) [4]. No conclusive evidence for significant differences in the primary HRQL endpoints (pain palliation and global HRQL [GHRQL]) was found [5]. Trial design, therapeutic results, and HRQL analyses are described below.

Moinpour et al. [3] illustrated use of multivariate growth curve methods [6-9] to examine treatment arm differences in HRQL outcomes measured over time in SWOG9916. This earlier analysis highlighted the presence of substantial individual differences in HRQL change trajectories, an important finding given the usual focus on average effects. In addition, the longitudinal analysis described how *relationships* between HRQL outcomes can differ as a function of treatment. The current paper focuses on detailed explication of causal effects at a single time point (adjusted for baseline levels) rather than on descriptive longitudinal summaries.

Methods

SWOG9916 [4, 5]

Trial design. Both regimens, docetaxel + estramustine (D+E) and mitoxantrone plus prednisone (M+P), were known to palliate pain in hormone refractory metastatic prostate cancer [10, 11]. In S9916, the D+E arm was hypothesized to have greater clinical efficacy as well as equivalent or better palliation of disease-related symptoms. Men with stage D1 or D2 prostate cancer were randomized to D+E or M+P; therapeutic results were reported by Petrylak et al. in 2004 [4] and additional detail regarding the therapeutic design can be found in this manuscript. Statistically significant differences favored D+E for median overall survival and time to progression as well as for the proportion of patients with at least a 50% decrease in prostate specific antigen (PSA). More grade 3/4 toxicities were observed in the D+E arm (neutropenic fevers, nausea and vomiting, and cardiovascular events).

For this illustrative application, we designate the M+P arm as "Standard or Control" (coded 0) and the D+E arm as "Experimental Treatment" (coded 1). At pre-randomization, patients were evaluated on the SWOG Performance Status (PS) Grading Scale. For this re-analysis of the S9916 HRQL data, we designate patients who were "fully or somewhat active" with a binary code of 0 (PS codes of 0 or 1), and those patients who were "relatively inactive" with a binary code of 1 (PS codes of 2 or 3); this direction maintains consistency with the SWOG scoring and the categorization matches the PS stratification variable for the therapeutic trial [4]. The S9916 trial enrolled 674 patients eligible for analysis in the original reports; 629 patients had HRQL data.

HRQL assessment. HRQL was assessed with the McGill Pain Questionnaire (MPQ) [12] and the European Organization for Research and Treatment of Cancer (EORTC) Quality of Life Questionnaire – Core 30 (QLQ-C30) [13, 14] with its prostate module, the PR-25 [15]. The Present Pain Intensity (PPI) item from the MPQ and the GHRQL item from the QLQ-C30 were the two pre-designated primary HRQL outcomes for the trial. To facilitate interpretation, both outcomes were scaled from 0 (best possible score) to 100 (worst possible score).

For this application, we analyze the pain and GHRQL outcomes, and the relationship between them, at Week 10. Six months (Cycle 8) was the pre-specified primary time point for the primary analyses. However, we observed substantial missing data at the later time points (6 months: 56% for the EORTC QLQ-C30, 58% for the McGill Pain Questionnaire; 12 months: 49% for the EORTC QLQ-C30, 48% for the McGill Pain Questionnaire). Therefore, this application examines pain and GHRQL at week 10 (Cycle 4) (adjusted for baseline levels),when both questionnaires were completed and submission rates were acceptable (82% EORTC QLQ-C30; 89% McGill Pain Questionnaire) [5]. The D+E arm had better submission rates than the M+P arm, consistent with the D+E arm's statistically longer survival. Patients who submitted fewer forms over time had worse scores at baseline and worsening HRQL at time of drop-out. A series of pattern mixture models suggested no consistent statistically significant differences in global HRQL or pain by treatment arm [5].

Statistical methods

We illustrate application of DYNAMO principles using pain and GHRQL data from the clinical trial described above. In this population, pain is a prominent symptom and we start from the premise that pain is a critical contributor to GHRQL; if severe enough,

pain is almost certain to depress one's GHRQL. The regression of GHRQL on pain is thus an index of the centrality or impact of pain on the more general HRQL rating. A causal interpretation is also defensible: increases in pain should cause poorer GHRQL because pain is a constituent of GHRQL. The converse, however, is not necessarily true; many aspects of GHRQL may worsen without affecting pain. On either interpretation, it is reasonable to inquire whether the randomized intervention led to a change in the degree of association between pain and GHRQL, as indexed by the regression of GHRQL on pain.

Consistent with our emphasis on model conceptualization, the key results appear in Figure 1, in a manner designed to resemble the generic model of Figure 5 in Donaldson et al. (this issue). Omitted from Figure 1 are coefficients corresponding to standard baseline adjustments that regress later versions of the measured variables on their baseline counterparts[1]. All coefficients in Figure 1 attained statistical significance ($p < .05$) using robust standard errors [16]. Figure 1 contains the key findings illustrating the approach, but complete syntax and technical estimation results using the MPlus multilevel structural equation modeling program [16] are available upon request.

Results

We first summarize the major results of Figure 1, then provide additional interpretation guided by the four steps recommended in the companion DNYAMO paper (Donaldson et al., this issue).

[1] These standard adjustments have no effect on the interpretation of the key features of the model, but merely condition responses on baseline values.

The direct causal effects of treatment

The Direct Causal Effect *(DCE)* of the intervention on Pain (Z) was not significantly different from zero (p=.964, likelihood ratio test), and therefore coefficient *a* was set to zero in the model and in Figure 1. This, by definition, defined the mediating path X->Z->Y as zero. The *DCE* on the relational outcome (corresponding to *b* in Figure 5 from the companion paper) was important, however. The coefficient of .20 combines additively with the intercept coefficient of .17 to signify that the effect of Pain on GHRQL is roughly twice as strong in the D+E group as in the M+P group. For D+E, a unit worsening of 1 Pain point led, on average, to a .37 (.17 + .20) worsening of GHRQL, compared with a .17 value for M+P. To put it in a more realistic context, a 50-point increase on a 0 to 100 pain scale would lead, on average, to an 18.5-point worsening in GHRQL in D+E, but only an 8.5-point worsening in M+P. This is the causal reading of the coefficients. An alternative interpretation is that the intervention led to a less central role for pain in evaluating GHRQL in the M+P group. The *DCE* of treatment on GHRQL (Y, the designated endpoint), was *moderated* by pre-randomization Performance Status. The D+E treatment intervention directly caused an average improvement of 5.89 GHRQL points in the good performance status stratum (PS=0), but only a negligible improvement of .25 GHRQL points in the poor performance status stratum (PS=1).

Treatment effect mediation

In this study there was no direct effect of therapy on the mediator Pain, and hence no Average Mediating Effect on GHRQL via Pain. There are nonetheless nonzero and differing Individual Mediating Effects, because patients in both arms differ in their mediating pain scores. Patients with pain scores Z_i that lie far from the average (who

have extreme values of U_Z) would experience greater mediation than patients with pain scores near the average, since $IME = (\beta_1 - \beta_0)U_{Zi}$ even when the means of Z are equal (see Table 1 in the companion paper).

Average causal effect

Because there was no Average Mediated Effect (since the *DCE* of treatment on Pain was zero), the Average Causal Effect (*ACE*) within each performance stratum equals the *DCE* within that stratum. Although it is possible to calculate an overall *ACE* across performance strata, the number does not represent the expected causal effect for patients in either stratum.

What made GHRQL happen in one randomized controlled trial? A second look

The strategies and guidelines suggested in the companion paper provide an approach to more comprehensive analysis of the full spectrum of causes and responses operating during a clinical trial.

1. Examine and report the diversity of individual responses, treated separately from "error."

Significant individual differences, systematic and distinct from random error, modulated the relational outcome. The random regression of GHRQL on Pain had a variance estimate of .024. This variance appears small in absolute terms, but converting to the standard deviation scale helps place it in proper context. The corresponding standard deviation of .15 is nearly as large as the .20 average difference in slopes between treatment arms, hence there is considerable individual variability in the strength of association between Pain and GHRQL in this population. The regression coefficients for two randomly selected patients from the same treatment

arm would be expected to differ by .21 (.15 x square root of 2), which exceeds the expected average difference between arms. Figure 2 represents the range of individual relational outcomes relative to treatment arm average differences. Each individual's relational outcome is a personal attribute indicating a dispositional sensitivity to pain, conceptually distinct, and estimated separately, from "error." More complex longitudinal designs would permit full characterization of modulators generating individual differences in direct causal effects on the Pain and GHRQL outcomes as well as on the relational outcome.

2. Evaluate whether treatment has affected the relationships as well as the levels of outcome variables.

In this trial there were two important *DCE*s that operated separately. The intervention (D+E *instead of* M+P) tended to improve the level of GHRQL in one performance stratum but also to increase the sensitivity of GHRQL to pain in both strata. Both causal aspects are crucial to understanding whether patients receive benefit from an intervention. Increased sensitivity to Pain leads to increased volatility in GHRQL, a negative result that may be offset by direct improvement in GHRQL. In Step 4, we illustrate how questions of benefit depend on the details of patient mediation, moderation, and modulation.

3. Include and evaluate moderators of the treatment intervention.

The *DCE* of treatment depended on the level of the pre-randomization Performance Status measure. The benefit of understanding moderated interventions is great, because it allows prediction of who will benefit from therapy. In our substantive

example, patients with poor performance status tend to receive little benefit from the D+E treatment beyond that offered by M+P.

4. *Consider individual causal inference. Evaluate the extent to which the treatment was responsible for the observed changes in individual patients.*

Although a full presentation of these procedures is beyond this paper's scope, we would like to introduce the expanded inference possible under a comprehensive causal model like DYNAMO. The key idea involves estimating the Individual Causal Effects (ICEs) for each patient on the trial, and combining this information with observed change to deduce what role the therapy must have had in effecting change.

Because causal models are modular, it is meaningful to ask what would happen to Y if we could intervene to change the treatment a patient received while holding constant that patient's other causes of Y, Z, and β. This is the ICE. Under the modularity assumption, the remaining (non-treatment) causes of Y can be directly calculated, once the ICE is estimated from the model (since Y is known, as suggested in Figure 3). Holding background variables constant, one can compare the estimated ICE with the observed change and deduce whether change in Y happened because of, despite, or regardless of the intervention [17]. Table 1 presents this process for three representative patients from the SWOG trial. The first data column presents the estimated ICE from the model (reflected in the "good" direction, for ease of interpretation), while the second data column contains the adjusted gain in GHRQL from baseline (also reflected in the "good" direction). The third column then computes the value for Uy, the sum of other causes, by subtraction. Conditional on the modular causal assumptions of Figure 3, the final column, labeled Y*, calculates the (counterfactual) value expected for GHRQL if we could intervene to set the patient's

ICE to zero (receiving no benefit from treatment). Y* is the expectation for what would have happened to Y had the ICE been zero instead of what it really was.

Figure 4 portrays these results on a coordinate scheme cross-classifying observed change and ICEs. All three patients were observed to improve (positive change) on GHRQL, and hence are plotted above the abscissa. We now consider why the patients improved. Patient A experienced an ICE of +10 from therapy, but had this effect been zero, A would have worsened by 5 points instead, since other causes were negative. Therapy was thus necessary for A's improvement; legal and common language reasoning describe A as improving *because of* therapy, since A would not have improved otherwise. Now consider Patient B, who had an ICE of +5 and was observed to improve by 17. Had B's ICE been zero, B would still have improved (though by +12 instead of +17); we say B improved *regardless of* therapy, since other causes would still allow B to experience a positive change. Finally, consider Patient C, who improved *despite* having a negative (harmful) causal effect of therapy. Had C not received therapy, C would have improved by 19 points instead of only 7.

As the geometry of Figure 4 makes clear, any patient falling into the same sector as A, B, or C shares the respective causal attribution for that sector. Patients observed to worsen instead of improve fall into sectors below the abscissa having corresponding, though reversed, interpretations.

Figure 5 represents a comprehensive cross-classification for all patients on the trial in the manner of Figure 4 for both M+P (left panel) and D+E (right panel). Each point in the causal attribution plots represents an observed HRQL change and an inferred (estimated from the model) ICE. The side-by-side comparison is telling. Only one

patient in the M+P group improved because of therapy[2], while 15 patients in the D+E arm improved because of therapy. Three D+E patients worsened because of therapy, but over 25 M+P patients worsened because of therapy. The bivariate distributions of Figure 5 present compelling as well as innovative evidence supporting the general efficacy of the D+E intervention. At the same time, the comprehensive cross-classifications identify the smaller number of patients who benefited from the M+P therapy, as well as some of their characteristics. (In this analysis, those benefiting from M+P were primarily low performance status patients having extreme pain who would receive large mediating effects from M+P's advantage in the relational outcome but who would not receive the benefit of D+E's direct causal effect on GHRQL, which operated only within the high performance stratum.) Almost as many patients improved on M+P as on D+E. However, reasons for improvement differed. D+E patients improved in large part because of treatment whereas M+P patients improved despite or regardless of treatment. Therefore, other causes are at least as important as treatment in explaining why GHRQL changes happen.

In more definitive analyses, one could incorporate standard errors of estimation and pragmatic effect sizes to allow regions of uncertainty as well as practical importance. The simplified example presented here collapses all other causes of Y into a single category. This is mathematically accurate, but unsatisfactory in that the set of all other causes subsumes random measurement error. In longitudinal analyses, it would be possible, and highly desirable, to distinguish systematic causes of individual differences from random measurement error.

[2] In this and all statements drawn from Figures 3 and 4, the conclusions about therapeutic impact pertain to receiving one therapy instead of the other. The attributed causal impacts are relative, not absolute.

Discussion

A "standard" analysis of SWOG9916 would indicate that, adjusted for baseline GHRQL, the D+E Cycle 4 GHRQL mean was 3.35 points lower than the M+P mean, a difference that does not quite reach statistical significance (p=.078), in line with the nonsignificant GHRQL results reported in the primary publication [5]. According to traditional guidelines, the "treatment did not work" better than the comparator for GHRQL. Yet this summary conclusion is both inaccurate and incomplete. In fact the study provides an interesting combination of findings, leading to a more nuanced interpretation of treatment effects. Consider first the moderated intervention. Whether the treatment "works" depends on which kind of patient you are. For good performance status patients, D+E had a beneficial *Direct Causal Effect* (*DCE*) on GHRQL, improving it on average relative to the M+P group. Since the average mediated effect was zero, the DCE is equivalent to the *Average Causal Effect* (*ACE*): the D+E treatment appeared to work, on the average, for patients with good performance status. For the poor performance stratum, the *DCE,* and hence the *ACE,* were negligible.

Several familiar statistical approaches can work well to evaluate moderators, and one need not rely on the full DYNAMO framework to address the critical question of differential treatment benefit. Conventional moderated regression (i.e., incorporation of an interaction term or the conduct of separate subgroup analyses) [18] or a two-group structural equation approach [1, 9, 16] would yield similar conclusions for the moderating effect of performance status. The difficulties with evaluating moderators are less technical than substantive and psychological: one must know what to measure and maintain a certain equanimity in the face of complexity. With moderated interventions, randomized controlled trials cannot yield a single answer to how well a

treatment works. This reality resides in the diversity of individual attributes, not in the statistics. An important prerequisite is to understand the clinical mechanisms, which in turn suggest which moderators to measure. These may then be included in statistical models such as the DYNAMO approach proposed here to provide important insights into improved clinical management.

In the SWOG trial, understanding benefit depends on reconciling causal effects that may not work in tandem. The D+E arm provided an average direct benefit to patients with good performance status, but not with poor performance status (moderated intervention). Yet D+E also led to greater sensitivity of GHRQL to Pain than did M+P (relational outcome). These two interactions, the relational outcome and the moderated intervention, are to a considerable extent in opposition, with conclusions depending on particular combinations of values for Pain and GHRQL. Although common statistical practice sanctions transforming variables to eliminate interactions, this may be counterproductive when these interactions are themselves of primary clinical interest. Taking interactions seriously requires that the metrics of clinically interacting variables be treated somewhat consistently across studies.

Standard approaches to the mediation problem [18-20] assume that a nonzero association between X and Z is a necessary condition for Z to mediate the effect of X on Y. Though this assumption seems reasonable, it may inappropriately rule out many interesting patterns of mediational results [21-24]. In S9916, the *DCE* of X on Z was zero, and hence the Average Mediating Effect of X on Y via Z was zero. Traditional methods, such as those expressed in Baron and Kenny's [18] regression rules, take this finding as evidence that Z is "not a mediator." The Individual Mediating Effects (modulation), however, are not zero. Consider a Control patient with a Pain score one standard deviation, about 23 points, above the Control mean. The

Individual Mediating Effect for that patient is an expected improvement of $23(\beta_1 - \beta_0)$ = 23(.20) = 4.6 HRQL points by virtue of the treatment change in the relational outcome. A full range of Individual Mediating (and hence causal) Effects may arise even when the population average relationship is zero. By any method, the relational outcomes in the SWOG trial differ between treatment arms. Ordinary least squares regression with observed variables including an interaction term yields results that agree with the DYNAMO average relational outcomes, estimating the treatment arm regression coefficients as .37 and .17 for D+E and M+P, respectively (see also [3]).

The relational outcome captures a unique aspect of the treatment's effect. Independent of performance status, the intervention led to a reduced average impact of – a less central role for – pain in the M+P group. Important individual differences modulated the relational outcome, however. Figure 2 presents boxplots showing the relational outcome distributions separately by treatment arm. The boxplots indicate that the modulation by individual differences (the spread or length of the boxplots) is at least as important as the direct causal effect of the intervention (the distance between the median values). The dependence of GHRQL on pain varied widely across patients within each treatment arm, using model-based estimates that are theoretically purged of measurement error. Nonetheless the direct causal effect of the intervention was substantial: roughly speaking, the 25th percentile of the D+E arm corresponds to the 75th percentile of the M+P arm. The treatment effect on the relational outcome may reflect the anti-inflammatory role of the prednisone component of the M+P combination [5]. The presence of prednisone may reduce the functional consequences of inflammatory pain, even though the direct causal effects of therapy on pain were equivalent. That is, for the same degree of pain, we speculate

that the extent of mobility and activity possible could be greater in the M+P arm, over against the direct benefit of D+E (in the good PS stratum) on GHRQL.

If this interpretation is correct, it presents an interesting example of the clinical tradeoffs that can be considered using the DYNAMO approach. In patients for whom severe pain is the primary risk to quality of life, the M+P arm might be the better choice when weighing the risks and benefits of treatment (recalling that there was a therapeutic benefit for D+E) regarding HRQL. In high performance patients with less severe pain, but with broader-based GHRQL limitations, the D+E therapy would generally be superior.

A key objective of this analysis has been to express the conditional dependence of the Z->Y relationship on X. It is of course completely equivalent statistically to express the conditional dependence of X->Y on Z. Or, in many situations, one may only wish to consider the relational outcome with respect to the symmetrical association of Y and Z, and consider this covariance as the outcome. Rather than the directed graph motivating the present analyses, one could as well consider graphical models, such as chain graphs, interaction graphs, and conditional Gaussian graphs, that are at least partly undirected [25]. In fact, we have conducted all these approaches using the general graphical modeling program MIM [26], and these have led to similar statistical conclusions.

Still, we believe there are advantages for the directed approach when justifiable. It provides a natural interpretation for the real and hypothetical experiments about what would happen if variables were manipulated in a certain sequence. A secondary consideration is that a fully directed model permits random effect components that correspond naturally to individual differences.

Acknowledgements The authors would like to thank the patients who contributed HRQL data to S9916 and the Clinical Research Associates at Southwest Oncology Group institutions who monitored the submission of the HRQL forms. We recognize the contributions of Dr. Donna L. Berry, the HRQL Study Coordinator for S9916 and Dr. Daniel P. Petrylak, the therapeutic trial study coordinator.

Funding Sources This investigation was supported in part by the following PHS Cooperative Agreement grant numbers awarded by the National Cancer Institute, DHHS: CA38926, CA32102, CA37135, CA25224, CA46441, CA37981, CA45808,CA27057, CA12644, CA68183, CA22433, CA35261, CA58861, CA20319, CA46113, CA58882, CA76447, CA04919, CA16385, CA35090, CA03096, CA67663, CA45450, CA35431, CA45807, CA58416, CA14028, CA45377, CA63845, CA42777, CA46136, CA11083, CA35119, CA58658, CA46282, CA76129, CA46368, CA35176, CA86780, CA46462, CA35192, CA35178, CA67575, CA63844, CA12213, CA74647, CA35128, CA35996, CA58686, CA13612, CA45461, CA58723, CA63848, CA35281, CA63850, CA76132, CA74811, and supported in part by Aventis.

Table 1 Causal attributions for individual patients

Patient	ICE	Y (GHRQL)	Uy= Y - ICE	Y* $= Y \mid ICE \to 0$ $= U_Y$
A	10	5	-5	-5
B	5	17	12	12
C	-12	7	19	19

References

1. Donaldson, G. (2003). General linear contrasts on latent variable means: Structural equation hypothesis tests for multivariate clinical trials. *Statistics in Medicine*, 22, 2893-2917
2. Donaldson, G. W. & Moinpour, C. M. (2002). Individual differences in quality-of-life treatment response. *Medical Care*, 40, (6 Suppl), III39-53
3. Moinpour, C. M., Donaldson, G. W. & Redman, M. W. (2007). Do general dimensions of quality of life add clinical value to symptom data? Journal of the National Cancer Institute *Monographs*, 37, 31-38
4. Petrylak, D. P., Tangen, C. M., Hussein, M. H., Lara, P. N., Jones, J. A., Ellen, T. M., et al. (2004). Docetaxel and estramustine compared with mitoxantrone and prednisone for advanced refractory prostate cancer. *New England Journal of Medicine*, 351, (15), 1513-1520
5. Berry, D. L., Moinpour, C. M., Jiang, C. S., Ankerst, D. P., Petrylak, D. P., Vinson, L. et al. (2006). Quality of life and pain in advanced stage prostate cancer: Results of a southwest oncology group randomized trial comparing docetaxel and estramustine to mitoxantrone and prednisone. *Journal of Clinical Oncology*, 24, (18), 2828-2835
6. Laird, N. M. & Ware, J. W. (1982). Random-effects models for longitudinal data. *Biometrika*, 38, 963-974
7. Curran, P. & Hussong, A. (2003). The use of latent trajectory models in psychopathology research. *Journal of Abnormal Psychology*, 112, (4), 526-544
8. Singer, J. & Willett, J. (2003), *Applied longitudinal data analysis: Modeling changes and event occurrence*. (New York: Oxford University Press)
9. Kline, R. (2005), *Principles and practice of structural equation modeling*. (New York: Guilford Press)
10. Sinibaldi, V. J., Carducci, M. A., Moore-Cooper, S., Laufer, M., Zahurak, M. & Eisenberger, M. A. (2002). Phase ii evaluation of docetaxel plus one-day oral estramustine phosphate in the treatment of patients with androgen independent prostate carcinoma. *Cancer*, 94, (5), 1457-1465
11. Tannock, I. F., Osoba, D., Stockler, M. R., Ernst, D. S., Neville, A. J., Moore, M. J., et al. (1996). Chemotherapy with mitoxantrone plus prednisone or prednisone alone for symptomatic hormone-resistant prostate cancer: A canadian randomized trial with palliative end points. *Journal of Clinical Oncology*, 14, (6), 1756-1764
12. Melzack, R. (1987). The short-form mcgill pain questionnaire. *Pain*, 30, 191-197
13. Aaronson, N. K., Ahmedzai, S., Bergman, B., Bullinger, M., Cull, A., Duez, N. J., et al. for the European Organization for Research and Treatment of Cancer Study Group on Quality of Life. (1993). The European Organization for Research and Treatment of Cancer QLQ-C30: A quality of life instrument for use in international clinical trials in oncology. *Journal of the National Cancer Institute*, 85, (5), 365-373
14. Fayers, P. M., Aaronson, N., Bjordal, K., Groenvold, M., Curran, D., Bottomley, A. on behalf of EORTC Quality of Life Group. (2001), *EORTC QLQ-C30 scoring manual*. (Brussels: EORTC)

15. Borghede, G. & Sullivan, M. (1996). Measurement of quality of life in localized prostatic cancer patients treated with radiotherapy. Development of a prostate cancer-specific module supplementing the EORTC QLQ-C30. *Quality of Life Research*, 5, 212-222
16. Muthén, L. & Muthén, B. (1998-2005), *Mplus user's guide*. (Los Angeles: Muthén & Muthén)
17. Pearl, J. (2000), *Causality: Models, reasoning, and inference*. (Cambridge: Cambridge University Press)
18. Baron, R. & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173-1182
19. Judd, C. & Kenny, D. A. (1981). Process analysis: Estimating direction in evaluation research. *Evaluation Research*, 9, 602-618
20. MacKinnon, D. P., Fairchild, A. J. & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology*, 58, 593-614
21. Albert, J. M. (2008). Mediation analysis via potential outcomes models. *Statistics in Medicine*, 27, (8), 1282-304
22. Kraemer, H. C., Wilson, G. T., Fairburn, C. G. & Agras, W. S. (2002). Mediators and moderators of treatment effects in randomized clinical trials. *Archives of General Psychiatry*, 59, (10), 877-83
23. Kraemer, H. C., Lowe, K. K. & Kupfer, D. J. (2005), *To your health: What research tells us about risk*. (New York City: Oxford University Press)
24. Kraemer, H. C., Stice, E., Kazdin, A., Offord, D. & Kupfer, D. (2001). How do risk factors work together? Mediators, moderators, and independent, overlapping, and proxy risk factors. *American Journal of Psychiatry*, 158, (6), 848-56
25. Edwards, D. (1995), *Introduction to graphical modeling*. (New York: Springer-Verlag)
26. Edwards, D. Mim 3.2, (2004). Free Software Foundation: Boston

Figure Legends

Figure 1. Key results of the relational outcome model applied to the SWOG 9916 example. The Direct Cause Effect (DCE) of Treatment (TX, 0=M+P, 1=D+E) on Pain was zero, so the Average Mediated Effect was also zero. The DCE of Treatment on the relational outcome β showed that the average dependence of GHRQL on Pain was roughly twice as strong in the D+E arm (.17+.20) as in the M+P arm (.17). The β_i values varied across patients, with within-group variance equal to .024 (SD=.15). The DCE of treatment depended on the level of the pre-randomization Performance Status measure. The randomized treatment intervention caused an average improvement of 5.89 GHRQL points in the good performance status stratum, but only an improvement of .25 GHRQL points in the poor performance status stratum. Since the Average Mediated Effect was zero, the Average Causal Effects equaled the Direct Causal Effects (5.89 GHRQL improvement for PS=0, .25 GHRQL improvement for PS=1).

Figure 2. Boxplots of the distributions of the relational outcome slope coefficient for individuals in the M+P and D+E treatment arms. The relational outcome is an individual's expected change in GHRQL given a one-unit increase in Pain, E(GHRQL|Patient=*i*,Pain=*p*+1) – E(GHRQL|Patient=*i*,Pain=*p*), if other variables could be held at fixed values. This is a systematic attribute of a person, distinct from measurement error. The boxes show the 75th and 25th percentiles, with the central lines denoting the medians. The whiskers include the non-outlying values, while the isolated dots represent outliers. The interquartile range (75th percentile – 25th percentile) was approximately equal to the average treatment arm difference, so both population and individual relational effects were important.

Figure 3. Partitioned causes of observed GHRQL change in individual patients. Causal models are modular, representing the expected changes from controlled manipulations holding constant other variables. The Individual Causal Effect (ICE) summarizes relationships in Figure 1 to show how GHRQL (Y) would change for a particular patient because of treatment (if he or she were to receive the other randomized intervention, holding constant all other causes). Under these assumptions of modularity and unchanged other causes, the total cause of Y must equal the sum of a patient's ICE and all other causes U_Y. Therefore one can calculate U_Y by simple subtraction once the ICE has been estimated.

Figure 4. Model-based estimation and counterfactual inference for individual patients. The abscissa in this scatterplot represents the model-based Individual Causal Effect (ICE) estimate, while the ordinate is the observed Y (GHRQL) adjusted for baseline (a residual "change" score). Patients A and B had both positive (beneficial) causal effects from therapy and positive (beneficial) observed change in Y. Patient A had an estimated positive ICE of 10, but an observed improvement of only 5, hence the other causes of Y amounted to -5. Holding constant the other causes, the value expected for Y if the ICE were modified to zero would be -5. Hence A's improvement would not have happened without therapy: A improved *because* of therapy. Patient B's observed improvement was 17, of which only 5 resulted from therapy. If the ICE had been zero for B, we still would have observed a positive change of 12 for B; hence B improved *regardless* of therapy. Patient C had a negative (harmful) ICE of -12, but a positive observed change of 7. Thus C improved *despite* therapy; C would have improved by even more without the harmful effect of therapy. Any patient falling in the same sector as patients A, B, or C

shares the causal attribution appropriate for that sector. The causal attributions for patients who worsen (below the abscissa) follow similarly.

Figure 5. Causal attribution plots for the M+P (left panel) and D+E (right panel) treatment arms. The plots contain the specific model-based and observed data points for all patients. The sectors of the plots correspond to the definitions in Figure 4. For example, the upper left quadrants of both panels contain data for patients who improved despite therapy. The lines of identity are not represented diagonally in these plots because the ranges are greater for observed data (the ordinates) than for the Individual Causal Effects (ICEs) (the abscissas); this is a graphical display choice made to enhance point visibility and has no effect on the interpretations, which are identical to those of Figure 4.

Figure 1

Click here to download line figure: Application Figure 1.ppt

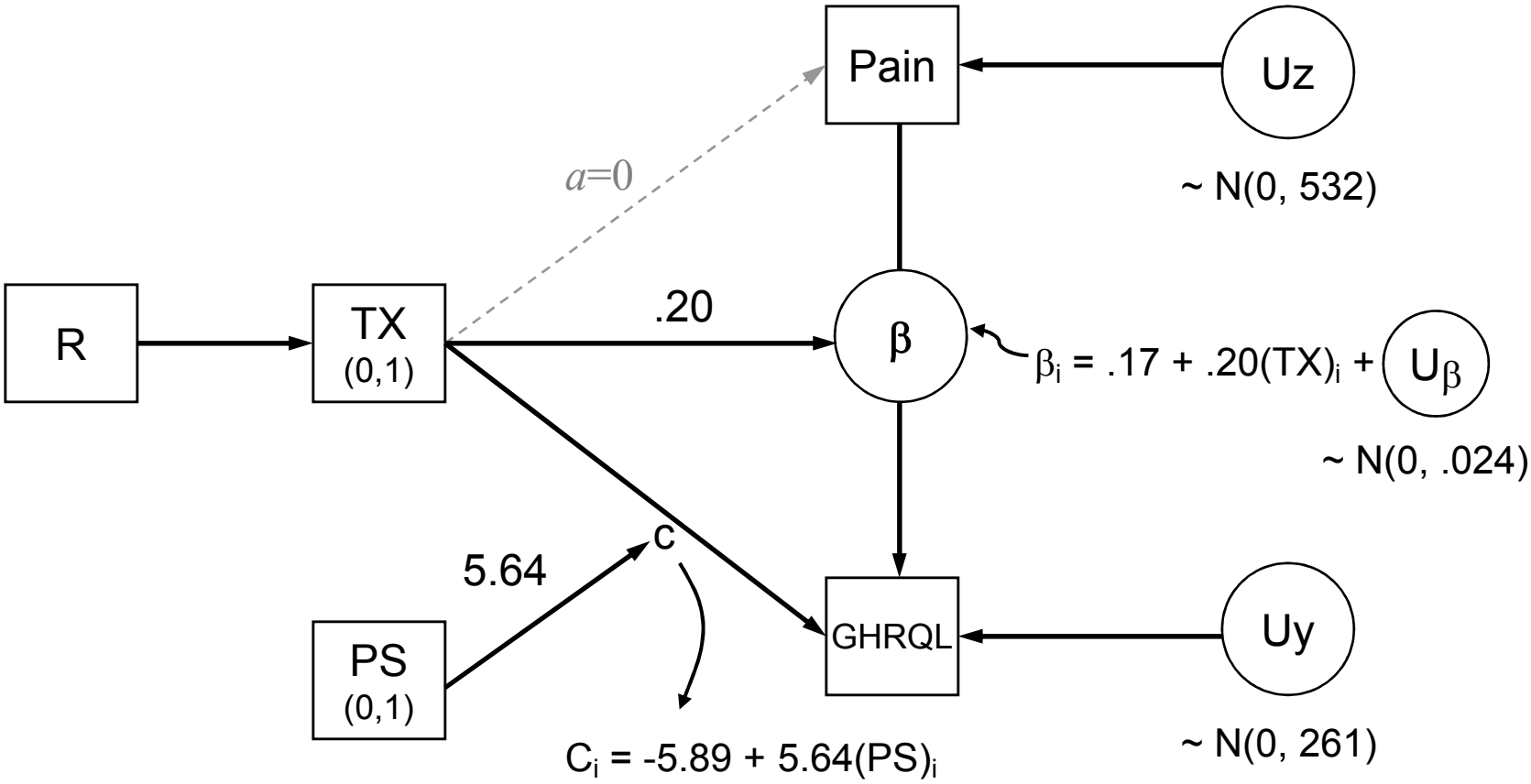


Figure 2

Click here to download line figure: Application Figure 2.ppt

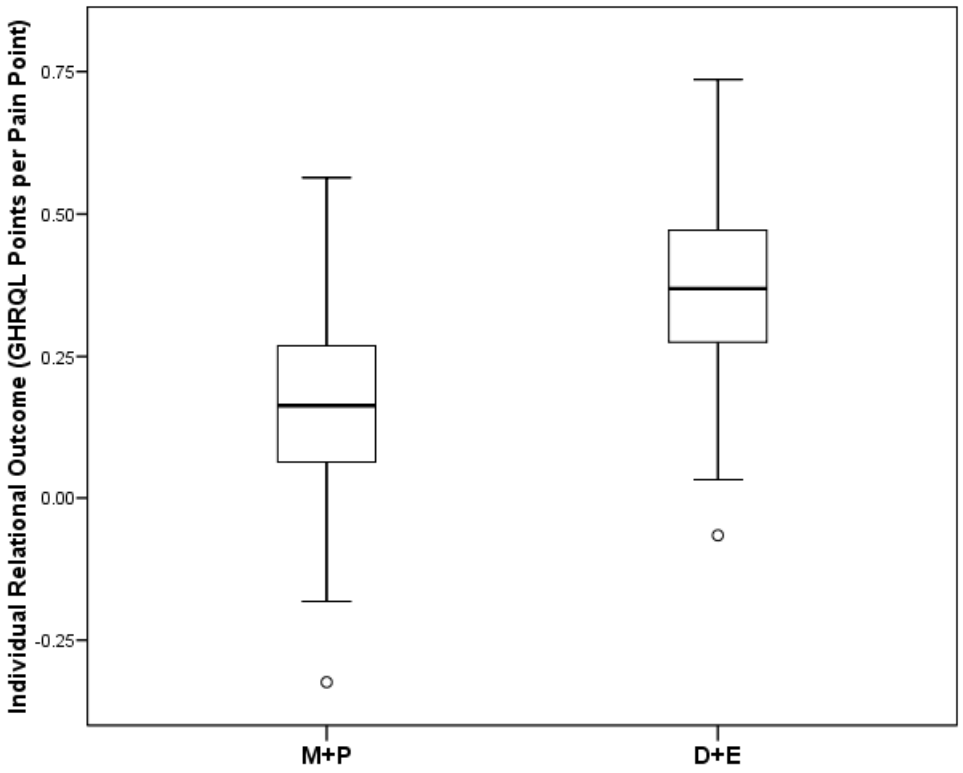
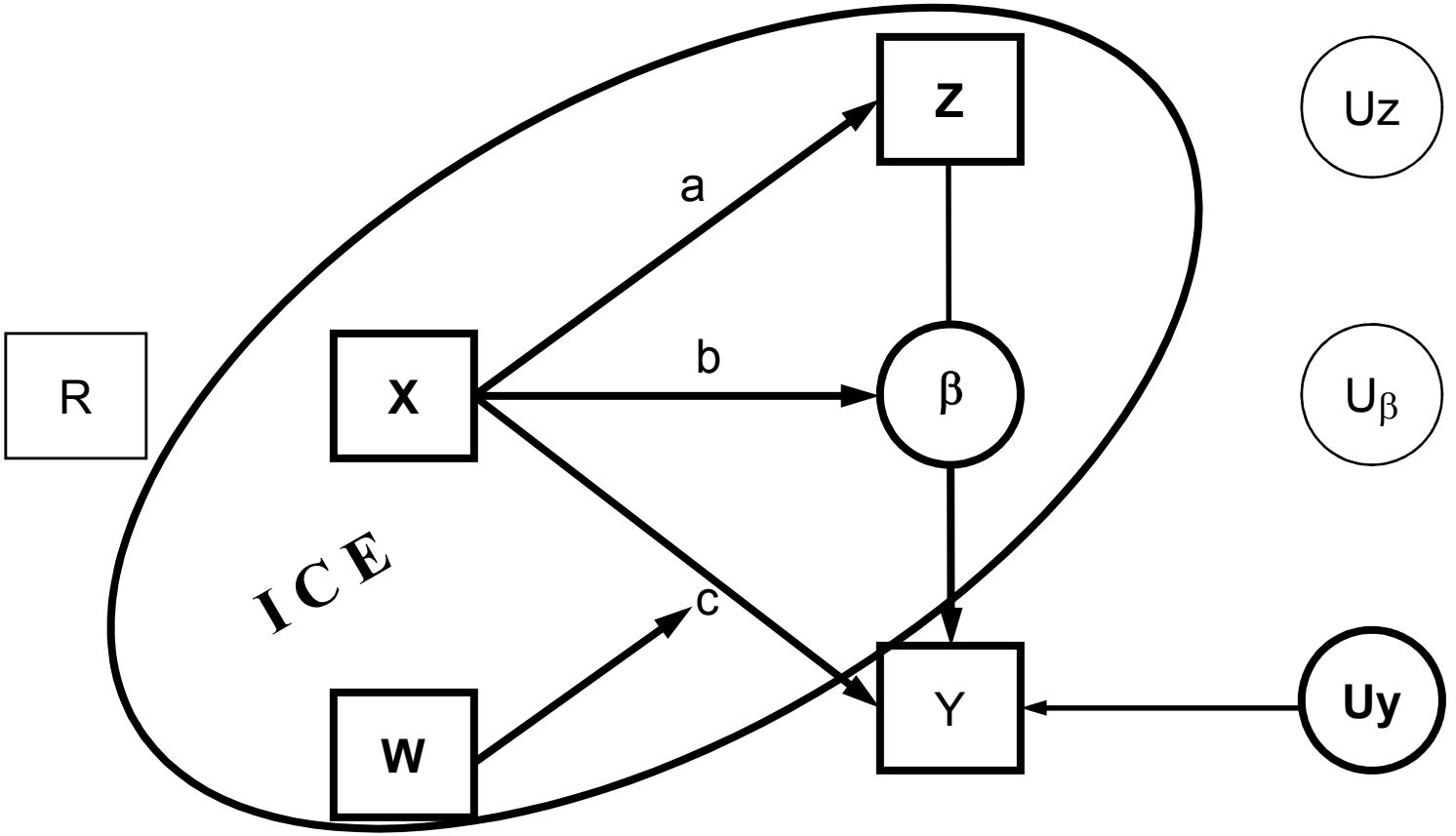


Figure 3

Click here to download line figure: Application Figure 3.ppt



$$Y = ICE + Uy$$

Figure 4

Click here to download line figure: Application Figure 4.ppt

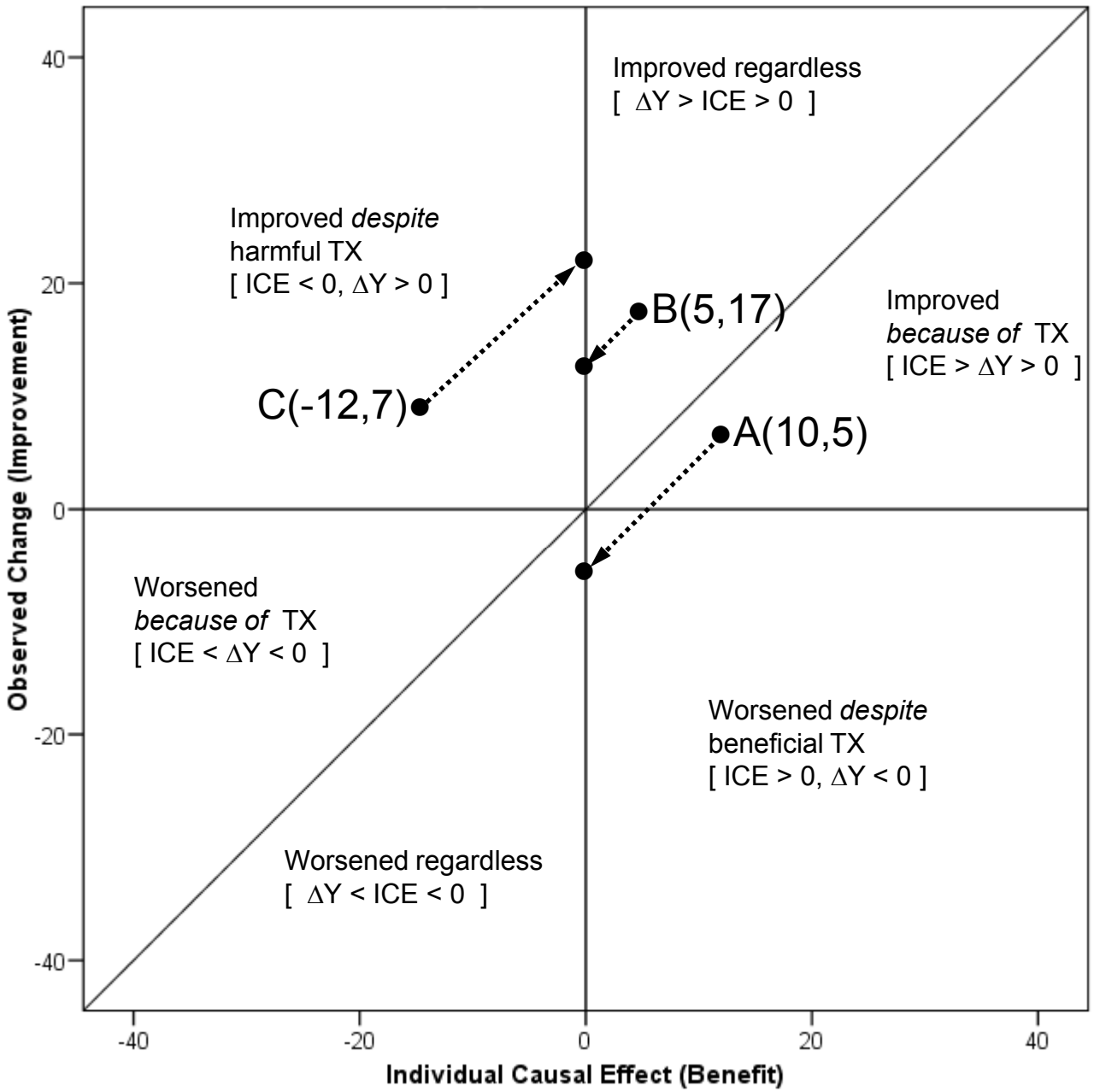


Figure 5

Click here to download line figure: Application Figure 5.ppt