# Predicting Outcomes of Hospitalization for Heart Failure Using Logistic Regression and Knowledge Discovery Methods

**Kirk T. Phillips, MSW, MS, PhD (candidate)[1]**
**William Nick Street, PhD[2]**
**[1] Interdisciplinary Health Informatics Program, University of Iowa**
**[2]Tippie College of Business, University of Iowa**

**Abstract**
The purpose of this study is to determine the best prediction of heart failure outcomes, resulting from two methods -- standard epidemiologic analysis with logistic regression and knowledge discovery with supervised learning/data mining. Heart failure was chosen for this study as it exhibits higher prevalence and cost of treatment than most other hospitalized diseases. The prevalence of heart failure has exceeded 4 million cases in the U.S.. Findings of this study should be useful for the design of quality improvement initiatives, as particular aspects of patient comorbidity and treatment are found to be associated with mortality. This is also a proof of concept study, considering the feasibility of emerging health informatics methods of data mining in conjunction with or in lieu of traditional logistic regression methods of prediction. Findings may also support the design of decision support systems and quality improvement programming for other diseases.

**Description**
An integrated data set was developed from three sources of administrative data representing 2,500 hospitalized heart failure patients treated in eight Iowa hospitals. Insurance claims were used to describe patient demographics, diagnoses and treatments. ORYX data submitted to the Joint Commission and CMS were used to describe care processes including use of left ventricular contractile function tests, appropriate administration of ACE inhibitors and performance of specific elements of discharge education. Death records were used to ascertain 30 day post discharge mortality. A survival distribution was formed with mortality records, revealing that 43% of all heart failure deaths in this study occurred within 30 days after discharge from the hospital.

All diagnoses were pre-classified with single level Clinical Classification Software (CSS)[(1)] as a means of reducing more than 12,000 ICD-9-CM diagnostic codes to fewer than 260 mutually exclusive categories. These diagnoses could appear in any of 25 fields in the data set. Comorbidities were tested with special algorithms developed by Elixhauser et al, designed to predict mortality [(2)]. Up to 25 procedure fields were available on the hospital bill. Use of the CCS reduced these codes from a possible 3,500 procedure codes to 231 categories. Several steps were applied to prepare the data for analysis, in addition to integration and pre-classification noted above. Univariate tables were constructed for each predictor and odds ratios were calculated with the outcome variable. Variables were recoded and dummy variables were created, per standard logistic regression methods[(3)]. Logistic regression was performed with a parsimonious set of predictors selected with statistical criteria using odds ratios, maximum likelihood estimates and others, as well as their biologic or practical association with outcomes. Strong predictors included age, number of prior hospitalizations, type of admission, appropriate use of ACE inhibitors, and discharge instructions including appropriate activity level. Predictors also included comorbidities of neurological disorders, renal failure, coagulopathy, obesity, weight loss, fluid/electrolyte disorders, and blood loss anemia.

Data mining routines of decision trees, neural networks and others are currently being executed to compare with predictions derived from established epidemiologic methods described earlier. An evaluation process will include selection of resulting models with lower levels of error in predicting hospital readmission and mortality.

Knowledge discovery databases with data mining present opportunities for organizing and analyzing large databases, including data from existing administrative sources. Findings of this research are intended to contribute to an emerging body of literature, which may suggest that data mining methods outperform multiple logistic regression and traditional epidemiological methods. This study will address trade-offs between the two methods, including technical accuracy, ease of interpretation, and clinical usefulness.

[(1)] Elixhauser A, Steiner C, Palmer L. *Clinical Classifications Software (CCS), 2004*. February 6, 2004. U.S. Agency for Healthcare Research and Quality.

[(2)] Elixhauser A. Steiner C. Harris DR. Coffey RM. Comorbidity measures for use with administrative data. *Medical Care. 36(1):8-27, 1998 Jan.*

[(3)] Hosmer, D.W., Lemeshow, S. *Applied Logistic Regression, 1989.* John Wiley & Sons, Inc.