Access to health services, health insurance, and regular providers have been shown to vary significantly across sociodemographic groups within the United States.[1, 2] Marital status, employment, and job characteristics have been found to predict health service receipt beyond insurance status,[3] and differences in health care access across racial and ethnic lines are also well-documented.[1] Improved health outcomes and reduced mortality have been linked to increased primary care usage as compared to emergency services.[4, 5] Variation in environmental conditions across communities has been shown to correspond to differences in trajectory of medical conditions among residents[6, 7] Environment characteristics common among disadvantaged populations have been associated with greater frequency of negative physical health outcomes[8, 9] In addition, common mental health conditions such as depression can affect medical care receipt and treatment adherence.

Varying relationships across individual and environmental characteristics as related to missed needed medical services suggests the utility for recognition of complexity in relationships corresponding to potentially related risk factors. To do this, we consider the roles for sociodemographic factors, insurance status, neighborhood characteristics and depression in estimating the probability of missed needed care. Emphasizing prediction of missed needed care through neighborhood and individual characteristics, and the likely occurrence of interaction effects, we use and assess the gradient boosted regression trees[13] (BRT) method. BRT combines the tree-based model fitting process for CART[14] with boosting,[15] an adaptive stagewise model fitting procedure, to improve predictive accuracy through an ensemble of many simple decision tree models.[16, 17] Through regularization methods, cross-validation, subsampling, and shrinkage,[18] we carry out estimation to emphasize generalizability in estimation, caution against overfitting relationships in training data ,while measuring for relative influence[13] of variables in estimation error reduction and estimation of interaction effects.[19] Analyses were carried out using the 'gbm' package[20] within the open-source statistical software environment R version 3.0.1.[21] Analyses were carried out in the full sample and stratified by health insurance status. BRT specifications were evaluated for predictive performance using ROC curve figures and calculation of AUC values.

Data were collected in New York City from 18,552 participants surveyed through the 2009 (n=9,900) and 2010 (n=8,622) Community Health Survey (CHS).[22] The CHS is an annual telephone survey, carried out by the NYC Department of Health and Mental Hygiene, designed to identify health behaviors and conditions among non-institutionalized adults age 18 and older living in NYC's 34 United Hospital Fund neighborhoods. Response rates for the 2009 and 2010 CHS were 37.7% and 39.0% respectively, with cooperation rates of 89.5% and 89.4%. Sampling weights were generated based on respondent age, race, and gender as proportionate to United Hospital Fund (UHF) neighborhood demographics. CHS data were de-identiified and publicly available.[22] Additional weighting accounted for selection probability by ratio of household adults to phone lines. Mobile phones were sampled with nominal differences found between mobile and landline samples.[23] Neighborhood characteristics were collected from the Census' American Community Survey 2007-2012 estimates at zip code level. UHF neighborhood characteristics were then evaluated as population weighted values aggregated among zip codes encompassed by each UHF. To improve measurement of environment corresponding to each respondent, we will conduct analyses with individuals linked to zip code.

As missed needed care occurs outside the scope of the medical system, we defined 'missed needed care' as an adult's perception of having a medical condition requiring treatment for effective recovery, coupled with non-receipt of medical services. CHS respondents responded 'yes' or 'no' to the survey item: "Was there a time in the past 12 months when you needed medical care but did not get it?" Our dependent variable was defined categorically; values of 1 were given where the respondent reported not getting needed care in the previous 12 months and 0 when an occasion of missed needed care did not occur. Independent variables were included for both individual characteristics and neighborhood characteristics. Respondents reported insurance status, income, employment status, age, cohabitation with adults and children, marital status, education, foreign born status, and whether they had previously been told they had depression. Neighborhood characteristics matched to zip code and UHF include poverty rate, crowding, median income and housing vacancy rates.

Our results suggest that tree-based adaptive statistical learning methods like BRT can improve model prediction accuracy through estimation of complex relationships and interactions in both location demographic characteristics. Through use of BRT, neighborhood characteristics were shown to have high degree of relative influence on optimal prediction of missed necessary medical treatment among New York City residents. Health insurance enrollment, while important, is not the only factor useful for predicting which individuals missed needed medical care. Recognition that additional factors, including prior depression, socioeconomic characteristics, and neighborhood characteristics predict missed care among insured adults may be a significant consideration toward continued health policy refinement.

Figure 1. Relative influence of ten most influential predictor variables for BRT estimation.

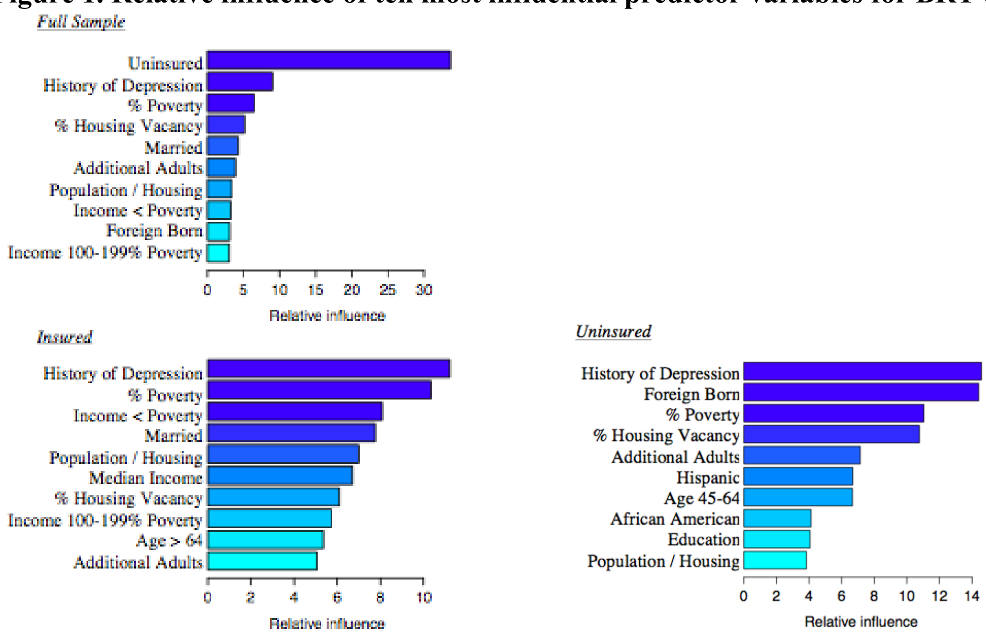
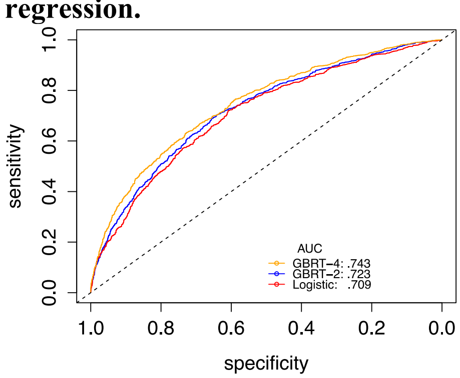


Figure 2. Comparison of classification performance for BRT with 2 and 4 tree-splits and logistic regression.



References

1. Weinick RM, Zuvekas SH, Cohen JW. Racial and ethnic differences in access to and use of health care services, 1977 to 1996. Medical Care Research and Review. 2000;57(suppl 1):36-54.
2. Weinick RM, Zuvekas S, Drilea SK. Access to health care: sources and barriers, 1996: US Department of Health and Human Services, Public Health Service, Agency for Health Care Policy and Research; 1997.
3. Zuvekas SH, Taliaferro GS. Pathways to access: health insurance, the health care delivery system, and racial/ethnic disparities, 1996–1999. Health Affairs. 2003;22(2):139-53.
4. Shi L, Macinko J, Starfield B, Politzer R, Xu J. Primary care, race, and mortality in US states. Social Science & Medicine. 2005;61(1):65-75.
5. Starfield B, Shi L, Macinko J. Contribution of primary care to health systems and health. Milbank quarterly. 2005;83(3):457-502.
6. Diez-Roux AV, Nieto FJ, Muntaner C, Tyroler HA, Comstock GW, Shahar E, et al. Neighborhood environments and coronary heart disease: a multilevel analysis. American Journal of Epidemiology. 1997;146(1):48-63.
7. Lovasi GS, Quinn JW, Neckerman KM, Perzanowski MS, Rundle A. Children living in areas with more street trees have lower prevalence of asthma. Journal of Epidemiology and Community health. 2008;62(7):647-9.
8. Adler NE, Boyce T, Chesney MA, Cohen S, Folkman S, Kahn RL, et al. Socioeconomic status and health: the challenge of the gradient. American psychologist. 1994;49(1):15.
9. Lovasi GS, Bader MD, Quinn J, Neckerman K, Weiss C, Rundle A. Body mass index, safety hazards, and neighborhood attractiveness. American journal of preventive medicine. 2012;43(4):378-84.
10. Pratt LA, Brody DJ. Depression in the United States household population. Age. 2008;18:39.
11. Benton T, Staab J, Evans DL. Medical co-morbidity in depressive disorders. Annals of Clinical Psychiatry. 2007;19(4):289-303.
12. Wells KB, Stewart A, Hays RD, Burnam MA, Rogers W, Daniels M, et al. The functioning and well-being of depressed patients: results from the Medical Outcomes Study. Jama. 1989;262(7):914-9.
13. Friedman JH. Greedy function approximation: a gradient boosting machine. Annals of Statistics. 2001:1189-232.
14. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees: CRC press; 1984.
15. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The annals of statistics. 2000;28(2):337-407.
16. Schapire RE. The boosting approach to machine learning: An overview. Nonlinear estimation and classification: Springer; 2003. p. 149-71.
17. Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. Journal of Animal Ecology. 2008;77(4):802-13.
18. Hastie T, Tibshirani R, Friedman J, Hastie T, Friedman J, Tibshirani R. The elements of statistical learning: Springer; 2009.
19. Friedman JH, Popescu BE. Predictive learning via rule ensembles. The Annals of Applied Statistics. 2008:916-54.
20. Ridgeway G. gbm: generalized boosted regression models. R package version 1.6-3.1. 2010.
21. R Development Core Team. R: A language and environment for statistical computing. 2012.
22. New York City Department of Health and Mental Hygiene. Community Health Survey [2009, 2010].
23. Corey C, Eisenhower D, Immerwahr S, Konty K, Norton JM, Sanderson M. Including New Yorkers Who Can Only Be Reached by Cell Phones in the Survey: Results from the 2008 Cell Phone Pilot Survey. Epi Research Report 2010. 2010.