

Predictive Modeling in Long-Term Care Insurance

Nathan Lally and Brian Hartman
University of Connecticut

Overview

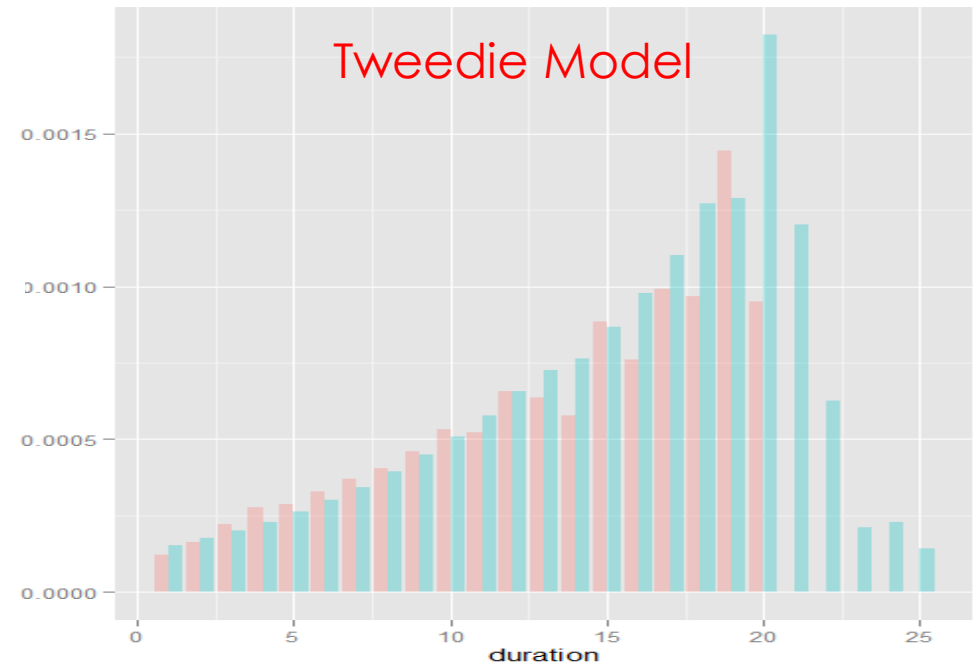
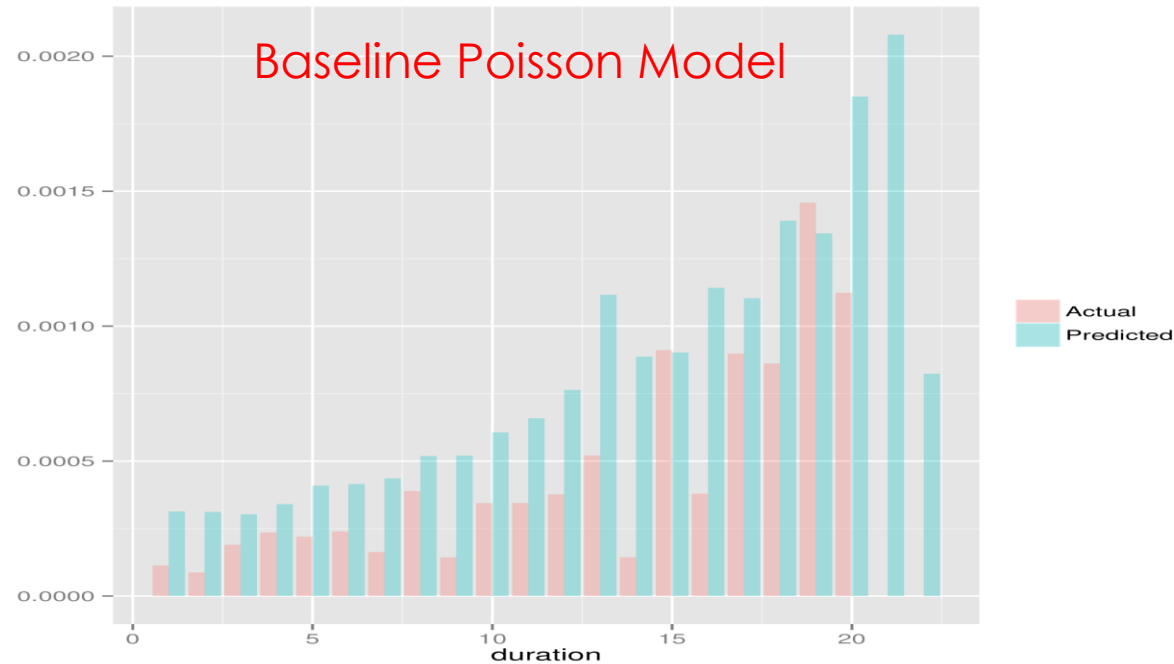
1. Project Background
2. The Data
3. Current Common Industry Methods: Poisson Regression
4. Problems with Poisson Regression for LTC Rate Data
5. Improved Methods
6. Metrics of Interest and Model Selection
7. Results
8. Discussion

Project Background

- Last year, the Goldenson Actuarial Research Center developed predictive models for long-term care insurance (LTCI) claims, mortality and lapse rates.
- Under the guidance of industry actuaries several Poisson regression models were constructed to predict the aforementioned rates.

Project Background

- Our Objective: Utilizing generalized linear and or additive models (GLM, GAM) construct models for LTCI rates that outperform the baseline Poisson models in terms of predictive accuracy.



The Data

Summary

- A major U.S. LTCI provided a large data set containing 13 years of aggregated LTCI policy information.
- The dataset is massive and contains 9,429,590 observations.
- After some cleanup at the suggestion of the insurer, we were left with approximately 7,750,000 observations.

Response Variables

- Lapse Count
- Mortality Count

Predictor Variables (Covariates)

- 22 Predictor Variables Included
 - 16 Categorical
 - 6 Continuous

Exposure Variables

- Exposure to lapse or mortality risk in months

Common Industry Method for LTC Rate Data: Poisson Regression with Log Link and Offset

- What is Poisson Regression?

$$Y \sim \text{Pois}(\mu) \quad , \quad g(\mu) = x' \beta$$

$$\ln(\mu) = x' \beta + \ln(t) \quad \dots \quad \mu = t e^{x' \beta}$$

- Poisson regression models are generalized linear models (GLM) designed to model count data.
- These models assume the response variable Y follows a Poisson distribution
- The log link function is typically used to relate the linear predictors to the mean
- We are interested in modeling $E[Y/t] = \mu/t$ where Y is a count of events and t is an exposure variable (in our case representing time in months). Therefore $\ln(t)$ is used as an offset in our models.

Common Industry Method for LTC Rate Data: Poisson Regression with Log Link and Offset

- 3 Main Assumptions Required for Poisson Regression
 1. **Perfect homogeneity** throughout the sample (the rate parameter is the same for each unit of exposure in a given observation).
 2. Each unit of exposure generates events (e.g. claims, lapses, or deaths) in accordance with a **Poisson process**.
 3. Response variable outcomes are **mutually independent** for all observations

Problems with Poisson regression for LTCl Rate Data

1. Over-abundance of zeros

- The data displays more zeros than one would expect given they come from a Poisson process

2. Overdispersion

- The Poisson regression model is a single parameter model where $E(Y_i) = \text{Var}(Y_i)$.
- Often with real data sets $E(Y_i) < \text{Var}(Y_i)$ (overdispersion).
- In these cases we could use models with more free parameters that relate the expectation and the variance.
 - Ex: $E(Y_i) = \theta * \text{Var}(Y_i)$, where θ is a dispersion parameter

Problems with Poisson regression for LTCI Rate Data

- Testing for overdispersion with Lagrange multiplier test

Mortality Rates		
Sample Through Year N	Test Statistic*	P-Value
2002	11.19693	0.000819
2003	4.376259	0.036443
2004	3.007823	0.082864
2005	1.84618	0.174228
2006	2.076497	0.149583
2007	4.034079	0.04459
2008	2.198942	0.138105
2009	3.343193	0.067484
2010	2.399046	0.121409
2011	2.773524	0.095835
2012	2.106376	0.146686

Lapse Rates		
Sample Through Year N	Test Statistic*	P-Value
2002	669190.6	0
2003	777822.2	0
2004	821966.8	0
2005	863701.5	0
2006	1125427	0
2007	1072929	0
2008	1009507	0
2009	1003570	0
2010	1004291	0
2011	1028645	0
2012	1103502	0

*The Test Statistic for the Lagrange multiplier test is distributed χ^2 with $df = 1$

Problems with Poisson regression for LTCl Rate Data

3. Properties of the Poisson Distribution

- The Poisson regression model can predict counts from zero to infinity.
- This means our predicted rate defined by $\left(\frac{\text{Predicted Count}}{\text{Exposure time}}\right)$ can exceed 100%.

Improved Methods: Multi-parameter GLM & GAM Models designed to handle excess zeros and overdispersion

GLM & GAM Error Structures Considered

- Negative-Binomial Regression
- Zero-Inflated Poisson Regression
- Tweedie Regression
- Generalized Additive Models (GAMs) with Above Error Structures

Statistical Learning Techniques Considered

- Random Forest Regression

Negative Binomial Regression

- Negative binomial regression is similar to Poisson regression in that it models count data therefore we also require an offset variable to model rates.
- However, negative binomial models include a shape parameter θ that helps address the problem of overdispersion.

$$p(y_i) = \frac{\Gamma(y_i + \theta)}{\Gamma(\theta)y_i!} \mu_i^\theta (1 - \mu_i)^{y_i}, \quad \theta > 0, \quad y_i = \{0, 1, 2, \dots, \infty\}, \quad \mu_i = \frac{\theta}{\theta + \lambda_i}$$

where

$$E[y_i] = \lambda_i \quad \text{and} \quad \text{Var}[y_i] = \lambda_i \left(1 + \frac{1}{\theta} \lambda_i \right)$$

consider $\frac{\text{Var}[y_i]}{E[y_i]} = 1 + \frac{1}{\theta} \lambda_i > 1$ as a measure of overdispersion

Zero-Inflated Poisson Regression (ZIP)

- The Zero-Inflated model (ZIP) accounts for extra-Poisson zeros by assuming there are two processes at work that can generate zeros in a sample.
- One process generates only zeros and occurs with probability p .
- the other process, occurring with probability $(1 - p)$, generates events according to a Poisson distribution with mean λ . The result is a distribution in the form,

$$P(Y_i = 0) = p_i + (1 - p_i)e^{-\lambda_i}$$

$$P(Y_i = k) = \frac{(1-p_i)e^{-\lambda_i}\lambda_i^k}{k!}, \text{ where } k = \{1, 2, \dots, \infty\}$$

Tweedie Regression

- The Tweedie family are exponential dispersion models which include a set of compound Poisson-gamma distributions.
- Suggested for modeling semi-continuous data (positive point mass at zero).
- A convenient parameterization of the Tweedie distribution is given below where μ is the location parameter, σ^2 is the diffusion parameter, and p is the power parameter.

$$f(y|\mu, \sigma^2, p) = a(y|\sigma^2, p)e^{-\frac{1}{2\sigma}d(y|\mu, p)}$$

where

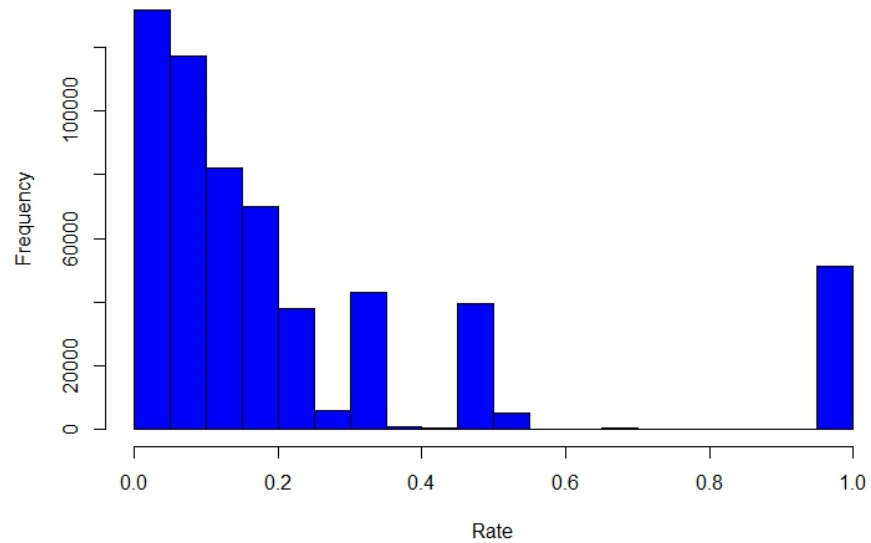
$$\text{Var}[Y] = \sigma^2 \text{E}[Y]^p = \sigma^2 \mu^p$$

- When $1 < p < 2$ this corresponds to a compound Poisson-gamma distribution

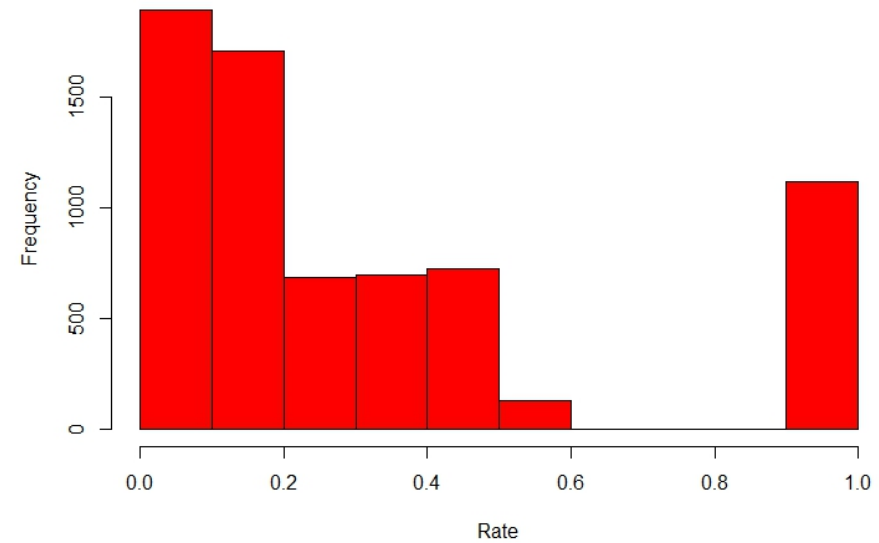
Tweddie Regression

- We take the arrival of events (lapse or death) to be Poisson distributed while the non-zero rates are assumed to be gamma distributed.

Histogram of Non-Zero Lapse Rates

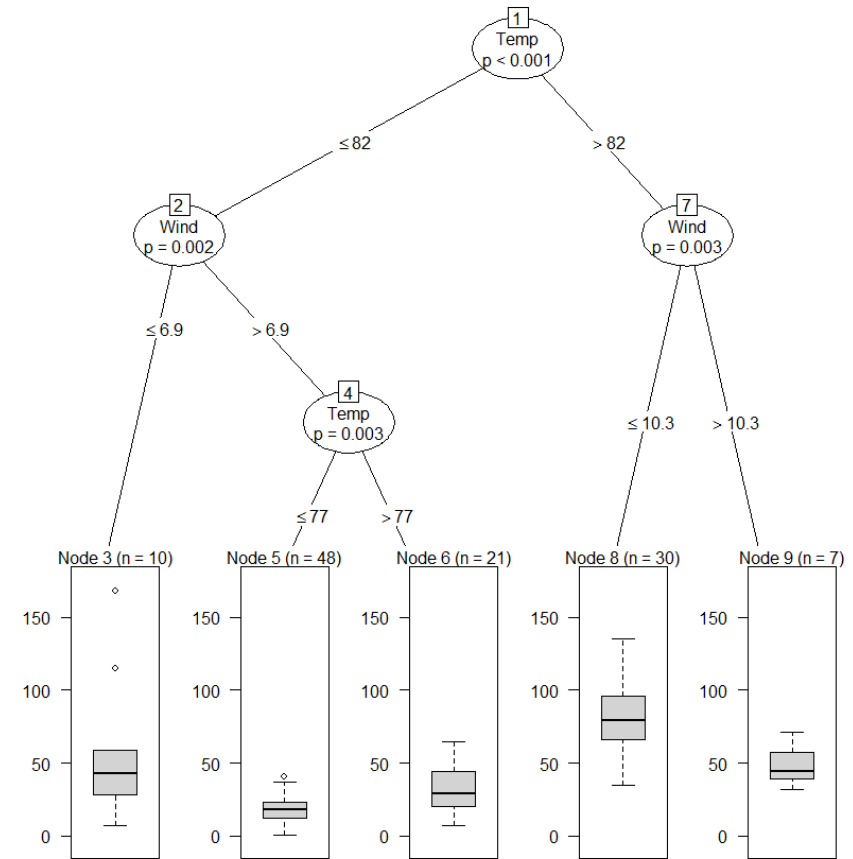


Histogram of Non-Zero Mortality Rates



Random Forest Regression

- Random forests are an ensemble learning method for classification and regression
- Based on decision trees
- Many trees are grown on random subsets of the training data.
- The trees select random subsets of the features in the data.
- Predictions from each individual tree are averaged through a process called bagging (bootstrap aggregation)
- Research suggests random forests do not over fit (Breiman 2001, Biau 2012)



Metrics of interest & Model Selection

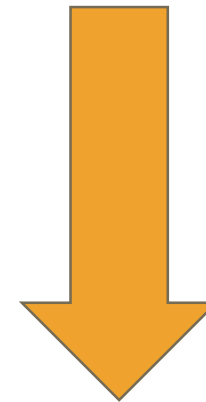
Let \mathbf{x} be a vector such that $\mathbf{x} = [(\hat{y}_1 - y_1), (\hat{y}_2 - y_2), \dots, (\hat{y}_n - y_n),]$
where \hat{y}_i and y_i are the i th predicted and observed responses respectively

- Weighted Median Absolute Prediction Error
 - *Weighted.Median*($|\mathbf{x}|, \boldsymbol{\omega}$) where $\boldsymbol{\omega} = \text{exposure}_1, \text{exposure}_2, \dots, \text{exposure}_n$
- Weighted Median Squared Prediction Error
 - *Weighted.Median*($\mathbf{x}^2, \boldsymbol{\omega}$) where $\boldsymbol{\omega} = [\text{exposure}_1, \text{exposure}_2, \dots, \text{exposure}_n]$
- Weighted Mean Absolute Prediction Error
 - $= \frac{\sum_{i=1}^n \omega_i |x_i|}{\sum_{i=1}^n \omega_i}$
- Weighted Mean Squared Prediction Error
 - $= \frac{\sum_{i=1}^n \omega_i x_i^2}{\sum_{i=1}^n \omega_i}$

Metrics of interest & Model Selection

- How models were compared and selected?
 - GLM models were fit to the first n years of data and used to predict the $(n+1)^{st}$ year to test out of sample performance.
 - Models were chosen which minimized the prediction error statistics and which improved when applied to larger subsets of the data.

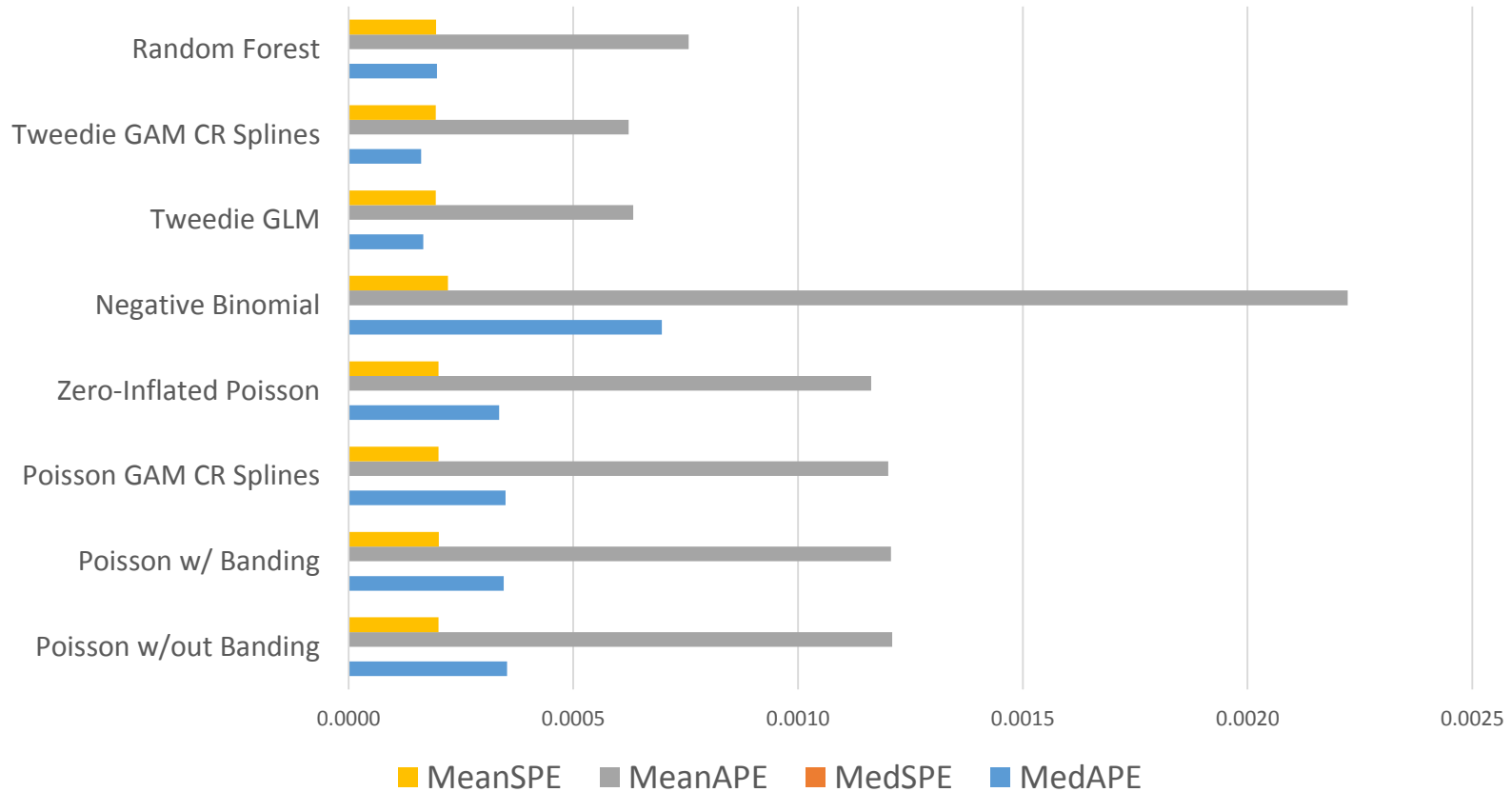
	Year	Poisson	Zero-Inflated Poisson	Negative Binomial	Tweedie
MedianAPE	Year 3	0.04657917	0.04807680	0.05274269	0.01902117
MedianSPE	Year 3	0.00216962	0.00231138	0.00278179	0.00036181
MeanAPE	Year 3	0.14742651	0.14774762	0.16277423	0.03787409
MeanSPE	Year 3	0.28991497	0.27422015	0.38704965	0.01017436
MedianAPE	Year 4	0.04579718	0.04714716	0.05035809	0.01756052
MedianSPE	Year 4	0.00209738	0.00222285	0.00253594	0.00030837
MeanAPE	Year 4	0.13383555	0.13410427	0.14535156	0.03625792
MeanSPE	Year 4	0.19016869	0.18644882	0.24994383	0.01040087
MedianAPE	Year 5	0.04342123	0.04480840	0.04772306	0.01673459
MedianSPE	Year 5	0.00188540	0.00200779	0.00227749	0.00028005
MeanAPE	Year 5	0.12449863	0.12501742	0.13540464	0.03554705
MeanSPE	Year 5	0.15985148	0.15752921	0.21831618	0.01038227



Results

Mortality Rate Models

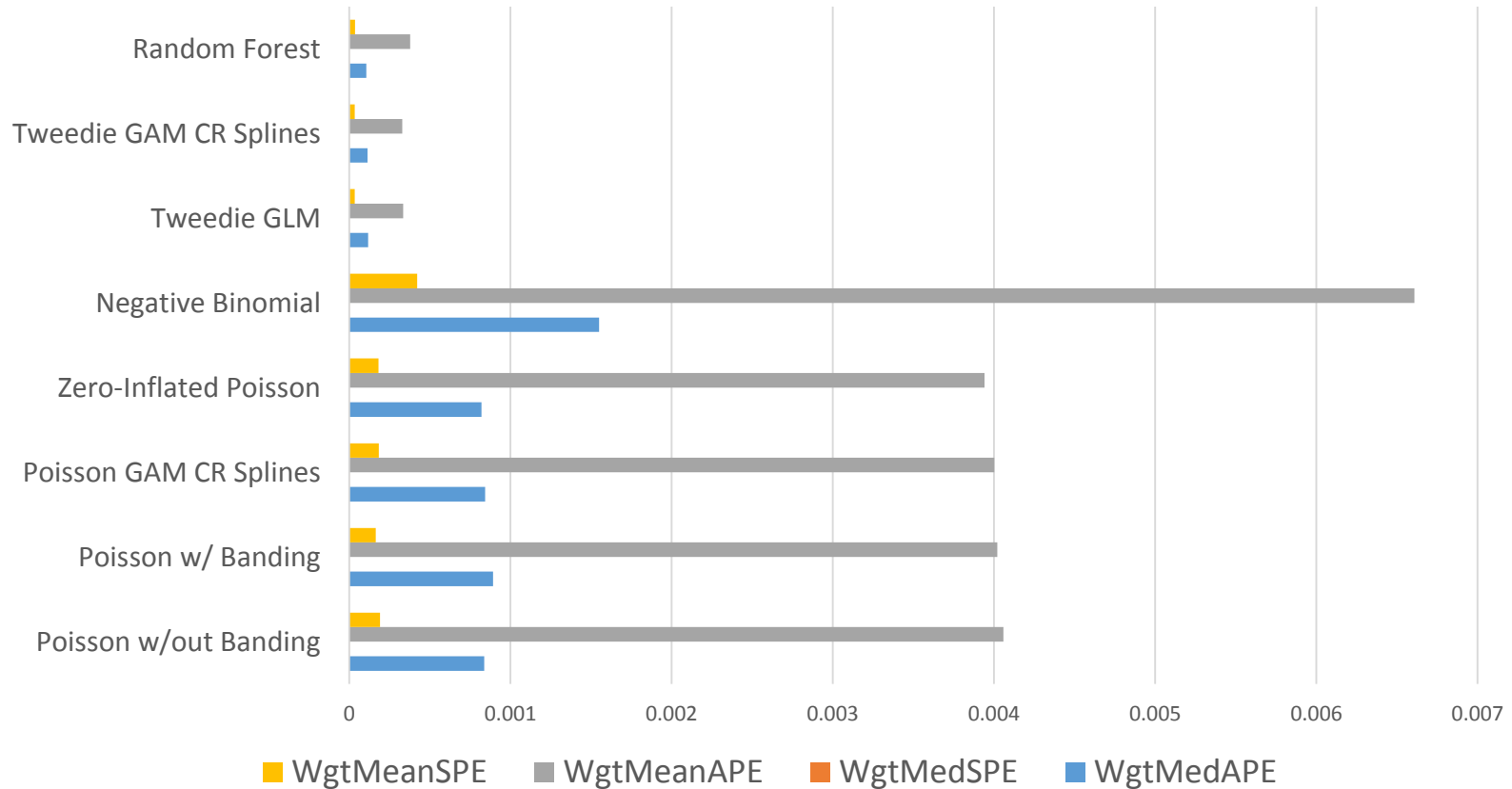
Un-Weighted Mortality Prediction Errors Average for All Years



Results

Mortality Rate Models

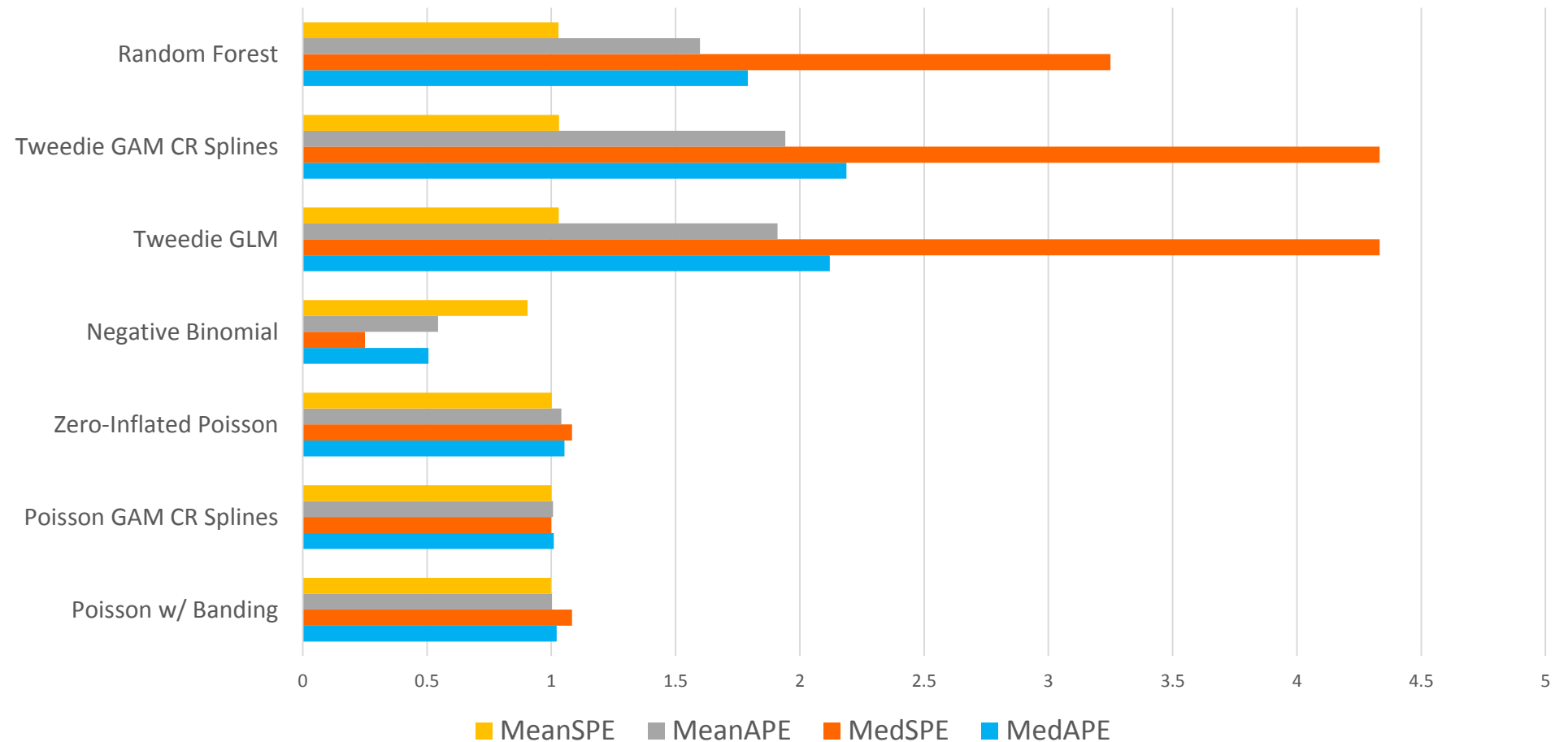
Weighted Mortality Prediction Errors Average for All Years



Results

Mortality Rate Models

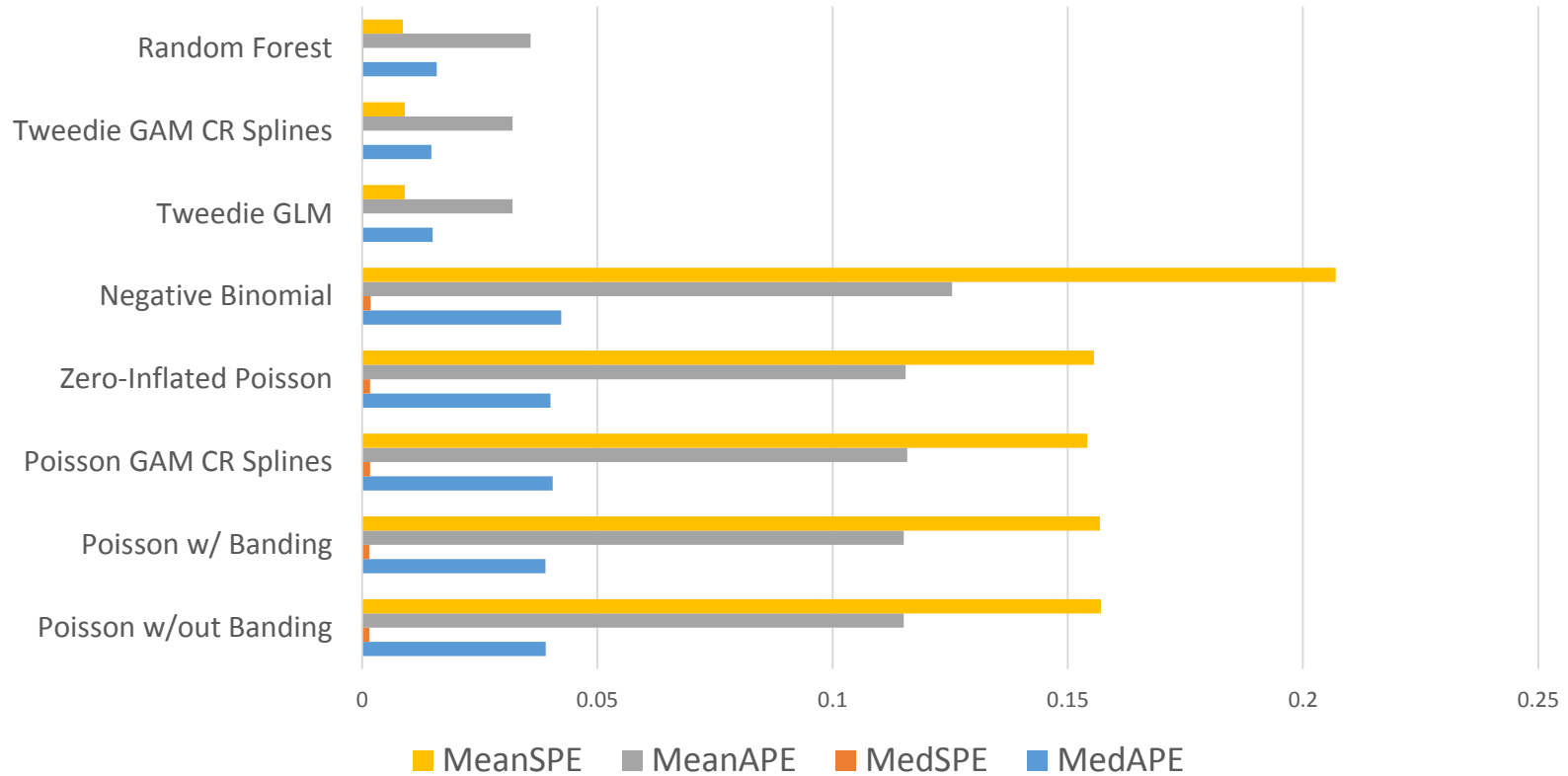
Mortality Predictive Accuracy Improvement Factor Over Baseline Poisson Model



Results

- Lapse Rate Models

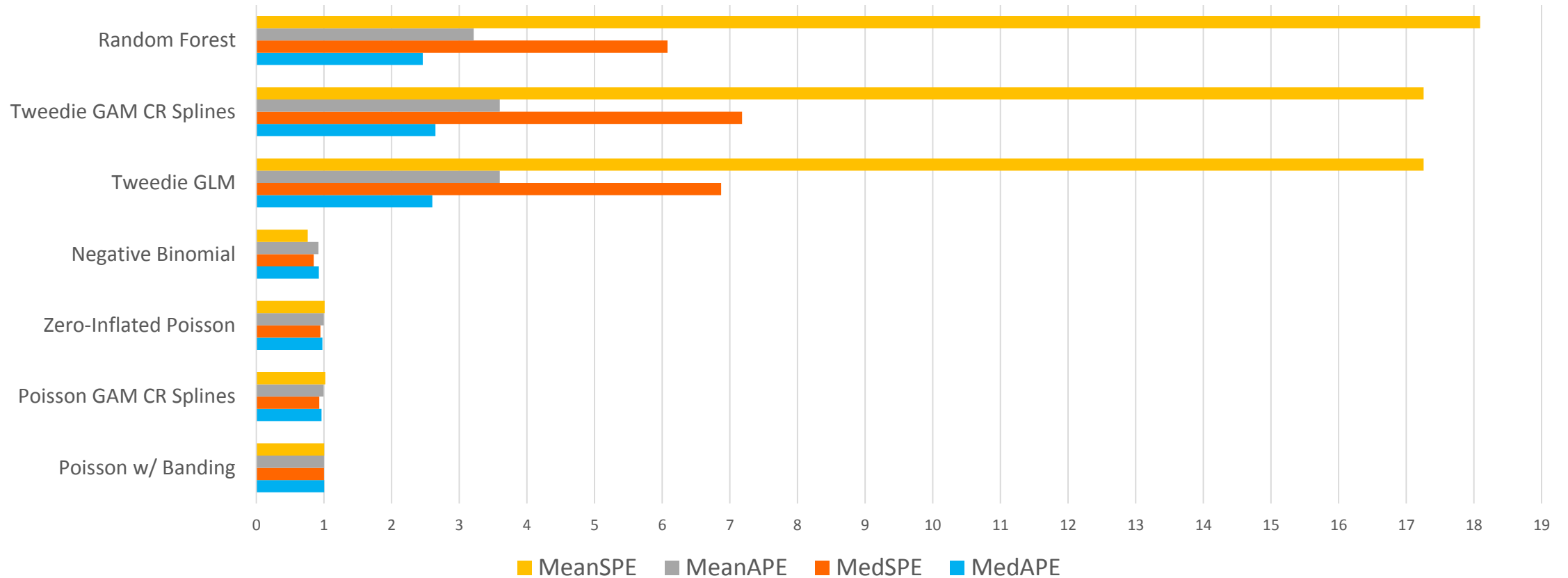
Un-Weighted Lapse Prediction Errors Average for All Years



Results

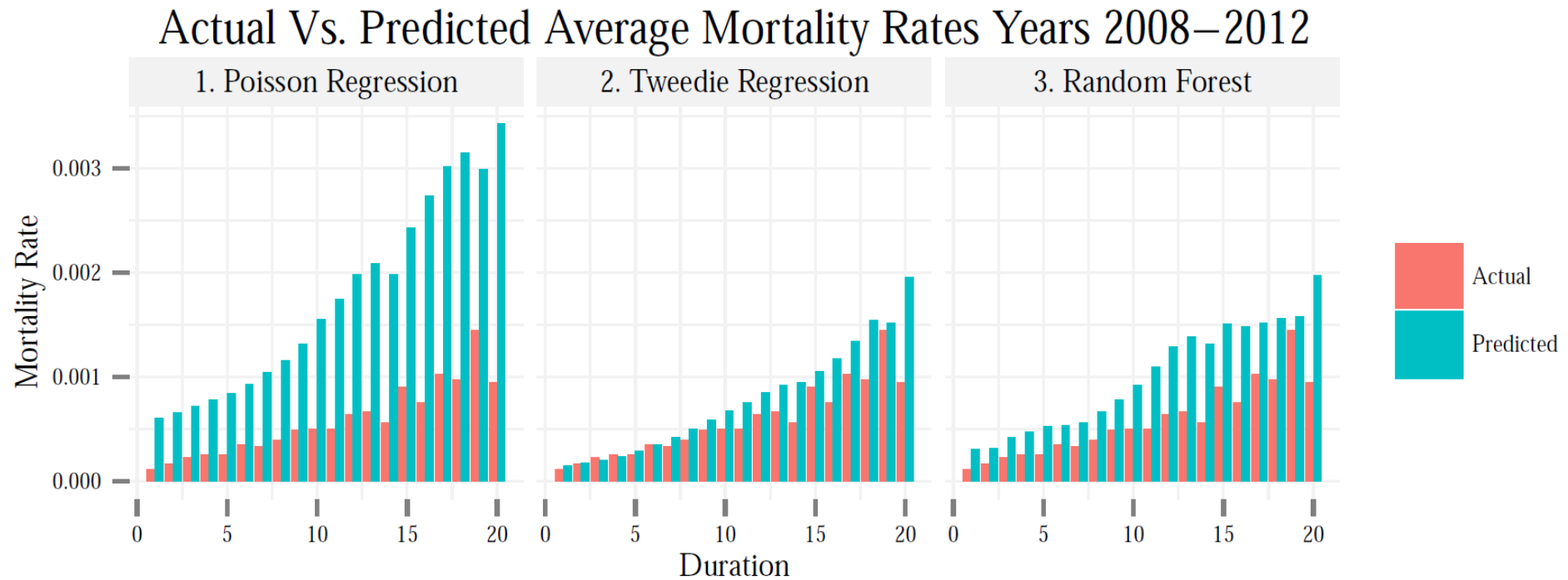
○ Lapse Rate Models

Lapse Predictive Accuracy Improvement Factor Over Baseline Poisson Model



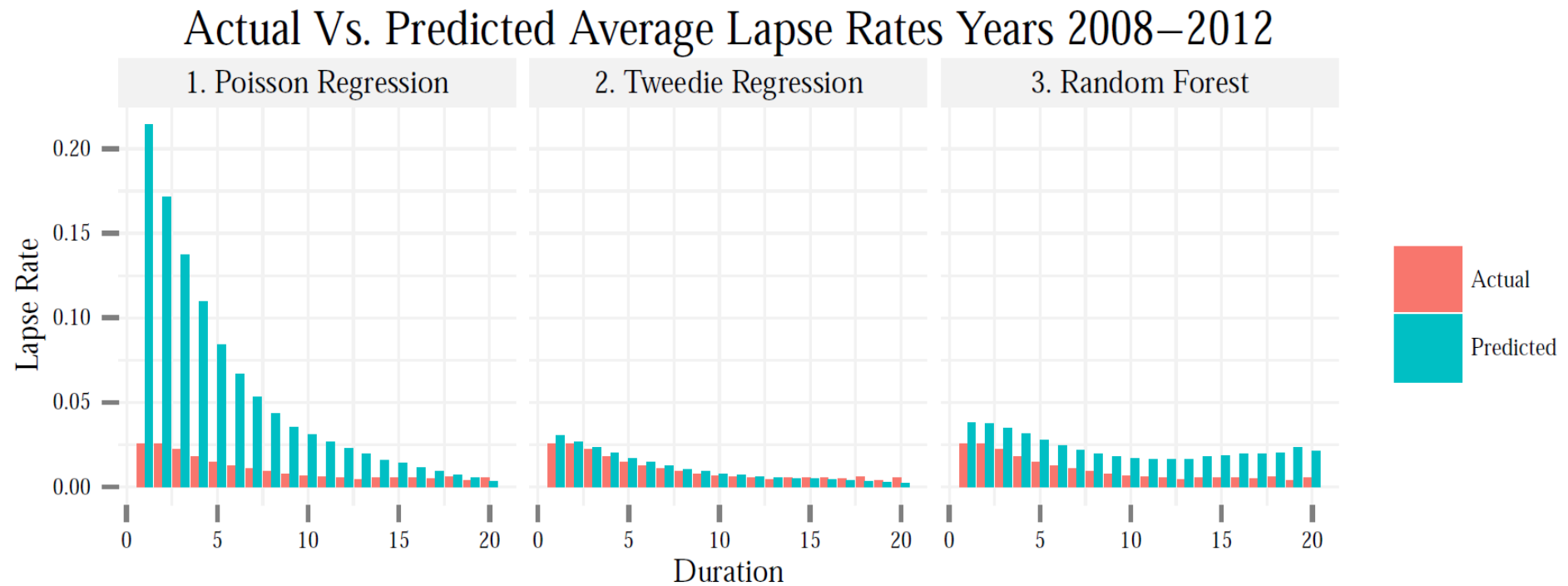
Results

- Evidence of overestimation with Poisson regression




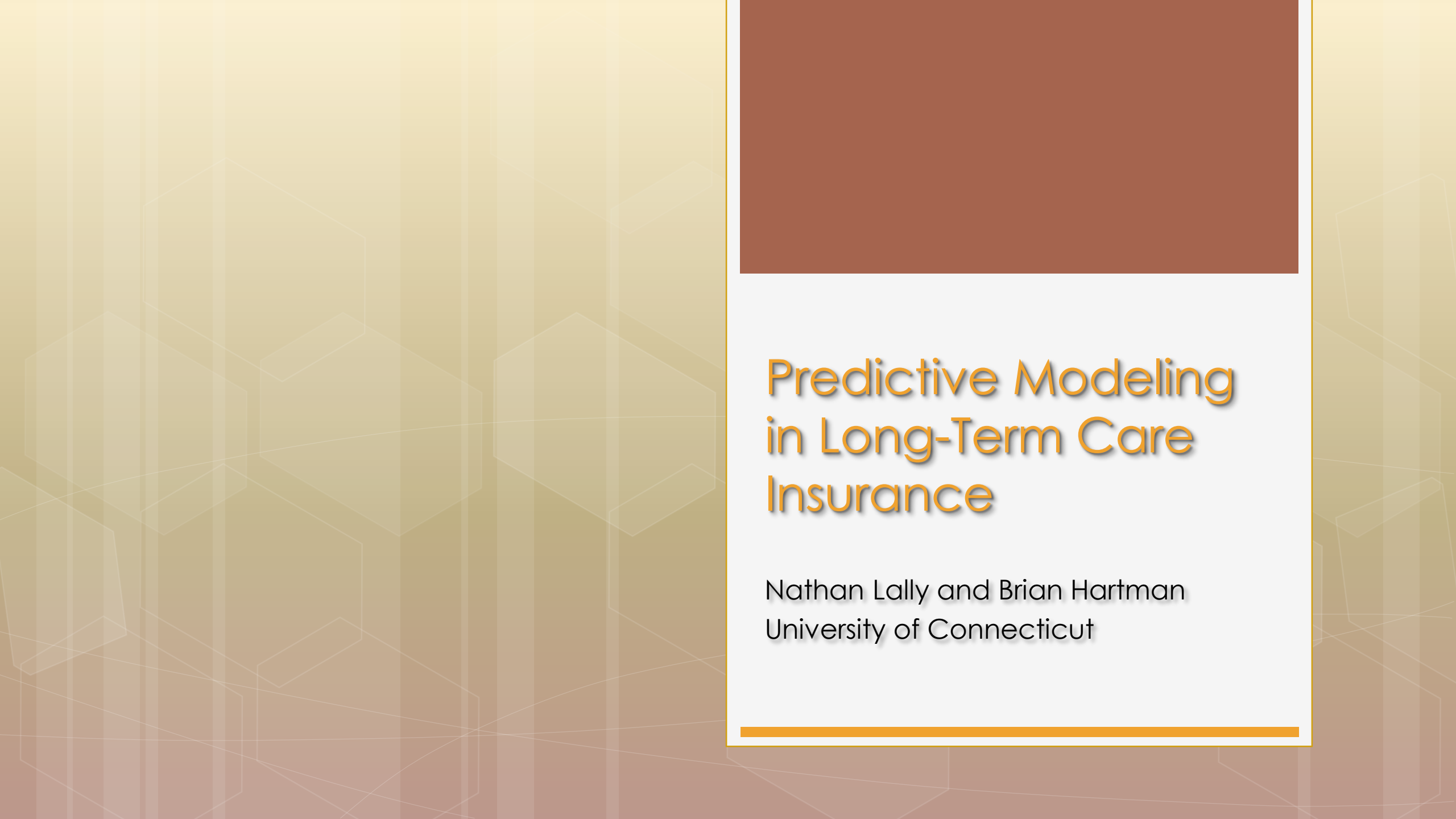
Results

- Evidence of overestimation with Poisson regression



Discussion

- Without accounting for overdispersion, Poisson models will have deflated standard errors for model parameters and therefore inflated t-statistics.
- For LTCI rate data GLMs with Tweedie errors outperform all other models by a substantial margin.
- Tweedie GAM models show no substantial improvement over the GLM models and are harder to interpret.
- Random forests are a promising method but we are currently restricted by computational performance.
- Negative binomial models performed surprisingly poorly.



Predictive Modeling in Long-Term Care Insurance

Nathan Lally and Brian Hartman
University of Connecticut
