Modeling Incidence and Severity of Disease using Healthcare Data

Lauren Nicole Geiser[1], Beverly Anne Setzer[2], Jason Cory Brunson PhD[3], Reinhard Laubenbacher PhD[3]
[1]University of Dayton, [2]North Carolina State University, [3]Center for Quantitative Medicine, UConn Health

Question

What can we learn about heart attack risk from routinely-collected administrative and billing records?

Introduction

The risk of having a myocardial infarction (MI) is dependent on numerous variables unique to each patient. However, in a clinical setting only a few are used to evaluate risk. Increasingly, new variables are considered, creating the potential to incorporate a patient's entire unique medical history into prediction for diseases.

We aim to build modeling infrastructure to incorporate these data to assist physicians in evaluation of patient risk. With these models we would also be able to simulate and forecast health trajectories for patients so that they can adapt their lifestyle in response.

MIMIC-III Database

Collection	Pre-Processing	Limitations
• Beth Israel Deaconess Medical Center • ICU admissions 2001-2012 • With admissions and billing data	• Maintained by MIT • De-identified • Accessible to researchers after training course	• Time window • Patient population • Date-shifting • Sparsity

Methodology

We use the generalized linear regression framework:

$$MI_i \sim Binom(\pi_i, n)$$
$$g(\pi_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon_i$$

MI_i : number of MI diagnoses
g(x) : log-odds link function
x : predictor
ε_i :error term
π_i : probability of MI
β_k : coefficient
n : number of trails

We use autoregression time series models for risk progression:

$$g(\pi_t) = z_t\,, \quad z_t = z_{t-1} + w_t, \quad w_t \sim wn(\sigma^2)$$

w_t: white noise

We use a Bayesian parameter estimation:

$$p(\theta|MI) = p(MI|\theta) * p(\theta)$$

$p(\theta)$: prior distribution
$p(MI|\theta)$: likelihood function
$p(\theta|MI)$:posterior distribution

Admission-Level Models

Predicting MI Occurrence

Goal: Predict MI diagnosis for a patient admitted to an ICU
Framework: Generalized linear regression
Observational Unit: Hospital admission (n=50,806)
Response Variable: Diagnosis of MI during admission

$$(1): g(\pi_i) = \beta_1 + \beta_H H_i + \beta_{H`} H`_i + \beta_L L_i + \beta_{L`} L`_i + \beta_D D_i + \beta_{D`} D`_i + \beta_O O_i + \beta_{O`} O`_i + \beta_A A_i + \beta_G G_i + \beta_E E_i + \varepsilon_i$$
$$(2): g(\pi_i) = \text{Model 1 predictors} + \beta_{Labs} Labs_i$$
$$(3): g(\pi_i) = \text{Model 2 predictors} + \beta_{CPK} CPK_i + \beta_{LDH} LDH_i + \beta_{Trop} Trop_i$$

Concurrent and Prior (') Diagnoses	Demographics	MI-Related Lab Tests
Hypertension, Lipid metabolism disorders, Diabetes, and Obesity	Age, Gender, and Ethnicity	Creatine Phosphokinase, Lactate Dehydrogenase, and Troponin

Model	Predictors	AUC	Odds Ratio(s)
(1)	Diagnoses, demographics	0.684	
(2)	(1) + indication of lab testing	0.730	OR(*Labs*) = 18.2
(3)	(2) + test flags	0.751	OR(*CPK*) = 2.2 OR(*LDH*) = 1.0 OR(*Trop*) = 2.6

Predicting In-Hospital Mortality for MI Patients

Goal: Conduct follow-up analysis to a study [1] examining the impact of lab tests and ethnicity data on predicting in-hospital mortality for MIs
Framework: Generalized linear regression
Observational Unit: Hospital admission (n=5,148)
Response Variable: In-hospital mortality during admission

$$(1): g(\pi_i) = \beta_1 + \beta_{DRG} DRG_i + \beta_A A_i + \beta_G G_i + \varepsilon_i$$
$$(2): g(\pi_i) = \text{Model 1 predictors} + \beta_{Sev} Sev_i$$
$$(3): g(\pi_i) = \text{Model 2 predictors} + \beta_E E_i$$

	Lim et. al. [1]	Present Study
Data	20 hospitals across HI	MIMIC-III
Conventional risk score	Risk of Mortality class	Disease-Related Group mortality score
Lab severity score	25 lab tests	3 lab tests

Model	Lim et. al. [1]	AUCs	Present Study	AUCs
(1)	ROM class, age, gender	0.845	DRG score, age, gender	0.740
(2)	(1) + lab severity score OR(*Sev*)=5.5	0.869	(1) + lab severity score OR(*Sev*)=3.1	0.744
(3)	(2) + ethnicity	0.872	(2) + ethnicity	0.742

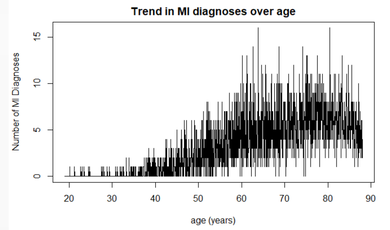
[1] Lim, E., Cheng, Y., Reuschel, C., Mbowe, O., Ahn, H. J., Juarez, D. T., Chen, J. J. (2015). Risk-Adjusted In-Hospital Mortality Models for Congestive Heart Failure and Acute Myocardial Infarction: Value of Clinical Laboratory Data and Race/Ethnicity. Health Services Research, 50(Suppl 1), 1351–1371. http://doi.org/10.1111/1475-6773.12325

Funded by the National Science Foundation award #1460967

Time Series Models

Modeling MI Risk Progression with Age

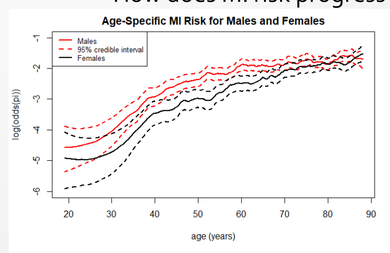
Goal: Predict MI risk progression with age
Observational Unit: Population ICU admissions at each month of life
Response Variable: Diagnosis of MI during admission



Data:
- Aggregated patient data by month of age
- Lose patient-level correlations

Framework: Generalized linear regression and time series modeling

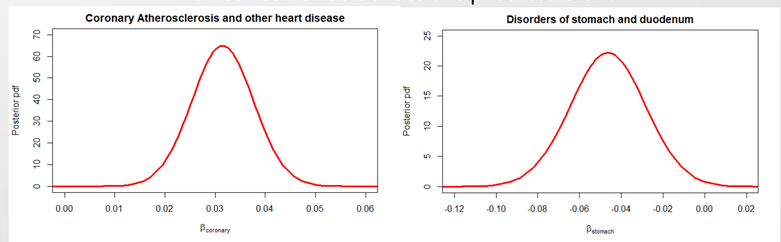
How does MI risk progress with age for different genders?



$$g(\pi_t) = z_t\,, \quad z_t = z_{t-1} + w_t, \quad w_t \sim wn(\sigma^2)$$

- Males are at a higher risk of MI for most of their lives
- After age 80 genders have a similar risk
- Risk progresses and varies at a similar rate

Which concurrent diseases are predictors of an MI?



$$g(\pi_t) = \beta_1 x_1 + \cdots + \beta_k x_k + z_t, \quad z_t = z_{t-1} + w_t, \quad w_t \sim wn(\sigma^2)$$

- Capturing diseases with similar progression over age rather than risk factors because of aggregated data
- Need data to be longer-term and not randomly date-shifted to detect patient-level effects and population-level effects over time

Conclusion

We were able to create data-driven risk models and model pipelines that can be generalized and applied to subsequent projects, but were limited by the scope and sparsity of the data. Future work will include fitting time series models to longer-term, more consistent data sets; expanding to multivariate time series models; and combining diagnosis-specific risk models into a visual reference tool.