JOURNAL OF THE AMERICAN COLLEGE OF CARDIOLOGY
© 2015 BY THE AMERICAN COLLEGE OF CARDIOLOGY FOUNDATION
PUBLISHED BY ELSEVIER INC.

VOL. 66, NO. 17, 2015
ISSN 0735-1097/$36.00
http://dx.doi.org/10.1016/j.jacc.2015.08.037

EDITORIAL COMMENT

Merits and Pitfalls of Using Observational "Big Data" to Inform Our Understanding of Socioeconomic Outcome Disparities*

David A. Alter, MD, PhD

The evolution and infusion of "big data" in our abilities to track personal health information across multiple registries and administrative databases longitudinally has fostered contemporary health services and outcomes-based research worldwide (1). Big data have been particularly useful in evaluating social determinants of health, where studies rely heavily on observational data. Yet, the merits and pitfalls of using large systems of observational data to inform our understanding of socioeconomic status (SES) and outcomes warrant critical appraisal.

SEE PAGE 1888

With this context in mind, we should read the study by Dalén et al. (2) with interest. This large population-based SES-outcome study was conducted within Sweden's universal health care system and explored the relationship between household disposable income and all-cause mortality among 100,534 patients who underwent cardiac surgery between 1999 and 2012. The authors used the national registry known as SWEDEHEART (Swedish Web-System for Enhancement and Development of Evidence-Based Care in Heart Disease Evaluated According to Recommended Therapies). After adjusting for various factors—including age, sex, birth region, education, marital status, and comorbid diseases—patients with household disposable incomes in the highest quintiles experienced a 30% lower relative risk for post-operative all-cause mortality compared to their lowest household income quintile counterparts. The study tracked outcomes over a mean follow-up of 7.3 years and demonstrated similar relative income-mortality associations in early and late post-operative time periods. The authors speculate that household disposable income may be serving as a surrogate for a host of unmeasured lifestyle behaviors and, accordingly, advocate for better implementation of secondary prevention strategies in low-income patient groups.

One of this observational study's unique strengths was in the use of big data to acquire individual-level household disposable income. Utilizing their national health insurance and labor market registry, investigators were able to determine household disposable income (which represented taxable and tax-free income minus final tax and other negative transfers) for each patient throughout multiple years leading up to, and including, the calendar year of surgery. While details around the accuracy and validity of such data were not provided, SES information was ascertained more rigorously than in previous studies of similar size and scope. Historically, SES-outcome studies have relied on either neighborhood census data or individual self-reports to characterize SES. Such ascertainment methods carry numerous biases. In using actual patient-level labor market data, Dalén et al. (2) illustrated a new application of big data in a comprehensive population of patients.

Yet, has this study's use of national registry data helped to inform, unravel, or address the mysteries underlying SES-outcome associations? Or, alternatively, has this study simply reaffirmed associations that we know already exist?

Dalén et al. (2) justify their study in part by arguing that most previous research demonstrating

*Editorials published in the *Journal of the American College of Cardiology* reflect the views of the authors and do not necessarily represent the views of *JACC* or the American College of Cardiology.

From the University Health Network-Toronto Rehabilitation Institute, Institute for Clinical Evaluative Sciences, Toronto, Ontario, Canada. Dr. Alter has reported that he has no relationships relevant to the contents of this paper to disclose.

associations between SES and mortality has done so in populations without universal tax-financed health care where affordability issues may impede health care access. While publicly funded health care programs remove significant financial burden, studies have demonstrated pervasive SES-outcome differences even in the presence of universal health care. For example, Mackenbach et al. (3) explored 22 European countries, each of which had some level of national health insurance, using multiple national registries, surveys, and census data to determine socioeconomic health and outcome gradients in all countries examined. Similarly in Canada, where all citizens receive access to medical care without user fees or out-of-pocket payments, SES was shown to be independently associated with both outcomes and utilization of specialty services (4,5). Our group has previously demonstrated that SES-mortality associations following acute myocardial infarction (AMI) were explained predominantly by differences in age, baseline cardiovascular risk profiles, and functional recovery during the first year following AMI hospitalization. In contrast, neither variations in access, health care utilization, or quality of care meaningfully accounted for SES-mortality associations post-AMI (6,7). In the end, many have drawn similar conclusions to those by Dalén et al. (2), with a focus on preventive or social health policies and significantly less attention on health care delivery per se (8,9).

While some may infer that such conclusions have demonstrated the merits of big data in helping to inform and unravel the SES-outcome paradigm, we should take pause. Though intuitive, the implications of such conclusions are nontrivial, and more importantly, potentially misleading. Dalén et al.'s (2) study once again sheds light.

Per Table 1 in Dalén et al. (2), household disposable income correlated with several factors. For example, the average age between lowest and highest income quintiles differed by >7 years, which is of considerable magnitude for a population-based study. Age may have directly affected income ascertainment; income opportunities among older individuals would have been less partly due to retirement. Additionally, income may have been biased by "reverse causality," where a higher burden of symptoms and disease may have led to lower income earnings because of greater workforce absenteeism and disability. Patients in lower-income quintiles were also more likely to be women, and have achieved <10 years of education; experienced greater social deprivation from being single, divorced, or widowed; and had higher burden of comorbid diseases (including pre-existing heart failure, diabetes, and prior myocardial infarction) than their highest-income counterparts. While the authors used Cox proportional hazards with multivariable analyses, the extent to which such traditional risk-adjustment techniques can "negate" or "adjust" for such large differences in "measured" baseline characteristics remains debatable.

Furthermore, a myriad of other potential "unmeasured" confounders likely also existed and could not be taken into account in the study, including functional capacity, mobility, and depression. Moreover, there was no information on pharmacological treatments, self-management, cardiac rehabilitation, or disease progression over time. No information was available on patients' behavioral risks such as smoking, diet, physical activity, or health care-seeking behaviors, to determine with certainty whether such lifestyle preventive behaviors may be important in the income-mortality differences observed in the study. Short of a large natural history study that prospectively tracks SES, behaviors, health status, risk factors, and disease longitudinally over time, it is unlikely that researchers would ever be able to temporarily disentangle potential root causes from their downstream disease sequelae. In sum, notwithstanding the impressive data source used to ascertain individual income, SES remains a complex and heavily confounded variable.

Even if such data-intensive comprehensive datasets existed, no observational study could ever determine the extent to which such SES-outcome gradients were modifiable. Residual confounding will always remain the "weakest link" of big data observational research. Only interventions and clinical trials can test the potential modifiability of SES mortality-associated gradients. Yet, the efficacy, effectiveness, and cost effectiveness of preventive strategies targeting lower SES groups who are less behaviorally engaged in the self-management of their own health remain unclear (10).

What then is the future role for big data in SES outcomes research? First, these growing repositories of data are uniquely positioned to evaluate the impact of natural health-system experiments, temporal changes in health care delivery, and public-health policy initiatives (e.g., SES-outcome gradients before and after antitobacco legislation). Second, big data can contribute as a screening and surveillance tool to identify high-risk lower SES regions, communities, hospitals, or individuals for risk-stratification targeted interventions. Third, these large databases can be further leveraged with clinical trials to improve the efficiency and comprehensiveness of patient follow-up to track outcomes. In so doing, clinical

trials may be better positioned to evaluate the impact of pharmacological and nonpharmacological interventions across SES to determine if treatment effects systematically vary between socioeconomically advantaged and disadvantaged populations.

In conclusion, observational studies using large multilinked registries will never fully elucidate intermediary causal mechanisms and/or the modifiability of SES-outcome gradients. For this to be accomplished, researchers will require intervention-based research and implementation science. However, as long as health systems continue to value the importance of SES as a health equity metric, big data will continue to play an important role in the future of social epidemiology. The study by Dalén et al. exemplifies the leveragability of such data for social epidemiological health outcomes research. It is now up to all researchers to optimize the application of big data in order to determine the extent to which such SES-outcome-associated gradients are modifiable moving forward.

REPRINT REQUESTS AND CORRESPONDENCE: Dr. David A. Alter, University Health Network-Toronto Rehabilitation Institute, Institute for Clinical Evaluative Sciences, 2075 Bayview Avenue, G Wing, 1-06, Toronto, Ontario M4N 3M5, Canada. E-mail: david.alter@ices.on.ca.

REFERENCES

1. Tu JV, Chu A, Donovan LR, et al. The Cardiovascular Health in Ambulatory Care Research Team (CANHEART): using big data to measure and improve cardiovascular health and healthcare services. Circ Cardiovasc Qual Outcomes 2015;8:204–12.

2. Dalén M, Ivert T, Holzmann MJ, Sartipy U. Household disposable income and long-term survival after cardiac surgery: a Swedish nationwide cohort study in 100,534 patients. J Am Coll Cardiol 2015;66:1888–97.

3. Mackenbach JP, Stirbu I, Roskam AJ, et al. Socioeconomic inequalities in health in 22 European countries. N Engl J Med 2008;358:2468–81.

4. Pilote L, Tu JV, Humphries K, et al. Socioeconomic status, access to health care, and outcomes after acute myocardial infarction in Canada's universal health care system. Med Care 2007;45:638–46.

5. Booth GL, Bishara P, Lipscombe LL, et al. Universal drug coverage and socioeconomic disparities in major diabetes outcomes. Diabetes Care 2012;35:2257–64.

6. Alter DA, Chong A, Austin PC, et al. Socioeconomic status and mortality after acute myocardial infarction. Ann Intern Med 2006;144:82–93.

7. Alter DA, Franklin B, Ko DT, et al. Socioeconomic status, functional recovery, and long-term mortality among patients surviving acute myocardial infarction. PLoS One 2014;8:e65130.

8. Irwin A, Valentine N, Brown C, et al. The commission on social determinants of health: tackling the social roots of health inequities. PLoS Med 2006;3:e106.

9. Friel S, Marmot MG. Action on the social determinants of health and health inequities goes global. Annu Rev Public Health 2011;32:225–36.

10. Korczak D, Dietl M, Steinhauser G. Effectiveness of programmes as part of primary prevention demonstrated on the example of cardiovascular diseases and the metabolic syndrome. GMS Health Technol Assess 2011;7:Doc02.

KEY WORDS health data, outcomes, socioeconomic status