Residency Exams

Background & Motivation

Prior to practicing full-time medicine, internal medicine students must complete a three year training known as a medical residency. At the end of their medical residency, they are required to pass an examination to obtain their MD. These residency programs are known for their extensive work weeks and rigorous workload.

In 2003 and 2011, two reform policies were passed to change the structure of internal medicine residency. The 2003 reform, passed on July 1, sought to limit the number of hours worked each week by capping it at 80 hours. The 2011 reform, also passed on July 1, placed stricter limits on which types of activities are permitted during the 80-hour work period.

These reforms were passed with the goal of improving patient care by decreasing the stress placed on internal medicine residents. In other words, with more time away from the hospital to rest and rejuvenate, students will be more successful at their jobs when they are on their shift. However, limiting the number of hours medical residents can work each week brings about several concerns regarding their performance of the examination at the end of their residency. While some believed a work time limit granted them more time to study, others thought that less hands on work would result in decreased pass rates.

These concerns bring about out motivating question: is there empirical evidence of an association between the reforms and the rate at which the medical residents passed the exam?

Data Manipulation and Exploratory Data Analysis

The provided data set contains three columns and twenty rows, containing a column for year, number of exam takers, and pass rate as a decimal. Each rows represents a single year in which internal medicine residents completed their residency and took the examination.

We manipulated the data by adding three additional columns for the number of students that passed the examination, the number of students that failed the examination, and the time period based on the year. The first time period is from 1996-2002, when there was no reform policy in place. The second time period is 2003-2010, when the first reform policy was in place. The third time period is 2011-2015, when the second reform policy was in place.

```
# Output Table with All Columns
kable(head(data, n = 3))
```

Year	N	Pct	Pass	Fail	timeperiod
1996	6964	0.82	5710	1254	tp1
1997	7173	0.85	6097	1076	tp1

Year	N	Pct	Pass	Fail	timeperiod
1998	7348	0.84	6172	1176	tp1

We further manipulated the data by releveling to time period two. Releveling to time period two enabled us to more easily compare time period one to time period two (no reform policy to reform policy one) and time period two to time period three (reform policy one to reform policy two).

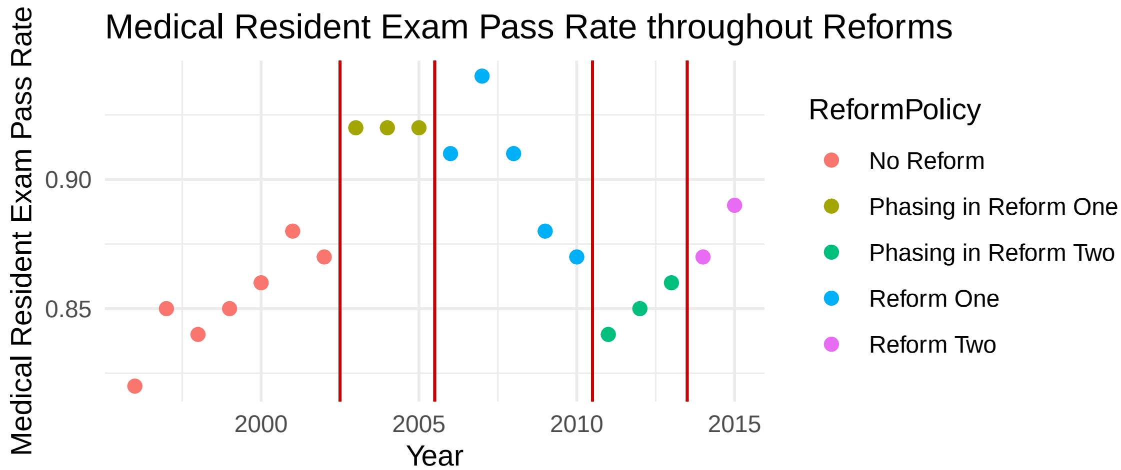


Figure 1: Internal medicine resident exam pass rate over time periods for no reform (1996−2002), phasing in reform one (2003−2005), reform one (2006−2010), phasing in reform two (2011−2013), and reform two (2014−2015).

The plot shown above displays the medical resident exam pass rate over the years 1996 to 2015. The plot was further split into five sections. The initial three time periods - no reform policy, reform policy one, and reform policy two - were further broken down into additional time periods. Each reform policy was split into a phasing-in time period a reform policy time period. The phasing-in time period contains the three years in which students taking the exam took some combination of the previous reform policy and the new reform policy during their three-year medical residency.

Students that fall in the section "Phasing in Reform One" (green dots on the plot), had experienced both no reform policy and reform policy one. Students that fall in the section "Phasing in Reform Two" (teal dots on the plot), had experienced both reform policy one and reform policy two during their residency.

As can be seen from the plot, students that took the examination in years where they experienced some amount of reform policy one appeared to perform better than students that took the examination in years where they experience no reform policy or some amount of reform policy two.

Model Implementation Details

Model 1 - Binomial Mixture

We reasonably assumed that exam pass rates contain unobserved year-to-year variation, such as differences in exam difficulty or cohort quality. To capture this extra source of randomness, we fit a binomial mixture model by including a random intercept for each year. This approach allowed us to separate systematic effects of reform policies from idiosyncratic annual noise.

We fit a generalized linear mixed model (GLMM) with a logit link, specifying pass/fail outcomes as the response and reform time period as the fixed effect, with year as a random effect:

$$\mathrm{Pass}_{iy} \sim \mathrm{Binomial}(n_{iy}, p_{iy}), \quad \mathrm{logit}(p_{iy}) = \alpha + \beta \cdot \mathrm{timeperiod}_{iy} + u_y, \quad u_y \sim N(0, \sigma^2).$$

Model 2 - Beta-binomial

The binomial assumption may underestimate variability in exam outcomes because pass probabilities likely vary within each period. To allow for overdispersion, we modeled the yearly pass probabilities as Beta-distributed. This yields a hierarchical structure where exam outcomes are drawn from a binomial conditional on the latent Beta-distributed rate.

We fit a beta-binomial regression with time period as the predictor using the glmmTMB package with a logit link. The model structure was:

$$\pi_y \sim \mathrm{Beta}(\alpha, \beta), \quad \mathrm{Pass}_y \sim \mathrm{Binomial}(n_y, \pi_y).$$

Here, dispersion was estimated directly from the data, capturing unmodeled heterogeneity.

Model 3 - Beta-binomial on subset

Because policy reforms were phased in gradually, exam cohorts in transition years likely had mixed exposure. To avoid contamination, we fit the beta-binomial model on a restricted dataset excluding these phase-in years. This design isolates the effect of reforms once they were fully implemented.

We restricted the dataset to years 1996–2002, 2006–2010, and 2014–2015. The same beta-binomial specification was used, with time period as the predictor and a logit link for the mean pass probability.

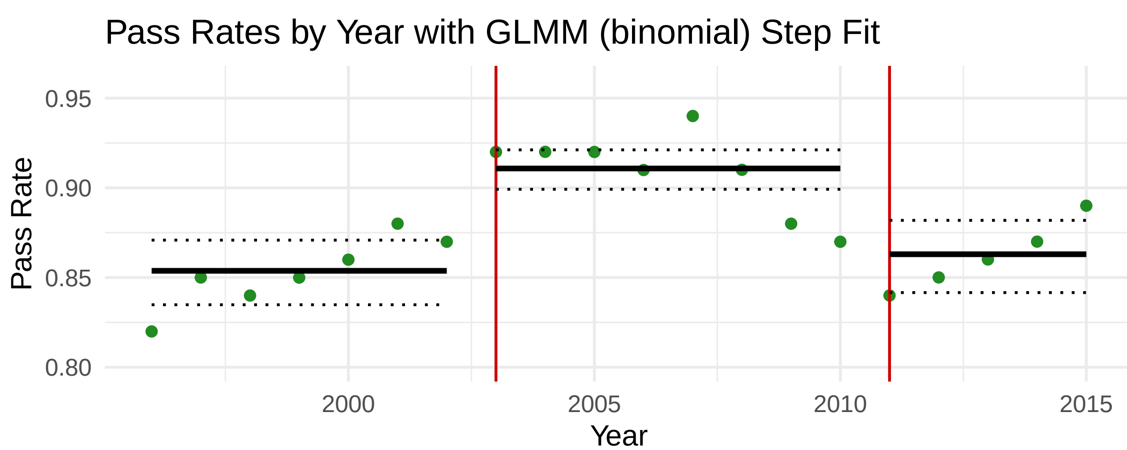


Figure Three: Internal medicine examination pass rates are plotted against year. Horizontal solid lines represent pass rates chosen by the binomial mixture model for no reform, reform one, and reform two. Dotted lines represent 95% confidence interval for pass rate.

Model 2 - Beta-binomial

Rationale

The binomial assumption may underestimate variability in exam outcomes because pass probabilities likely vary within each period. To allow for overdispersion, we modeled the yearly pass probabilities as Beta-distributed. This yields a hierarchical structure where exam outcomes are drawn from a binomial conditional on the latent Beta-distributed rate.

Implementation Details

We fit a beta-binomial regression with time period as the predictor using the glmmTMB package with a logit link. The model structure was:

$$\pi_y \sim \text{Beta}(\alpha, \beta), \quad \text{Pass}_y \sim \text{Binomial}(n_y, \pi_y).$$

Here, dispersion was estimated directly from the data, capturing unmodeled heterogeneity.

Result

```
$cond
                Estimate Std. Error   z value      Pr(>|z|)
(Intercept)    2.3061300 0.07465315 30.891263 1.565055e-209
timeperiodtp1 -0.5521503 0.09862549 -5.598454  2.162715e-08
timeperiodtp3 -0.4769413 0.10817339 -4.409045  1.038276e-05

$zi
NULL

$disp
NULL
```

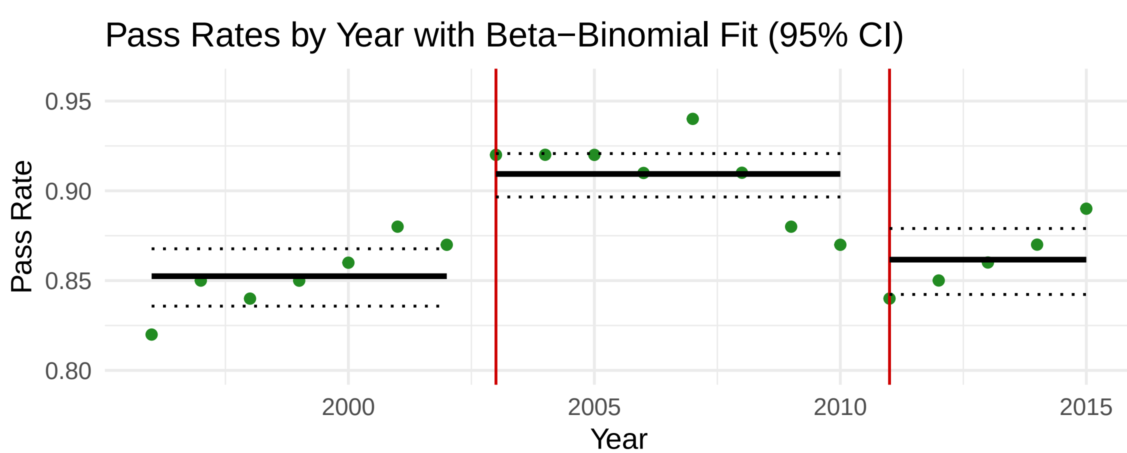


Figure Four: Internal medicine examination pass rates are plotted against year. Horizontal solid lines represent pass rates chosen by the beta−binomial model on full data for no reform, reform one, and reform two. Dotted lines represent 95% confidence interval for pass rate.

The estimated dispersion parameter was 280, confirming overdispersion relative to a pure binomial. Both time period 1 and time period 3 had significantly lower pass rates than time period 2 ($p < 0.001$).

Model 3 - Beta-binomial on subset

Rationale

Because policy reforms were phased in gradually, exam cohorts in transition years likely had mixed exposure. To avoid contamination, we fit the beta-binomial model on a restricted dataset excluding these phase-in years. This design isolates the effect of reforms once they were fully implemented.

Implementation Details

We restricted the dataset to years 1996–2002, 2006–2010, and 2014–2015. The same beta-binomial specification was used, with time period as the predictor and a logit link for the mean pass probability.

Result

```
$cond
                Estimate Std. Error   z value      Pr(>|z|)
(Intercept)    2.2345867 0.09658272 23.136507 1.987766e-118
timeperiodtp1 -0.4817205 0.11798326 -4.082956  4.446642e-05
timeperiodtp3 -0.2528107 0.16833072 -1.501869  1.331310e-01

$zi
NULL

$disp
NULL
```

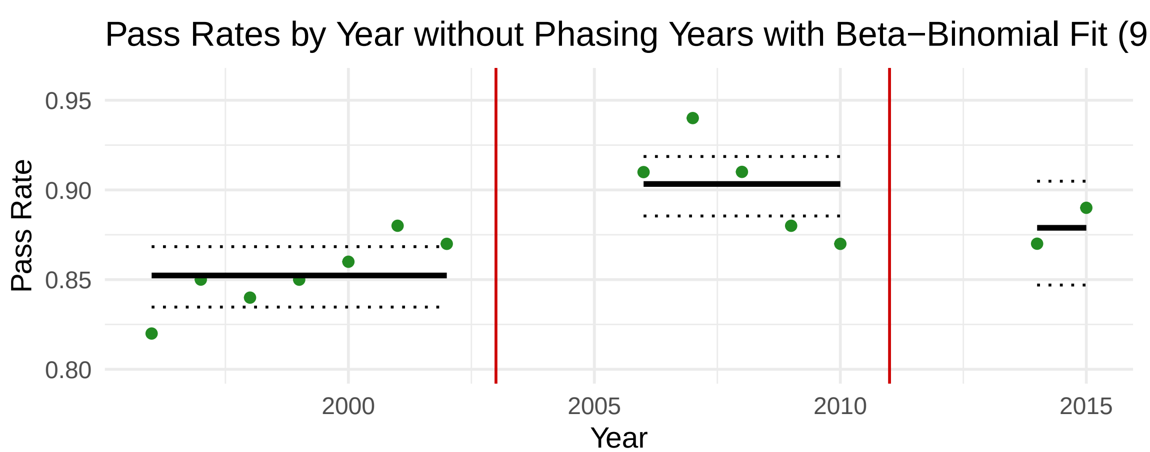
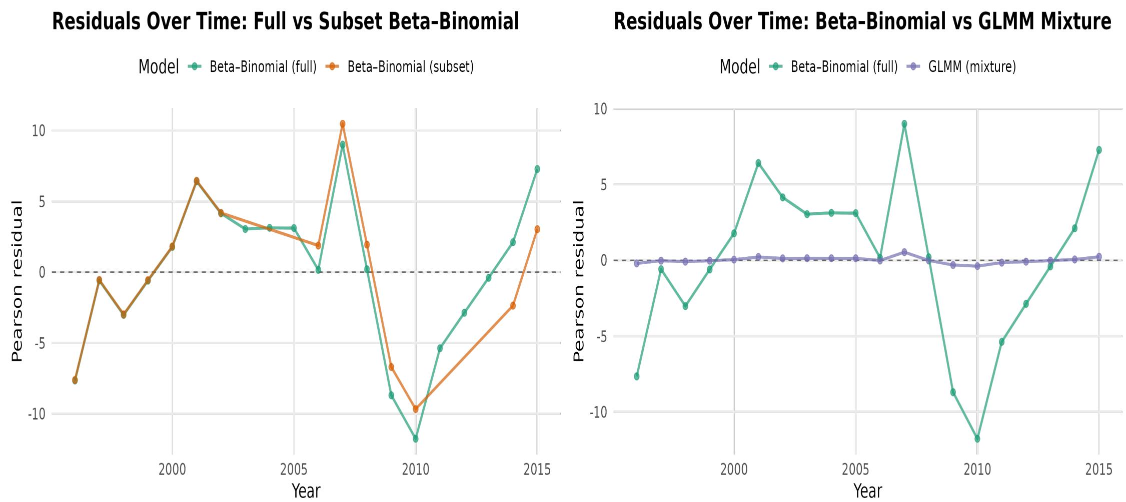


Figure Five: Internal medicine examination pass rates are plotted against year. Horizontal solid lines represent pass rates chosen by the beta−binomial model on non−phasing period data for no reform, reform one, and reform two. Dotted lines represent 95% confidence interval for pass rate.

The dispersion parameter was estimated at 251, again supporting overdispersion. Results showed that time period 1 had significantly lower pass rates than time period 2 ($p < 0.001$). However, time period 3 was no longer significantly different from time period 2 ($p = 0.133$).

Model Evaluation

To check goodness of fit, we plotted Pearson residuals over time. We appear to minimize residuals over time in the mixture model, as seen below. Putting this together, we concluded that the full beta-binomial mixture model is the best fit.



lotted by Year for Full Beta−Binomial Model Compared to Subsetted Model Fits and Mixture Binomial Model

Shortcomings

One challenge was finding evaluation metrics that could be compared across models. We first tried using AIC, but this was not reliable since different packages calculate it differently, and AIC depends on the likelihood, which changes when models use different numbers of observations.

For future work, Bayesian models could be helpful since they would let us incorporate additional information into the analysis. We also plan to explore other comparable measures of model fit and parsimony.

Conclusion

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.3234138	0.06885337	33.744374	1.292817e-249
timeperiodtp1	-0.5591105	0.10041764	-5.567851	2.579000e-08
timeperiodtp3	-0.4831369	0.11034058	-4.378597	1.194459e-05

We concluded that the binomial mixture model is the best fit. With that model, we found that the first reform increased the odds of passing, while the second reform was associated with a decline in performance. Both of these were confirmed by significant coefficients.