Using the SAS® System to Develop Risk Prediction Models for Patient Re-admission Reduction in VAMCs

Issac Shams

Veteran Engineering Resource Center (VERC)-VA-CASE
Healthcare Systems Engineering Research Group
Department of Industrial and Systems Engineering
Wayne State University

MISUG, October 2012

OVERVIEW

- Motivation and Statement
- Preliminary Analysis
- Modeling Strategy and details
- Result and outcomes

BACKGROUND AND MOTIVATION

- 17 percent of Medicare patients discharged from the hospitals have a *readmission* (*rehospitalization*) within 30 days of discharge, accounting for $15 billion in spending (Medicare Payment Advisory Commission Report, 2007).

- Under Obama Care Rule (Patient Protection and Affordable Care Act or PPACA), about two-thirds of U.S. hospitals stand to be penalized for excess readmissions starting Oct. 1, 2012.

- Based on the formula developed by CMS (Centers for Medicare & Medicaid Services), this will account for 2,211 hospitals a cumulative $280 million cut in Medicare funds (1%) due to high rates of 30-day readmission after discharge. For hospitals that don't improve, penalties will grow to a maximum of 2% for the 2014 program year and 3% for 2015 (American Medical News, Aug. 27, 2012).

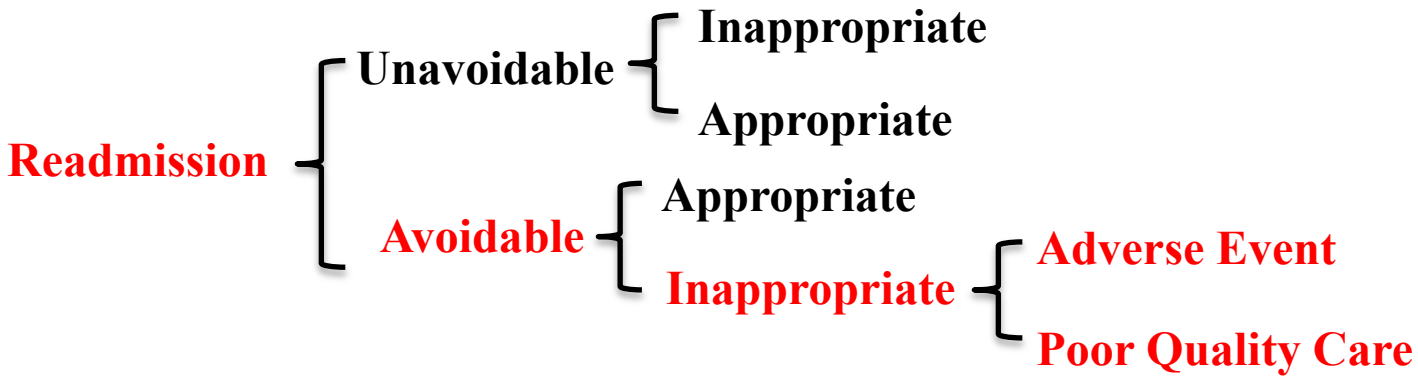
Definition and Categorization

- Readmission (in VAMCs):

The proportion of patients who were readmitted (for any cause) to the acute care wards of the hospital within 30 days following discharge from the hospital.

$$\frac{\text{\# 30 Day Readmits}}{\text{\# Discharges in 30 Day Window}}$$

- Some categories of potential readmission:



STATEMENTS AND OBJECTIVES

- Data were collected from John D. Dingell VAMC during September and October 2006 through 2011.
- There were 3108 records, 2449 patient id (subject), and 50 prognostic factors. Except for 'DOB', 'admdate' and 'disdate', all others were in nominal scale.
- Investigate the relationships between set of patient prognostic factors and his/her readmission risk.
- Utilize the predictive models to filter out significant variables for internally categorizing patients in terms of their hazard to rehospitalization and adjust the budget allocation accordingly.

PRELIMINARY STEPS

- Data collection and extraction using SAS SQL
- Data structure and exploration using Base SAS

```
PROC FREQ < DATA= > ;
TABLES  requests < / chisq> ;
RUN;
```

Results: 94.72% male, 89.74% without insurance, 99.26% veterans, 96.59% inpatients, etc.

'Admsource' ⟶ 93.85% Hospital, 6.15% NHCU

Tentative clusters of features:

1. Demographic **2.** Financial Situation

3. War-Related **4.** Admission-Related

5. Health Care Related

PRELIMINARY STEPS (CONT'D)

- Data Quality problem handling with Base SAS

1. **Duplications**

```
DATA readmission;
MERGE radmsep06--radmoct11;
BY ssn;
run;
```

2. **Outliers** (for 'los' only)

```
PROC MEANS < DATA= > <min max  median std  p1 p5 p95 p99> ;
VAR  variable(s) < >;
run;
```

3. **Missing Values** (again with **PROC FREQ**)

High-missing-value attributes:

'agentl'(93.27%), race(52.70%), 'pct'(24.55%), 'providerpclass'(22.10%)

DATA PRE-PROCESSING

- Data cleansing with Base SAS

'viet'=YES and 'vet'=NO ! – 'opatientind'=YES and 'pows'=YES !

- Feature creation with Base SAS

los = admdate – disdate;

disage = (disdate – dob)/365.25;

seq: if first.id then seq=1; else seq+1;

rtime:

```
data want (drop=_:);
set have;
set have (firstobs = 2 keep = admdate rename = (admdate = _admdate2))have(obs = 1 drop = _all_);
rtime=ifn(missing(_admdate2) or last.id,31,_admdate2-disdate);
run;
```

rstatus: if rtime=31 then rstatus=0; else rstatus=1;

ANALYSIS APPROACHES

Since

1. The response variable 'rtime' is discrete and display timing of event (readmission),
2. There were lots of (near 88%) patients who had not re-admitted, and we did not want to discard all of them, and
3. There were so many repeated admissions for some specific patients;

we used survival analysis in SAS/STAT.

'rtime' is a response: [0,31]

'rstatus' is 0-1 censoring ind. (right censored, Singly Type I)

Origin of time ⟶ 'disdate'

- Relevant procedures in SAS :

lifetest, lifereg, phreg, nlmixed

MODELING CHARACTERISTICS

Nonlinear Mixed Effect Model aka Hierarchical Nonlinear Model

- Widely used in pharmacokinetics (Row ;1997) , HIV Dynamics (Wu and Ding;1999), Forestry (Fang and Bailey;2001) and etc.
- Response evolves over *time, within* individual profiles of *repeated events* and these do *vary* in the cohort.
- *Inter-patient variations* of re-admission process and how these *vary* across subjects (unique SSN) can, *simultaneously*, be elucidated.
- Both *fixed* and *random* effects can enter *nonlinearly*, and response (given the random effects) can have lots of conditional distributions.
- Inference is based on *marginal likelihood* and maximizing its approximation over the random effects.

MODELING FRAMEWORK

- $\mathbf{y}_i$: observed data vector for subject $i = 1, \ldots, s$
- $\mathbf{u}_i$: latent random vector for within-subject covariance

$\mathbf{y}_i$ and $\mathbf{u}_i$ are independent across i

- There is an appropriate model linking $\mathbf{y}_i$ and $\mathbf{u}_i$ which leads to joint density function

$$p(\mathbf{y}_i|\mathbf{X}_i, \phi, \mathbf{u}_i)q(\mathbf{u}_i|\xi)$$

- $\mathbf{X}_i$: matrix of observed explanatory variables
- ϕ and ξ are vectors of unknown parameters

Let $\theta = [\phi, \xi]$, then inference on θ is based on the marginal likelihood function

$$m(\theta) = \prod_{i=1}^{s} \int p(\mathbf{y}_i|\mathbf{X}_i, \phi, \mathbf{u}_i)q(\mathbf{u}_i|\xi)\, d\,\mathbf{u}_i$$

Then $f(\theta) = -\log m(\theta)$ is minimized over θ numerically in order to estimate θ, and the inverse Hessian matrix at the estimates provides an approximate variance-covariance matrix for the estimate of θ.

MODELING ASSUMPTIONS (FRAILTY MODEL)

Hazard for the *j*th re-admission for *i*th patient at time *t* is governed by

$$\log h_{ij}(t) = \alpha(t) + \boldsymbol{\beta X}_{ij} + \varepsilon_i$$

$\boldsymbol{\varepsilon_i}$ represents unobserved heterogeneity, and is subscripted by *i* not by *j*.

It has, here, a normal density with a mean of 0 and a variance σ^2.

When events are repeated, such models are not highly sensitive to the choice of a distribution for ε (Kelin,1992; McGilchrist,1993).

$\alpha(t)$ represents the baseline hazard function, describes how the risk of re-admission changes over time at baseline levels of covariates ($\boldsymbol{\beta} = \mathbf{0}$).

$\boldsymbol{\beta}$ represents the effect parameters, uncovers how the re-admission risk varies in response to explanatory covariates.

SETTINGS AND IMPLEMENTATIONS

NLMIXED procedure

- It has MLE of non-linear mixed models.
- In MODEL statement, there are bunches of conditional distribution to choose or you can define a general log likelihood function.
- It's better to have some initial guesses for parameter estimates in PARMS.
- It has different integral approximation methods as well as lots of optimization techniques for dealing with minimizing the marginal log-likelihood function.

How to find out $\alpha(t)$?

1. Non-parametric: KDE with bootstrap, NPEB (Robbins, Herbert; 1956)
2. Parametric:

Exponential: $\log h(t) = \mu + \beta \mathrm{X}$ 

Gompertz: $\log h(t) = \mu + \alpha t + \beta \mathrm{X}$

Weibull: $\log h(t) = \mu + \alpha \log t + \beta \mathrm{X}$

SETTINGS AND IMPLEMENTATIONS (CONT'D)

How to implement Frailty model into HNLMMs framework? (Allison, 2010)

Let t_i be the event (or censoring) time, and δ_i be the censoring indicator

Define $\lambda_i = \exp(\boldsymbol{\beta X}_i + \varepsilon_i)$

Based on the Weibull baseline hazard, the single patient's contribution to the log-likelihood function (condition on ε) is:

$$\log L_i = -\lambda_i t_i^{(\alpha+1)} + \delta_i \{\log(\alpha+1) + \alpha \log t_i + \log \lambda_i\}$$

and $\alpha = 0$ for the exponential distribution.

In SAS we have:

```
ll= – lambda*nrtime**(alpha+1)+rstatus*(LOG(alpha+1)+alpha*LOG(nrtime)
+LOG(lambda));
MODEL nrtime~GENERAL(ll);
```

DIFFICULTIES AND SOLUTIONS

D1. Near all covariates were categorical and in nominal scale, and proc nlmixed, until now, does not have CLASS statement.

S1. Generate a non-singular one with PROC LOGISTIC.

PROC LOGISTIC data=< > outdesign=< > outdesignonly;

CLASS admsource(ref='NHCU') mar(ref='NEVER MARRIED') userenrollee (ref='NO') pheartind(ref='YES');

D2. Overspecification: 'pct' had 54 values, 'DRG' had 87 values, 'pdiagnosis' had 65 values, etc. In a fully specified survival model, there should be 10-15 outcomes (readmission) per degree of freedom (Harrell, 2010). So we needed more than 2000 readmissions in the data set but we had only 372.

S2. Combine levels of 'pdiagnosis' to a higher level and transform 'DRG' to Major Disease Classification 'mdc' , based on ICD09 index.

- DRG: 570-585 and 592-607 belongs to MDC09, then new levels are 25.
- 'pdiagnosis': 715.96, 541, 250.12, etc. are associated with 'DISEASES OF THE BLOOD AND BLOOD-FORMING ORGANS' , resulting in only 15 new levels.

DIFFICULTIES AND SOLUTIONS (CONT'D)

D3. Model-selection methods (like Forward Selection, Backward Elimination, etc.) are not provided in proc nlmixed.

S3. Since there were no continuous attributes, Correlation Feature Selection (CFS) and minimum-redundancy-maximum-relevance (mRMR) techniques were used to shrink the pool of attributes to less than 10.

A. Explore class variables interactions and continuous-by-class interactions

PROC FREQ <*data*= >;
tables (admsource ward--elig enrollp enrolls)*(pct--pdiagnosis providerpclass) / chisq;
run;

PROC GLM <*data*= >;
class mtest;
model disage=mtest/solution;
run;

B. Drop redundant effects that carries the same meaning and functionality

'enrollp' is determined by 'elig'; 'enrolls' is digged out from 'mtest' and 'elig'

DIFFICULTIES AND SOLUTIONS (CONT'D)

D4. The final Hessian matrix is full rank but has at least one negative eigenvalue. Second-order optimality condition violated.

S4. There are some suggestions provided (**SAS Global Forum 2012;** Paper 332-2012):

1. Rescale your data so the values are on a similar scale

```
slos=(los-7.653)/2.8814713;
nrtime=rtime/10;
```

2. Reparameterize the model. Use a standard deviation rather than the variance.

```
RANDOM e~normal(0,sd*sd) SUBJECT=id;
```

D5. Find the most appropriate functional form of each effect.

S5. Did some plotting efforts and best-guess approaches. For example 'sloglos' is used instead of 'los'.

RESULTS

THE NLMIXED PROCEDURE

Specifications

Data Set	WORK.X
Dependent Variable	nrtime
Distribution for Dependent Variable	General
Random Effects	e
Distribution for Random Effects	Normal
Subject Variable	id
Optimization Technique	Dual Quasi-Newton
Integration Method	Adaptive Gaussian Quadrature

NOTE: GCONV CONVERGENCE CRITERION SATISFIED

Fit Statistics

-2 Log Likelihood	2973.9
AIC (smaller is better)	2987.9
AICC (smaller is better)	2987.9
BIC (smaller is better)	3028.5

RESULTS

PARMS b0=0 badmsource=0 bseqadmsource=0 bseqsloglos=0 b1maruser=0 b2maruser=0 sd=1 ;

Parameter Estimates

Parameter	Estimate	St. Error	DF	t Value	Pr > |t|	Lower	Upper	Gradient
b0	-8.7573	0.4963	2443	-17.65	<.0001	-9.7304	-7.7841	0.000423
badmsource	4.7175	0.4359	2443	10.82	<.0001	3.8628	5.5723	0.000473
bseqadmsource	-1.1979	0.1176	2443	-10.19	<.0001	-1.4284	-0.9673	0.000622
bseqsloglos	-0.1795	0.0533	2443	-3.37	0.0008	-0.284	-0.07495	-0.00064
b1maruser	-0.1581	0.1464	2443	-1.08	0.2805	-0.4452	0.1291	-0.0001
b2maruser	0.3321	0.1229	2443	2.7	0.0069	0.0911	0.573	-0.00037
sd	2.7614	0.1794	2443	15.39	<.0001	2.4096	3.1132	0.000738

- ✓ 'sd'→ There is definitely an unobserved heterogeneity across patients, or there is undoubtedly high dependence among repeated re-admissions.
- ✓ 'badmsource'→ Hazard of re-admission, controlling for other covariates, for those admitted in 'Hospital' is only about 4.7 % of the hazard for those admitted in 'NHCU'

RESULTS (CONT'D)

Empirical Correlation Matrix of Parameter Estimates

Parameter	b0	badmsource	bseqadmsource	bseqsloglos	b1maruser	b2maruser	sd
b0	1	-0.884	0.3763	0.08008	0.0191	-0.09193	-0.627
badmsource	-0.884	1	-0.512	-0.06851	-0.00529	0.0136	0.3243
bseqadmsource	0.3763	-0.512	1	0.1232	0.0138	-0.01987	-0.431
bseqsloglos	0.08008	-0.06851	0.1232	1	0.02835	0.009294	-0.0876
b1maruser	0.0191	-0.00529	0.0138	0.02835	1	-0.3486	0.00092
b2maruser	-0.09193	0.0136	-0.01987	0.009294	-0.3486	1	0.00578
sd	-0.627	0.3243	-0.431	-0.08761	0.000919	0.005782	1

Contrast (Delta Method; Cox 1998)

Label	Num DF	Den DF	F Value	Pr > F
mar*userenrollee	2	2443	4.14	0.0160

How well did the model work?

- Generalized R-Squared (Cox and Snell; 1989)

$$R^2 = 1 - \exp(\frac{-G^2}{n})$$

G^2 is the likelihood ratio chi-square statistics for null hypothesis that all covariates are set to zero.

For our modeling, it was **79.79%** !

- Model evaluation and robustness

7-Fold cross-validation

Thanks for your patience

Questions
Comments