Things to discuss/show in a paper using propensity score methods (and examples)

Elizabeth Stuart
Johns Hopkins Bloomberg School of Public Health
January 7, 2013

(Note: The term "matching" is used below to mean any sort of propensity score adjustment, which could be matching, weighting, or subclassification)

In Methods section:

1. Sample size in treatment and comparison groups: Initially, and after matching
2. Propensity score model estimation procedure used (e.g., random forests/gbm/logistic regression)
3. Matching method used (e.g., 1:1 matching without replacement, full matching, IPTW weighting, etc.)
4. Covariates included in the matching process
5. Whether any covariates were treated in a "special" way, e.g., by an exact or Mahalanobis match
6. The estimand of interest (e.g., ATT or ATE)
7. Diagnostics:
    a. Some measure of the success of the procedure, in terms of covariate balance before and after matching. The length of this will depend on the journal and space constraints.
    b. Most common is a table showing means and standardized biases before and after matching.
    c. Figures are not usually shown, although sometimes there are nice figures showing the standardized biases before and after matching (see below for an example of a “Love” plot; named because it was invented by Thomas Love).
    d. [Note: Sometimes people put this information in the Results section.]
8. Description of how effects estimated following the matching (e.g., regression adjustment on matched samples, with X, Y, and Z as predictors).
9. Optional: Some discussion of why a particular matching method was selected (e.g., to retain the full sample, because it led to the best balance, because we wanted to estimate the ATE)
10. Optional: Mention of other matching methods used as sensitivity analyses, or discussion of other approaches considered but not used for some reason. (This is not very common).
11. Optional: Description of sensitivity analysis done to assess sensitivity to unobserved confounding

In Results or Conclusion section:

1. State the effects clearly, accounting for the estimand (ATT vs. ATE); i.e., if estimating the ATT don't have statements that imply the effect applies to everyone.
2. Provide the effect estimates for each outcome and matching method selected.
3. The discussion and tables should make it clear that effects estimated in the

matched sample (vs. the original data)
4. Discuss limitation of method to only adjust for observed confounders; there may still be bias due to unobserved confounding
5. Optional: Discuss the analysis of sensitivity to an unobserved confounder

Examples:

1. Werner, R., Polsky, D., and Stuart, E.A. (2010). Public Reporting Drove Quality Gains At Nursing Homes. *Health Affairs* 29(9): 1706-1713.

Breaking the changes down proceeded in three steps. The first step was to estimate nursing home–specific quality improvements by holding case-mix and market share constant. To do this, we used one-to-one propensity-score matching, in which each postacute care patient in the pre–NursingHomeCompare period was matched to patient with similar characteristics from the same nursing home in the post–Nursing Home Compare period. We conducted this match "with replacement," meaning that a single individual in the post period could be selected as a match for up ten patients in the pre period. This process kept case-mix and market share for each nursing home constant at pre–Nursing Home Compare levels. We then recalculated the postacute care quality measures on the propensity score– matched subsample, estimating nursing home–specific quality improvements.

2. **Werner, R.M., Konetzka, R.T., Stuart, E.A., Norton, E.C., Polsky, D., and Park, J. (2009). The impact of public reporting on quality of postacute care. *Health Services Research* 44(4): 1169-1187. PMCID PMC2739023.**

Because public reporting of patient outcomes may cause providers to select (or cherry-pick) their patients based on their illness severity (Dranove et al. 2003; Werner and Asch 2005), patient characteristics may systematically vary with the launch of public reporting. To account for this possibility, we use propensity score matching to ensure the similarity of patients being compared (Rosenbaum and Rubin 1983; Heckman,Urzua, and Vytlacil 2006), matching patients from before Nursing Home Compare was released to patients with similar characteristics after Nursing Home Compare was released. By matching we assure similar distributions in observable characteristics between the patients being compared before and after Nursing Home Compare and limit our analyses to the units over which comparisons can be reliably made. This thereby reduces the possibility of model extrapolations and the biasing effects of model misspecification (Hill, Reiter, and Zanutto 2004; Ho et al. 2007).

Propensity scores were constructed using a large set of variables measured before or at nursing home admission. In particular, we conducted Mahalanobis matching on three variables (age, Cognitive Performance Scale [Morris et al. 1994], and RUG-III ADL Scale [Morris, Fries, and Morris 1999]) within propensity score calipers (Rubin and Thomas 2000). Each patient in the preperiod was matched to a patient in the postperiod within the same nursing home, and the matching was conducted ''with replacement'' in the sense that individuals in the post period could be selected as a match for up to 10 patients in the preperiod; the final analyses of the outcomes used weights to adjust for

this. Propensity score matching thus constrains patients risk profiles to be the same before and after Nursing Home Compare was launched and also constrains each nursing home's market share to be the same. Thirty-three variables were included in the propensity score and an additional 20 variables were assessed and were included in the propensity score if they were out of balance after an initialmatch. (A full list of variables included in the propensity score and details of the propensity score matching are included in Appendix SA2.)

3. **Slade, E.P., Stuart, E.A., Salkever, D.S., Karakus, M., Green, K.M., and Ialongo, N. (2008). Impacts of age of onset of substance use disorders on risk of adult incarceration among disadvantaged urban youth: A propensity score matching approach. *Drug and Alcohol Dependence* 95: 1-13. PMCID PMC2387099.**

Introduction:
Here, we use a common version of propensity score matching (Dehejia andWahba, 2002; Rosenbaum and Rubin, 1983a), one-to-one matching, to guard against the influence of factors that could confound the association between substance use disorders and incarceration risk. Mechanistically, propensity score matching generates treated and untreated comparison samples that closely resemble one another, or are "balanced," with respect to their observed characteristics and estimated propensity for exposure to the "treatment" (Dehejia and Wahba, 2002). Once matched, the outcomes of individuals in the comparison sample and those of individuals in the treatment sample – in this case having a substance use disorder – are compared, using mean comparisons or multivariable regression to estimate the marginal influence of exposure to the treatment. The internal validity of propensity score analyses depends critically on the available covariate measures. Even if the matched samples have similar observed characteristics, other unobserved characteristics could be correlated with both exposure to the treatment and with the outcome of interest. The propensity score analysis of the present study is enhanced by an unusually rich set of covariate predictors of participation in crime and substance use. Consequently, the resulting estimates of the association of substance use disorders to incarceration risk may be more robust to residual unobserved differences between the substance use disordered and comparison groups than is normally the case in studies based on observational data.

In Methods:
Each of the 279 young men who met DSM-IV criteria for a lifetime substance use disorder (American Psychiatric Association, 2000) was matched one-to-one with a young man who did not meet lifetime criteria but had similar propensity for these disorders. Data from the resulting matched sample (n = 558) were used in the analyses. Results for these 558 individuals may not generalize to the entire PIRC sample of 780 males, as they represent a non-random minority selected for their high propensity for substance use disorders. To describe the potential impact on generalizability, we report comparisons of the characteristics of those who were matched to those who were excluded from the analysis. In our discussion of results, we also discuss the results of sensitivity analyses using data from all 780 individuals and using an alternative matching technique.

Propensity scores were estimated in a multivariable logistic regression model in which

the dependent variable was a binary indicator for any lifetime substance use disorder. Propensity scores were used to identify from among potential comparison sample subjects the individual with the most similar characteristics, measured by absolute propensity score distance (Dehejia andWahba, 2002). To allow the nearest match possible for each individual with a disorder, we allowed individuals with no disorder to be matched multiple times. The resulting matched samples (treatment and comparison) were then compared and tested for balance on covariates.

Covariates in the logistic regression model for substance use disorder were chosen in three steps. First, we selected covariates in our database that previous studies suggested might have an association with substance use disorders and with incarceration (McBride et al., 2003; White and Gorman, 2000). These covariates included family background and demographic characteristics and the scale and school records measures of behavioral and emotional problems and school performance. Second, we estimated bivariate logistic regressions, using each potential covariate as the sole predictor of a disorder. Any variable that had a *P*-value less than 0.30 in the bivariate model was included in the full multivariable logistic regression model that was used to estimate propensity scores. Third, following procedures used in Dehejia and Wahba (2002), we iteratively added to the logistic regression specification higher order polynomials of covariates that did not perfectly balance in the previous specification, and again matched the sample. After an iteration of the propensity score matching, covariate values in the treatment and comparison samples were tested for balance across samples using *t*-tests.

As a result of matching, individuals in the matched nosubstance- use-disorder sample resembled (*t*-test values less than 1.96) those in the substance-use-disorder sample with respect to all covariates (shown in Table 1), including school performance, conduct disorder symptoms, proportion of school days missed in 7th grade, total disciplinary removals from school, juvenile court appearances, delinquency of peers, peer alcohol and marijuana use, age, employment status of childhood caregiver, caregiver substance use, race, and free lunch status. Within particular strata of the propensity scores, minor differences between treatment and matched comparison sample covariate means were identified. Mean sample differences within each quintile of the estimated propensity scores were tested for each variable, using *t*-tests. Differences at the 5% level were considered significant. Within the highest and lowest propensity score quintiles, significant mean differences could not be eliminated for age, the number of conduct disorder symptoms, and teacher-rated school performance measures.We therefore included linear and quadratic measures of these variables in the subsequent incarceration outcome regressions to adjust for residual imbalances in the extremes of their distributions.

Propensity score matching estimates of the “average treatment effect” depend critically on the so-called “conditional independence assumption” (Rosenbaum and Rubin, 1983a). Conditional independence means that, conditional on the observed covariates, the “treatment,” which in the present analysis is having a substance use disorder, is assigned independently of the outcome of interest. If the conditional independence assumption is violated, propensity score estimates of the average treatment effect may be biased and inconsistent. Although this assumption is not directly testable, methods for assessing the sensitivity of propensity score matching estimates to violations of the

conditional independence assumption have been proposed (Rosenbaum and Rubin, 1983b; Rosenbaum, 1987), and more recently have been developed for applications (Ichino et al., in press; Nannicini, 2007). These methods can be used to quantify the potential magnitude of bias in the propensity score matching estimate of the average treatment effect. In the present study we use this approach to assess the plausibility of the notion that one or more unmeasured confounders could fully explain the estimated marginal association of onset of substance use disorders with incarceration outcomes.

4. **Harder, V.S., Stuart, E.A., & Anthony, J. (2008). Adolescent cannabis problems and young adult depression: Male-female stratified propensity score analyses.** ***American Journal of Epidemiology*** **168: 592-601. PMCID: PMC2727198**

This study focuses on estimating the causal effect of adolescent- onset cannabis problems on the odds of young adult depression by using propensity score techniques, with stratification to capture possible male-female variation in the link between cannabis problems and depression. Results from more traditional epidemiologic analyses (multivariable logistic regressions) are also presented. Two parametric models and one nonparametric model were used to estimate the propensity score. We then applied the estimated propensity score to the final outcome logistic regression using three different application models. Details of these propensity score estimation and application techniques are included in the Models and software section below.

Decision criteria. Although this study builds and tests nine combinations of estimation and application techniques for the propensity score-adjusted models, the reported results are limited to estimates from the better performing propensity score techniques, as determined through decision criteria based on the assessment of covariate effect sizes (37). The propensity score techniques that perform well may vary for other data sets and research questions. These decision criteria can help researchers select which method is better for their particular study. The ‘‘effect size’’ for a particular covariate is the difference in average covariate values between the exposed and comparison groups divided by the standard error in the exposed group. In brief, the decision criteria identify the techniques that yield the smallest effect size across the majority of the covariates and across a few theoretically critical confounding covariates, while minimizing the extreme values of effect size for all covariates. Of importance, the propensity score techniques that meet the decision criteria are chosen prior to running the final outcome regressions, thus preventing bias through the selection of a method that yields a desired result.

Average causal effect. This article presents the estimated average causal effects of the ‘‘treatment on the treated.’’ In the causal inference methodology literature, the exposure variable is referred to as a ‘‘treatment,’’ but in this article we retain the epidemiologic terminology of exposure. The ‘‘treatment on the treated’’ is an estimate of the average causal effect that would be seen if everyone in the exposed group had been exposed versus no one in the exposed group being exposed. The other commonly reported average causal effect is referred to simply as the ‘‘average treatment effect’’ and is described elsewhere (38). In this article, we present the ‘‘treatment on the treated’’ estimate.

Models and software. Multivariable logistic regression (MLR), MLR with critically chosen interaction terms (39, 40), and generalized boosted modeling (GBM), a nonparametric regression tree technique (41), were used to estimate the propensity score. Each of these techniques models cannabis problem use as a function of the measured covariates. The propensity scores are the resulting predicted probabilities of cannabis problems for each individual. One to one (1:1) matching (18), full matching (42, 43), and weighting by the odds (44) were used to apply the propensity score to the final regression. Prior to running the final logistic regressions predicting young adult depression, we compared the resulting covariate effect sizes from each of the nine combinations of estimation and application techniques utilizing the aforementioned decision criteria. For females, the two propensity score techniques that performed well were MLR paired with full matching and MLR paired with weighting by the odds. For males, GBM paired with weighting by the odds and MLR paired with weighting by the odds both performed well. For the combined sample, GBM paired with weighting by the odds performed well with regard to the decision criteria. All statistical analyses were conducted in the R language (45). Two propensity score packages written for the R environment were used: MatchIt (46) and Twang (47). The two parametric propensity score estimation techniques used MatchIt, while the nonparametric estimation technique used Twang, which utilized the GBM package in R (48). The final logistic regression models were adjusted for the preexposure covariates used in the propensity score models to account for residual confounding. The results presented below are the propensity score-adjusted odds ratios from these logistic regressions, run for males and females separately, as well as for the combined sample.

RESULTS:
Cannabis problem users (the ‘‘exposed’’ group) are different from comparison individuals on many measured preexposure covariates. Across males, females, and the combined sample, the cannabis problem users and comparison individuals do not appear to have markedly different preexposure depression or anxiety levels. However, a higher percentage of the cannabis problem users were daily tobacco users, had problem alcohol use, or had slightly higher concentration and behavior problems than the comparison individuals (tables 1 and 2). The application of the propensity score corrected for these imbalances, as evidenced by the decrease in all measured covariate effect sizes below 0.25 and by nonsignificant chi-squared test statistics for all covariates after propensity score adjustment (table 2).

5. Green, K.M., Doherty, E.E., Stuart, E.A., and Ensminger, M.E. (2010). Does Heavy Adolescent Marijuana Lead to Criminal Involvement in Adulthood? Evidence from a Multiwave Longitudinal Study of Urban African Americans. *Drug and Alcohol Dependence* 112: 117-125. PMCID PMC2950879.

Methods:

2.4.1. Propensity score matching. After some descriptive statistics were run to compare arrest rates and ages between heavy adolescent marijuana users and light/non-users, the multivariate analysis was conducted in stages, as suggested by Ho et al. (2007). Propensity score matching was conducted using the MatchIt Program (Ho et al., 2006), a component of the R Statistical package. We utilized the full matching approach, as

described by Rosenbaum (1991) and Hansen (2004).
This approach allows us to retain all adolescents in our data analysis sample, and has been shown to be particularly effective at reducing bias due to observed confounding variables (Stuart and Green, 2008). Unlike k:1 matching, full matching is a more flexible approach. It creates a series of matched sets grouping together, in an optimal way, individuals with similar propensity scores, with each set including at least one exposed individual (i.e., heavy marijuana user) and at least one comparison individuals (i.e., light/non-user).

The purpose of this approach was to preprocess the data before the parametric analysis in order to reduce the association between heavy adolescent marijuana use and the confounding variables. This approach assumes that after conditioning on the observed covariates, there are no other differences between the heavy marijuana users and light/non-users; i.e., that all confounding is taken into account by the observed covariates. Therefore, bias (to the extent which the observed variables capture confounding) is removed and potential causal impacts of heavy adolescent marijuana can be estimated. Within the propensity score analysis, the first step is to estimate the probability of being a heavy adolescent marijuana user for each individual using logistic regression in which heavy adolescent marijuana use is the outcome and the matching variables are the covariates. After estimating the propensity score, full matching uses these propensity scores to group all of the individuals into a series of matched sets based on their overall likelihood of being a heavy marijuana user, with similar individuals (defined by the propensity score) placed into the same set.

Next, we assessed the adequacy of matching by performing a series of diagnostic checks as described in Stuart and Green (2008). The assessment included an examination of the balance of each covariate, its square, and every two-way interaction as determined by standardized bias. Standardized biases of less than .25, that is less than a quarter of a standard deviation difference in means between heavy users and light/non-users, were considered good matches (Ho et al., 2007). To improve the matching, we included various squared terms and interactions in the matching equation and re-estimated the propensity scores. The final model included a squared term of adolescent delinquency and of shyness and interactions between female-headed household and maternal substance use and between poverty status and maternal substance use. Once adequate sets were formed, each individual was then assigned a weight based on the ratio of heavy users to light/non-users within a set. In this analysis, we created 142 matched sets based on the propensity score. Propensity scores ranged from .01 (very low) to .95 (high). Each matched set contained an average of five individuals. Although sets varied in terms of the number of marijuana users and comparison individuals, each included at least one of the 185 heavy marijuana users (mean 1.30, median 1.00) and at least one of the 517 light/non-users (mean 3.64, median 1.00). The average difference in propensity scores within a set ranged from 0 to .05, with a mean of .002, demonstrating similarity of propensity scores within sets.

2.4.2. Weighted logistic regression. After matching, we first used weighted logistic regression in StataIC 10, to estimate the association of heavy adolescent marijuana use with each criminal outcome. All regression models include the matching variables as controls in order to further adjust for small differences remaining in the matched samples after matching (Ho et al., 2007). We also included the interaction of shyness and family

history of substance use since the standardized bias of this interaction was greater than .25 in the final model.

6. **Example from BMJ:**

Wijeysundera, D.N., Beattie, W.S., Karkouti, K., Neuman, M.D., Austin, P.C., and Laupacis, A. (2011). Association of echocardiography before major elective non-cardiac surgery with postoperative survival and length of hospital stay: population based cohort study. *British Medical Journal* 342:d3695.

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3127454/?tool=pubmed

Methods:
We developed a non-parsimonious multivariable logistic regression model to estimate a propensity score for echocardiography.27 Clinical significance guided the initial choice of covariates: age, sex, year, surgery, fifth of income, residence (urban *v*rural), hospital's characteristics (teaching status and volume of procedures), comorbid disease, preoperative medical consultation (general internist, cardiologist), preoperative anaesthesia consultation, epidural anaesthesia, and invasive monitoring. The comorbidities included in the model were coronary artery disease, congestive heart failure, atrial fibrillation, cardiac valvular disease, mechanical heart valve, cerebrovascular disease, peripheral vascular disease, hypertension, diabetes, pulmonary disease, renal disease, previous venous thromboembolism, liver disease, peptic ulcer disease, rheumatological disease, hemiplegia or paraplegia, malignancy, and dementia. We used previously described methods to categorise hospitals into quarters,28 on the basis of the total volume of included procedures. We used a structured iterative approach to refine this logistic regression model to achieve balance of covariates within the matched pairs.29 We used the standardised difference to measure covariate balance, whereby an absolute standardised difference above 10% represents meaningful imbalance.29 We then matched (without replacement) patients who had echocardiography to those who did not, by using a greedy matching algorithm with a calliper width of 0.2 SD of the log odds of the propensity score. We used statistical methods appropriate for paired data to compare outcomes.29

Results:
We matched approximately 89% (n=35498) of patients who had echocardiography to similar controls. The covariate balance was considerably improved (tables 3 and 4); the median absolute standardised difference decreased from 12.2% (range 0.5-83.4%) to 0.4% (0-1.9%). Within the matched cohort, preoperative echocardiography was associated with a small and statistically significant increase in postoperative mortality, both at 30 days (relative risk 1.14, 95% confidence interval 1.02 to 1.27; P=0.02; number needed to harm (NNH) 423) and at one year (1.07, 1.01 to 1.12; P=0.02; NNH 222) (table 5). It was also associated with an increase in mean hospital stay (0.31 (95% confidence interval 0.17 to 0.44) days; P<0.001), but not surgical site infection (relative risk 1.03, 0.98 to 1.06; P=0.18).

Discussion:
Influence of unmeasured confounding
Especially in the context of a non-randomised study, considering whether the increased mortality after echocardiography was simply due to residual unmeasured confounding is

important. Specifically, patients who had echocardiography may have been sicker and therefore at increased risk for postoperative complications. Despite our use of statistical methods to adjust for these differences and excellent covariate balance within the matched pairs, our data sources may have lacked sufficient detail to allow adequate adjustment for risk. However, several factors indicate that any residual confounding was small in magnitude and unlikely to have masked important benefits from preoperative echocardiography. Firstly, we found no association between echocardiography and the tracer outcome, surgical site infection. Surgical site infection is unlikely to be influenced by echocardiography. However, it is associated with markers of increased perioperative risk,32 which include characteristics not captured by administrative data. These risk factors include some comorbidities, such as smoking or obesity,33 34 as well as the severity of pre-existing diseases. Thus, if residual confounding within the matched pairs was minimal, rates of surgical site infection should be similar, as was the case in our study. Secondly, we found no association between echocardiography and mortality in patients who had concurrent stress testing. Echocardiography is unlikely to provide additional information that will change perioperative management for such patients. Thus, if it was associated with increased mortality in this subgroup, echocardiography may have been a marker for unmeasured risk factors, such as poor exercise capacity. As would be expected in the presence of minimal residual confounding, we found no difference in mortality. Finally, previous studies using these same data sources have found that perioperative interventions (such as epidural anaesthesia or stress testing) that are preferentially used in patients at high risk are nevertheless associated with improved outcomes after adjustment for risk.614 These results suggest that our data sources contain sufficient detail for adjustment for confounding.

7. **Example from the Journals of Gerontology:**

Williams, B.R., Zhang, Y., Sawyer, P., Mujib, M., Jones, L.G., Feller, M.A., Ekundayo, J., Aban, I.B., Love, T.E., Lott, A., and Ahmed, A. Intrinsic Association of Widowhood With Mortality in Community-Dwelling Older Women and Men: Findings From a Prospective Propensity-Matched Population Study. ***The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*** **66A(12): 1360-1368.**

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3252210/

Methods:
Because of significant imbalances in baseline characteristics between married and widowed individuals (Table 1), we used propensity score matching to assemble a cohort of participants so that the two groups would be well balanced on all measured baseline characteristics (13,14). We estimated propensity scores for widowhood for each of the 5,256 participants using a nonparsimonious multivariable logistic regression model (15–20). In the model, widowhood was the dependent variable, and the 74 baseline characteristics displayed in Figure 1 were entered as covariates. The propensity score for widowhood for a participant would be the conditional probability of that participant being a widow, given his or her measured baseline characteristics (13,14).

Because propensity score models are sample-specific adjusters and are not intended to be used for out-of-sample prediction or estimation of coefficients, measures of fitness

and discrimination are not important for the assessment of the model's effectiveness (21–24). The efficacy of propensity score models is best assessed by estimating postmatch absolute standardized differences between baseline covariates that directly quantifies the bias in the means (or proportions) of covariates across the groups, expressed as a percentage of the pooled standard deviations. We therefore calculated pre- and postmatch absolute standardized differences and presented those findings as Love plots (25,26). An absolute standardized difference of 0% would indicate no residual bias and less than 10% would be considered of inconsequential bias. Using a greedy matching protocol described elsewhere, we were able to match 819 pairs of married and widowed participants who had similar propensity scores (21–24).

Statistical Analysis

For descriptive analyses, we used Pearson chi-square and Wilcoxon rank-sum tests for the prematch data and McNemar's test and paired sample t test for postmatch comparisons as appropriate. Kaplan–Meier survival and Cox proportional hazard analyses were used to determine the association of widowhood with outcomes during 13 years of follow-up. We conducted a formal sensitivity analysis to quantify the degree of a hidden bias due to potential imbalance of an unmeasured covariate that would need to be present to invalidate our main conclusions (27). We then repeated our analysis in all 5,256 prematch participants using four different Cox regression models: (a) unadjusted; (b) age–sex–race adjusted; (c) multivariable adjusted, using all covariates used in the propensity score model; and (d) propensity score adjusted. Selected subgroup analyses were conducted to determine the heterogeneity of the association of baseline widowhood and mortality. All statistical tests were two sided, and tests with p value < .05 were considered significant. SPSS for Windows (Version 18) was used for all data analysis.

[Note: Nice figure showing the standardized biases.]

8. From the Journal of Adolescent Health:
Thornberry, T.P., Henry, K.L., Ireland, T.O., and Smith, C.A. (2011). The causal impact of childhood-limited maltreatment and adolescent maltreatment on early adult adjustment. *Journal of Adolescent Health* 46(4): 359-365.
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2871696/?tool=pubmed

Introduction:

The strongest design for assessing causality, a true experiment, is obviously unethical and illegal for this topic. The best approach currently available to assess causality in non-experimental designs is propensity score matching [7]. All sample members have some underlying propensity to be exposed to the "treatment," which in this case is to be maltreated. But, for youth with similar propensities, only some actually experience maltreatment. By comparing them to non-maltreated subjects with similar propensities, selection effects are greatly reduced and the rigor of experimental design more closely approximated. Causal inferences can be made with greater confidence because propensity score matching ensures that there is balance on all observed covariates prior to estimation of treatment effects, thereby more thoroughly and appropriately controlling for pretreatment selection bias. Our first purpose in this paper is to use propensity score matching to examine whether maltreatment causes subsequent problems in four domains of adjustment during the early adult years: involvement in crime and violence,

substance use, health-risking sex behaviors, and internalizing problems.

Methods:
The propensity score model uses 19 variables that reflect risk factors for child maltreatment [28]. These include child's age at baseline, gender, race/ethnicity, mother's age at first birth, neighborhood arrest rate, neighborhood proportion of families living in poverty, family structure, and family socio-economic status; parent measures, including their education, alcohol use, drug use, depressive symptoms, level of stress, incidence of stressful life events, social support, and harsh parenting; as well as family history of maltreatment, of substance use, and/or of mental health problems.

Analysis
Two separate propensity score models were estimated by regressing each of the developmentally specific maltreatment indicators on the covariates just described. These models estimate the log odds that an individual would be maltreated and the resultant score is the propensity score. The sample of non-maltreated individuals was restricted to those in the region of common support of the maltreated individuals (i.e., non-maltreated youth with a propensity score more than .25 of a standard deviation outside the range of maltreated youth were excluded), a technique recommended when the average causal effect among the treated, which estimates the predicted difference between the observed outcomes for maltreated youth (i.e., the observed young adult outcomes for each maltreated youth) and the outcomes that would have been observed if each maltreated youth was not maltreated (i.e., if the maltreatment had been prevented), is desired [29]. Restriction of the data to the range of common support reduced the sample size for the child match from 835 to 749 (with 645 control cases and 104 maltreated cases) and for the adolescent match from 803 to 674 (with 602 control cases and 72 maltreated cases). Next, a full matching approach [30] to casual inference, which makes use of all available data by matching treated individuals with as many similar controls as possible, weighting each individual by the number of individuals in the set, was used [29].

To assess the balancing of covariates across groups, we first used t-tests and chi-square tests; no significant differences exist on any of the covariates between the groups in the matched datasets (p-values ranged from .22 to 1.00 for the childhood match and from .48 to 1.00 for the adolescent match). The standardized bias [30] should be less than .25 [29]; across all matches, the largest was .17 for the childhood match and .09 for the adolescent match and most (84% childhood; 72% adolescent) were .05 or less. The standardized bias for the propensity score was drastically reduced in both models to -.01 for childhood-limited maltreatment and to .01 for adolescent maltreatment. For the childhood match, the full matching procedure created 99 subclasses, ranging in size from 2 (1 maltreated and 1 non-maltreated youth) to 61 (1 maltreated and 60 non-maltreated youth). For the adolescent match, it created 71 subclasses, ranging from 2 (1 maltreated and 1 non-maltreated youth) to 88 (1 maltreated and 87 non-maltreated youth). After matched datasets were obtained, we employed either negative binomial (for count outcomes with over dispersion), logistic, or ordinary least squares regression models, depending on the measurement of the outcome variable. All regressions were weighted to account for the full matching subclasses. In addition to the maltreatment predictor, age at baseline, gender, race/ethnicity, family structure, neighborhood arrest rate, neighborhood poverty, and family socio-economic status were included in the

regression models to adjust for any residual bias and to increase precision [31]. To account for missing data in the outcome variables, we created 10 multiply imputed datasets using a multiple imputation program that allows for categorical and count variables [32]. Missing data across the outcomes ranged from 3.6% to 10.5% for the childhood analyses and from 3.9% to 10.5% for the adolescent analyses. Analyses were run on each of the 10 imputed datasets and the estimates were combined using the procedures outlined by Rubin [33].

DOI: 10.1111/j.1475-6773.2009.00967.x
RESEARCH ARTICLE

Policies that Impact Care

Impact of Public Reporting on Quality of Postacute Care

Rachel M. Werner, R. Tamara Konetzka, Elizabeth A. Stuart, Edward C. Norton, Daniel Polsky, and Jeongyoung Park

Objective. Evidence supporting the use of public reporting of quality information to improve health care quality is mixed. While public reporting may improve reported quality, its effect on quality of care more broadly is uncertain. This study tests whether public reporting in the setting of nursing homes resulted in improvement of reported and broader but unreported quality of postacute care.

Data Sources/Study Setting. 1999–2005 nursing home Minimum Data Set and inpatient Medicare claims.

Study Design. We examined changes in postacute care quality in U.S. nursing homes in response to the initiation of public reporting on the Centers for Medicare and Medicaid Services website, Nursing Home Compare. We used small nursing homes that were not subject to public reporting as a contemporaneous control and also controlled for patient selection into nursing homes. Postacute care quality was measured using three publicly reported clinical quality measures and 30-day potentially preventable rehospitalization rates, an unreported measure of quality.

Principal Findings. Reported quality of postacute care improved after the initiation of public reporting for two of the three reported quality measures used in Nursing Home Compare. However, rates of potentially preventable rehospitalization did not significantly improve and, in some cases, worsened.

Conclusions. Public reporting of nursing home quality was associated with an improvement in most postacute care performance measures but not in the broader measure of rehospitalization.

Key Words. Quality of care, postacute care, nursing home quality, public reporting

Recent evidence about the poor quality of health care delivered in the United States (Kohn, Corrigan, and Donaldson 1999; Institute of Medicine 2001; McGlynn et al. 2003) has caused an outcry among health care consumers, providers, and policymakers. In an effort to improve quality of care, policymakers have turned to market-based reforms across the health care system. For example, the Centers for Medicare and Medicaid Services (CMS) have begun reporting health

was released. The MDS contains detailed clinical data collected at regular intervals for every resident in a Medicare- or Medicaid-certified nursing home. Data on patients' health, physical functioning, mental status, and psycho-social well-being have been collected electronically since 1998. These data are used by nursing homes to assess the needs and develop a plan of care unique to each resident and by the CMS to calculate Medicare prospective reimbursement rates. Because of the reliability of these data (Gambassi et al. 1998; Mor et al. 2003) and the detailed clinical information contained therein, they are considered the best available data for measuring nursing home clinical quality and thus are the source for the quality measures reported on Nursing Home Compare. Calculating the quality measures directly from the MDS allows us to measure quality both before and after Nursing Home Compare was released.

We also used the 100 percent MedPAR data (with all Part A claims) to calculate rates of potentially preventable rehospitalizations over the same time period, an accepted indicator of SNF quality (MedPAC 2005) that broadly measures a major goal of postacute care—stabilization following acute hospitalization (Donelan-McCall et al. 2006). Rehospitalization has thus far not been included in Nursing Home Compare's quality measures because data on rehospitalization are unavailable using MDS, the current data source for the quality measures. However, we linked the MDS data to MedPAR data using unique patient identifiers to determine rehospitalizations, providing an opportunity to examine whether changes in the reported quality measures were associated with broader changes in the quality of care provided at SNFs.

Propensity Score Matching

Because public reporting of patient outcomes may cause providers to select (or cherry-pick) their patients based on their illness severity (Dranove et al. 2003; Werner and Asch 2005), patient characteristics may systematically vary with the launch of public reporting. To account for this possibility, we use propensity score matching to ensure the similarity of patients being compared (Rosenbaum and Rubin 1983; Heckman, Urzua, and Vytlacil 2006), matching patients from before Nursing Home Compare was released to patients with similar characteristics after Nursing Home Compare was released. By matching we assure similar distributions in observable characteristics between the patients being compared before and after Nursing Home Compare and limit our analyses to the units over which comparisons can be reliably made. This thereby reduces the possibility of model extrapolations and the biasing effects of model misspecification (Hill, Reiter, and Zanutto 2004; Ho et al. 2007).

Propensity scores were constructed using a large set of variables measured before or at nursing home admission. In particular, we conducted Mahalanobis matching on three variables (age, Cognitive Performance Scale [Morris et al. 1994], and RUG-III ADL Scale [Morris, Fries, and Morris 1999]) within propensity score calipers (Rubin and Thomas 2000). Each patient in the preperiod was matched to a patient in the postperiod within the same nursing home, and the matching was conducted "with replacement" in the sense that individuals in the postperiod could be selected as a match for up to 10 patients in the preperiod; the final analyses of the outcomes used weights to adjust for this. Propensity score matching thus constrains patients risk profiles to be the same before and after Nursing Home Compare was launched and also constrains each nursing home's market share to be the same. Thirty-three variables were included in the propensity score and an additional 20 variables were assessed and were included in the propensity score if they were out of balance after an initial match. (A full list of variables included in the propensity score and details of the propensity score matching are included in Appendix SA2.)

Dependent Variables: Quality Measures

Using the propensity score-matched cohorts, we applied the technical definitions of the quality measures provided by CMS (Morris et al. 2003; Ho et al. 2007) to calculate quality measures for postacute care patients over the time period of the study, 1999–2005. We calculated all postacute care quality measures that were publicly reported on Nursing Home Compare when it was launched in 2002: percent of short-stay patients who did not have moderate or severe pain; percent of short-stay patients without delirium; and percent of short-stay patients whose walking improved. Following the conventions of the CMS quality measures we calculated the postacute care quality measures only on those patients who stay in the facility long enough to have a 14-day assessment, and limit the SNFs to those that were included in Nursing Home Compare (i.e., those SNFs that had at least 20 eligible postacute care patients over 6 months). Our calculations of SNF-level quality measures on the full sample of postacute care patients consistently matched the quality measures reported by CMS on Nursing Home Compare. We rescaled all reported quality measures so that a higher score indicates higher quality of care. The means and standard deviations for these measures over the study period were as follows: no pain 76.3 (19.3); no delirium 96.4 (7.5); and improved walking 6.9 (10.5).

We also measured rates of potentially preventable rehospitalizations as a broader indicator of SNF quality. Whereas Nursing Home Compare quality

measures are defined based only on patients who stay in postacute care for at least 14 days, we measured rehospitalizations for all postacute care patients regardless of their length of stay. Potentially preventable rehospitalizations were defined based on the Agency for Healthcare Research and Quality Prevention Quality Indicators (Agency for Healthcare Research and Quality 2004) that were applicable to patients aged 65 and older (bacterial pneumonia, chronic obstructive pulmonary disease, dehydration, heart failure, hypertension, short-term diabetic complications, uncontrolled diabetes, and urinary infection), occurring within 30 days of admission to postacute care (White and Seagrave 2005). Higher rehospitalization rates indicate lower quality of care. Over the study period, an average of 7.0 percent of patients (standard deviation 5.9) had a potentially preventable rehospitalization.

Main Independent Variable: Nursing Home Compare Indicator Variables

We examined changes in the quality measures with the launch of Nursing Home Compare using two separate independent variables. First, we used a set of year indicator variables (2000–2005, omitting 1999) to examine changes in quality over the study period. These year indicator variables were used to examine changes in quality at the launch of Nursing Home Compare (in November 2002) by testing the difference between the coefficients on the 2002 and 2003 indicator variables. Second, we used a pre–post indicator variable equaling 1 in the period after Nursing Home Compare was launched (after April 22, 2002, in pilot states and November 12, 2002, in nonpilot states) and 0 otherwise. This pre–post indicator variable was used to test whether the quality level differed in the 3-year period after Nursing Home Compare was launched compared with the pre-Nursing Home Compare period.

Covariates

In all analyses we included the variables included in the propensity scores as covariates (see Appendix SA2) to adjust for any remaining small differences between the groups after the matching (Ho et al. 2007). In addition, to estimate changes in delirium we include prior residential history as a covariate, as specified by the CMS quality measure (Moore et al. 2005); and when estimating changes in rehospitalization rates, we include previously developed variables for risk adjustment of potentially preventable rehospitalization from postacute care (Donelan-McCall et al. 2006). Finally, because the Nursing Home Compare quality measures are calculated only on patients who remain in postacute care for at least 14 days, in all regressions we control for the "censoring" rate at

each facility using quarterly measures of the proportion of all postacute care admissions that remain in postacute care for 14 days at each SNF.

Empirical Specifications

To examine the effect of publicly reporting nursing home quality on nursing home quality of care, we test for within-SNF changes in facility-level quality. Because the propensity score matched cohorts explicitly constrains changes in market share (through 1:1 pre–post matching of patients within SNFs) we empirically isolate one way that Nursing Home Compare was designed to affect quality of care—through provider-driven quality improvements—while controlling for any consumer-driven changes in care.

First, we describe changes in postacute care quality of care using a pre–post specification. Within this pre–post specification, we test for within-SNF improvements in quality of care using individual-level linear probability models, where changes in quality for patient i in SNF j at time t were estimated as a function of Nursing Home Compare indicator variables, patient- and SNF-level covariates, and SNF fixed effects:

$$Quality_{i,j,t} = \boldsymbol{\beta}_{\mathbf{j}} + \beta_1 NHC_t + \boldsymbol{\beta}\mathbf{X}_{\mathbf{i,j,t}} + \boldsymbol{\varepsilon}_{i,j,t} \quad (1)$$

Second, we test whether the estimated changes in postacute care quality are attributable to Nursing Home Compare using a difference-in-differences specification, using small SNFs not subject to public reporting as a control group. Although one third of SNFs are excluded from Nursing Home Compare at any given time, many of these small SNFs are only intermittently excluded from Nursing Home Compare when their census drops below the 20-patient threshold. We include the 15 percent of SNFs that are *never* included in Nursing Home Compare as our control group. Empirically, we estimate changes in quality as a function of a Nursing Home Compare indicator variable, an SNF size indicator variable ($Large_j$, equaling 1 if a large SNF is included in Nursing Home Compare; 0 if an SNF is never included), its interaction with $NHC_{j,t}$, patient- and SNF-level covariates, and SNF fixed effects:

$$Quality_{i,j,t} = \boldsymbol{\beta}_{\mathbf{j}} + \beta_1 NHC_t + \beta_2 Large_j \times NHC_t + \boldsymbol{\beta}\mathbf{X}_{\mathbf{i,j,t}} + \boldsymbol{\varepsilon}_{i,j,t} \quad (2)$$

Although small SNFs are clearly different from larger SNFs on average, this is not a problem in that these average differences are controlled through the SNF fixed effects. This model does rely on the assumption that small and large SNFs would exhibit similar trends over time in the absence of Nursing Home Compare. To test the validity of this assumption, we test whether trends in quality improvement in the pre-Nursing Home Compare period were the same for

small and large SNFs using multiple *F*-tests and find that for all four measures there were no significant differences in quality trends between small and large SNFs before Nursing Home Compare. Nonetheless, because small SNFs are likely to be different than large SNFs in their response to quality improvement initiatives, we test the sensitivity of our results by limiting the treatment group to mid-sized SNFs, or those that are large enough to be included in Nursing Home Compare but more similar in size and other characteristics to the small SNFs (mid-sized SNFs were defined as those included in Nursing Home Compare with fewer than 110 total beds and <30 percent Medicare residents). In addition, because nursing homes with few postacute care residents may have larger numbers of long-stay residents, and thus be included in Nursing Home Compare for the long-stay measures, we tested the correlation between the short-stay and long-stay measures and found it to be low (range 0.04–0.06). Robust standard errors were used to account for nonindependence of observations from the same facility in all regressions (Huber 1967; White 1980).

RESULTS

A total of 8,137 SNFs from Nursing Home Compare were included in the study, covering 9,390,930 postacute care stays and 5,899,327 postacute care stays of at least 14 days. An additional 2,277 small SNFs, covering 442,952 postacute care stays and 214,094 postacute care stays of at least 14 days, were not included in Nursing Home Compare and serve as a control group. Characteristics of these SNFs, stratified by size, are summarized in Table 1. By definition, large and mid-sized SNFs (those included in Nursing Home Compare) had more beds. Mid-sized SNFs were more similar to small SNFs in size and percent of Medicare residents, but were more likely to be part of a chain and be for-profit facilities than small and large SNFs. Mid-sized facilities were also less likely to be hospital based and had fewer staff hours per resident day compared with large and small SNFs. These characteristics of mid-sized SNFs are often associated with lower SNF quality.

All three reported measures of quality improved over time, as did rates of potentially preventable rehospitalizations. (See Figure 1 for risk-adjusted trends in postacute quality of care.) Multivariate regression results testing within-SNF changes in quality of care associated with the launch of Nursing Home Compare are summarized in Table 2. The three reported quality measures were better in the 3 years after Nursing Home Compare was launched compared with before. Over those years, the percent of patients without moderate to severe pain improved by 2.0 percentage points (on a base of 76

control for changing case mix at SNFs using propensity score matching, and control for differential rates of discharge before day 14 across SNFs, we are unable to adequately control for rates of selective discharge in this analysis. Disentangling the effects of selective discharge from true changes in quality is an important area for future research.

Our results should be interpreted in light of the study's limitations. First, results based on propensity score-matched cohorts may not be representative of the changes in quality that occurred as we do not include all patients in our analyses. (Rather, we limit our analyses to the subset of patients with similar characteristics for whom we can best estimate the effects of Nursing Home Compare.) Nonetheless, these results are important from a policy standpoint, as they provide more precise estimates of the effect of public reporting on quality of care. Second, although we extensively control for patient selection based on observed differences between patients before versus after the launch of Nursing Home Compare, unobserved differences that are uncorrelated with observed differences remain a threat to the validity of our findings. Nonetheless, prior work suggests that using propensity score matching in the setting of an exogenous treatment decision, such as the launch of Nursing Home Compare, is a valid approach to account for differences across groups (Heckman, Urzua, and Vytlacil 2006). Finally, the quality changes we demonstrate may be due to changes in data accuracy rather than true quality changes, particularly given the subjective nature of the Nursing Home Compare quality measures. While other work has found that data changes explain some quality improvement (Green and Wintfeld 1995; Roski et al. 2003), this is less likely in the nursing home quality measures based on MDS. Electronic MDS data collection started in 1998, well before Nursing Home Compare was launched, and has been used to determine Medicare payment since that time, increasing nursing homes' incentive to accurately report these data for several years before the launch of Nursing Home Compare.

Despite these limitations, our study offers important new findings with regard to the role of public reporting in quality improvement. We find that most quality measures improve in response to public reporting even after controlling for secular trends. However, the clinical significance of these improvements may be limited given that improvements in narrow measures of quality of care may not translate into broader quality improvement. To achieve more robust quality improvement, stronger incentives to improve quality may be needed. One possible way to do this is to combine public reporting with pay for performance. While public reporting of quality information is important for reasons beyond quality improvement, such as

Rubin, D. B., and N. Thomas. 2000. "Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates." *Journal of the American Statistical Association* 95 (450): 573–85.

Shorr, R. I., R. L. Fought, and W. A. Ray. 1994. "Changes in Antipsychotic Drug Use in Nursing Homes during Implementation of the OBRA-87 Regulations." *Journal of the American Medical Association* 271 (5): 358–62.

Snowden, M., and P. Roy-Byrne. 1998. "Mental Illness and Nursing Home Reform: OBRA-87 Ten Years Later. Omnibus Budget Reconciliation Act." *Psychiatric Services* 49 (2): 229–33.

Werner, R. M., and D. A. Asch. 2005. "The Unintended Consequences of Publicly Reporting Quality Information." *Journal of the American Medical Association* 293 (10): 1239–44.

Werner, R. M., and E. T. Bradlow. 2006. "Relationship between Medicare's Hospital Compare Performance Measures and Mortality Rates." *Journal of the American Medical Association* 296 (22): 2694–702.

White, C., and S. Seagrave. 2005. "What Happens When Hospital-Based Skilled Nursing Facilities Close? A Propensity Score Analysis." *Health Services Research* 40 (6, part 1): 1883–97.

White, H. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48: 817–30.

Williams, S. C., S. P. Schmaltz, D. J. Morton, R. G. Koss, and J. M. Loeb. 2005. "Quality of Care in U.S. Hospitals as Reflected by Standardized Measures, 2002–2004." *New England Journal of Medicine* 353 (3): 255–64.

Wunderlich, G. S., and P. Kohler. 2000. *Improving the Quality of Long-Term Care.* Washington, DC: Division of Health Care Services, Institute of Medicine.

Zinn, J., W. Spector, L. Hsieh, and D. B. Mukamel. 2005. "Do Trends in the Reporting of Quality Measures on the Nursing Home Compare Web Site Differ by Nursing Home Characteristics?" *Gerontologist* 45 (6): 720–30.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix SA1: Author Matrix.

Appendix SA2: The Impact of Public Reporting on Quality of Postacute Care: Propensity-Score Methods.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Appendix

The impact of public reporting on quality of post-acute care

Rachel M. Werner, R. Tamara Konetzka, Elizabeth A. Stuart, Edward C. Norton, Daniel Polsky, Jeongyoung Park

For within-SNF comparisons of quality the propensity score was estimated and the matching was conducted separately within each SNF. For the small SNFs used as a control group, because these SNFs were too small to reliably estimate the propensity score within each SNF, the propensity score was estimated within each SNF's Healthcare Service Area (HSA) as defined in the Dartmouth Atlas. Prior to matching, between 12.5 and 21.4% of the variables used in the matching were significantly different in the pre- vs. post-NHC period, defined by an absolute standardized difference in means greater than 0.2 (Ho et al. 2007). In each of the matched samples, none of the variables were significantly different.

Ho, D. E., K. Imai, G. King and E. A. Stuart (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Political Analysis 15(3): 199-236.

Appendix Table. Variables used in propensity score matching. All variables listed under "Propensity score" were included in all propensity scores. In addition, "Balance check variables" were also included in the propensity score if they were out of balance after an initial match based only the propensity score variables.

Mahalanobis variables
Age
Cognitive performance scale
RUG-ADL scale
Propensity score variables
Pain on day 5
Delirium on day 5
Pressure sores on day 5
Independent in walking on day 5
log (total Medicare charges over past year)
Number of hospitalizations over past year
Number of SNF admissions over past year
RUG group
Sex
Race
Body mass index
Bedfast
Urinary incontinence
Use of bladder catheter
Surgical wounds
Open skin lesions
End stage disease
History of stroke
History of vision loss
History of dehydration
History of peripheral vascular disease
History of diabetes
History of resolved pressure ulcer
Balance check variables
Depression rating scale
Clinically complex scale
Infection
Bed mobility support
Transfer supported
Weight loss
History of heart failure
Hemiplegia/hemiparesis
Quadriplegia
Recent hip fracture
Mood persistence
Wandering
Verbally abusive
Physically abusive
Socially inappropriate
Independent in dressing
Independent in personal hygiene
Hallucinations
Unsteady gait
Discharge potential

Published in final edited form as:
Int J Cardiol. 2010 May 28; 141(2): 167–174. doi:10.1016/j.ijcard.2008.11.195.

Oral Potassium Supplement Use and Outcomes in Chronic Heart Failure: A Propensity-Matched Study

O. James Ekundayo, MD, DrPH[a], **Chris Adamopoulos, MD**[b], **Mustafa I. Ahmed, MD**[a], **Bertram Pitt, MD**[c], **James B. Young, MD**[d], **Jerome L. Fleg, MD**[e], **Thomas E. Love, PhD**[f], **Xuemei Sui, MD, MPH**[g], **Gilbert J. Perry, MD**[a,h], **David S. Siscovick, MD, MPH**[i], **George Bakris, MD**[j], and **Ali Ahmed, MD, MPH**[a,h,*]

[a]University of Alabama at Birmingham, Birmingham, AL, USA

[b]Papageorgiou General Hospital, Thessaloniki, Greece

[c]University of Michigan, Ann Arbor, Michigan, USA

[d]Cleveland Clinic Foundation, Cleveland Ohio, USA

[e]National Heart, Lung, and Blood Institute, Bethesda, Maryland, USA

[f]Case Western Reserve University, Cleveland, Ohio, USA

[g]University of South Carolina, Columbia, South Carolina, USA

[h]VA Medical Center, Birmingham, AL, USA

[i]University of Washington, Seattle, Washington, USA

[j]University of Chicago, Chicago, Illinois, USA

Abstract

Background—Hypokalemia is common in heart failure (HF) and is associated with increased mortality. Potassium supplements are commonly used to treat hypokalemia and maintain normokalemia. However, their long-term effects on outcomes in chronic HF are unknown. We used a public-use copy of the Digitalis Investigation Group (DIG) trial dataset to determine the associations of potassium supplement use with outcomes using a propensity-matched design.

Methods—Of the 7788 DIG participants with chronic HF, 2199 were using oral potassium supplements at baseline. We estimated propensity scores for potassium supplement use for each patient and used them to match 2131 pairs of patients receiving and not receiving potassium supplements. Matched Cox regression models were used to estimate associations of potassium supplement use with mortality and hospitalization during 40 months of median follow-up.

Results—All-cause mortality occurred in 818 (rate, 1327/10000 person-years) and 802 (rate, 1313/10000 person-years) patients respectively receiving and not receiving potassium supplements (hazard ratio {HR} when potassium supplement use was compared with nonuse, 1.05; 95% confidence interval {CI}, 0.94–1.18; P=0.390). All-cause hospitalizations occurred in 1516 (rate,

*Corresponding author: Ali Ahmed, MD, MPH, University of Alabama at Birmingham, 1530 3rd Ave South, CH-19, Ste-219, Birmingham AL 35294-2041, USA. **Tel.:** +1-205-934-9632; **fax:** +1-205-975-7099; aahmed@uab.edu.

Conflict of Interest Disclosure: None

4777/10,000 person-years) and 1445 (rate, 4120/10,000 person-years) patients respectively receiving and not receiving potassium supplements (HR, 1.15; 95% CI, 1.05–1.26; P=0.004). HR (95% CI) for hospitalizations due to cardiovascular causes and worsening HF were respectively 1.19 (95% CI, 1.08–1.32; P=0.001) and 1.27 (1.12–1.43; P<0.0001).

Conclusion—The use of potassium supplements in chronic HF was not associated with mortality. However, their use was associated with increased hospitalization due to cardiovascular causes and progressive HF.

Keywords

Heart failure; potassium supplement; mortality; hospitalization; propensity score

1. Introduction

Hypokalemia is common in heart failure (HF) and is associated with poor outcomes [1]. Oral potassium supplements are often used to treat hypokalemia and maintain normokalemia in HF patients with low serum potassium levels. However, the long-term effects of potassium supplement use on outcomes in chronic HF are unknown. The objective of this study was to examine the associations of potassium supplement use with mortality and hospitalization in a propensity-matched cohort of ambulatory chronic HF patients.

2. Materials and methods

2.1. Study patients

The Digitalis Investigation Group (DIG) trial was a multi-center randomized clinical trial, the design and results of which have been reported previously [2,3]. Briefly, 7788 ambulatory chronic HF patients (6800 had left ventricular ejection fraction ≤45%) in normal sinus rhythm receiving angiotensin-converting enzyme inhibitors and diuretics were randomized to receive digitalis and placebo. Overall, 2199 (28%) patients were receiving oral potassium supplements at baseline and 5589 (72%) patients were not receiving potassium supplements. Data on the use of potassium supplements were available from all 7788 participants.

2.2. Study design: propensity score matching

We focus our current analysis to a subset of 4262 patients, who were assembled through propensity score matching [4–7]. Because patients in the DIG trial were not randomized to receive potassium supplements, the probabilities of actually receiving potassium supplements varied according to the baseline characteristics of those patients. The propensity matching approach allows the assembly of a cohort who would be well-balanced in all measured baseline covariates. Importantly, this can be done without access to the outcomes data, thus maintaining a degree of blindness, which is a key feature of randomized clinical trials [4–7].

The propensity score for potassium supplement use for a patient is the conditional probability of receiving these drugs given that patient's baseline characteristics [4–7]. We estimated propensity scores for the use of potassium supplements for each of the 7788 patients with a non-parsimonious multivariable logistic regression model using baseline characteristics presented in Figure 1, and checking for plausible interactions [1,8–10]. We then matched patients who were receiving potassium supplements with those who were not receiving potassium supplements but had similar propensity to receive them [1,8–10]. Using a greedy matching protocol, we were able to match 97% (2131 of 2199) of patients receiving potassium supplements, yielding a matched cohort of 4262 patients. We then estimated absolute standardized differences to assess pre-match imbalances and post-match balance in baseline

covariates and presented those findings as a Love plot [1,8–12]. An absolute standardized difference of 0% would suggest no residual bias, and those below 10% suggest negligible bias.

2.3. Study outcomes

The primary outcomes for the current analysis were all-cause mortality and all-cause hospitalization, and secondary outcomes included mortality and hospitalizations due to cardiovascular causes and HF. DIG participants were followed for a median of 38 months and vital status data were complete for 99% of the patients [13].

2.4. Statistical analysis

For descriptive analyses, we used Pearson Chi square and Wilcoxon rank-sum tests for the pre-match, and McNemar's test and paired sample t-test for the post-match comparisons, as appropriate. For the pre-match comparison, of the 5589 patients not receiving potassium supplements, a random sample of 2131 patients were selected and were compared with 2131 matched patients receiving potassium supplements. This was done to have similar pre- and post-match sample sizes (n=4262), and to avoid overestimation of significant p values from a larger sample size (n=7788). We then used Kaplan-Meier plots and matched Cox regression analysis to estimate associations of potassium supplement use with total and cause-specific deaths and hospitalizations in the matched cohort [1,8]. Formal sensitivity analyses were conducted to determine the effect of a potential hidden confounder on our study findings [14]. Subgroup analyses were conducted to determine the homogeneity of the association between potassium supplement use and mortality. All statistical tests were evaluated using two-tailed 95% confidence levels, using SPSS-15 for Windows [15].

3. Results

3.1. Patient characteristics

Patients (n=4262) had a mean (±SD) age of 64 (±11) years, 28% were women and 16% were nonwhites. There were significant pre-match imbalances in key baseline covariates including gender, race, comorbidity, disease severity and baseline serum potassium, all of which were balanced after matching (Table 1). Absolute standardized differences for all measured covariates were below 5% indicating negligible post-match bias (Figure 1).

3.2. Potassium supplement use and mortality

During a median follow up of 40 months, 1620 (38%) patients died from all causes, 1290 (30%) due to cardiovascular causes and 607 (14%) due to progressive HF. All-cause mortality occurred in 818 (rate, 1327/10000 person-years) and 802 (rate, 1313/10000 person-years) patients receiving and not receiving potassium supplements respectively (hazard ratio {HR} when potassium supplement use was compared with its nonuse, 1.05; 95% confidence interval {CI}, 0.94–1.18; P=0.390; Figure 2a and Table 2). This association was homogeneous across a wide spectrum of HF patients (Figure 3). Associations between potassium supplement use and cause-specific mortalities are displayed in Table 2. In the pre-match cohort of 4262 patients, 1503 (35%) died from all causes. All-cause mortality occurred in 818 (38%) and 685 (32%) patients receiving and not receiving potassium supplements respectively (HR, 1.30; 95% CI, 1.17–1.43; P<0.0001).

3.3. Potassium supplement use and hospitalization

Overall, hospitalizations due to all causes, cardiovascular causes, and worsening HF occurred respectively in 2961 (69%), 2362 (55%) and 1483 (35%) patients. All-cause hospitalizations occurred in 1516 (rate, 4777/10000 person-years) and 1445 (rate, 4120/10000 person-years) patients receiving and not receiving potassium supplements respectively (HR for potassium

supplement use, 1.15; 95% CI, 1.05–1.26; P=0.004; Figure 2b and Table 3). Results of our sensitivity analysis suggest that in the absence of hidden bias, a sign-score test for matched data with censoring provides strong evidence (two-tailed p=0.003) that the use of potassium supplement was associated with an increased risk of all-cause hospitalization. A hidden covariate that would increase the odds of potassium supplement use by only 4.9% may potentially explain away this association [14]. Associations between potassium supplement use and cause-specific hospitalizations are displayed in Table 3.

4. Discussion

Findings from the current analysis demonstrate that oral potassium supplement use was not associated with all-cause mortality in chronic HF, but was associated with increased all-cause hospitalization, which was mostly driven by cardiovascular and HF hospitalizations. Considering that the use of potassium supplements is a marker of hypokalemia, the lack of an independent association between potassium supplement use and all-cause mortality suggest that potassium supplements may have eliminated the increased mortality associated with hypokalemia [1]. However, the association between potassium supplement use and increased cardiovascular hospitalization, which was primarily driven by an increase in HF hospitalization, is somewhat puzzling. Even though the underlying mechanism for this latter association remains unclear, data from animal studies suggest a possible association between plasma potassium and salt appetite [16,17]. To the best of our knowledge this is first report of an association between potassium supplement use and outcomes in a propensity-matched cohort of chronic HF patients in which those receiving and not receiving potassium supplements were well-balanced in all measured baseline covariates.

Hypokalemia is associated with increased mortality without increase in hospitalization suggesting that most hypokalemia-associated deaths may be sudden and caused by fatal arrhythmias [1]. We find a reversal of these associations in those using oral potassium supplements: no association with mortality and an association with increased hospitalization. Potassium supplement use is only indicated to treat and prevent hypokalemia and thus may be considered a marker of hypokalemia. Yet, our findings suggest that hypokalemia-associated mortality was eliminated in those receiving potassium supplements [1]. This is likely mediated via the correction of hypokalemia. It is unknown whether potassium supplements have intrinsic survival benefit in HF. In patients with hypertension, potassium-sparing diuretics, but not potassium supplements, have been shown to reduce the risk of cardiac arrest [18–20]. Findings from the current analysis suggest that in patients with chronic HF potassium supplements may eliminate hypokalemia-associated deaths. However, potassium supplements do not correct other electrolyte imbalances such as hypomagnesemia, and may not also efficiently replenish tissue potassium [21]. Aldosterone antagonists, such as spironolactone, on the other hand, may replenish body potassium, and also prevent disease progression and mortality in HF [22–25].

Our findings suggest that potassium supplements neutralize the excess deaths associated with hypokalemia, and support their use to correct hypokalemia. However, our findings also suggest that potassium supplements may not provide any additional mortality reduction. Therefore, once hypokalemia has been corrected, aldosterone antagonists may be preferable over chronic potassium supplements for the maintenance of normokalemia [23,24,26]. When aldosterone antagonists are used, potassium supplements are often not needed and may even be deleterious [27]. This is also suggested by our subgroup analyses in which the subgroup of patients receiving potassium-sparing diuretics were the only subgroup in which the use of potassium supplements was associated with a trend toward increased mortality (Figure 3). For patients who fail to maintain normokalemia despite the use of aldosterone antagonists in appropriate doses, potassium supplements may be used judiciously to maintain serum potassium between 4 and 5 mEq/L [1,28]. HF patients, who are elderly, have diabetes or renal insufficiency, or

are receiving inhibitors of the renin-angiotensin system or beta-blockers, or non-steroidal anti-inflammatory drugs, are at increased risk of hyperkalemia [29,30]. Aldosterone antagonists and potassium supplements should be used with caution along with close monitoring of serum potassium levels in these patients.

A rather unexpected finding of our study was increased hospitalizations associated with the use of potassium supplements. While we do not have a clear mechanistic insight into this association, data from animal studies suggest that potassium supplements may increase salt appetite [16,17]. This notion is also supported by the fact that hyperkalemia-associated hospitalizations were mostly due to cardiovascular causes and worsening HF. We had no data on diuretic dosages and it is possible that use of potassium supplements was a marker of disease severity and use of higher doses of diuretics. Symptomatic HF patients receiving higher doses of diuretics are more likely to develop hypokalemia and receive potassium supplements. However, our matched patients were well balanced in all measured baseline characteristics that also included markers of symptoms and disease severity. Finally, the association between potassium supplement use and hospitalization was relatively sensitive to a potential unmeasured covariate. However, sensitivity analysis cannot determine if such a hidden confounder exists. For a covariate to become a confounder it should be a strong predictor of hospitalization and not be strongly correlated with any of the measured covariates in our study.

Our study has several limitations. Patients in our study were predominantly white, male, relatively young, and in normal sinus rhythm, from the pre-beta-blocker era of HF therapy, which may limit generalizability. Thus, the results of this study should be interpreted with caution, and need to be replicated in contemporary propensity-matched HF populations, and preferably in prospective randomized clinical trials.

In conclusion, the use of potassium supplements was not associated with mortality but was associated with increased hospitalization in chronic HF patients. Once hypokalemia has been corrected with potassium supplements, aldosterone antagonists may be preferable for the maintenance of normokalemia in those with recurrent hypokalemia requiring chronic oral potassium supplement use.

Acknowledgments

The Digitalis Investigation Group (DIG) study was conducted and supported by the NHLBI in collaboration with the DIG Investigators. This Manuscript was prepared using a limited access dataset obtained by the NHLBI and does not necessarily reflect the opinions or views of the DIG Study or the NHLBI.

Funding/Support: Dr. Ahmed is supported by the National Institutes of Health through grants from the National Heart, Lung, and Blood Institute (5-R01-HL085561-02 and P50-HL077100), and a generous gift from Ms. Jean B. Morris of Birmingham, Alabama.

References

1. Ahmed A, Zannad F, Love TE, et al. A propensity-matched study of the association of low serum potassium levels and mortality in chronic heart failure. Eur Heart J 2007;28:1334–1343. [PubMed: 17537738]
2. The Digitalis Investigation Group. Rationale, design, implementation, and baseline characteristics of patients in the DIG trial: a large, simple, long-term trial to evaluate the effect of digitalis on mortality in heart failure. Control Clin Trials 1996;17:77–97. [PubMed: 8721804]
3. The Digitalis Investigation Group. The effect of digoxin on mortality and morbidity in patients with heart failure. N Engl J Med 1997;336:525–533. [PubMed: 9036306]
4. Rosenbaum PR, Rubin DB. The central role of propensity score in observational studies for causal effects. Biometrika 1983;70:41–55.

5. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. J Am Stat Asso 1984;79:516–524.
6. Rubin DB. Using propensity score to help design observational studies: Application to the tobacco litigation. Health Services and Outcomes Research Methodology 2001;2:169–188.
7. Rubin DB. On principles for modeling propensity scores in medical research. Pharmacoepidemiol Drug Saf 2004;13:855–857. [PubMed: 15386710]
8. Ahmed A, Husain A, Love TE, et al. Heart failure, chronic diuretic use, and increase in mortality and hospitalization: an observational study using propensity score methods. Eur Heart J 2006;27:1431–1439. [PubMed: 16709595]
9. Ahmed A, Perry GJ, Fleg JL, Love TE, Goff DC Jr, Kitzman DW. Outcomes in ambulatory chronic systolic and diastolic heart failure: a propensity score analysis. Am Heart J 2006;152:956–966. [PubMed: 17070167]
10. Ahmed A, Aban IB, Vaccarino V, et al. A propensity-matched study of the effect of diabetes on the natural history of heart failure: variations by sex and age. Heart 2007;93:1584–1590. [PubMed: 17488764]
11. D'Agostino RB Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med 1998;17:2265–2281. [PubMed: 9802183]
12. Normand ST, Landrum MB, Guadagnoli E, et al. Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: a matched analysis using propensity scores. J Clin Epidemiol 2001;54:387–398. [PubMed: 11297888]
13. Collins JF, Howell CL, Horney RA. Determination of vital status at the end of the DIG trial. Control Clin Trials 2003;24:726–730. [PubMed: 14662278]
14. Rosenbaum, PR. Sensitivity to Hidden Bias. In: Rosenbaum, PR., editor. Observational Studies. 2 ed.. New York: Springer-Verlag; 2002. p. 110-124.
15. SPSS for Windows, Rel. 15. program. Chicago, IL: SPSS Inc; 2008.
16. Michell AR. Relationships between individual differences in salt appetite of sheep and their plasma electrolyte status. Physiol Behav 1976;17:215–219. [PubMed: 996157]
17. Michell AR. Plasma potassium and sodium appetite; the effect of potassium infusion in sheep. Br Vet J 1978;134:217–224. [PubMed: 667588]
18. Siscovick DS, Raghunathan TE, Psaty BM, et al. Diuretic therapy for hypertension and the risk of primary cardiac arrest. N Engl J Med 1994;330:1852–1857. [PubMed: 8196728]
19. Laragh JH, Sealey JE. K() depletion and the progression of hypertensive disease or heart failure. The pathogenic role of diuretic-induced aldosterone secretion. Hypertension 2001;37:806–810. [PubMed: 11230377]
20. Siegel D, Hulley SB, Black DM, et al. Diuretics, serum and intracellular electrolyte levels, and ventricular arrhythmias in hypertensive men. JAMA 1992;267:1083–1089. [PubMed: 1735925]
21. Delgado, MC.; Delgado-Almeida, A. Abnormal potassium. In: Mohler, IER., editor. Advanced Therapy in Hypertension and Vascular Disease. Elsevier: Decker, B.C. Inc; 2006. p. 261-299.
22. Packer M. Potential role of potassium as a determinant of morbidity and mortality in patients with systemic hypertension and congestive heart failure. Am J Cardiol 1990;65:45E–51E.
23. Pitt B, Remme W, Zannad F, et al. Eplerenone, a selective aldosterone blocker, in patients with left ventricular dysfunction after myocardial infarction. N Engl J Med 2003;348:1309–1321. [PubMed: 12668699]
24. Pitt B, Zannad F, Remme WJ, et al. Randomized Aldactone Evaluation Study Investigators. The effect of spironolactone on morbidity and mortality in patients with severe heart failure. N Engl J Med 1999;341:709–717. [PubMed: 10471456]
25. Weber KT, Villarreal D. Aldosterone and antialdosterone therapy in congestive heart failure. Am J Cardiol 1993;71:3A–11A.
26. Gennari FJ. Hypokalemia. N Engl J Med 1998;339:451–458. [PubMed: 9700180]
27. Hunt SA, Abraham WT, Chin MH, et al. ACC/AHA 2005 Guideline Update for the Diagnosis and Management of Chronic Heart Failure in the Adult. A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Writing Committee to Update the 2001 Guidelines for the Evaluation and Management of Heart Failure). Developed in Collaboration

With the American College of Chest Physicians and the International Society for Heart and Lung Transplantation. Endorsed by the Heart Rhythm Society. Circulation 2005;112:e154–e235. [PubMed: 16160202]

28. Macdonald JE, Struthers AD. What is the optimal serum potassium level in cardiovascular patients? J Am Coll Cardiol 2004;43:155–161. [PubMed: 14736430]
29. Petch MC, McKay R, Bethune DW. The effect of beta, adrenergic blockade on serum potassium and glucose levels during open heart surgery. Eur Heart J 1981;2:123–126. [PubMed: 6115752]
30. Ahuja TS, Freeman D Jr, Mahnken JD, Agraharkar M, Siddiqui M, Memon A. Predictors of the development of hyperkalemia in patients using angiotensin-converting enzyme inhibitors. Am J Nephrol 2000;20:268–272. [PubMed: 10970978]

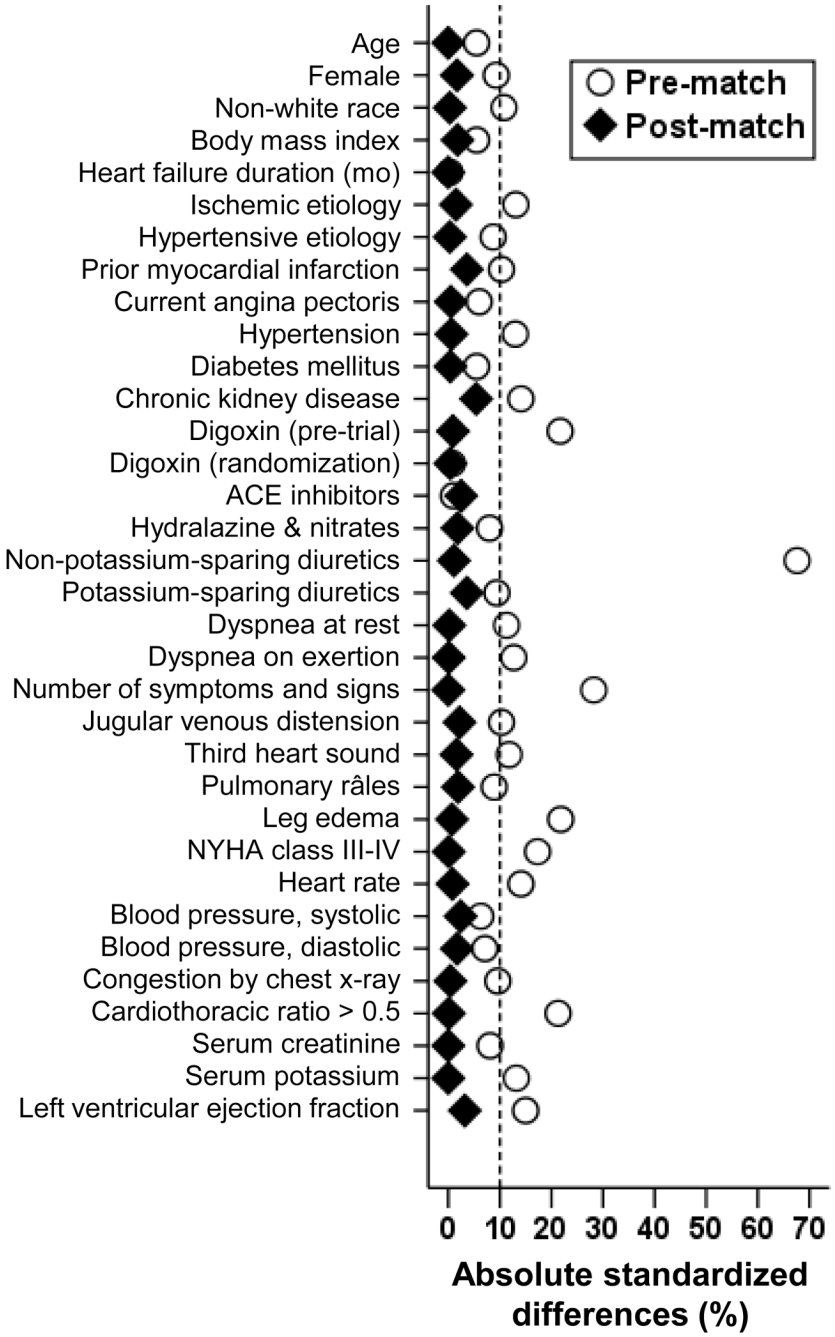


Fig. 1.

Love plots for absolute standardized differences in covariates between patients receiving and not receiving potassium supplements, before and after propensity score matching. (ACE=angiotensin-converting enzyme; NYHA=New York Heart Association)

Figure 2a

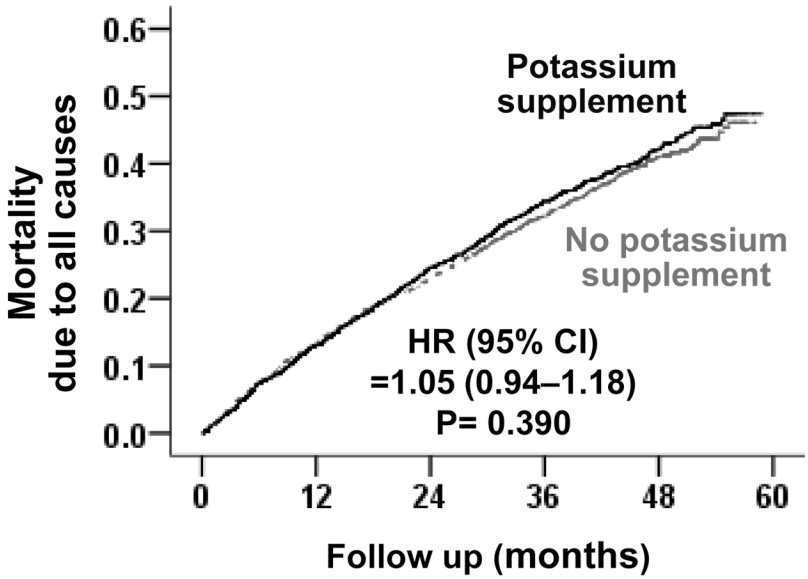


Figure 2b

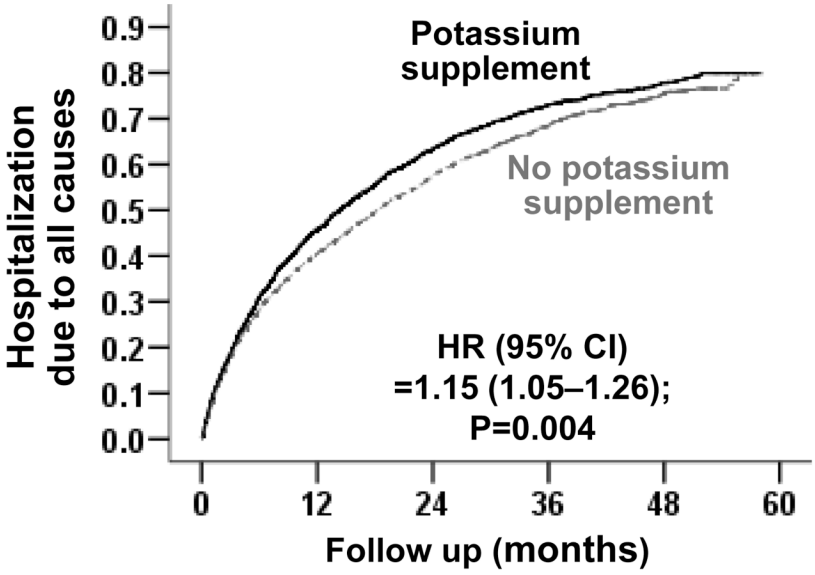


Fig. 2.
Kaplan-Meier plots for all-cause mortality and all-cause hospitalization by the use of oral potassium supplements.

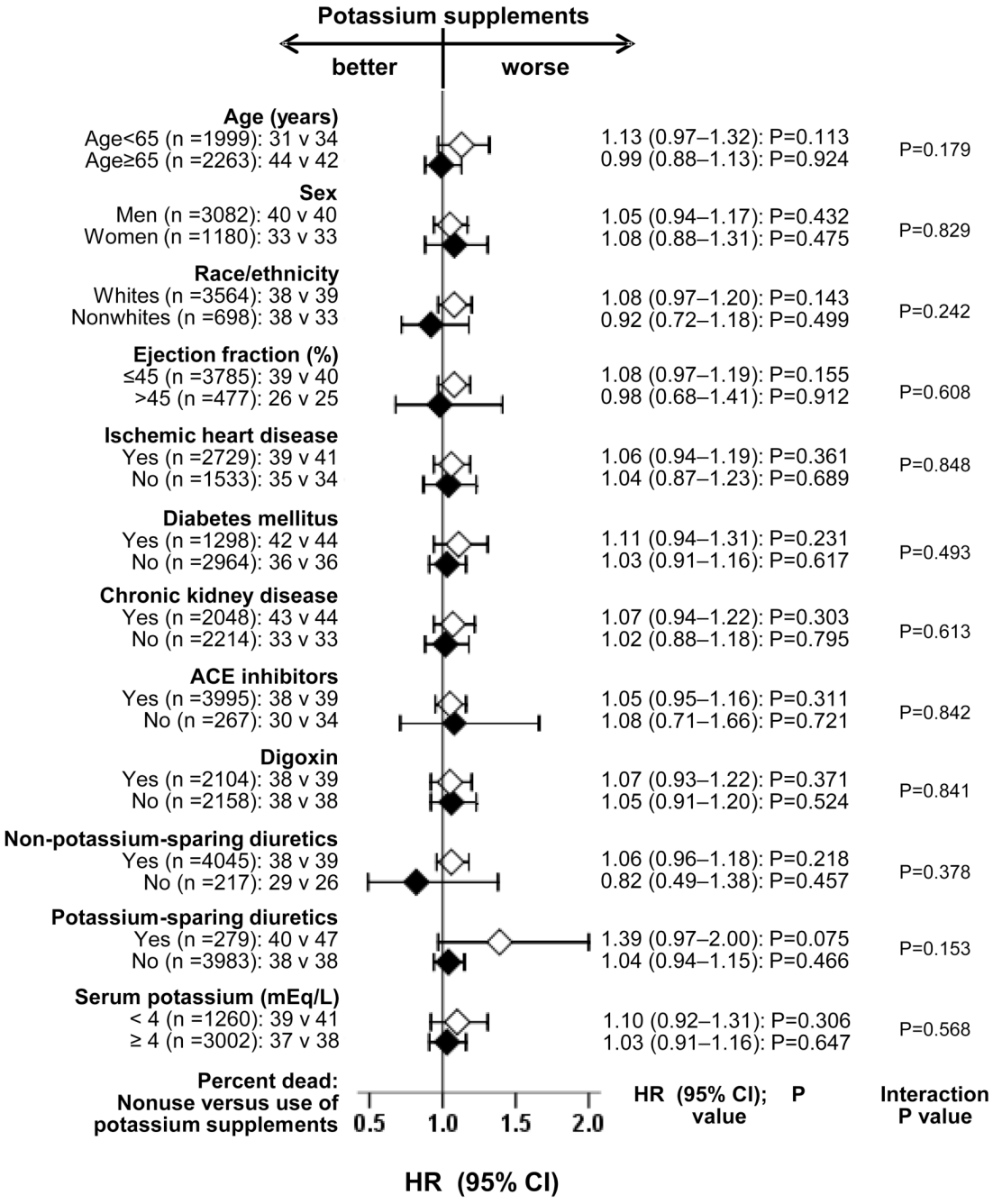


Fig. 3.

Hazard ratios (HR) and 95% confidence intervals (CI) for all-cause mortality associated with potassium supplement use in subgroups of patients with chronic heart failure. (ACE=angiotensin-converting enzyme)

Int J Cardiol. Author manuscript; available in PMC 2011 May 28.

Table 1

Baseline patient characteristics

Variables	Before matching*			After matching		
	No potassium supplement use (N = 2131)	Potassium supplement use (N = 2131)	P values	No potassium supplement use (n = 2131)	Potassium supplement use (n = 2131)	P values
Age,(yrs)	63.6 (±11)	64.2 (±11)	0.086	64.2 (±11)	64.2 (±11)	0.858
Female	496 (23%)	582 (27%)	0.002	598 (28%)	582 (27%)	0.605
Non-white	269 (13%)	350 (16%)	<0.0001	348 (16%)	350 (16%)	0.967
Body mass index (kg/m^2)	27.3 (±5.3)	27.6 (±5.7)	0.088	27.5 (±5.6)	27.6 (±5.7)	0.567
Duration of HF (mo)	30 (±39)	30 (±36)	0.976	30 (±36)	30 (±36)	0.975
Primary cause heart failure						
Ischemic	1503 (71%)	1372 (64%)	<0.0001	1357 (64%)	1372 (64%)	0.646
Hypertensive	203 (10%)	261 (12%)	0.004	259 (12%)	261 (12%)	0.963
Idiopathic	281 (13%)	338 (16%)	0.013	346 (16%)	338 (16%)	0.771
Others	144 (7%)	160 (8%)	0.341	169 (8%)	160 (8%)	0.649
Prior myocardial infarction	1361 (64%)	1254 (59%)	0.001	1216 (57%)	1254 (59%)	0.237
Current angina	598 (28%)	541 (25%)	0.048	536 (25%)	541 (25%)	0.886
Hypertension	964 (45%)	1102 (52%)	<0.0001	1108 (52%)	1102 (52%)	0.876
Diabetes	594 (28%)	647 (30%)	0.074	651 (31%)	647 (30%)	0.921
Chronic kidney disease	904 (42%)	1053 (49%)	<0.0001	995 (47%)	1053 (49%)	0.082
Medications						
Digoxin (pretrial use)	839 (39%)	1068 (50%)	<0.0001	1078 (51%)	1068 (50%)	0.769
Digoxin (trial use)	1063 (50%)	1054 (50%)	0.783	1050 (49%)	1054 (50%)	0.927
ACE inhibitors	1999 (94%)	2004 (94%)	0.749	1991 (93%)	2004 (94%)	0.436
Hydralazine & nitrates	15 (1%)	33 (2%)	0.009	38 (2%)	33 (2%)	0.583
Diuretics	1512 (71%)	2025 (95%)	<0.0001	2020 (95%)	2025 (95%)	0.332
PS diuretics	182 (9%)	130 (6%)	0.002	149 (7%)	130 (6%)	0.252
Symptoms and signs of heart failure						
Dyspnea at rest	440 (21%)	541 (25%)	<0.0001	543 (26%)	541 (25%)	0.972

Variables	Before matching*			After matching		
	No potassium supplement use (N = 2131)	Potassium supplement use (N = 2131)	P values	No potassium supplement use (n = 2131)	Potassium supplement use (n = 2131)	P values
Dyspnea on exertion	1572 (74%)	1687 (79%)	<0.0001	1688 (79%)	1687 (79%)	1.000
Jugular venous distension	253 (12%)	328 (15%)	0.001	345 (16%)	328 (15%)	0.499
Third heart sound	456 (21%)	563 (26%)	<0.0001	578 (27%)	563 (26%)	0.631
Pulmonary râles	312 (15%)	382 (18%)	0.004	398 (19%)	382 (18%)	0.551
Lower extremity edema	374 (18%)	565 (27%)	<0.0001	558 (26%)	565 (27%)	0.815
NYHA class III–IV	608 (29%)	780 (37%)	<0.0001	779 (37%)	780 (37%)	1.000
Heart rate (beats per minute),	78 (±13)	80 (±13)	<0.0001	80 (±12)	80 (±13)	0.836
Systolic BP (mmHg)	128 (±21)	127 (±21)	0.034	127 (±21)	127 (±21)	0.416
Diastolic BP (mmHg)	76 (±11)	75 (±11)	0.033	75 (±12)	75 (±11)	0.694
Chest radiograph findings						
Pulmonary congestion	281 (13%)	354 (17%)	0.002	357 (17%)	354 (17%)	0.934
Cardiothoracic ratio >0.5	1191 (56%)	1411 (66%)	<0.0001	1410 (66%)	1411 (66%)	1.000
Serum potassium (mg/dL)	4.4 (±0.5)	4.3 (±0.5)	<0.0001	4.3 (±0.5)	4.3 (±0.5)	0.554
Serum creatinine (mg/dL)	1.3 (±0.4)	1.3 (±0.4)	0.002	1.3 (±0.4)	1.3 (±0.4)	0.696
Ejection fraction (%)	33 (±12)	31 (±13)	<0.0001	30 (±12)	31 (±13)	0.370

ACE=angiotensin converting enzyme; BP=blood pressure; HF=heart failure.

*Of the 5589 patients not receiving potassium supplements, a random sample of 2131 patients were selected and paired with 2131 matched patients receiving potassium supplements. This was done to have similar pre- and post-match sample sizes, and to avoid overestimation of significant p values from a larger sample size.

Int J Cardiol. Author manuscript; available in PMC 2011 May 28.

Table 2

Mortality by potassium supplement use

Outcomes	Rate / 10000 person-years follow up (total number of events / follow-up in years)		Rate difference (per 10,000 person-years)*	Hazard ratio (95% CI)†	P value
	No potassium supplement use (N = 2131)	Potassium supplement use (N = 2131)			
All-cause	1313 (802 / 6109)	1327 (818 / 5928)	+ 67	1.05 (0.94–1.18)	0.390
Cardiovascular	1044 (638 / 6109)	1100 (652 / 5928)	+ 56	1.07 (0.95–1.22)	0.275
Worsening heart failure‡	476 (291 / 6109)	533 (316 / 5928)	+ 57	1.18 (0.98–1.42)	0.081
Other cardiac§	521 (318 / 6109)	516 (306 / 5928)	− 5	0.98 (0.82–1.17)	0.784
Other vascular¶	47 (29 / 6109)	51 (30 / 5928)	+4	1.15 (0.63–2.09)	0.648
Non-cardiac, non-vascular	196 (120 / 6109)	214 (127 / 5928)	+ 18	1.08 (0.81–1.43)	0.611
Unknown	72 (44 / 6109)	66 (39 / 5928)	− 6	0.69 (0.42–1.16)	0.161

*Absolute rate differences were calculated by subtracting the rates of death in the potassium-supplement group from the rates of death in the no potassium-supplement group (before values were rounded).

†Hazard ratios and confidence intervals (CI) were estimated from matched Cox proportional-hazards models.

‡This category includes patients who died from worsening heart failure, even if the final event was an arrhythmia.

§This category includes deaths presumed to result from arrhythmia without evidence of worsening heart failure and deaths due to atherosclerotic coronary disease, bradyarrhythmias, low-output states, and cardiac surgery.

¶This category includes deaths due to stroke, embolism, peripheral vascular disease, vascular surgery, and carotid endarterectomy.

Int J Cardiol. Author manuscript; available in PMC 2011 May 28.

Table 3

Hospitalization by potassium supplement use

Cause for hospitalization*	Rate / 10000 person-years follow up (total number of events / follow-up in years)		Rate difference (/ 10,000 person-years)†	Matched hazard ratio (95% CI)‡	P value
	No potassium Supplement use (N = 2131)	Potassium supplement use (N = 2131)			
All-cause	4120 (1445 / 3507)	4777 (1516 / 3173)	+ 657	1.15 (1.05–1.26)	0.004
Cardiovascular	2690 (1127 / 4189)	3284 (1235 / 3761)	+ 594	1.19 (1.08–1.32)	0.001
Worsening heart failure	1318 (676 / 5129)	1717 (807 / 4701)	+ 399	1.27 (1.12–1.43)	<0.0001
Ventricular arrhythmia, cardiac arrest	147 (88 / 5990)	132 (77 / 5812)	− 15	1.07 (0.75–1.53)	0.715
SV arrhythmias§	170 (101 / 5931)	195 (112 / 5745)	+ 25	1.06 (0.78–1.44)	0.697
AV block, bradyarrhythmia	15 (9 / 6098)	19 (11 / 5906)	+ 4	1.14 (0.41–3.15)	0.796
Suspected digoxin toxicity	58 (35 / 6052)	53 (31 / 5887)	− 5	0.92 (0.53–1.61)	0.923
Myocardial infarction	215 (128 / 5947)	191 (111 / 5807)	− 24	0.93 (0.69–1.24)	0.603
Unstable angina	417 (236 / 5661)	483 (262 / 5429)	+ 66	1.16 (0.95–1.41)	0.157
Stroke	155 (93 / 5985)	215 (124 / 5755)	+ 60	1.26 (0.93–1.71)	0.140
Coronary revascularization¶	65 (39 / 6034)	70 (41 / 5846)	+ 5	1.07 (0.64–1.79)	0.793
Cardiac transplantation	18 (11 / 6093)	31 (18 / 5896)	+ 13	2.60 (0.93–7.29)	0.069
Other cardiovascular**	468 (265 / 5665)	521 (283 / 5427)	+ 53	1.09 (0.90–1.32)	0.400
Respiratory infection	290 (170 / 5860)	320 (181 / 5665)	+ 30	1.18 (0.92–1.51)	0.189
Other non-cardiovascular	1509 (743 / 4924)	1547 (730 / 4720)	+ 38	1.05 (0.93–1.19)	0.414

Cause for hospitalization*	Rate / 10000 person-years follow up (total number of events / follow-up in years)		Rate difference (/ 10,000 person-years)†	Matched hazard ratio (95% CI)‡	P value
	No potassium Supplement use (N = 2131)	Potassium supplement use (N = 2131)			
Unspecified	20 (12 / 6095)	27 (16 / 5910)	+ 7	1.83 (0.68–4.96)	0.232
Number of hospitalizations	9714 (986 / 1015)	11893 (1093 / 919)	+ 2179		

*Data shown include the first hospitalization of each patient due to each cause.

†Absolute differences were calculated by subtracting the percentage of patients hospitalized in the potassium supplement group from the percentage of patients hospitalized in the no potassium supplement group (before values were rounded).

‡Hazard ratios and confidence intervals (CI) were estimated from a Cox proportional-hazards models that used the first hospitalization of each patient for each reason.

§Supraventricular (SV) arrhythmias include Atrioventricular (AV) block and bradyarrhythmias

¶This category includes coronary-artery bypass grafting and percutaneous transluminal coronary angioplasty

**This category includes embolism, venous thrombosis, peripheral vascular disease, hypertension, other vascular surgery, cardiac catheterization, other types of catheterization, pacemaker implantation, installation of automatic implantable cardiac defibrillator, electrophysiologic testing, transplant-related evaluation, nonspecific chest pain, atherosclerotic heart disease, hypotension, orthostatic hypotension, and valve operation

ELSEVIER

Journal of Adolescent Health ■ (2010) 1–7

JOURNAL OF ADOLESCENT HEALTH

Original Article

The Causal Impact of Childhood-Limited Maltreatment and Adolescent Maltreatment on Early Adult Adjustment

Terence P. Thornberry, Ph.D[a,*], Kimberly L. Henry, Ph.D[b], Timothy O. Ireland, Ph.D[c], and Carolyn A. Smith, Ph.D[d]

[a]*Department of Criminology and Criminal Justice, University of Maryland, College Park, Maryland*
[b]*Department of Psychology, Colorado State University, Fort Collins, Colorado*
[c]*Department of Criminology and Criminal Justice, Niagara University, Niagara University, New York*
[d]*School of Social Welfare, University at Albany, State University of New York, New York*

Manuscript received May 1, 2009; manuscript accepted September 30, 2009

Abstract

Purpose: We use full-matching propensity score models to test whether developmentally specific measures of maltreatment, in particular childhood-limited maltreatment versus adolescent maltreatment, are causally related to involvement in crime, substance use, health-risking sex behaviors, and internalizing problems during early adulthood.

Methods: Our design includes 907 participants (72% male) in the Rochester Youth Development Study, a community sample followed from age 14 to age 31 with 14 assessments, including complete maltreatment histories from Child Protective Services records.

Results: After balancing the data sets, childhood-limited maltreatment is significantly related to drug use, problem drug use, depressive symptoms, and suicidal thoughts. Maltreatment during adolescence has a significant effect on a broader range of outcomes: official arrest or incarceration, self-reported criminal offending, violent crime, alcohol use, problem alcohol use, drug use, problem drug use, risky sex behaviors, self-reported sexually transmitted disease diagnosis, and suicidal thoughts.

Conclusions: The causal effect of childhood-limited maltreatment is focused on internalizing problems, whereas adolescent maltreatment has a stronger and more pervasive effect on later adjustment. Increased vigilance by mandated reporters, especially for adolescent victims of maltreatment, along with provision of appropriate services, may prevent a wide range of subsequent adjustment problems.

Keywords: Child maltreatment; Adolescent maltreatment; Crime; Substance use; Health-risking sex behaviors; Suicidal thoughts

Since Kempe et al.'s article identified the ''battered child syndrome'' as a unique threat to the well-being of a child [1], maltreatment has been viewed as a significant health problem in American society. Attention has focused on the potential negative consequences of exposure to maltreatment, which is a risk factor for several health-compromising outcomes [2–6]. Nevertheless, the extent to which maltreatment actually *causes* such outcomes remains largely unknown. This is unfortunate because a risk factor is simply an antecedent characteristic associated with an increase in the likelihood of an outcome, whereas a cause, when altered, *changes* the likelihood of the outcome occurring. Thus, if maltreatment actually causes health-related problems, then preventing it, or changing the mechanisms through which it operates, will to some extent reduce those outcomes.

The strongest design for assessing causality, a true experiment, is obviously unethical and illegal for this topic. The best approach currently available to assess causality in nonexperimental designs is propensity score matching [7]. All sample members have some underlying propensity to be exposed to the ''treatment,'' which in this case is to be

Disclaimer: Points of view or opinions in this document are those of the authors and do not necessarily represent the official position or policies of the funding agencies.

*Address correspondence to: Terence P. Thornberry, Ph.D., Department of Criminology and Criminal Justice, University of Maryland, 2220 LeFrak Hall, College Park, MD 20742.

E-mail address: tthornberry@crim.umd.edu

1054-139X/10/$ – see front matter
doi:10.1016/j.jadohealth.2009.09.011

maltreated. But, for youth with similar propensities, only some actually experience maltreatment. By comparing them to nonmaltreated subjects with similar propensities, selection effects are greatly reduced and the rigor of experimental design more closely approximated. Causal inferences can be made with greater confidence because propensity score matching ensures that there is a balance on all observed covariates before estimation of treatment effects, thereby more thoroughly and appropriately controlling for pretreatment selection bias. Our first purpose in this article is to use propensity score matching to examine whether maltreatment causes subsequent problems in four domains of adjustment during the early adult years: involvement in crime and violence, substance use, health-risking sex behaviors, and internalizing problems.

We also examine the developmental specificity of the maltreatment effect. Ontogenetic developmental models [8] hypothesize that maltreatment occurring in childhood ''…should have stronger and more enduring negative effects on future adaptation than later exposure'' because it disrupts the early course of human development [4, page 285], see also [2,9]. In contrast, sociogenetic developmental models [8] hypothesize that maltreatment that occurs in adolescence is likely to be more damaging because of the person's increased autonomy, cognitive ability, and heightened reaction to stress [10,11], as well as the proximity of the maltreatment to the outcomes. Our second purpose, therefore, is to consider whether the causal effect of childhood-limited maltreatment differs from that of maltreatment that occurs during adolescence.

Prior Studies

Longitudinal studies have shown that experiencing maltreatment at some point between birth and 18 years of age is a significant risk factor for crime and violence [5,12], alcohol and drug use [3,13], risky sex behaviors and early pregnancy [6,14], and depression and suicidality [4,15]. Few studies, however, have investigated differences by the developmental stage at which the maltreatment occurred and most of those only examine antisocial behavior. Three studies only measured maltreatment occurring before age 12. They found that maltreatment at younger ages is related to externalizing problems [16], that maltreatment at older ages has a stronger effect on externalizing problems, whereas maltreatment at younger ages is related to depression and anxiety [9]; and that there are no age differences [17]. These studies did not extend the measurement of maltreatment into adolescence, however, and could not examine broader developmental differences. When this is done, we generally see that youth with no substantiated maltreatment and those with childhood-limited maltreatment are not statistically different with respect to externalizing problems, delinquency, and violence [18–21]. In contrast, youth maltreated in adolescence have significantly higher rates of those behaviors than those who were never maltreated. Maltreatment in adolescence is also more consistently and strongly related to early pregnancy, internalizing problems, drug use [19], and smoking [22] than is childhood-limited maltreatment.

Risk-factor studies typically control for several confounding variables to ward off spurious conclusions, but such designs fall far short of meeting the criteria for establishing causality. Three recent studies, however, suggest that maltreatment may be a cause of subsequent problems. Using a genetically-informed design, one study [23] found that childhood maltreatment increases antisocial behavior even after genetic influences are controlled, but this study did not rule out a spurious relationship because of prior environmental influences. Two additional studies used propensity score modeling and found a significant effect of maltreatment on criminal behavior [24] and on health-related quality of life [25]. These studies, however, relied on long-term retrospective measures of maltreatment and did not examine the timing of maltreatment. The present study extends this research by using prospective data and developmentally-specific measures of maltreatment, and by examining multiple outcomes during the early adult years when temporal order is firmly established.

Methods

Sample

We used data from the Rochester Youth Development Study begun in 1988 with 1,000 seventh- and eighth graders. Males and students from neighborhoods having high arrest rates were oversampled because they are at greater risk for problem behaviors. Because gender and arrest rates were used to formulate the probability of selection, they are predictors in all models.

Since 1988, sample members and an adult caregiver have been interviewed (14 and 11 times respectively), and data from school, police, and Child Protective Services records have been collected. Here, we rely on data from the first 12 interviews. Interview waves 1-9 were conducted at 6-month intervals (ages 14-18 years) and waves 10-12 at annual intervals (ages 21-23 years). At wave 1, the average age was 13.9 (SD = .78) and at wave 14 it was 22.7 (SD = .81). The sample had 73% males and 27% females; 68% were African American, 17% Hispanic, and 15% white. At age 23, retention was 85% for the focal subjects and 83% for the caregivers, with no evidence of differential subject loss. Because of missing data on the covariates in the propensity score model, the final sample size was 907. All study procedures were approved by the Institutional Review Board at the University at Albany. The study was explained and written informed consent obtained from adult participants (for themselves and their minor children); assent was obtained from minor children.

Measurement

Maltreatment. The maltreatment measure is based on all incidents of substantiated maltreatment (physical abuse, sex abuse, and neglect) from a county-wide search of Child

Protective Services records through 1992 when participants were completing high school. Most incidents involve multiple types of abuse [19], with the most common being neglect, followed by physical abuse, and sex abuse. On the basis of developmental theory and prior research with this sample [19,26], we examined three groups. *Never maltreated* denotes subjects who never had a substantiated incident of maltreatment from birth through age 17 (n = 731, 80.6%). *Childhood-limited maltreatment* refers to participants who had at least one substantiated incident from birth through age 11, but none after that (n = 104; 11.5%). The final category combines participants who had at least one substantiated incident between ages 12 and 17 with those who had an incident in both childhood and in adolescence (n = 72; 7.9%). Because all of these participants were maltreated in adolescence and the majority (61%) was maltreated only during adolescence, we refer to this group as *any adolescent maltreatment*. Sample size and the requirements of matching at different propensity score levels preclude analyzing these groups separately. Combining them is based on previous analyses [19] that found that these groups had very similar relationships to outcomes and that neither group drove the observed relationships. Nevertheless, the reader should bear in mind that 39% of the any adolescent maltreatment group was maltreated both in childhood and in adolescence.

Antisocial behavior. At the annual interviews of sample members aged 21–23 years, we asked respondents to report their criminal offending during the previous year. Two self-report indices were used: *general offending* (26 offenses ranging in seriousness from minor offenses like petty theft to serious offenses like robbery) and *violent crime* (a 6-item sub-index of violent offenses). Cumulative frequency measures were used. A dichotomous measure of *official arrest or incarceration* is based on a statewide search of New York State records covering the same ages.

Substance use. *Alcohol use* is a 3-item index of the cumulative frequency of drinking beer, wine, and liquor during this 3-year period. The *drug use* inventory covers 10 illicit drugs from marijuana to heroin. Two measures, *problem alcohol use* and *problem drug use*, are 11-item inventories of the number of problems associated with the use of these substances modeled after the diagnostic content of the Diagnostic and Statistical Manual for Mental Disorders (DSM-IV).

Health-risking sex behaviors. *Risky sex* is a 14-item inventory asking about the respondent's risky sex behavior, such as whether they ever had sex in exchange for money, been forced to have sex, had partners who were bisexual, or had HIV-positive partners. The second measure, *STD diagnosis*, is a single item that asks whether the respondent had ever been diagnosed with an STD.

Internalizing problems. *Depressive symptoms* is a 14-item scale derived from the Center for Epidemiologic Studies Depression Scale (CES-D) [27] assessing the frequency of symptoms such as feeling nervous, stressed, or lonely. *Suicidal thoughts* is a single item asking how often during the past year the respondent "thought seriously about suicide." Categories for depressive symptoms and suicidal thoughts ranged from one (never) to four (often).

Risk Factors. The propensity score model uses 19 variables that reflect risk factors for child maltreatment [28]. These include child's age at baseline, gender, race/ethnicity, mother's age at first birth, neighborhood arrest rate, neighborhood proportion of families living in poverty, family structure, and family socioeconomic status; parent measures, including their education, alcohol use, drug use, depressive symptoms, level of stress, incidence of stressful life events, social support, and harsh parenting; as well as family history of maltreatment, of substance use, and/or of mental health problems.

Analysis

Two separate propensity score models were estimated by regressing each of the developmentally specific maltreatment indicators on the covariates described earlier. These models estimate the log odds that an individual would be maltreated and the resultant score is the propensity score. The sample of nonmaltreated individuals was restricted to those in the region of common support of the maltreated individuals (i.e., nonmaltreated youth with a propensity score more than .25 of a standard deviation outside the range of maltreated youth were excluded), a technique recommended when the average causal effect among the treated, which estimates the predicted difference between the observed outcomes for maltreated youth (i.e., the observed young adult outcomes for each maltreated youth) and the outcomes that would have been observed if each maltreated youth was not maltreated (i.e., if the maltreatment had been prevented), is desired [29]. Restriction of the data to the range of common support reduced the sample size for the child match from 835 to 749 (with 645 control cases and 104 maltreated cases) and for the adolescent match from 803 to 674 (with 602 control cases and 72 maltreated cases). Next, a full matching approach [30] to casual inference, which makes use of all available data by matching treated individuals with as many similar controls as possible, weighting each individual by the number of individuals in the set, was used [29].

We first used t-tests and chi-square tests to assess the balancing of covariates across groups; no significant differences exist on any of the covariates between the groups in the matched datasets (childhood match, $p = .22$–1.00; adolescent match, $p = .48$–1.00). The standardized bias [30] should be less than .25 [29]; across all matches, the largest was .17 for the childhood match and .09 for the adolescent match and most (84% childhood; 72% adolescent) were .05 or less. The standardized bias for the propensity score was drastically reduced in both models, to −.01 for childhood-limited maltreatment and to .01 for adolescent

maltreatment. For the childhood match, the full matching procedure created 99 subclasses, ranging in size from two (one maltreated and one nonmaltreated youth) to 61 (one maltreated and 60 nonmaltreated youth). For the adolescent match, it created 71 subclasses, ranging from two (one maltreated and one nonmaltreated youth) to 88 (one maltreated and 87 nonmaltreated youth). After matched datasets were obtained, we used either negative binomial (for count outcomes with over-dispersion), logistic, or ordinary least squares regression models, depending on the measurement of the outcome variable. All regressions were weighted to account for the full matching subclasses. In addition to the maltreatment predictor, age at baseline, gender, race/ethnicity, family structure, neighborhood arrest rate, neighborhood poverty, and family socioeconomic status were included in the regression models to adjust for any residual bias and to increase precision [31]. To account for missing data in the outcome variables, we created 10 multiply-imputed datasets using a multiple imputation program that allows for categorical and count variables [32]. Missing data across the outcomes ranged from 3.6% to 10.5% for the childhood analyses and from 3.9% to 10.5% for the adolescent analyses. Analyses were run on each of the 10 imputed datasets and the estimates were combined using the procedures outlined by Rubin [33].

Results

The average causal effect of childhood-limited maltreatment is rather moderate (Table 1). There are no significant effects for any of the criminal behaviors or the health-risking sex behaviors. There is, however, evidence of a causal impact of childhood-limited maltreatment on the frequency of drug use and on problem drug use. Childhood-limited maltreatment victims are also significantly more likely to report suicidal thoughts and more depressive symptoms.

In contrast, there is a more consistent effect of adolescent maltreatment on these outcomes (Table 1). Indeed, the only outcome without a significant effect is depressive symptoms. Study participants who were maltreated during adolescence have significantly higher levels of general offending and greater involvement in violent crime and official arrests or incarcerations. They exhibit significantly higher levels of alcohol use, drug use, problem alcohol use, and problem drug use. They engage in more risky sex behaviors and are more likely to report a diagnosis of an STD. Finally, they are significantly more likely to report suicidal thoughts than those who were never maltreated.

To indicate the size of these effects, Tables 2 and 3 present the average predicted score for the maltreated individuals under the observed condition as well as the unobserved counterfactual condition. These results were obtained by using the regression equation for each outcome and estimating the predicted score for each maltreated youth under their observed status and their counterfactual status. Only contrasts for statistically significant relationships are presented.

The effect of childhood-limited maltreatment (Table 2) on drug use and on problem drug use is pronounced. For example, the frequency of drug use in the observed condition is 173 as compared with 84 in the counterfactual condition. The predicted probability of having suicidal thoughts is .24 under the observed status and .15 under the counterfactual status. Likewise, depression scores are, on average, higher under the observed condition.

Experiencing maltreatment either during adolescence or in both childhood and adolescence has a sizeable effect across these outcomes (Table 3). The average annual frequency of general offending under the observed, maltreated condition is 226 as compared with 110 under the counterfactual condition. The effects for drug use are particularly large – maltreated adolescents report rates of drug use and drug problems that are about four times higher than nonmaltreated youth. Almost half of the adolescent maltreatment victims report engaging in risky sex behavior as opposed to one-third under the counterfactual condition, and the prevalence of STD diagnosis and of suicidal thoughts is about twice as high.

Discussion

On the basis of these results it seems, first, that maltreatment is not merely a risk factor for later outcomes, but also a causal agent, and, second, that its effect is conditioned by the developmental stage at which the maltreatment occurs. Childhood-limited maltreatment significantly affects drug use, problem drug use, suicidal thoughts, and depressive symptoms – reactions to stress that are more inwardly directed. In contrast, maltreatment that occurs in adolescence has a more pervasive effect on early adult development, affecting 10 of the 11 outcomes including involvement in criminal behavior, substance use, health-risking sex behaviors, and suicidal thoughts. The consistent effect of any adolescent maltreatment suggests that there may be important recency effects (the adolescence-limited subgroup) and dose-response effects (the persistent maltreatment group) that future research needs to address.

These results also highlight the importance of using developmentally specific measures of maltreatment in assessing its subsequent effect. When a global measure of any maltreatment was used (analysis not shown), only three of the 11 relationships were statistically significant; this would mistakenly imply that, in general, maltreatment is not causally related to early adult outcomes. When developmentally specific measures are used, however, we see the strong, pervasive effect of adolescent maltreatment on these outcomes.

Implications

Regarding childhood-limited maltreatment, prior studies report that childhood maltreatment has negative consequences on child development in the short term [2,17] and the present findings extend that to a causal impact on selected consequences during the early adult years. It is possible that

Table 1
Effect of maltreatment on early adult outcomes (matched samples)

	A. Childhood-limited maltreatment				B. Any adolescent maltreatment			
	Est	SE	exp(Est)	Est/SE	Est	SE	exp(Est)	Est/SE
Crime								
General offending[a]	.04	.25	1.04	.14	.72	.30	2.06	2.41*
Violent crime[a]	−.19	.25	.83	−.77	.60	.26	1.83	2.28*
Official arrest or incarceration[b]	.44	.23	1.55	1.88	.74	.28	2.10	2.63**
Substance use								
Alcohol use[a]	.31	.17	1.37	1.84	.40	.18	1.49	2.18*
Problem alcohol use[a]	.30	.21	1.35	1.41	.85	.27	2.33	3.10**
Drug use[a]	.72	.31	2.05	2.32*	1.36	.35	3.89	3.94**
Problem drug use[a]	.73	.28	2.08	2.60*	1.32	.33	3.73	4.00**
Health-risking sex behaviors								
Risky sex[b]	.07	.28	1.07	.26	.58	.27	1.78	2.10*
STD diagnosis[b]	.06	.39	1.06	.15	.85	.31	2.34	2.76**
Internalizing problems								
Suicidal thoughts[b]	.67	.28	1.95	2.34*	.90	.32	2.46	2.80**
Depressive symptoms[c]	.12	.06		2.14*	.08	.05		1.47

Est = regression coefficient; SE = standard error; exp(Est) = exponentiated regression coefficient.
[a] Negative binomial regression coefficient.
[b] Logistic regression coefficient.
[c] Linear regression coefficient.
* $p < .05$ (2-tailed test).
** $p < .01$ (2-tailed test).

there are fewer negative effects for the childhood-limited group because they received services. Although our data cannot rule out this possibility, the published data on treatment provide very little evidence of typical services being effective in reducing underlying maltreatment risk or in preventing recurrence of maltreatment among those with substantiated maltreatment. These results highlight the importance of developing *effective* services, both to prevent childhood-limited maltreatment and to reduce its negative effect on the outcomes to which it is causally related. On the basis of our results, services for childhood victims of maltreatment should pay particular attention to more inwardly directed reactions. At the same time, the null findings with respect to criminal behavior and health-risking sex behaviors remind us that childhood maltreatment does not necessarily cause any or all subsequent negative outcomes.

Results for adolescent maltreatment, which includes both adolescence-limited maltreatment and maltreatment that begins in childhood and extends into adolescence, greatly extend those of earlier risk factor studies. Given the breadth of its effect on early adult functioning, it is imperative both to identify the mechanisms by which adolescent maltreatment generates those consequences and to understand why adolescent maltreatment differs so substantially from childhood-limited maltreatment in this regard. Adolescents face more adjustment demands from the intense emotional experiences of puberty and complex peer and romantic interactions [34], and they have greater cognitive sophistication that leads to new appraisals of maltreatment that are likely to increase negative emotions such as shame and anger [35]. All of this may heighten oppositional behavior and promote further victimization at home and on the streets [36], leading to long-term adjustment problems.

It is also essential to develop effective *and* developmentally appropriate programs for adolescent victims [37]. But as scholars have noted [10,21], there are fewer treatment programs for adolescent victims than child victims, and many adolescent interventions are either downward extensions of adult programs or upward extensions of child programs [38]. Also, there are numerous barriers to enrolling adolescent victims and their families in any program, partly because maltreated adolescents are more apt to leave home and act out and, as a result, are less likely to be offered or to accept preventive and supportive services [37], and partly because of unintentional biases of mandated reporters who assume that maltreatment at younger ages has more severe consequences and therefore requires more extensive intervention resources. Obviously, childhood victims require and deserve excellent services. Nevertheless, given the substantial causal effect that adolescent maltreatment has on later negative outcomes, redoubling our efforts in this area is warranted.

Table 2
Predicted mean scores for young adult outcomes by childhood-limited maltreatment status

	Observed status (maltreated)	Counterfactual status (not maltreated)
Substance use		
Drug use	173.02	84.22
Problem drug use	.60	.29
Internalizing problems		
Suicidal thoughts[a]	.24	.15
Depressive symptoms	2.02	1.90

[a] Estimate is a predicted probability.

Table 3
Predicted mean scores for young adult outcomes by any adolescent maltreatment status

	Observed status (maltreated)	Counterfactual status (not maltreated)
Crime		
General offending	225.57	109.67
Violent crime	2.10	1.15
Official arrest or incarceration[a]	.54	.39
Substance use		
Alcohol use	208.59	139.54
Problem alcohol use	1.58	.68
Drug use	355.00	91.28
Problem drug use	.90	.24
Health-risking sex behaviors		
Risky sex[a]	.47	.33
STD diagnosis[a]	.28	.15
Internalizing problems		
Suicidal thoughts[a]	.24	.12

[a] Estimate is a predicted probability.

Limitations

Given the uneven gender distribution in the Rochester study (male, 73%; female, 27%), it was impossible to conduct gender-specific analyses. We recognize that there are gender differences across these areas of functioning, and we controlled for gender both in the propensity score and regression analyses; however, this issue cannot be addressed further with these data. It would be helpful to examine these relationships when maltreatment is disaggregated by type and by finer developmental distinctions. We relied entirely on official Child Protective Services records to measure maltreatment. These records have well-known biases [39], but also have substantial validity [40] and enable the construction of developmentally specific measures. Finally, although propensity score models more closely approximate the features of randomized experiments, the lack of random assignment to treatment conditions limits our ability to definitively test a causal hypothesis.

Conclusion

Despite limitations, this study presents compelling evidence about the causal status of developmentally specific measures of maltreatment. Childhood-limited maltreatment affects a somewhat narrower range of early adult outcomes, primarily affecting internalizing problems. Adolescent maltreatment has a much more pervasive influence affecting all four areas of adjustment that we investigated: criminal behavior, substance use, health-risking sex behaviors, and suicidal thoughts. These results underscore the importance of having all mandated reporters, including health care providers, show the same level of vigilance for cases of adolescent maltreatment that have historically been shown for childhood maltreatment. Preventing maltreatment and providing services to reduce its negative sequelae are likely to have major benefits for society given the extensive damage to later functioning that maltreatment seems to cause.

Acknowledgments

Support for the Rochester Youth Development Study has been provided by the Office of Juvenile Justice and Delinquency Prevention (86-JN-CX-0007, 96-MU-FX-0014, 2004-MU-FX-0062; 2006-JW-BX-0074), the National Institute on Drug Abuse (DA005512, DA017810), and the National Science Foundation (SBR-9123299, SES-9123299). Work on this project was also aided by grants to the Center for Social and Demographic Analysis at the University at Albany from NICHD (P30-HD32041) and NSF (SBR-9512290).

References

[1] Kempe CH, Silverman FN, Steele BF, et al. The battered child syndrome. JAMA 1962;181:17–24.
[2] Dodge KA, Bates JE, Pettit GS. Mechanisms in the cycle of violence. Science 1990;250:1678–83.
[3] Ireland TO, Widom CS. Childhood victimization and risk for alcohol and drug arrests. Int J Addict 1994;29:235–74.
[4] Sternberg KJ, Lamb ME, Guterman E, et al. Effects of early and later family violence on children's behavior problems and depression: A longitudinal, multi-informant perspective. Child Abuse Negl 2006; 30:283–306.
[5] Widom CS. The cycle of violence. Science 1989;244:160–6.
[6] Wilson HW, Widom CS. An examination of risky sexual behavior and HIV in victims of child abuse and neglect: A 30-year follow-up. Health Psychol 2008;27:149–58.
[7] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41–55.
[8] Dannefer D. Adult development and social theory: A paradigmatic appraisal. Am Soc Rev 1984;49:100–16.
[9] Kaplow J, Widom CS. Age of onset of child maltreatment predicts long-term mental health outcomes. J Abnorm Psychol 2007;116:176–87.
[10] Garbarino J. Troubled youth, troubled families: The dynamics of adolescent maltreatment. In: Cicchetti D, Carlson V, eds. Child Maltreatment: Theory and Research on the Causes and Consequences of Child Abuse and Neglect. New York, NY: Cambridge University Press, 1989:685–706.
[11] Larson R, Hamm M. Stress and 'storm and stress' in early adolescence: The relationship of negative events with dysphoric affect. Dev Psychol 1993;29:130–41.
[12] English DJ, Widom CS, Brandford C. Childhood victimization and delinquency, adult criminality, and violent criminal behavior. Final report presented to the National Institute of Justice under grant 97-IJ-CX-0017; 2001.
[13] Widom CS, Ireland TO, Glynn PJ. Alcohol abuse in abused and neglected children followed-up: Are they at increased risk? J Stud Alcohol 1995;56:207–17.
[14] Herrenkohl EC, Herrenkohl RC, Egolf BP, et al. The relationship between early maltreatment and teenage parenthood. J Adol 1998;21: 291–303.
[15] Brown J, Cohen P, Johnson JG, et al. Childhood abuse and neglect: Specificity of effects on adolescent and young adult depression and suicidality. J Am Acad Child Adolesc Psychiatry 1999;38:1490–6.
[16] Keiley MK, Howe TR, Dodge KA, et al. The timing of child physical maltreatment: A cross-domain growth analysis of impact on adolescent externalizing and internalizing problems. Dev Psychopathol 2001;13: 891–912.

[17] English DJ, Graham JC, Litrownik AJ, et al. Defining maltreatment chronicity: Are there differences in child outcomes? Child Abuse Negl 2005;29:575–95.
[18] Eckenrode J, Sielinski D, Smith E, et al. Child maltreatment and the early onset of problem behaviors: Can a program of nurse home visitation break the link? Dev Psychopathol 2001;13:873–90.
[19] Thornberry TP, Ireland TO, Smith CA. The importance of timing: The varying impact of childhood and adolescent maltreatment on multiple problem outcomes. Dev Psychopathol 2001;13:957–79.
[20] Jonson-Reid M, Barth R. From maltreatment report to juvenile incarceration: The role of child welfare services. Child Abuse Negl 2000; 24:505–20.
[21] Stewart A, Livingston M, Dennison S. Transitions and turning points: Examining the links between child maltreatment and juvenile offending. Child Abuse Negl 2008;32:51–66.
[22] Jun HJ, Rich-Edwards W, Boynton-Jarrett R, et al. Child abuse and smoking among young women: The importance of severity, accumulation, and timing. J Adolesc Health 2008;43:55–63.
[23] Jaffee SR, Caspi A, Moffitt TE, et al. Physical maltreatment victim to antisocial child: Evidence of an environmentally-mediated process. J Abnorm Psychol 2004;113:34–55.
[24] Currie J, Tekin E. Does Child Abuse Cause Crime? Andrew Young School of Policy Studies. Research Paper Series. Working Paper 06–31. Atlanta, GA: Georgia State University; 2006.
[25] Corso PS, Edwards VJ, Fang X, et al. Health-related quality of life among adults who experienced maltreatment during childhood. Am J Public Health 2008;98:1094–100.
[26] Ireland TO, Smith CA, Thornberry TP. Developmental issues in the impact of child maltreatment on later delinquency and drug use. Criminology 2002;40:359–99.
[27] Radloff LS. The CES-D scale: A self-report depression scale for research in the general population. Appl Psychol Meas 1977;1: 385–401.
[28] Tolan PH, Gorman-Smith D, Henry D. Family violence. Annu Rev Psychol 2006;57:557–83.
[29] Ho DE, Imai K, King G, et al. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Politic Analysis 2007;15:199–236.
[30] Stuart EA, Green KM. Using full matching to estimate causal effects in non-experimental studies: Examining the relationship between adolescent marijuana use and adult outcomes. Dev Psychol 2008;44:395–406.
[31] Stuart EA, Rubin DB. Best practices in quasi-experimental designs: Matching methods for causal inference. In: Osborne J, ed. Best Practices in Quantitative Social Science. Thousand Oaks, CA: Sage Publications, 2008:155–76.
[32] Raghunathan TE, Lepkowski J, VanHoewyk J, et al. A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology 2001;27:85–95.
[33] Rubin DL. Multiple Imputation for Nonresponse in Surveys. Hoboken, NJ: John Wiley & Sons, 1987.
[34] Lee V, Hoaken MS. Cognition, emotion, and neurobiological development: Mediating the relation between maltreatment and aggression. Child Maltreat 2008;12:281–98.
[35] Feiring C, Miller-Johnson S, Cleland CM. Potential pathways from stigmatization and internalizing symptoms to delinquency in sexually abused youth. Child Maltreat 2007;12:220–32.
[36] Tyler KA, Johnson KA. A longitudinal study of the effects of early abuse on later victimization among high-risk adolescents. Violence Vict 2006;21:287–306.
[37] Garbarino J, Eckenrode J, Powers JL. The maltreatment of youth. In: Garbarino J, Eckenrode J, eds. Understanding Abusive Families. San Francisco, CA: Jossey-Bass Publishers, 1997:145–65.
[38] Weisz JR, Hawley KM. Developmental factors in the treatment of adolescents. J Consult Clin Psychol 2002;70:21–43.
[39] Smith CA, Ireland TO, Thornberry TP, et al. Childhood maltreatment and antisocial behavior: Comparison of self-reported and substantiated maltreatment. Am J Orthopsychiatry 2008;78:173–86.
[40] Widom CS, Raphael KG, DuMont KA. The case for prospective longitudinal studies in child maltreatment research: Commentary on Dube, Williamson, Felitti, & and Anda. Child Abuse Negl 2004;28:715–22.

tein synthesis. Although a small number of proteins appeared toxic (Fig. 3, inset), the vast majority had only a limited effect on cell growth. Overexpression levels do not correlate with—and hence cannot be predicted by—obvious sequence characteristics such as codon usage, protein size, hydrophobicity, and number of transmembrane helices (table S2). The C-terminal His_8 tag and the tobacco etch virus (TEV) protease site present in the GFP fusions (*8*) make it possible to use an efficient, standardized purification protocol for the whole clone collection; yields of purified fusion protein are typically ≥1 mg per liter of culture (*25*). This sets a lower limit for what can be expected for individual proteins expressed, for example, without a GFP tag or using other expression vectors and growth conditions (*26*).

In conclusion, by analyzing a library of *E. coli* inner membrane proteins fused to PhoA and GFP, we have derived an experimentally based set of topology models for the membrane proteome and provide a large-scale data set on membrane protein overexpression. Our results provide an important basis for future functional studies of membrane proteomes and will facilitate the identification of well-expressed targets for structural genomics projects.

References and Notes

1. A. Krogh, B. Larsson, G. von Heijne, E. Sonnhammer, *J. Mol. Biol.* **305**, 567 (2001).
2. S. H. White, *Protein Sci.* **13**, 1948 (2004).
3. K. Melén, A. Krogh, G. von Heijne, *J. Mol. Biol.* **327**, 735 (2003).
4. C. Manoil, J. Beckwith, *Science* **233**, 1403 (1986).
5. B. J. Feilmeier, G. Iseminger, D. Schroeder, H. Webber, G. J. Phillips, *J. Bacteriol.* **182**, 4068 (2000).
6. D. Drew *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 2690 (2002).
7. M. Rapp *et al.*, *Protein Sci* **13**, 937 (2004).
8. Materials and methods are available as supporting material on *Science* Online.
9. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **215**, 403 (1990).
10. J. Wu, L. S. Tisa, B. P. Rosen, *J. Biol. Chem.* **267**, 12570 (1992).
11. A. Kihara, Y. Akiyama, K. Ito, *EMBO J.* **18**, 2970 (1999).
12. A. Sääf, M. Johansson, E. Wallin, G. von Heijne, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 8540 (1999).
13. I. T. Paulsen *et al.*, *Mol. Microbiol.* **19**, 1167 (1996).
14. K. Nishino, A. Yamaguchi, *J. Bacteriol.* **183**, 5803 (2001).
15. I. Ubarretxena-Belandia, C. G. Tate, *FEBS Lett.* **564**, 234 (2004).
16. I. Ubarretxena-Belandia, J. M. Baldwin, S. Schuldiner, C. G. Tate, *EMBO J.* **22**, 6175 (2003).
17. C. Ma, G. Chang, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 2852 (2004).
18. G. von Heijne, *EMBO J.* **5**, 3021 (1986).
19. B. van den Berg *et al.*, *Nature* **427**, 36 (2004).
20. D. Fu *et al.*, *Science* **290**, 481 (2000).
21. R. Dutzler, E. B. Campbell, M. Cadene, B. T. Chait, R. MacKinnon, *Nature* **415**, 287 (2002).
22. S. Khademi *et al.*, *Science* **305**, 1587 (2004).
23. T. Shimizu, H. Mitsuke, K. Noto, M. Arai, *J. Mol. Biol.* **339**, 1 (2004).
24. D. Drew, G. von Heijne, P. Nordlund, J. W. L. de Gier, *FEBS Lett.* **507**, 220 (2001).
25. D. Drew *et al.*, *Protein Sci*, in press.
26. S. Eshaghi *et al.*, *Protein Sci.* **14**, 676 (2005).

27. Supported by grants from the Swedish Research Council, the Marianne and Marcus Wallenberg Foundation, the Swedish Foundation for Strategic Research, and the Swedish Cancer Foundation to G.v.H., by the Swedish Knowledge Foundation to K.M., and by a European Molecular Biology Organization Long-Term Fellowship to D.O.D.

Supporting Online Material

www.sciencemag.org/cgi/content/full/308/5726/1321/DC1
Materials and Methods
Figs. S1 and S2
Tables S1 and S2
References

13 January 2005; accepted 14 March 2005
10.1126/science.1109730

Firearm Violence Exposure and Serious Violent Behavior

Jeffrey B. Bingenheimer,[1]* Robert T. Brennan,[2] Felton J. Earls[2]

To estimate the cause-effect relationship between exposure to firearm violence and subsequent perpetration of serious violence, we applied the analytic method of propensity stratification to longitudinal data on adolescents residing in Chicago, Illinois. Results indicate that exposure to firearm violence approximately doubles the probability that an adolescent will perpetrate serious violence over the subsequent 2 years.

Within the past few decades, the popular notion that violence begets violence has come under scientific scrutiny. Early research by psychologists, criminologists, and others focused on the impact of being physically abused as a child on subsequent delinquency, community violence, and spouse and child abuse. Simple comparisons of violent offenders and nonoffenders showed that the former were more likely to report having been abused during childhood (*1*, *2*). More carefully controlled prospective studies comparing abused and nonabused children confirmed these basic relationships (*3*) and provided insights into the cognitive and neurological mechanisms involved (*4*, *5*).

Recently, interest has expanded to encompass exposure to violence occurring in community settings such as neighborhoods and schools. This change was spurred in part by elevated rates of violent crime, including firearm homicide, in American cities in the early 1990s (*6*, *7*). In several studies conducted around that time, urban children and adolescents reported alarmingly high levels of exposure to community violence, both as witnesses and as victims (*8*–*10*). These findings raised troubling questions about the possible developmental ramifications of such widespread experience with violence.

Numerous recent investigations have revealed statistical associations between children's and adolescents' self-reports of exposure to community violence and concurrent or subsequent assessments of violence and aggression (*11*–*14*). Available estimates of these associations, however, do not adequately control for the possibility that a common set of personal characteristics and environment circumstances may jointly influence who is exposed to community violence and who becomes a perpetrator of violent acts. The extent to which these statistical associations are attributable to cause-effect relationships therefore remains uncertain (*15*).

The randomized experiment is the scientific gold standard for causal inference, but in the instance of community violence is neither technically nor ethically feasible. We used the method of propensity score stratification (*16*–*18*) to approximate a randomized experiment in which exposure to firearm violence was the treatment variable and subsequent perpetration of serious violence was the outcome. This method is based upon counterfactual thinking and the framework of potential outcomes described by Rubin (*19*) and others (*20*). Investigators in economics (*21*), medicine (*22*), and other fields (*23*) are increasingly using propensity score matching and stratification to improve the credibility of estimates of cause-effect relationships obtained from observational data.

Propensity stratification views exposure allocation as a process involving both systematic and random components. First, personal and environmental characteristics of the individual determine systematically her or his probability π of exposure, called the propensity score. The individual then participates in a lottery in which the exposure is assigned with probability π, or nonexposure is assigned with probability $1 - \pi$. In theory, comparing individuals with identical propensity scores but different realized exposures is analogous to conducting a randomized experiment, and therefore provides a valid basis for measuring a cause-effect relationship between expo-

[1]Department of Health Behavior and Health Education, 1420 Washington Heights, University of Michigan School of Public Health, Ann Arbor, MI 48109–2029, USA. [2]Department of Social Medicine, Harvard Medical School, 1430 Massachusetts Avenue, 4th Floor, Cambridge, MA 02138, USA.

*To whom correspondence should be addressed. E-mail: bartbing@umich.edu

sure and outcome. The analytic strategy, then, has four stages: (i) Use all available preexposure information to derive an estimate $\hat{\pi}$ of each subject's propensity score; (ii) divide subjects into strata on the basis of $\hat{\pi}$; (iii) within each stratum, confirm that exposed and unexposed subjects are balanced on all measured preexposure covariates; and (iv) compute the mean difference between exposed and unexposed subjects on the outcome measure within these strata. This produces an unbiased estimate of the average causal effect of exposure under the assumption, termed "strongly ignorable treatment assignment" (*16*), that no measured or unmeasured preexposure characteristic predicts both exposure and outcome independent of estimated propensity scores. The greater the quantity and quality of preexposure information available to the analyst, the more precisely $\hat{\pi}$ will estimate π, and the more plausible the strongly ignorable treatment assignment assumption will become. Sensitivity analyses (*24*) enable the investigator to study the robustness of results to plausible violations of this assumption.

We analyzed data from a longitudinal cohort study of adolescents residing in 78 neighborhoods of Chicago, Illinois (Fig. 1) (*25*). The subjects (N = 1517) were aged 12 or 15 years at the beginning of the study (table S1). Subjects and their primary caregivers were each interviewed on three occasions over a period of 5 years. At Assessment 1, subjects and their caregivers provided detailed information about themselves. From these data we derived measures of 139 preexposure covariates falling into 10 domains: demographic background, family history and home environment, temperament, health and physical development, social support, peer influences, vocabulary and reading proficiency, school-related factors, behavioral patterns, and previous exposure to violence (*25*) (table S2). Additionally, 14 neighborhood social and economic characteristics were quantified by means of census data and an independent survey of a probability sample of adult residents (*26*, *27*), bringing the total number of preexposure covariates to 153 (*25*) (table S2).

At Assessment 2, subjects (n = 1239, 81.7% of the original sample) who could be located and who agreed to continue participating answered a series of questions regarding their exposure to firearm violence in the previous 12 months (*25*) (table S1). They were classified as exposed if they reported that they had been shot or shot at, or if they had seen someone shot or shot at, during that period. Those who reported none of these experiences were classified as unexposed. Fourteen subjects could not be classified due to missing or inconsistent data. Of those who could be classified, the majority (n = 942, 76.9%) were unexposed, and the remainder (n = 283, 23.1%) were exposed (*25*).

Subjects who reported exposure to firearm violence at Assessment 2 differed from unexposed subjects on many Assessment 1 covariates at the $\alpha = 0.01$ statistical significance level (Table 1; for more details see table S2). Compared with unexposed subjects, exposed subjects were more aggressive and reported committing more violent offenses. They tended to be non-white, male, from single-parent households, and receiving public assistance. In terms of temperament, exposed subjects were more impulsive and emotional and less inhibited than unexposed subjects. They were more likely to report having engaged in alcohol and drug use, truancy, general delinquency, and property crimes. Exposed subjects were more likely than unexposed subjects to have family members with criminal records or legal problems. They were also more likely to have been corporally punished and physically abused and to have witnessed domestic violence in their households. Their peer groups were characterized by higher levels of aggressive and delinquent behaviors compared with the peer groups of unexposed subjects. Exposed subjects had lower scores on standardized tests of vocabulary and reading proficiency. Their neighborhoods of residence were characterized by more anomie, physical and social disorder, perceived violence, and concentrated disadvantage, and by lower levels of informal social control and satisfaction with policing, compared with the neighborhoods in which unexposed subjects resided. Many of these correlates of exposure to firearm violence are also well-established predictors of violent behavior (*28*, *29*), substantiating the possibility that statistical associations between violence exposure and perpetration may be attributable in part to the joint influence of these factors on exposure and outcome status.

We used a maximum-likelihood logistic regression model to obtain an estimated propensity score $\hat{\pi}$ for each subject whose Assessment 2 exposure status was available. The model was constructed by an iterative stepwise selection procedure. The 153 Assessment 1 covariates, plus squared terms for the 92 covariates that were measured on continuous metrics, comprised the pool of candidate predictors. At each iteration, the procedure either (i) added to the model the single covariate that was most strongly associated with firearm violence exposure conditional upon the covariates already in the model, provided that the conditional association was statistically significant at the $\alpha = 0.10$ level, or (ii) removed from the model any covariate whose conditional association with gun violence exposure was no longer statistically significant at that level. The resulting model included 37 covariates. We modified this model by adding a squared term for each first-order continuous covariate selected by the procedure, and by adding a first-order term for each squared continuous covariate selected, bringing to 48 the total number of covariates included in the final model (*25*) (table S3).

Exposed and unexposed subjects had notably different probability densities of $\hat{\pi}$ (Fig. 2). The distribution for unexposed subjects was skewed sharply to the right, with nearly half (n = 459, 48.7%) having $\hat{\pi} < 0.10$ and none having $\hat{\pi} > 0.85$. In contrast, the probability density for exposed subjects was more nearly uniform, the only exception being that very few (n = 3, 1.1%) had $\hat{\pi} < 0.05$. We divided subjects into 12 strata on the basis of their estimated propensity scores (Table 2). Cut points were selected such that exposed and unexposed subjects had statistically indistinguishable mean estimated propensity scores

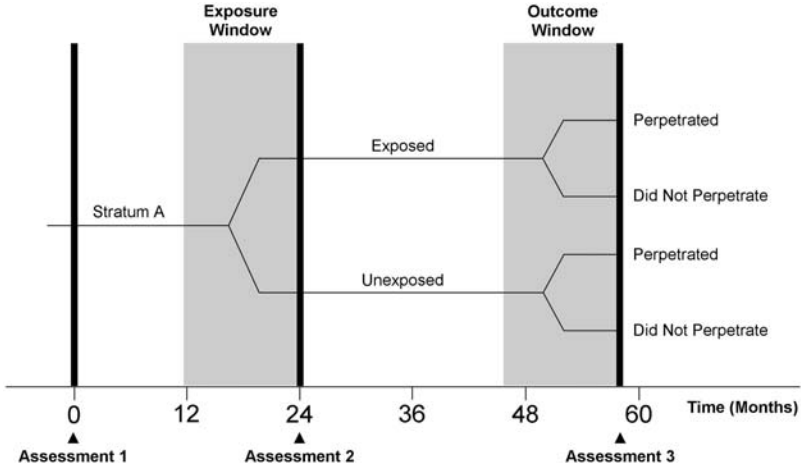


Fig. 1. Design of the propensity-stratified longitudinal study. This design was implemented for 12 propensity strata, 10 of which were used for estimating the cause-effect relationship between exposure to firearm violence (obtained at Assessment 2) and subsequent perpetration of serious violence (obtained at Assessment 3). Strata were defined on the basis of modeled probabilities of exposure at Assessment 2 conditional upon 153 covariates measured at Assessment 1. The placement of Assessments 2 and 3 on the horizontal axis is based upon the median time between interviews.

within each of the resulting strata. Satisfactory balance of unexposed and exposed subjects could not be achieved within the ranges $\hat{\pi} < 0.05$ and $\hat{\pi} > 0.75$. We therefore excluded the lowest and highest strata from the propensity-stratified analyses. Within each of the remaining 10 strata, exposed and unexposed subjects had nearly identical mean values of $\hat{\pi}$ [no statistical difference: $F(10,862) = 0.11$, $P = 0.9998$]. We also used statistical hypothesis tests to determine whether unexposed and exposed subjects had similar distributions of the Assessment 1 covariates within each of the 10 analytic propensity strata. We found no statistically significant ($\alpha = 0.01$) within-stratum differences between exposed and unexposed subjects on any of the 153 covariates (table S2), suggesting that our propensity model and stratification scheme adequately controlled selection on all measured preexposure covariates.

At Assessment 3, subjects who could again be located and who agreed to continue participating (n = 984, 80.3% of those whose Assessment 2 exposure status could be ascertained) answered questions about their perpetration of violent behavior in the previous 12 months (*25*) (table S1). Those reporting that they had carried a hidden weapon, attacked someone with a weapon, shot someone, shot at someone, or been in a gang fight in which someone was hurt or threatened with harm were classified as perpetrators of serious violence. Those reporting none of these five activities were classified as nonperpetrators. The majority of subjects (n = 856, 87.0%) were classified as nonperpetrators, and a minority (n = 122, 12.4%) were classified as perpetrators (*25*). A few (n = 6, 0.6%) could not be classified due to incomplete or inconsistent responses.

We used a series of maximum-likelihood logistic regression models to obtain estimates of the statistical association and the cause-effect relationship between exposure to firearm violence and subsequent perpetration of serious violence (*25*). A model with no covariates revealed a strong statistical association; subjects who were exposed to firearm violence at Assessment 2 were much more likely than unexposed subjects to report perpetration of serious violence at Assessment 3 [odds ratio (OR) = 3.71, $\chi^2(1) = 41.99$, $P < 0.0001$]. Adjustment for race/ethnicity, age, sex, family socioeconomic index, and neighborhood of residence by regression methods attenuated this association only slightly (adjusted OR = 3.57, $t_{970} = 5.29$, $P < 0.0001$). Further adjustment for previous violence exposure, self-reported violent crime, and self- and caregiver-reported delinquency attenuated the association more substantially (adjusted OR = 2.47, $t_{966} = 3.64$, $P = 0.0003$).

Within the 10 analytic propensity strata, we found that subjects who were exposed to firearm violence at Assessment 2 were more likely than unexposed subjects in the same stratum to report serious violence perpetration at Assessment 3 (common within-stratum OR = 2.43, $\chi^2(1) = 11.74$, $P = 0.0006$). Regression adjustment for race/ethnicity, age, sex, family socioeconomic index, and neighborhood of residence did not substantially alter this finding (adjusted common within-stratum OR = 2.62, $t_{677} = 3.039$, $P = 0.0025$), nor did further adjustment for previous violence exposure and violent and aggressive behaviors (adjusted common within-stratum OR = 2.76, $t_{673} = 2.721$, $P = 0.0067$). We also estimated a model in which the effect of firearm violence exposure was allowed to vary across the propensity strata, but comparison of this model with the original propensity-stratified model revealed that any heterogeneity in this effect was too small to be estimated reliably [no statistical difference: $\chi^2(9) = 15.12$, $P = 0.0877$].

Our estimate of the cause-effect relationship between firearm violence exposure and subsequent perpetration of serious violence represents an average treatment effect estimate. This estimate does not apply to individuals with very low or very high levels of propensity to be exposed. We did not estimate a cause-effect relationship outside the range $0.05 < \hat{\pi} < 0.75$ because of imbalance between exposed and unexposed subjects in the lowest and highest propensity strata.

Some sources of potential bias should also be noted. First, the subjects' exposure to firearm violence and subsequent perpetration of serious violence were both assessed by

Table 1. Comparison of exposed and unexposed subjects on select Assessment 1 covariates. Unless otherwise noted, covariates are measured on a continuous scale and standardized to unit variance; differences are standardized mean comparisons; and test statistics are $F(1,1223)$.

Preexposure covariate	Difference	Test statistic	*P* value
Demographic characteristics			
Male sex*	1.199	7.37	0.0066
Minority race/ethnicity†	–	38.55	<0.0001
Receiving public assistance*	1.664	24.34	<0.0001
Single-parent family structure†	–	28.52	0.0002
Temperament			
Impulsivity	0.298	19.63	<0.0001
Inhibitory control	0.218	10.43	0.0013
Sensation-seeking	0.341	25.91	<0.0001
Emotionality	0.203	9.00	0.0028
Antisocial behaviors			
Self-reported aggression	0.315	22.00	<0.0001
Caregiver-reported aggression	0.349	27.14	<0.0001
Violent offenses	0.619	89.19	<0.0001
Alcohol use*	1.540	19.31	<0.0001
Cigarette use*	1.508	15.48	<0.0001
Marijuana use*	2.550	39.41	<0.0001
Truancy	0.118	20.00	<0.0001
School drop-out*	0.977	12.14	0.0005
Self-reported delinquency	0.398	35.40	<0.0001
Caregiver-reported delinquency	0.471	50.27	<0.0001
Property crimes	0.375	31.36	<0.0001
Family environment			
Family members with criminal records*	1.407	12.98	0.0003
Family members with legal problems*	1.620	13.39	0.0003
Corporal punishment	0.244	13.10	0.0003
Physical abuse	0.299	19.80	<0.0001
Witnessing domestic violence	0.240	12.60	0.0004
Peer group characteristics			
Aggressive behaviors	0.567	74.13	<0.0001
Property crimes	0.358	28.52	<0.0001
Drug use	0.397	35.17	<0.0001
Drug selling	0.173	35.02	<0.0001
Sexual activity	0.373	81.76	<0.0001
Indicators of intelligence			
Vocabulary	–0.233	11.90	0.0006
Reading proficiency	–0.181	7.17	0.0075
Neighborhood characteristics			
Anomie	0.352	27.56	<0.0001
Social disorder	0.359	28.73	<0.0001
Perceived violence	0.265	15.45	<0.0001
Concentrated disadvantage	0.529	64.16	<0.0001
Informal social control	–0.190	7.95	0.0049
Satisfaction with policing	–0.269	15.92	<0.0001

*Measured dichotomously; differences are relative risks and test statistics are $\chi^2(1)$. †Measured nominally with more than two categories; reported differences are too complicated to tabulate here, but test statistics are $\chi^2(5)$ for race/ethnicity and $\chi^2(7)$ for family structure.

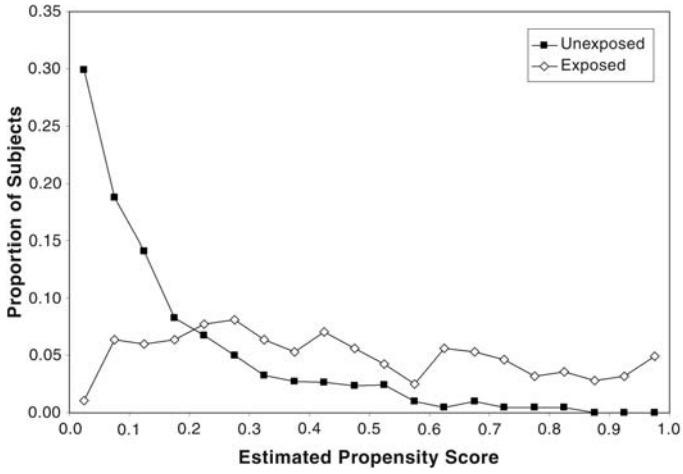


Fig. 2. Probability densities of estimated propensity scores.

Table 2. Distribution of Assessment 3 serious violence perpetration, by propensity stratum and Assessment 2 firearm violence exposure status. Columns labeled "No" contain subjects who denied serious violence perpetration at Assessment 3; columns labeled "Yes" contain subjects who reported serious violence perpetration; columns labeled "?" contain subjects who were lost to follow-up or whose responses to Assessment 3 questions about violence perpetration were inconsistent.

Propensity stratum	Unexposed (N = 942)			Exposed (N = 283)		
	No	Yes	?	No	Yes	?
0.0000 – 0.0500*	224	12	46	3	0	0
0.0500 – 0.0875	96	5	24	8	3	2
0.0875 – 0.1250	85	10	17	4	0	7
0.1250 – 0.1625	82	5	16	7	3	6
0.1625 – 0.2000	35	0	13	8	1	4
0.2000 – 0.2500	44	6	14	10	9	3
0.2500 – 0.3375	58	8	9	19	3	9
0.3375 – 0.4375	31	8	10	25	6	10
0.4375 – 0.5375	26	6	13	20	6	4
0.5375 – 0.5875	6	1	5	2	3	2
0.5875 – 0.7500	12	3	4	22	9	15
0.7500 – 1.0000*	2	3	3	27	12	11
Total	701	67	174	155	55	73

*These strata are excluded from propensity-stratified analyses of covariate balance and treatment effects.

self-report, which implies that correlated misclassification on exposure and outcome status may have occurred, but the magnitude and direction of the resulting bias cannot be determined. Differential attrition of subjects between assessments is another potential source of bias. We found, however, that attrition was not strongly related to baseline covariates (*25*) (table S1), suggesting that any bias resulting from this may have been of minor importance.

Furthermore, we cannot rule out the possibility that unmeasured preexposure characteristics may have jointly influenced both exposure status at Assessment 2 and perpetration at Assessment 3. Yet, the omission of such variables from our analyses would constitute a violation of the assumption of strongly ignorable treatment assignment only to the extent that their influences were independent of estimated propensity scores. This would be most likely in the case of an omitted variable that was uncorrelated with the 153 covariates used in developing and testing our propensity model, but could occur under other circumstances as well. Sensitivity analyses (*24*, *25*) (table S4) showed that these independent influences on both exposure and perpetration would need to be very strong to reduce substantially our estimate of the effect of firearm violence exposure on subsequent violent perpetration.

In conclusion, our focus on firearm violence exposure provided an operational definition of the treatment and control conditions and facilitated the assessment of each subject's exposure status. The longitudinal structure of the investigation established the temporal ordering of preexposure covariates, realized exposure status, and the behaviors comprising the focal outcome. We had access to a large and diverse set of well-measured preexposure covariates obtained from multiple sources, including subjects, their caregivers, neighborhood residents, and census data. We used this information to develop a model of how subjects' personal characteristics and environmental circumstances systematically influenced their probability of being exposed to firearm violence. Stratifying on estimated propensity scores derived from this model provided statistical balance on all 153 measured preexposure covariates, suggesting that we succeeded in isolating the random part of the exposure allocation process, and thereby adequately approximated a randomized experiment. Our results thus provide a more credible basis for the conclusion that exposure to violence is causally related to violent behavior. Specifically, we estimate that being exposed to firearm violence approximately doubles the probability that an adolescent will perpetrate serious violence over the 2 subsequent years.

References and Notes

1. L. B. Silver, C. C. Dublin, R. S. Lourie, *Am. J. Psychiatry* **126**, 404 (1969).
2. J. J. Spinetta, D. Rigler, *Psychol. Bull.* **77**, 296 (1972).
3. C. S. Widom, *Science* **244**, 160 (1989).
4. K. A. Dodge, J. E. Bates, G. S. Pettit, *Science* **250**, 1678 (1990).
5. A. Caspi *et al.*, *Science* **297**, 851 (2002).
6. L. A. Fingerhut, D. D. Ingram, J. J. Feldman, *JAMA* **267**, 3048 (1992).
7. L. A. Fingerhut, D. D. Ingram, J. J. Feldman, *JAMA* **267**, 3054 (1992).
8. J. E. Richters, P. Martinez, *Psychiatry* **56**, 7 (1993).
9. H. Schubiner, R. Scott, A. Tzelepis, *J. Adol. Health* **14**, 214 (1993).
10. M. E. Schwab-Stone *et al.*, *J. Am. Acad. Child Adol. Psychiatry* **34**, 1343 (1995).
11. D. Gorman-Smith, P. Tolan, *Dev. Psychopathol.* **10**, 101 (1998).
12. L. Song, M. I. Singer, T. M. Anglin, *Arch. Pediatr. Adol. Med.* **152**, 531 (1998).
13. L. S. Miller *et al.*, *J. Clin. Child Psychol.* **28**, 2 (1999).
14. C. A. Halliday-Boykins, S. Graham, *J. Ab. Child Psychol.* **29**, 383 (2001).
15. S. L. Buka, T. L. Stichick, I. Birdthistle, F. J. Earls, *Am. J. Orthopsychiatry* **71**, 298 (2001).
16. P. R. Rosenbaum, D. B. Rubin, *Biometrika* **70**, 41 (1983).
17. P. R. Rosenbaum, D. B. Rubin, *J. Am. Stat. Assoc.* **79**, 516 (1984).
18. D. B. Rubin, *Ann. Intern. Med.* **127**, 757 (1997).
19. D. B. Rubin, *J. Educ. Psychol.* **66**, 688 (1974).
20. P. W. Holland, *J. Am. Stat. Assoc.* **81**, 945 (1986).
21. D. J. Benjamin, *J. Public Econ.* **87**, 1259 (2003).
22. K. He *et al.*, *JAMA* **288**, 3130 (2002).
23. J. L. Skeem, E. P. Mulvey, *J. Consult. Clin. Psychol.* **69**, 358 (2001).
24. J. Cornfield *et al.*, *J. Natl. Cancer Inst.* **22**, 173 (1959).
25. Materials and methods are available as supporting material on *Science* Online.
26. R. J. Sampson, S. W. Raudenbush, F. J. Earls, *Science* **277**, 918 (1997).
27. S. W. Raudenbush, R. J. Sampson, *Soc. Methodol.* **29**, 1 (1999).
28. J. D. Hawkins *et al.*, in *Serious and Violent Juvenile Offenders: Risk Factors and Successful Interventions*, R. Loeber, D. P. Farrington, Eds. (Sage, Thousand Oaks, CA, 1998), chap. 7.
29. D. P. Farrington, in *Youth Violence*, M. Torny, M. H. Moore, Eds. (Chicago Univ. Press, Chicago, 1998), pp. 421–475.

30. We thank J. Brooks-Gunn, R. J. Sampson, and S. W. Raudenbush for insight and advice. This work was funded in part by grants from the John D. and Catherine T. MacArthur Foundation, the National Institute of Justice, and the National Institute of Mental Health.

Supporting Online Material

www.sciencemag.org/cgi/content/full/308/5726/1323/DC1
Materials and Methods
Tables S1 to S4

24 January 2005; accepted 30 March 2005
10.1126/science.1110096

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 APRIL 19, 2012 VOL. 366 NO. 16

Comparative Effectiveness of Revascularization Strategies

William S. Weintraub, M.D., Maria V. Grau-Sepulveda, M.D., M.P.H., Jocelyn M. Weiss, Ph.D., M.P.H., Sean M. O'Brien, Ph.D., Eric D. Peterson, M.D., M.P.H., Paul Kolm, Ph.D., Zugui Zhang, Ph.D., Lloyd W. Klein, M.D., Richard E. Shaw, Ph.D., Charles McKay, M.D., Laura L. Ritzenthaler, M.B.A., Jeffrey J. Popma, M.D., John C. Messenger, M.D., David M. Shahian, M.D., Frederick L. Grover, M.D., John E. Mayer, M.D., Cynthia M. Shewan, Ph.D., Kirk N. Garratt, M.D., Issam D. Moussa, M.D., George D. Dangas, M.D., and Fred H. Edwards, M.D.

ABSTRACT

BACKGROUND

Questions persist concerning the comparative effectiveness of percutaneous coronary intervention (PCI) and coronary-artery bypass grafting (CABG). The American College of Cardiology Foundation (ACCF) and the Society of Thoracic Surgeons (STS) collaborated to compare the rates of long-term survival after PCI and CABG.

METHODS

We linked the ACCF National Cardiovascular Data Registry and the STS Adult Cardiac Surgery Database to claims data from the Centers for Medicare and Medicaid Services for the years 2004 through 2008. Outcomes were compared with the use of propensity scores and inverse-probability-weighting adjustment to reduce treatment-selection bias.

RESULTS

Among patients 65 years of age or older who had two-vessel or three-vessel coronary artery disease without acute myocardial infarction, 86,244 underwent CABG and 103,549 underwent PCI. The median follow-up period was 2.67 years. At 1 year, there was no significant difference in adjusted mortality between the groups (6.24% in the CABG group as compared with 6.55% in the PCI group; risk ratio, 0.95; 95% confidence interval [CI], 0.90 to 1.00). At 4 years, there was lower mortality with CABG than with PCI (16.4% vs. 20.8%; risk ratio, 0.79; 95% CI, 0.76 to 0.82). Similar results were noted in multiple subgroups and with the use of several different analytic methods. Residual confounding was assessed by means of a sensitivity analysis.

CONCLUSIONS

In this observational study, we found that, among older patients with multivessel coronary disease that did not require emergency treatment, there was a long-term survival advantage among patients who underwent CABG as compared with patients who underwent PCI. (Funded by the National Heart, Lung, and Blood Institute.)

From the Christiana Care Health System, Newark, DE (W.S.W., P.K., Z.Z.); Duke Clinical Research Institute, Durham, NC (M.V.G.-S., S.M.O., E.D.P.); the American College of Cardiology, Washington, DC (J.M.W., L.L.R.); Advocate Illinois Masonic Medical Center (L.W.K.) and the Society of Thoracic Surgeons (C.M.S.) — both in Chicago; California Pacific Medical Center, San Francisco (R.E.S.); Harbor–UCLA Medical Center, Torrance, CA (C.M.); Beth Israel Deaconess Medical Center (J.J.P.), Massachusetts General Hospital (D.M.S.), and Children's Hospital Boston (J.E.M.) — all in Boston; the University of Colorado School of Medicine, Aurora (J.C.M., F.L.G.); Denver Veterans Affairs Medical Center, Denver (F.L.G.); Lenox Hill Heart and Vascular Institute of New York (K.N.G.) and Mount Sinai Medical Center (G.D.D.) — both in New York; and the Mayo Clinic (I.D.M.) and the University of Florida (F.H.E.) — both in Jacksonville. Address reprint requests to Dr. Weintraub at the Section of Cardiology, Christiana Care Health System, 4755 Ogletown-Stanton Rd., Newark, DE 19718, or at wweintraub@christianacare.org.

This article (10.1056/NEJMoa1110717) was published on March 27, 2012, at NEJM.org.

N Engl J Med 2012;366:1467-76.

The strategies of percutaneous coronary intervention (PCI) and coronary-artery bypass grafting (CABG) for revascularization have been compared in randomized clinical trials.[1,2] Although the best way to control for treatment-selection bias is to conduct a randomized trial, such trials often have limited power to evaluate subgroups, and the results may not be generalizable, since patients and centers are often highly selected. Nonrandomized, observational data from clinical databases can complement data from clinical trials, because observational data, if they are from a larger and more representative population, may better reflect real-world practice.

The American College of Cardiology Foundation (ACCF) and the Society of Thoracic Surgeons (STS) developed a partnership, the ACCF and STS Database Collaboration on the Comparative Effectiveness of Revascularization Strategies (ASCERT), to compare the outcomes of PCI and CABG, using information from records in their respective databases, with follow-up data from claims records of the Centers for Medicare and Medicaid Services (CMS).

Methods

Study Oversight

The authors designed the ASCERT study. The data were collected at the participating institutions of the STS and ACCF databases and were assembled and analyzed by the authors, who vouch for the accuracy of the data and all analyses. An independent institutional review board approved the study and waived the requirement for informed consent.

Study Population

The process of selecting the study population began with the identification of CMS claims for either CABG or PCI with hospital discharge dates between January 1, 2004, and December 31, 2007, from sites participating in both the ACCF PCI database (CathPCI Registry) and the STS Adult Cardiac Surgery Database (ACD). The CathPCI Registry and the ACD were linked to CMS claims files by probabilistic matching, thus circumventing the need for universal patient identifiers.[3] Records from the clinical and CMS databases were considered to represent the same patient if they were fully matched with respect to a set of indirect identifiers, including the patient's date of birth, sex, hospital identification number, admission date, and discharge date.

Patients were excluded from the study if they met any of the following criteria: single-vessel disease, left main coronary artery disease, cardiogenic shock within 24 hours before CABG or at the time of admission to the hospital for PCI, myocardial infarction within 7 days before CABG or before admission to the hospital for PCI, insertion of an intraaortic balloon pump before either procedure, or CABG or valve surgery or PCI within 180 days before the current admission. If a report of CABG and a report of PCI were both associated with the same CMS claim record, the patient was considered to have undergone PCI followed by CABG. Only the first eligible revascularization record for each patient was analyzed.

Adjustment for Differences between Groups

It was anticipated that the PCI study population and the CABG study population would differ substantially with respect to preprocedural characteristics. We therefore collected information on baseline variables that were available in both registries to make adjusted comparisons feasible. Variables common to both registries were identified from versions 2.41 and 2.52 of the data specifications for the ACD and versions 2 and 3 of the data specifications for the CathPCI Registry. We imputed missing values of continuous variables by stratifying patients according to treatment group and combinations of related risk factors and then imputing stratum-specific medians. Missing categorical variables were imputed to the most common category. Additional details regarding the approach to imputation are provided in the Supplementary Appendix, available with the full text of this article at NEJM.org.

Propensity scores to estimate the probability, on the basis of patient and hospital characteristics, that patients would be selected for CABG were developed with the use of logistic regression to adjust for between-group differences in baseline characteristics of the patients and hospitals.[4] Details of the individual variables included in the propensity model are provided in the Supplementary Appendix. Inverse probability weighting that was based on the propensity score was then used as the primary tool to adjust for differences between the two treatment groups.[5] This approach, which was implemented to create balance, involved weighting each patient who underwent CABG by the inverse of the probability that he or she would be selected for CABG and weighting each patient who underwent PCI by the inverse of the proba-

bility that he or she would be selected for PCI. We verified the performance of the propensity model by comparing the distribution of covariates and propensity scores between treatment groups both before and after inverse probability weighting.[6]

STATISTICAL ANALYSIS

Summary statistics are presented as percentages in the case of categorical variables and as means with standard deviations in the case of continuous variables. Baseline characteristics of the patients were compared between treatment groups with the use of the Pearson chi-square test for categorical variables and the Wilcoxon rank-sum test for continuous variables.

The primary end point was all-cause mortality, which was identified from information in the CMS database. Patients were followed from the date of the index revascularization through December 31, 2008. Unadjusted survival curves were estimated with the use of the Kaplan–Meier method[7]; adjusted survival curves were estimated with the use of the inverse-probability-weighting approach of Cole and Hernan.[8] For each treatment group, the survival curves adjusted with the use of inverse probability weighting represent the expected rate of survival if the treatment of interest (PCI or CABG) were applied to all study patients. Using estimated rates of survival among patients undergoing PCI and among those undergoing CABG, we calculated risk ratios at specific time points and used bootstrap methods to obtain 95% confidence intervals. The comparison of CABG with PCI was performed in the overall population and in prospectively defined subgroups.

Several sensitivity analyses were performed (as described in the Supplementary Appendix). Survival curves were re-estimated separately for CABG and PCI with the use of Cox proportional-hazard models without propensity scores.[9] Covariates for each model were identical to those in the propensity model described above. Using these models, we estimated the average survival curves that would be predicted if all the patients in the study were to undergo PCI and if all the patients were to undergo CABG. We also combined inverse probability weighting and model-based approaches for a "doubly robust" analysis.[10] In addition, we conducted a sensitivity analysis using data from patients matched with respect to the propensity score.

We also explored the effect of potential unmeasured confounders. We developed covariate-adjusted Cox models to estimate hazard ratios for CABG as compared with PCI. Even if the proportional-hazards assumption is not met for the treatment-group variable, the hazard ratio may be interpreted as an "average" over the observed event times.[9] We then used the method of Lin et al.[11] to assess whether the observed differences in the rate of death could be fully explained by an unmeasured confounder.

RESULTS

CHARACTERISTICS OF THE STUDY POPULATION

A total of 1,542,872 claims for PCI and 581,036 claims for CABG, for 1,943,653 unique patients, were recorded in the CMS database between January 1, 2004, and December 31, 2007, at 644 sites participating in both the CathPCI Registry and the ACD. After the exclusion criteria were applied, data from 103,549 patients who underwent PCI (7% of the total) and 86,244 patients who underwent CABG (15% of the total) were included in the analysis (Fig. 1 in the Supplementary Appendix).

Table 1 shows selected baseline characteristics of the study patients (a list of all variables is provided in Table 1 in the Supplementary Appendix). Before adjustment with the use of inverse probability weighting, the patients undergoing PCI, as compared with those undergoing CABG, were, on average, older, and more patients were women. More patients in the CABG group than in the PCI group had heart failure, diabetes, hypertension, chronic lung disease, cerebrovascular disease, a history of smoking, or peripheral arterial disease. More patients in the PCI group than in the CABG group had prior myocardial infarction or unstable angina or required urgent procedures. The ejection fraction was somewhat higher in the PCI group. The largest difference between the groups was in the distribution of the number of diseased vessels, with patients in the PCI group more often having two-vessel disease and patients in the CABG group more often having three-vessel disease. After adjustment with the use of inverse probability weighting, all the clinical covariates were well balanced (Table 1). Among patients who underwent PCI, 78% received drug-eluting stents, 16% received bare-metal stents, and 6% underwent the procedure without the placement of stents.

As expected, patients in the PCI group had a lower probability of being selected for CABG than did those in the CABG group (Fig. 1), with the median and interquartile range of the propensity

Table 1. Baseline Characteristics of the Patients.*

Characteristic	Unadjusted Data			Data Adjusted with the Use of Inverse Probability Weighting		
	CABG (N=86,244)	PCI (N=103,549)	P Value	CABG (N=86,244)	PCI (N=103,549)	P Value
Age (yr)	73.1±5.6	74.7±6.5	<0.001	74.0±9.2	74.0±8.3	0.49
Male sex (%)	68.6	57.8	<0.001	62.3	62.8	0.17
History of heart failure (%)	11.5	10.2	<0.001	11.2	10.8	0.07
History of myocardial infarction (%)	25.3	24.6	<0.001	24.5	24.7	0.51
Diabetes (%)						
Any	38.6	34.4	<0.001	35.8	35.8	0.97
Requiring insulin	10.2	9.8	0.007	9.7	9.9	0.35
Hypertension (%)	84.8	83.4	<0.001	83.9	83.8	0.58
Renal failure (%)	6.1	6.2	0.57	6.1	6.1	0.80
Chronic lung disease (%)	20.7	18.9	<0.001	19.4	19.6	0.50
Cerebrovascular disease (%)	17.6	15.8	<0.001	16.6	16.6	0.86
Peripheral arterial disease (%)	17.9	15.3	<0.001	16.4	16.4	0.97
Body-mass index†	28.7±5.8	28.7±5.9	0.78	28.8±8.6	28.7±7.9	0.97
Smoking status (%)						
Former smoker	44.0	42.5	<0.001	43.0	43.3	0.45
Current smoker	12.9	11.6	<0.001	11.9	12.0	0.74
Angina (%)						
None	21.8	30.8	<0.001	26.4	26.8	0.23
Stable	49.6	22.6	<0.001	34.6	34.9	0.46
Unstable	28.6	46.6	<0.001	39.0	38.3	0.07
Ejection fraction (%)	52.9±12.2	55.5±11.4	<0.001	54.4±17.6	54.4±16.2	0.58
Three-vessel disease (%)	80.3	32.1	<0.001	53.2	53.8	0.04
Urgent status (%)	68.6	57.8	<0.001	62.3	62.8	0.17

* Plus–minus values are means ±SD.
† The body-mass index is the weight in kilograms divided by the square of the height in meters.

scores for CABG reflecting this difference (PCI group: median, 20.3%; interquartile range, 9.9 to 44.7; CABG group: median, 71.3%; interquartile range, 50.1 to 85.1).

OUTCOMES

The follow-up time ranged from 1 to 5 years (average follow-up: overall, 2.72 years; CABG group, 2.82 years, and PCI group, 2.63 years; median follow-up: overall, 2.67 years; CABG group, 2.83 years, and PCI group, 2.53 years). Unadjusted survival curves are shown in Figure 2, and the survival curves adjusted with the use of inverse probability weighting are shown in Figure 3. At 1 year, there was no significant difference in adjusted mortality between the groups (6.2% in the CABG group as compared with 6.6% in the PCI group; risk ratio, 0.95; 95% confidence interval [CI], 0.90 to 1.00). The adjusted 4-year mortality was 16.4% in the CABG group and 20.8% in the PCI group (risk ratio, 0.79; 95% CI, 0.76 to 0.82). Sensitivity analyses performed with the use of the Cox model and with the use of data from propensity-matched groups yielded similar results (Table 1 and Fig. 2 in the Supplementary Appendix). The 4-year risk ratios showed a benefit of CABG across subgroups defined according to sex, age, presence or absence of diabetes, body-mass index, presence or absence of chronic lung disease, ejection fraction, and glomerular filtration rate (Fig.

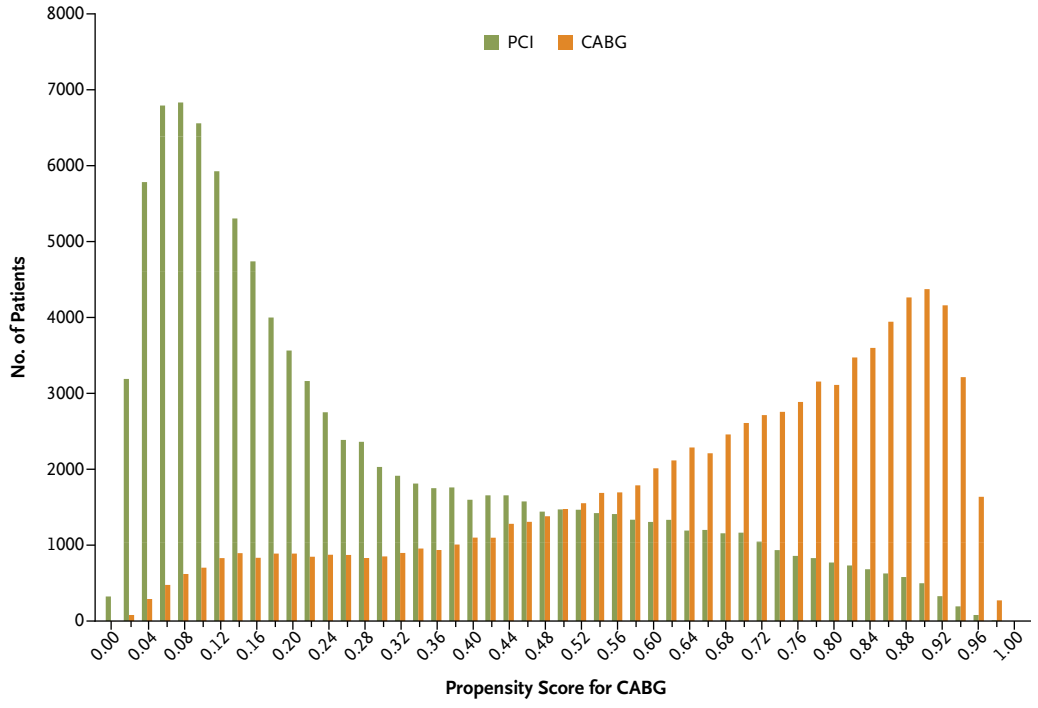


Figure 1. Propensity Scores for Coronary Artery Bypass Grafting (CABG) in the Percutaneous Coronary Intervention (PCI) and CABG Populations.

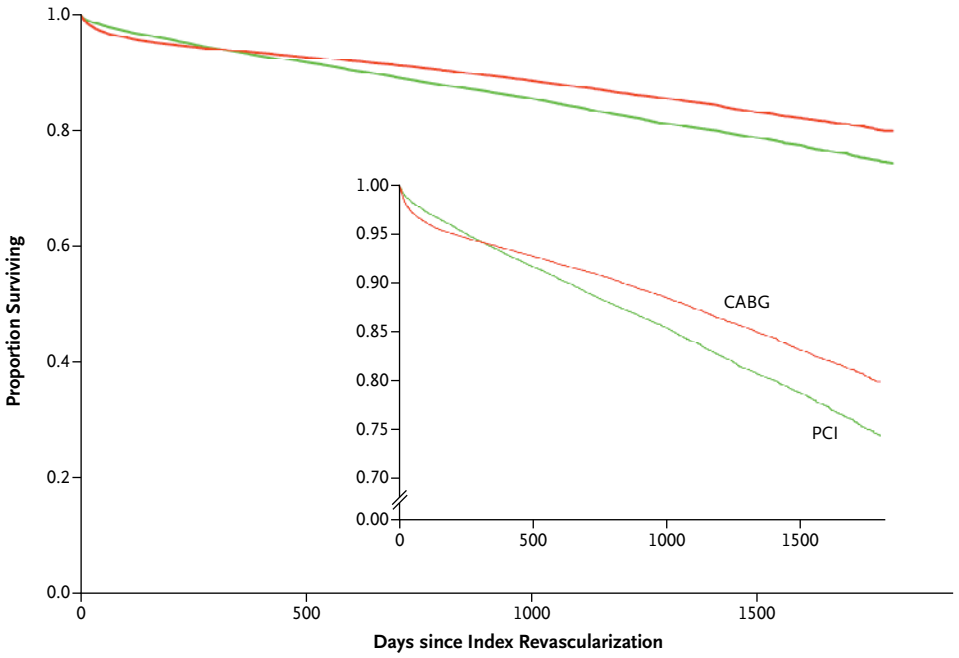
The propensity score for CABG is the probability given baseline variables that any patient in either group would be selected for CABG.

3 in the Supplementary Appendix) and in both a high-risk group and a low-risk group. CABG was also associated with a benefit across subgroups defined according to quintile of propensity score for CABG (from the lowest quintile, 0 to 20%, to the highest quintile, 80 to 100%). Thus, the rate of survival was better with CABG even among patients whose propensity scores were most consistent with selection for PCI.

EFFECT OF UNMEASURED CONFOUNDING

The estimated average hazard ratio with CABG as compared with PCI in the Cox-model analysis was similar to the estimated 4-year risk ratio derived with the use of inverse probability weighting (covariate-adjusted hazard ratio, 0.79; 95% CI, 0.76 to 0.82). Figure 4 shows the method that can be used to determine whether an unmeasured binary risk factor could explain a hazard ratio of this magnitude. The x axis represents the hypothetical prevalence of the unmeasured confounder in the PCI population, and the y axis represents the hypothetical hazard ratio for mortality associated with this confounder. The curved lines indicate the hypothetical prevalence (5%, 10%, 20%, 30%, or 40%) of the potential confounder in the CABG group. For example, if an unmeasured risk factor was present in 10% of the patients in the CABG group (green curved line) and in 20%, 35%, or 50% of the patients in the PCI group, then the hazard ratio that would be required for an unmeasured confounder to account for the observed decreased risk with CABG (i.e., to shift the upper 95% confidence interval from 0.80 to 1.00) would be 4.25, 2.09, and 1.65, respectively. Similarly, if an unmeasured risk factor was present in 20% of the patients in the CABG group (dark blue curved line) and in 30%, 45%, or 60% of the patients in the PCI group, then the hazard ratio that would be required for an unmeasured risk factor to account for the observed increased risk with PCI would be 5.82, 2.22, and 1.70, respectively.

A single unmeasured confounder could produce the observed survival differences only if it increased the long-term risk of death by a factor of approximately two or if the long-term risk of



	30-Day	1-Yr	2-Yr	3-Yr	4-Yr
Mortality after CABG, % (95% CI)	2.07 (1.98–2.17)	6.00 (5.58–6.17)	8.76 (8.56–8.94)	12.1 (11.9–12.4)	16.0 (15.7–16.3)
Mortality after PCI, % (95% CI)	1.21 (1.14–1.27)	6.36 (6.22–6.51)	11.2 (11.0–11.4)	16.0 (15.7–16.2)	20.9 (20.6–21.3)
Relative risk with CABG (95% CI)	1.72 (1.58–1.84)	0.94 (0.91–0.97)	0.78 (0.76–0.80)	0.76 (0.74–0.78)	0.76 (0.75–0.78)

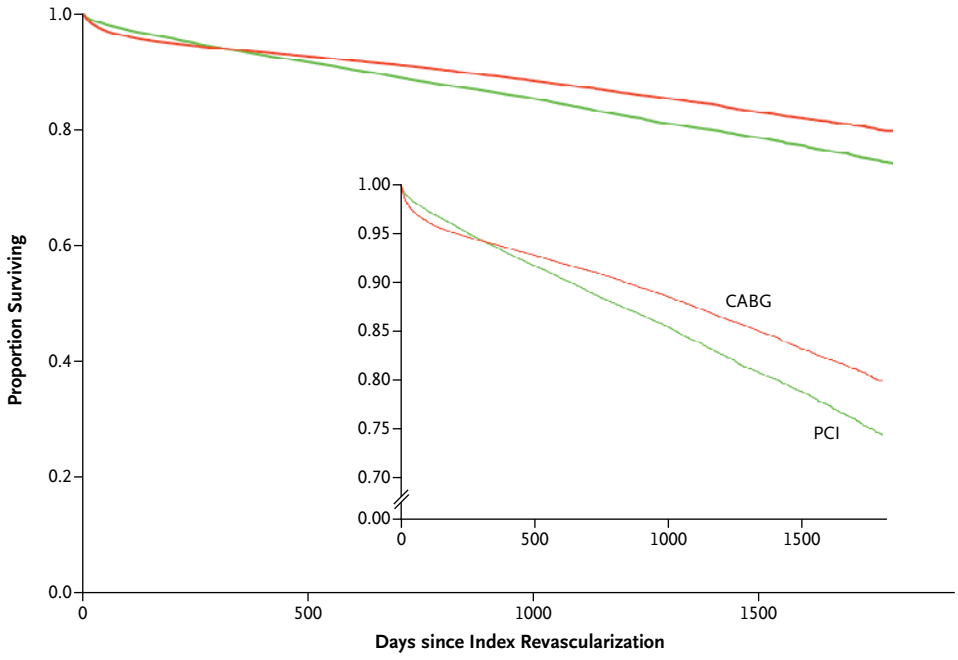
Figure 2. Rates of Survival in the CABG and PCI Populations, from an Unadjusted Analysis.
Cumulative mortality with CABG and with PCI and the relative risk of CABG as compared with PCI are shown.

death was three to five times as high in the PCI group as in the CABG group. As an example of a potential unmeasured confounder, suppose that patient frailty (yes or no) could be assessed in our study. If frailty was present in 10% of the patients in the CABG group (green curved line) but in 35% of patients in the PCI group (x axis), and if frailty increased the risk of death by a factor of slightly more than two (hazard ratio, 2.09), then frailty alone could itself account for the observed difference in mortality between the study groups.

DISCUSSION

The ASCERT study was a collaborative outcome study in which data from the STS and ACCF registries were used to evaluate the effectiveness of revascularization with CABG as compared with PCI. In this study, we found that among Medicare patients 65 years of age or older with ischemic heart disease that required revascularization on a nonemergency basis, there was no significant difference in adjusted mortality at 1 year between patients who had undergone CABG and those who had undergone PCI, but mortality at 4 years was lower in the CABG group than in the PCI group. These findings were noted in all subgroups. The survival rate was better with CABG even among patients whose propensity scores were most consistent with selection for PCI.

Our findings should be evaluated in the context of results from other studies. There have been seven randomized, controlled trials comparing CABG with balloon angioplasty,[12-18] four comparing CABG with PCI and placement of bare-metal stents,[19-22] and one comparing CABG with PCI and placement of drug-eluting stents.[2] A survival advantage with CABG was seen in the Stent or Surgery trial (SoS; ClinicalTrials.gov number, NCT00475449)[23] and in the subgroup with treated



	30-Day	1-Yr	2-Yr	3-Yr	4-Yr
Mortality after CABG, % (95% CI)	2.25 (2.09–2.41)	6.24 (5.97–6.50)	8.98 (8.68–9.29)	12.4 (12.0–12.8)	16.4 (15.9–16.9)
Mortality after PCI, % (95% CI)	1.31 (1.21–1.41)	6.55 (6.35–6.76)	11.3 (11.0–11.6)	15.9 (15.6–16.3)	20.8 (20.4–21.2)
Relative risk with CABG (95% CI)	1.72 (1.52–1.89)	0.95 (0.90–1.00)	0.79 (0.76–0.83)	0.78 (0.75–0.81)	0.79 (0.76–0.82)

Figure 3. Rates of Survival in the CABG and PCI Populations, from an Analysis Adjusted with the Use of Inverse Probability Weighting.
Cumulative mortality with CABG and with PCI and the relative risk of CABG as compared with PCI are shown. The inset shows the same data on an enlarged y axis.

diabetes in the Bypass Angioplasty Revascularization Intervention trial (BARI, NCT00000462).[24] A meta-analysis of these trials, which included 7812 patients, showed a trend toward a survival advantage with CABG.[1] In a subgroup analysis according to the presence or absence of diabetes, there was a survival advantage with CABG among patients who had diabetes, whereas there was no significant advantage among patients who did not have diabetes. Among patients younger than 55 years of age, there was a trend toward a benefit with PCI as compared with CABG, whereas among patients older than 65 years of age, mortality was significantly higher with PCI. This meta-analysis did not include the Synergy between PCI with Taxus and Cardiac Surgery trial (SYNTAX, NCT00114972), a randomized trial in which contemporary methods of revascularization were used. At 3 years, among patients in the SYNTAX trial who had three-vessel disease, both the overall rate of death and the rate of death from cardiac causes were significantly lower among patients who had undergone CABG than among those who had undergone PCI.[25] The mortality was substantially lower in the SYNTAX trial than in the present study, possibly because the patients in the SYNTAX trial were younger, may have had fewer coexisting conditions, or may have had less severe disease.

Six previous observational studies, three of which were multicenter studies[26-28] and three of which were single-center studies,[29-31] also showed results similar to those presented here. The consistent findings in these observational studies lend support to the finding of a survival advantage with CABG observed in our study. Whereas the multicenter studies focused on a single state[26,27] or region[28] in the United States, the national scope of the current study indicates that the favorable survival rates observed among those se-

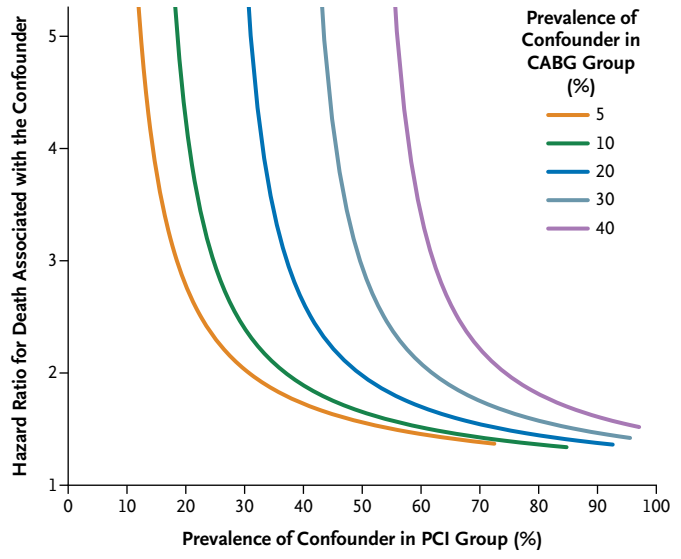


Figure 4. Effect of Unmeasured Confounding Factors.

Shown is a sensitivity analysis that illustrates how powerful a single confounder would have to be to account for the advantage of CABG over PCI that was detected in the adjusted analysis. A single unmeasured confounder could produce the observed survival differences only if it increased the long-term risk of death by a factor of approximately two or if the long-term risk of death was three to five times as high in the PCI group as in the CABG group. For example, if a confounder was present in 10% of the patients in the CABG group (green curved line) but in 35% of patients in the PCI group (x axis), and if it increased the risk of death by a factor of slightly more than two (hazard ratio, 2.09), then that confounder alone could itself account for the observed difference in mortality between the study groups.

lected for CABG extend across the entire United States.

The ASCERT study shows the potential benefits of linking large clinical and administrative databases to assess the comparative effectiveness of therapies in large patient populations. Perhaps the most compelling advantage of this approach is the ability to evaluate outcomes in broadly representative patient populations rather than the selected population of a randomized, controlled trial. Continually developing new randomized, controlled trials comparing PCI with CABG with each advancement in technology is not feasible, but data that are linked as they are in this project can be readily collected on an ongoing basis to provide continuity for subsequent studies.

This study also shows the specific advantages of linking clinical and administrative databases. Clinical databases are well suited to risk adjustment and the identification of clinically important subgroups but lack information on long-term outcomes. Administrative data sets have limited capacity for clinical considerations, but they provide information on long-term outcomes. Linking clinical data with administrative data capitalizes on the advantages of each.

There are several limitations of this study. Despite efforts by the STS and ACCF to provide coordinated, harmonized data records, there are limitations of the data that, in turn, limit the ability to match data across the registries and lead to some uncertainty regarding the interpretation of variables that could affect the results. The elapsed time from the onset of an event was tabulated differently for some of the variables, resulting in some small differences in definitions between the databases. The angiographic data in the STS and ACCF registries are not as detailed as they are in contemporary trials such as SYNTAX; therefore, the ability to establish balance with respect to angiographic variables was limited. No large clinical databases contain the level of anatomical detail required to calculate the SYNTAX score. Data on coronary flow reserve or other functional or anatomical data may supplant strictly angiographic data in the future.

The ASCERT trial is an observational study, and the unadjusted clinical profile and propensity scores for CABG differed between the treatment groups. Although adjustment with the use of inverse probability weighting resulted in excellent balance between the CABG and PCI populations, the potential remains for unmeasured confounders to have influenced the findings. Some variables that are known in clinical practice to have a profound effect on the choice of revascularization (e.g., extensive coronary disease, the presence of chronic total coronary occlusions, and patient frailty) were not available for this analysis. Although the sensitivity analysis suggests that either a single powerful variable or several confounding variables acting in concert could conceivably account for the between-group difference in the rate of survival, such confounders could also increase the difference.[32,33]

Our study also has analytic limitations. Data were missing for certain variables — in particular, the glomerular filtration rate and ejection fraction. In addition, probabilistic matching may not be as reliable a method for finding matches across databases as is the use of a universal patient identifier, which was not available. Finally, the study population consisted entirely of Medicare patients; there-

fore, the results may not be generalizable to younger patients.

In summary, the ASCERT study used data from the ACCF PCI database and the STS CABG database, with linkage to CMS claims records, to evaluate the comparative effectiveness of PCI and CABG. We found that among patients older than 65 years of age with multivessel coronary artery disease that did not require emergency treatment, there was a long-term survival advantage associated with CABG as compared with PCI.

The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung, and Blood Institute or the National Institutes of Health, the Society of Thoracic Surgeons, or the American College of Cardiology Foundation.

Supported by a grant (RC2HL101489) from the National Heart, Lung, and Blood Institute.

Dr. Popma reports receiving consulting fees from Abbott Vascular, Boston Scientific, and Covidien and grant support from Abbott Vascular, Abiomed, Boston Scientific, Cordis, and Medtronic; Dr. Messenger, receiving grant support from Medtronic; Dr. Mayer, receiving honoraria and reimbursement for travel expenses from CHMC Cardiovascular Surgical Foundation; Dr. Dangas, receiving consulting fees from Abbott Vascular, AstraZeneca, Eli Lilly, Johnson & Johnson, and Ogilvy, providing expert testimony regarding stroke in a patient after cardioversion for atrial fibrillation and regarding infection after heart-valve implantation, and receiving grant support from Bristol-Myers Squibb, Eli Lilly, Daichi-Sankyo, the Medicines Company, and Sanofi-Aventis, lecture fees from Abbott Vascular, AstraZeneca, Boston Scientific, Bracco, Bristol-Myers Squibb, Guerbet, Eli Lilly, Johnson & Johnson, the Medicines Company, and Sanofi-Aventis, royalties from Wiley and Informa, and reimbursement for travel expenses from the Cardiovascular Research Foundation; and Dr. Edwards, being an employee of the Society of Thoracic Surgeons. No other potential conflict of interest relevant to this article was reported.

Disclosure forms provided by the authors are available with the full text of this article at NEJM.org.

REFERENCES

1. Hlatky MA, Boothroyd DB, Bravata DM, et al. Coronary artery bypass surgery compared with percutaneous coronary interventions for multivessel disease: a collaborative analysis of individual patient data from ten randomised trials. Lancet 2009;373:1190-7.

2. Serruys PW, Morice MC, Kappetein AP, et al. Percutaneous coronary intervention versus coronary-artery bypass grafting for severe coronary artery disease. N Engl J Med 2009;360:961-72.

3. Douglas PS, Brennan JM, Anstrom KJ, et al. Clinical effectiveness of coronary stents in elderly persons: results from 262,700 Medicare patients in the American College of Cardiology-National Cardiovascular Data Registry. J Am Coll Cardiol 2009;53:1629-41.

4. Rosenbaum PR, Rubin D. The central role of propensity score in observation studies for causal effects. Biometrika 1983; 70:41-55.

5. Curtis LH, Hammill BG, Eisenstein EL, Kramer JM, Anstrom KJ. Using inverse probability-weighted estimators in comparative effectiveness analyses with observational databases. Med Care 2007;45: Suppl 2:S103-S107.

6. Rosenbaum PR. Model-based direct adjustment. J Am Stat Assoc 1987;82:387-94.

7. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc 1958;53:457-81.

8. Cole SR, Hernan MA. Adjusted survival curves with inverse probability weights. Comput Methods Programs Biomed 2004; 75:45-9.

9. Therneau TM, Grambsch PM. Modeling survival data: extending the Cox model. New York: Springer, 2000.

10. Hirano K, Imbens G. Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. Health Serv Outcomes Res Methodol 2001;2:259-78.

11. Lin DY, Psaty BM, Kronmal RA. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. Biometrics 1998;54: 948-63.

12. King SB III, Lembo NJ, Weintraub WS, et al. A randomized trial comparing coronary angioplasty with coronary bypass surgery. N Engl J Med 1994;331:1044-50.

13. Hamm CW, Reimers J, Ischinger T, Rupprecht HJ, Berger J, Bleifeld W. A randomized study of coronary angioplasty compared with bypass surgery in patients with symptomatic multivessel coronary disease. N Engl J Med 1994;331:1037-43.

14. First-year results of CABRI (Coronary Angioplasty versus Bypass Revascularisation Investigation). Lancet 1995;346:1179-84.

15. Coronary angioplasty versus coronary artery bypass surgery: the Randomized Intervention Treatment of Angina (RITA) trial. Lancet 1993;341:573-80.

16. Rodriguez A, Boullon F, Perez-Baliño N, Paviotti C, Liprandi MI, Palacios IF. Argentine randomized trial of percutaneous transluminal coronary angioplasty versus coronary artery bypass surgery in multivessel disease (ERACI): in-hospital results and 1-year follow-up. J Am Coll Cardiol 1993;22:1060-7.

17. The Bypass Angioplasty Revascularization Investigation (BARI) Investigators. Comparison of coronary bypass surgery with angioplasty in patients with multivessel disease. N Engl J Med 1996; 335:217-25. [Erratum, N Engl J Med 1997; 336:147.]

18. Carrié D, Elbaz M, Puel J, et al. Five-year outcome after coronary angioplasty versus bypass surgery in multivessel coronary artery disease: results from the French Monocentric Study. Circulation 1997;96: Suppl:II-1–II-6.

19. Serruys PW, Unger F, Sousa JE, et al. Comparison of coronary-artery bypass surgery and stenting for the treatment of multivessel disease. N Engl J Med 2001; 344:1117-24.

20. SoS Investigators. Coronary artery bypass surgery versus percutaneous coronary intervention with stent implantation in patients with multivessel coronary artery disease (the Stent or Surgery trial): a randomised controlled trial. Lancet 2002; 360:965-70.

21. Rodriguez A, Bernardi V, Navia J, et al. Argentine randomized study: coronary angioplasty with stenting versus coronary bypass surgery in patients with multiple-vessel disease (ERACI II): 30-day and one-year follow-up results. J Am Coll Cardiol 2001;37:51-8. [Erratum, J Am Coll Cardiol 2001;37:973-4.]

22. Hueb W, Lopes NH, Gersh BJ, et al. Five-year follow-up of the Medicine, Angioplasty, or Surgery Study (MASS II): a randomized controlled clinical trial of 3 therapeutic strategies for multivessel coronary artery disease. Circulation 2007; 115:1082-9.

23. Booth J, Clayton T, Pepper J, et al. Randomized, controlled trial of coronary artery bypass surgery versus percutaneous coronary intervention in patients with multivessel coronary artery disease: six-year follow-up from the Stent or Surgery Trial (SoS). Circulation 2008;118:381-8.

24. Seven-year outcome in the Bypass Angioplasty Revascularization Investigation (BARI) by treatment and diabetic status. J Am Coll Cardiol 2000;35:1122-9.

25. Kappetein AP, Feldman TE, Mack MJ, et al. Comparison of coronary bypass surgery with drug-eluting stenting for the treatment of left main and/or three-vessel

disease: 3-year follow-up of the SYNTAX trial. Eur Heart J 2011;32:2125-34.
26. Hannan EL, Racz MJ, Walford G, et al. Long-term outcomes of coronary-artery bypass grafting versus stent implantation. N Engl J Med 2005;352:2174-83.
27. Hannan EL, Wu C, Walford G, et al. Drug-eluting stents vs. coronary-artery bypass grafting in multivessel coronary disease. N Engl J Med 2008;358:331-41.
28. Malenka DJ, Leavitt BJ, Hearne MJ, et al. Comparing long-term survival of patients with multivessel coronary disease after CABG or PCI: analysis of BARI-like patients in northern New England. Circulation 2005;112:Suppl:I-371–I-376.
29. Brener SJ, Lytle BW, Casserly IP, Schneider JP, Topol EJ, Lauer MS. Propensity analysis of long-term survival after surgical or percutaneous revascularization in patients with multivessel coronary artery disease and high-risk features. Circulation 2004;109:2290-5.
30. Smith PK, Califf RM, Tuttle RH, et al. Selection of surgical or percutaneous coronary intervention provides differential longevity benefit. Ann Thorac Surg 2006; 82:1420-9.
31. van Domburg RT, Takkenberg JJM, Noordzij LJ, et al. Late outcome after stenting or coronary artery bypass surgery for the treatment of multivessel disease: a single-center matched-propensity controlled cohort study. Ann Thorac Surg 2005;79:1563-9.
32. McNulty EJ, Ng W, Spertus JA, et al. Surgical candidacy and selection biases in nonemergent left main stenting implications for observational studies. JACC Cardiovasc Interv 2011;4:1020-7.
33. Singh M, Rihal CS, Lennon RJ, Spertus JA, Nair KS, Roger VL. Influence of frailty and health status on outcomes in patients with coronary disease undergoing percutaneous revascularization. Circ Cardiovasc Qual Outcomes 2011;4:496-502.

Supplementary Appendix

This appendix has been provided by the authors to give readers additional information about their work.

Supplement to: Weintraub WS, Grau-Sepulveda MV, Weiss JM, et al. Comparative effectiveness of revascularization strategies. N Engl J Med 2012;366:1467-76. DOI: 10.1056/NEJMoa1110717.

Comparative Effectiveness of Revascularization Strategies

Supplementary Appendix

William S Weintraub MD, Maria V Grau-Sepulveda MD, MPH, Jocelyn M Weiss PhD, MPH, Sean M. O'Brien PhD, Eric D Peterson MD, MPH, Paul Kolm PhD, Zugui Zhang, PhD, Lloyd W Klein MD, Richard E Shaw PhD, Charles McKay MD, Laura L Ritzenthaler MBA, Jeffrey J Popma MD, John C. Messenger MD, David M Shahian MD, Frederick L Grover MD, John E Mayer MD, Cynthia M Shewan PhD, Kirk N Garratt MD, Issam D Moussa MD, George D Dangas MD, Fred H Edwards MD

Contents:

Imputation of Missing Variables

It was anticipated that the PCI study population and the CABG study population would differ substantially with respect to preprocedural characteristics. We therefore collected baseline variables available in both registries to make adjusted comparisons feasible. Variables common to both registries were identified from Versions 2.41 and 2.52 of the STS data specifications and Versions 2 and 3 of the ACCF CathPCI data specifications. Most variables were >99% complete in both groups. Exceptions were ejection fraction (missing 21% in PCI, 4% in CABG) and glomerular filtration rate (GFR) (missing 25% in PCI, 1% in CABG). Missing values of continuous risk factors were imputed by stratifying on treatment group and combinations of other related risk factors, and imputing stratum-specific medians. This approach was used for ejection fraction (stratification by sex, heart failure, prior myocardial infarction, and treatment group), GFR (stratification by age, race, gender, renal failure), and weight and height (stratification by gender and treatment group). Categorical variables had <1% missing data and were imputed to the most common category. Although a single imputation approach was used for our primary IPW analyses, additional analyses were performed using multiple imputation methodology, as described below.

Variables Included in the Propensity Model

Propensity scores to estimate the probability of receiving CABG were developed with logistic regression to adjust for between-group differences in baseline patient and hospital characteristics.[1] Patient-level covariates in the propensity model were: age, gender, race, height, BMI, smoking status, family history of coronary artery disease, GFR (defined as dialysis and/or GFR<=30), renal failure, hypertension, dyslipidemia, cerebrovascular disease, chronic lung disease, peripheral arterial disease, history of heart failure, prior PCI, prior myocardial infarction, angina prior to the procedure, ejection fraction, urgent procedure, number of diseased vessels, mitral insufficiency, mitral stenosis, aortic valve insufficiency and aortic stenosis. Hospital-level covariates were: hospital average annual PCI volume, hospital average annual CABG volume, academic hospital, and hospital location (rural/urban). For patients without renal failure, GFR was modeled as a linear trend between 30 and 90 and flat below 30 or above 90. Patients with renal failure were represented in the model by an indicator variable without further adjustment for GFR. The continuous variable ejection fraction was modeled as a linear trend. All other

continuous variables were modeled as a flexible polynomial with linear and quadratic components.

Sensitivity Analysis

Several sensitivity analyses were performed. First, to account for possible misspecification of the propensity model, survival curves were re-estimated using a regression-based approach that did not utilize propensity scores or inverse probability weighting (IPW). Briefly, we used the Cox proportional hazards model with time-varying hazard ratios to estimate the association between baseline covariates and subsequent survival separately within the PCI and CABG cohorts.[2] Covariates for each model were identical to the propensity model. Using these models, we estimated the average survival curves that would be predicted if all patients in the study were to undergo PCI and if all patients were to undergo CABG. Second, survival curves for PCI versus CABG were estimated using a double robust strategy of combining the IPW with regression-based estimation.[3] Finally, we used propensity matching to compare survival in a matched pairs cohort of CABG and PCI patients. Of the 103,549 PCIs and 86,244 CABGs in our data, 43,084 patients in each group had a match in the other group by at least 3 digits. The characteristics of the patients in the unadjusted, inverse probability weighted and matched pair groups are shown in Supplementary Appendix Table 1. Survival curves based on all of these alternative approaches were overlaid on those produced by the original IPW analysis and were found to be nearly identical (Supplementary Appendix Figure 2).

As a further sensitivity analysis, we estimated hazard ratios for CABG versus PCI using a series of covariate-adjusted Cox models. Although the proportional hazards assumption was not met for the treatment group variable (as evidenced by the crossing survival curves), the estimated hazard ratio may be interpreted as an "average" over the observed event times.[2] Model 1 included all of the hospital and patient-level covariates in the propensity model plus an indicator of treatment group (CABG versus PCI). Covariates other than treatment group were modeled with time-dependent hazard ratios in order to relax the proportional hazards assumption for these covariates. For Model 2, we removed hospital-level covariates and instead entered hospital ID as a stratification variable.[4] For Model 3, we accounted for missing data by using multiple imputation as implemented in the R (www.R-project.org) package Multivariate Imputation by Chained Equations (MICE).[5] The imputation model included covariates from Model 1 plus

mortality status (1=death, 0=censored), time to death or censoring (log-scale), and the interaction of mortality status and procedure date. Ten randomly imputed complete datasets were generated and analyzed individually using the methods described above. Regression coefficients from the 10 models were then combined using standard formulas.[6] Hazard ratios (HRs) for CABG versus PCI were similar to the IPW estimated 4-year risk ratios and were consistent for the 3 different versions of the Cox model (Model 1: HR 0.78 [95% CI: 0.75-0.80]; Model 2: 0.75 [95% CI: 0.73-0.78]; Model 3: 0.74 [95% CI: 0.72-0.76]). Based on Model 1 results, we employed the method of Lin et al[7] to assess whether an unmeasured binary risk factor could explain a hazard ratio of this magnitude (explained in the main text).

Supplementary Appendix Table 1

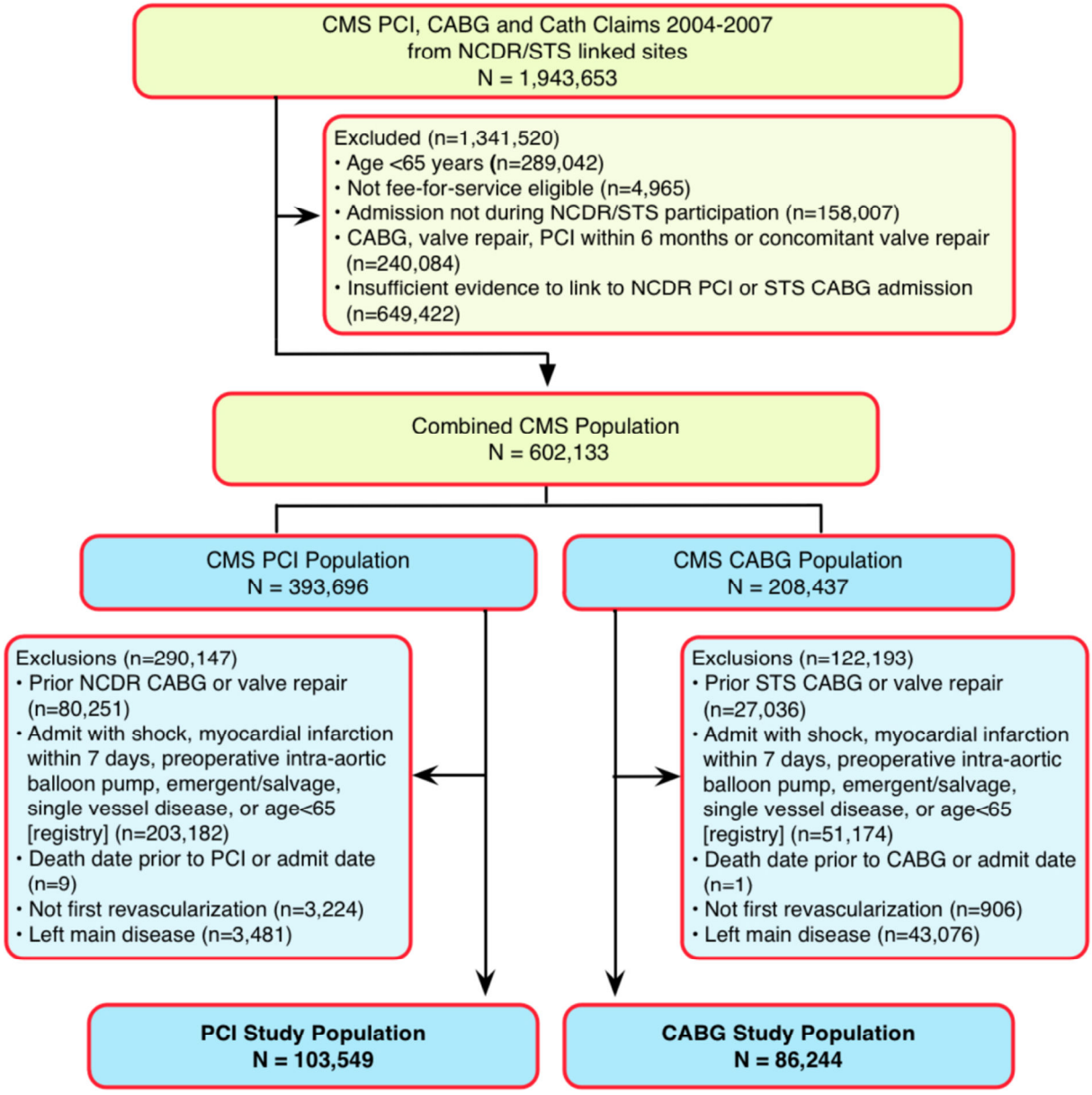
Unadjusted Inverse Probability Weighting Adjustment and Matched Pair Comparison

	Unadjusted			Adjusted			Matched Pair		
	CABG (n=86,244)	**PCI (n=103,549)**	**P Value**	**CABG (n=86,244)**	**PCI (n=103,549)**	**P Value**	**CABG (n=43,084)**	**PCI (n=43,084)**	**P Value**
Demographics									
Age (years)	73.1±5.6	74.7±6.5	<0.0001	74.0±9.2	74.0±8.3	0.49	73.8±5.9	73.9±5.9	0.62
Male	68.6	57.8	<0.0001	62.3	62.8	0.17	63.7	63.5	0.68
Race									
White	90.1	89.5	<0.0001	89.7	89.9	0.27	89.7	89.8	0.69
African American	4.67	4.77	0.30	4.78	4.64	0.35	4.78	4.83	0.73
Other	5.24	5.73	<0.0001	5.51	5.42	0.54	5.51	5.38	0.39
Risk Factors									
Height (cm)	171±10	169±11	<0.001	170±15.7	170.1±14.3	0.20	170±11	170±11	0.59
BMI ($kg/m^{2)}$	28.7±5.8	28.7±5.9	0.78	28.8±8.6	28.7±7.9	0.97	28.8±5.8	28.8±5.8	0.77
Smoking Status									
Current Smoker	12.9	11.6	<0.0001	11.9	12.0	0.74	12.3	12.1	0.38
Former Smoker	44.0	42.5	<0.0001	43.0	43.3	0.45	43.1	43.4	0.38
Never	43.1	45.9	<0.0001	45.0	43.3	0.45	44.6	44.5	0.75
Family History of CAD	33.0	21.9	<0.001	26.6	26.8	0.62	26.2	26.8	0.067

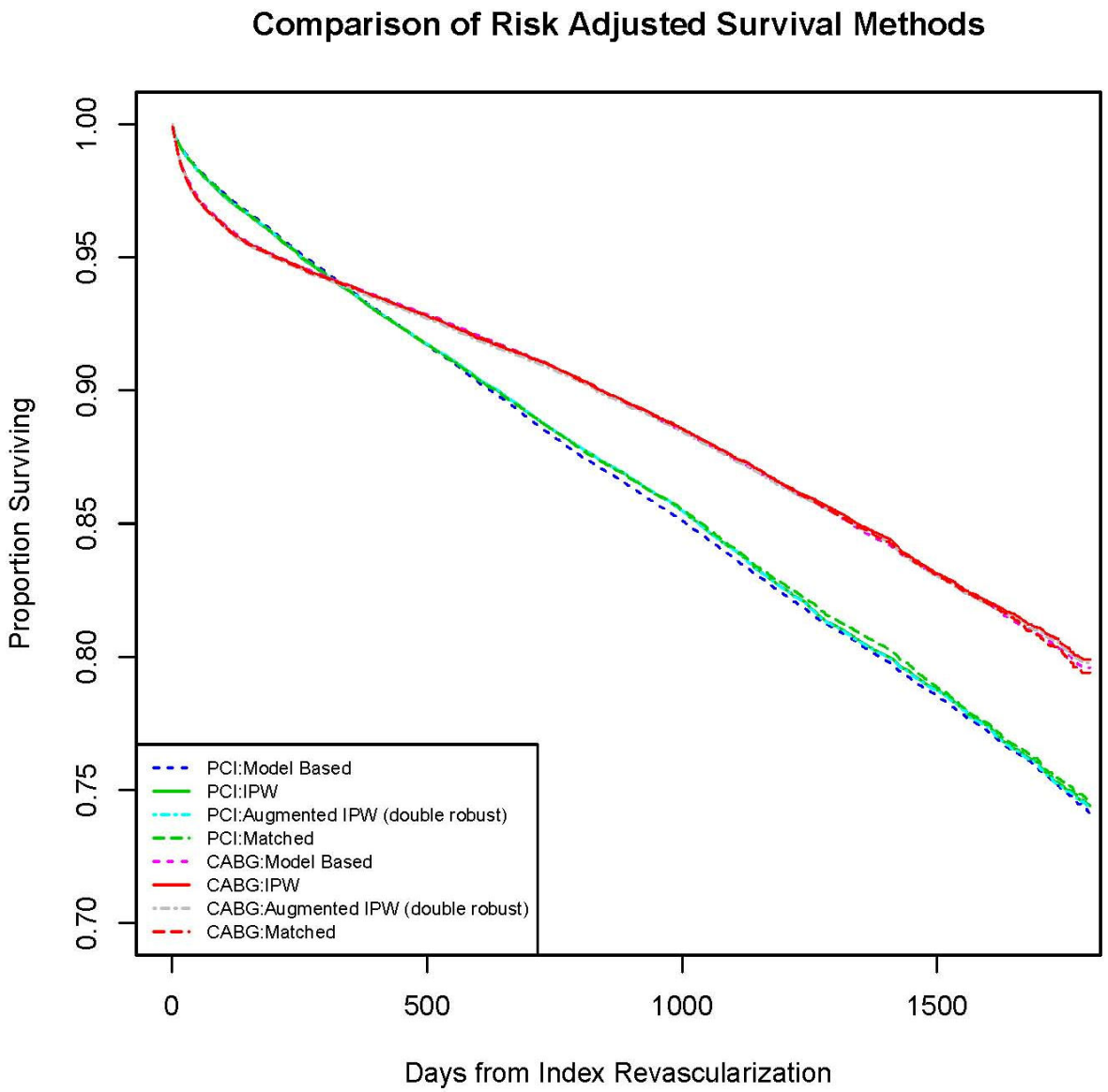
	Unadjusted			Adjusted			Matched Pair		
	CABG (n=86,244)	**PCI (n=103,549)**	**P Value**	**CABG (n=86,244)**	**PCI (n=103,549)**	**P Value**	**CABG (n=43,084)**	**PCI (n=43,084)**	**P Value**
Diabetes									
Insulin Requiring	10.2	9.8	0.0069	9.7	9.9	0.35	10.0	10.1	0.73
Not Insulin Requiring	28.4	24.6	<0.0001	26.8	25.9	0.56	26.6	26.4	0.44
No Diabetes	61.4	65.6	<0.0001	64.2	64.1	0.97	63.4	63.6	0.63
GFR (ml/min)	67.9±25.9	65.4±23.7	<0.0001	66.9 ±41.2	66.4 ±32	0.00063	66.9±24.3	66.5±24.0	0.58
Renal Failure	6.1	6.2	0.57	6.1	6.1	0.80	3.85	3.83	0.92
Hypertension	84.8	83.4	<0.0001	83.9	83.8	0.58	83.9	84.1	0.39
Dyslipidemia	77.7	74.9	<0.0001	75.9	76.0	0.61	75.8	75.8	0.87
Chronic Lung Disease	20.7	18.9	<0.0001	19.4	19.6	0.50	16.7	16.7	0.94
Cerebrovascular Disease	17.6	15.8	<0.0001	16.6	16.6	0.86	19.9	19.8	0.99
Peripheral Arterial Disease	17.9	15.3	<0.0001	16.4	16.4	0.97	16.4	16.6	0.45
Cardiac Status									
History of Heart Failure	11.5	10.2	<0.0001	11.2	10.8	0.067	10.8	11	0.30
Prior PCI	15.7	31.0	<0.0001	24.8	24.2	0.049	21.9	21.9	0.90
History of Myocardial Infarction	25.3	24.6	0.0001	24.5	24.7	0.51	24.1	24.2	0.88
Angina Prior to the Procedure									
No Angina	21.8	30.8	<0.0001	26.4	26.8	0.23	27.7	27.7	0.90
Stable Angina	49.6	22.6	<0.0001	34.6	34.9	0.46	33.9	34	0.64
Unstable Angina	28.6	46.6	<0.0001	39.0	38.3	0.066	38.4	38.3	0.73

	Unadjusted			Adjusted			Matched Pair		
	CABG (n=86,244)	**PCI (n=103,549)**	**P Value**	**CABG (n=86,244)**	**PCI (n=103,549)**	**P Value**	**CABG (n=43,084)**	**PCI (n=43,084)**	**P Value**
Ejection Fraction	52.9±12.2	55.5±11.4	<0.0001	54.4±17.6	54.4±16.2	0.58	54.2±11.6	54.3±11.9	<.0001
Vessels Diseased (3 vs 2)	80.3	32.1	<0.0001	53.2	53.8	0.043	62.7	62.6	0.88
Procedure Status Urgent	68.6	57.8	<0.0001	62.3	62.8	0.17	35.6	35.5	0.82
Valve Assessment									
Mitral Valve Insufficiency	2.56	1.44	<0.0001	1.95	1.95	0.72	1.88	1.9	0.76
Mitral Valve Stenosis	0.37	0.68	<0.0001	0.58	0.56	0.55	0.50	0.51	0.85
Aortic Valve Insufficiency	0.73	0.53	<0.0001	0.66	0.65	0.84	0.64	0.65	0.83
Aortic Valve Stenosis	1.98	2.16	0.0049	2.05	2.12	0.44	2.04	2.08	0.74
Hospital Variables									
Average CABG Volume/Year	215±161	191±144	<0.0001	201±224	201±202	0.87	201±151	201±151	0.053
Average PCI Volume/Year	512±371	530±379	<0.0001	516±547	518±496	0.65	513±367	513±365	0.27
Academic Institution	27.9	28.0	0.74	28.4	27.7	0.034	27.6	28	0.22
Rural (vs. Urban)	5.23	3.73	<0.0001	4.43	4.53	0.52	4.51	4.47	0.82

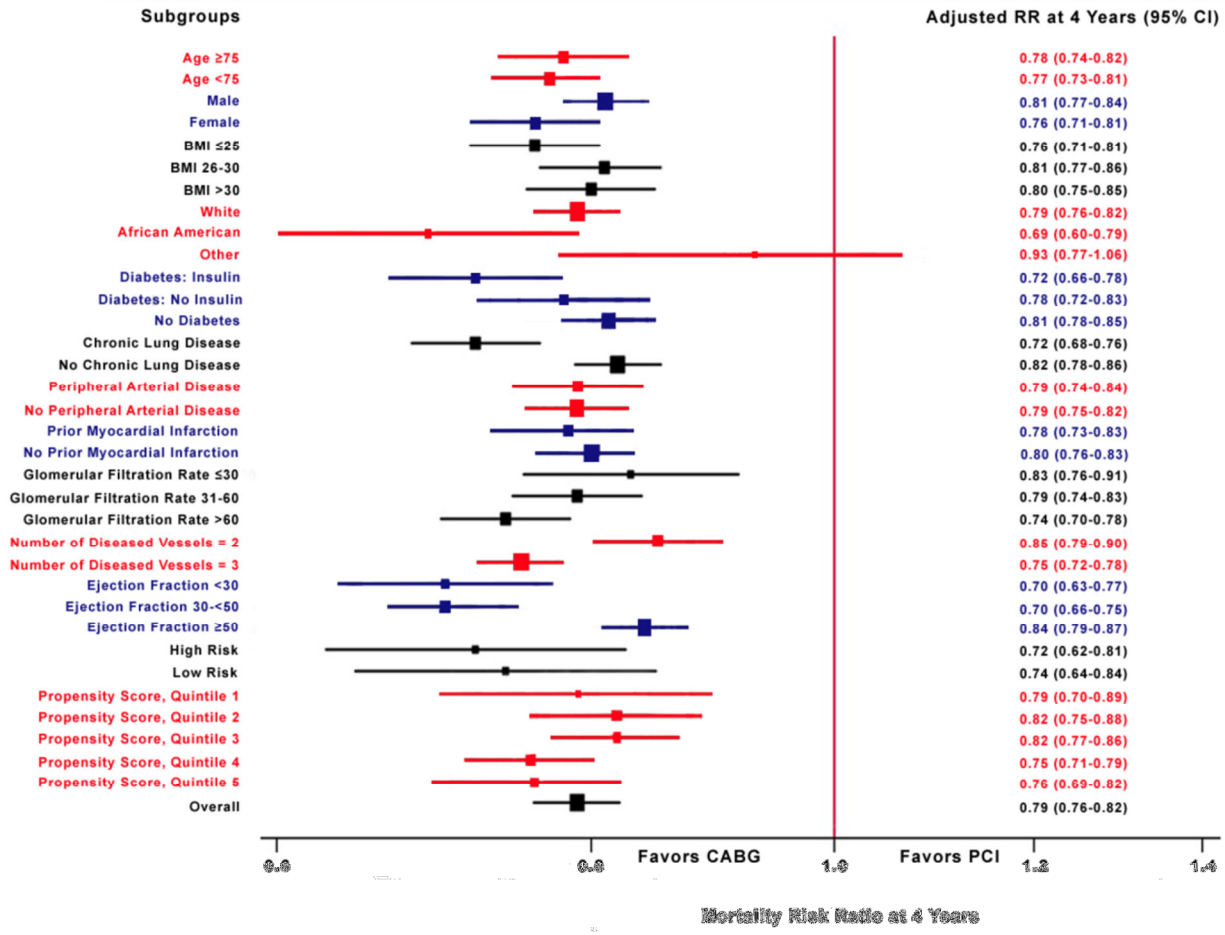
Numbers are in percentages or mean ± standard deviation. Abbreviations: CAD, coronary artery disease; GFR, glomerular filtration rate; CABG coronary artery bypass graft; PCI, percutaneous coronary intervention.



Supplementary Appendix Figure 1: Patient Selection Flow Diagram



Supplementary Appendix Figure 2: Survival in the PCI and CABG populations using different analytic methods: covariate-adjusted model, inverse probability weighted (IPW) analysis, augmented IPW double-robust analysis and propensity score matching analysis.



Supplementary Appendix Figure 3: Forest plot of hazard ratios for mortality by subgroup

References

1. Rosenbaum PR, Rubin D. The central role of propensity score in observation studies for causal effects. Biometrika 1983;70:41-55.
2. Therneau TM, Grambsch PM. Modeling Survival Data: Extending the Cox Model. New York: Springer; 2000.
3. Hirano K, Imbens G. Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart. Health Services and Outcomes Research Methodology 2001;2:259-78.
4. Localio AR, Berlin JA, Ten Have TR, Kimmel SE. Adjustments for center in multicenter studies: an overview. Ann Intern Med 2001;135(2):112-23.
5. Van Buuren S, Groothuis-Oudshoorn K. Multivariate imputation by chained equations: MICE V1.0 user's manual. TNO Quality of Life; 2000.
6. Rubin D. Multiple imputation for nonresponse in surveys. New York: Wiley; 1987.
7. Lin DY, Psaty BM, Kronmal RA. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. Biometrics 1998;54(3):948-63.