Paper 442-2013

Multilevel Reweighted Regression Models to Estimate County-Level Racial Health Disparities Using PROC GLIMMIX

Lucy D'Agostino
Melody S. Goodman, PhD
Division of Public Health Sciences, Department of Surgery, Washington University School of Medicine, St. Louis, MO, USA

ABSTRACT

The agenda to reduce racial health disparities has been set primarily at the national and state levels. These levels may be too far removed from the individual level where health outcomes are realized, and this disconnect may be slowing the progress made in reducing these disparities. This paper focuses on establishing county-level prevalence estimates of diabetes among Non-Hispanic Whites and Non-Hispanic Blacks. These estimates are produced using multilevel reweighted regression models through the GLIMMIX procedure with 2006-2010 Behavioral Risk Factor Surveillance System data and 2010 census data. To examine whether racial disparities exist at the county level, the paper estimates the risk difference of prevalence estimates between races. It subsequently ranks counties and states by the magnitude of disparities.

INTRODUCTION

Disparities in health by race have been consistently observed in mortality, morbidity and other indicators of health[1]. The agenda to reduce disparities has primarily been set on the national and state levels and is based on national and state level data. These levels, however, may be too far removed from the individual level where health outcomes are realized and this disconnect may be a contributing factor in the slow progress made in reducing racial health disparities. The lack of data on the local level hinders the ability to evaluate the effectiveness of local public health policy, programs, and interventions[2]. Public health data collection on the county level can be cost prohibitive; to obtain county-level estimates of prevalence of disease, small area analysis techniques can be applied to national and state level data[3,4]. Statistical methods that use small area analysis techniques provide estimates when the sample size for an area is too small or non-existent. This analysis fills the local data void, but there is a need for estimation techniques that allow for subgroup comparisons in the small areas. For the examination of disparities by race, it is necessary to have small area estimates by racial subgroup.

The Behavioral Risk Factor Surveillance System (BRFSS) is commonly used for estimating the prevalence of chronic disease. BRFSS collects uniform state-specific data on preventive health practices and risk behaviors that are linked to chronic diseases in adults[5]. Although BRFSS provides a wealth of information, valid direct estimation of prevalence can only be calculated at state and larger geographic levels because of the structure of the sampling design and weighting scheme[5,6].

At the national level, African Americans are 1.9 times more likely to have diabetes when compared to Whites.[7] We often assume that disparities seen on the national and state levels hold true at the county level, but for most counties we do not have the necessary data to support or dispute this assumption. This study re-weights BRFSS data and uses a multilevel regression model to produce county-level prevalence estimates by race in order to estimate racial disparities in diabetes on the county level. We are interested in examining the variability of racial disparities in diabetes within states, allowing for the targeting of interventions to the areas in most need.

METHODS

The prevalence estimates for diabetes were obtained from five years (2006-2010) of BRFSS survey data. Diabetes prevalence estimates were obtained from the following questions: (1) Have you ever been told by a doctor that you have diabetes? If "Yes" and respondent is female, they were then asked (2) "Was this only when you were pregnant?" If the answer to the first question is "Yes" and, if asked, the second answer is "No", then the respondent is considered to have chronic diabetes. Counties with less than 50 respondents per subgroup were excluded from this study. Only Non-Hispanic White and Non-Hispanic Black respondents were included. Additionally, those with missing information regarding age, sex, race, or diabetes status, were excluded from analysis.

In a previous study, it was established that multilevel logistic regression models show the least amount of discrepancy when estimating county-level prevalence of diabetes[4]. In order to make survey data more representative of the overall population, weights are often assigned to each respondent in complex survey analysis. BRFSS provides a direct weighting system that weights the prevalence data on the state level:

$$BRFSS\ Weight = W_{state} \times C_{state}$$

Where C_{state} is the number of people in an age-by-race-by-sex category in the population of the state divided by the sum of the products of the preceding weights for the respondents in that same age-by-race-by-sex category. It includes an adjustment for non-responders and households without telephones. W_{state} is the following design weight:

$$W_{state} = S \times \frac{1}{P} \times A$$

Where S accounts for differences in the probability of the respondent's telephone number selection, P is the number of residential telephone numbers in the respondent's household, and A is the number of adults in the respondent's household.

DATA ANALYSIS

Analysis was completed using SAS/STAT® version 9.3 (SAS Institute, Inc, Cary, North Carolina).

For this study, a modified weighting system was developed based on demographics at the county level. We used the above weight, replacing C_{state} with the following equation:

$$C_{county} = \frac{n_{ij}}{W_{county}}$$

Where n_{ij} is the number of people in county *i* that belong to demographic group *j* and W_{county} is defined as the following equation:

$$W_{county} = \sum_{j} \left(S \times \frac{1}{P} \times A \right)_{ij}$$

These modified weights were created using SAS® and subsequently applied to the multilevel regression model to estimate the prevalence of diabetes by race on the county level.

The following macro creates the modified weights, where the variable agegp is comprised of 2010 census-defined age groups and the variable age_race_sex is a county-level age by race by sex estimate:

```
%macro weight (data=, set=, race=, sex=, agegp=, age_race_sex=);
data &name;
 set &set;
  if race=&race;
   if sex=&sex;
    if agegp=&agegp;
 weight=(&age_race_sex/w_county)*w_state;
run;
%mend weight;
%weight(data=new1, set=brfss.pre_weight, race=1, sex=1, agegp=1,
       age_race_sex=white_male_18_19);
…
%weight(data=new2, set=brfss.pre_weight, race=1, sex=2, agegp=1,
       age_race_sex=white_female_18_19);
…
%weight(data=new3, set=brfss.pre_weight, race=2, sex=1, agegp=1,
       age_race_sex=black_male_18_19);
…
%weight(data=new4, set=brfss.pre_weight, race=2, sex=2, agegp=1,
       age_race_sex=black_female_18_19);
…

data brfss.data;
 set new1 new2 new3 new4 …;
run;
```

Using SAS® PROC GLIMMIX, we fit a multilevel reweighted regression model to obtain county- level prevalence estimates. The equation is of the form:

$$logit(p_{ij}) = X'\beta + \alpha_i$$

Where $x_{ij} = (x_{ij1}, \ldots x_{ijq})'$ is the vector of q covariates, $\beta = (\beta_1, \ldots \beta_q)'$ is the corresponding vector of fixed effects and α_i is the random effect for county. The model includes demographic variables: race, sex, and age group, as well as county-level data obtained from the 2010 census: county percent poverty and proportion of adults older than 25 with less than a high school education (county education). The variable 'age group' consists of nineteen age categories based on the manner by which census data was divided. The following significant interaction terms were also included: county percent poverty and age group, county percent poverty and race, county percent poverty and sex, county education and age group, county education and race, county education and sex, age group and race, age group and sex, and race and sex.

The following is the code for the multilevel reweighted regression model, where the variable agegp is comprised of 2010 census-defined age groups, nohs is the proportion of adults older than 25 with less than a high school education, and poverty is county percent poverty. The option ddfm=satther specifies a method, Satterthwaite, to adjust for denominator degree of freedom for tests of the fixed effects. The option tech-nrridg in the nloptions statement uses an optimization technique, Newton-Raphson, with ridging to help with the convergence of the model.

```
proc glimmix data=brfss.data;
 class county  sex race agegp nohs poverty;
 model diabetes (event=Last) = sex race agegp nohs poverty poverty*agegp poverty*race
  poverty*sex nohs*agegp nohs*race nohs*sex agegp*race agegp*sex race*sex/ solution
 dist=binary link=logit ddfm=satterth oddsratio;
 random county/solution;
 nloptions tech=nrridg;
 output out=brfss.prev pred=pred;
 weight weight;
run;
```

After calculating the regression parameter estimates, we estimated the county-level prevalence rates by race. The county-level age-by-race-by-sex estimated prevalence is calculated from the predictors given by the regression model as follows:

$$\hat{p}_{ijk} = \frac{e^{predictor}}{1 + e^{predictor}}$$

The county-level prevalence rates by race are then calculated with the following formula:

$$\hat{p}_{ij} = \sum_j \frac{n_{ijk}}{n_{ij}} \hat{p}_{ijk}$$

Where $\hat{p}_{ij}$ is the estimated prevalence of diabetes in county *i* of race *j*, $n_{ijk}$is the number of people in county *i* that are of race *j* and belong to age and sex demographic group *k*, $n_{ij} = \sum_j n_{ijk}$ is the total population in county *i* of race *j*, and $\hat{p}_{ijk}$ is the estimated prevalence of diabetes in county *i* for race *j* in demographic group *k*. We then compared these rates using risk difference and rate ratio.

Risk Difference= Estimated county-level prevalence rate $_{Black}$ *– Estimated county-level prevalence rate* $_{White}$
Rate Ratio=Estimated county-level prevalence rate $_{Black}$ *÷ Estimated county-level prevalence rate* $_{White}$

We ranked the counties based on both the magnitude of risk difference and the magnitude of rate ratio. Subsequently, we ranked the states based on the range of the race-stratified county-level diabetes prevalence estimates, focusing on the variability of racial disparities within each state. We compared the race-stratified within state (county-level) variance of estimated diabetes prevalence, to the overall (unstratified) within state (county-level) variance, and the between state variance of estimated diabetes prevalence. We then chose one state, North Carolina, to examine more closely. We compared the distribution of prevalence estimates of Blacks to Whites; using a two-sample test for proportions, we compared the mean estimated county-level prevalence of diabetes and tested the equality of the variances from the two distributions using an F-test.

RESULTS

After exclusion criteria, there were 592641 individual respondents included in our analysis. These respondents were from 372 counties; on average, there were 1593 respondents per county (Table 1). To assess county-level disparities in diabetes prevalence, an absolute disparity measure and a relative disparity measure were both calculated, risk difference and rate ratio respectively[8–10]. For the purposes of this analysis, the District of Columbia is included in county level analysis but excluded from state level analysis. The county with the largest estimated disparity based on an absolute disparity measure, risk difference, is Halifax County, Virginia. The county with the largest estimated disparity based on a relative calculation of disparity, rate ratio, is the District of Columbia (Table 2). The top five counties with the largest disparity measured by risk difference are from the following states: Virginia (two counties), North Carolina, Maryland, and South Carolina. The top five counties with the largest disparity measured by rate ratio are from the following states: New York, Colorado, North Carolina (two counties), and Maryland (Table 3).

Table 1. Summary Information

	N	Mean	Median	SD
Observations	592641			
Observations per county		1593.1	1031	1886.0
Counties	372			
Counties per state		11.3	4	13.6

Table 2. Disparity Information

	N	Mean	SD	Minimum (County)	Maximum (County)
Risk Difference†	372	0.023	0.013	-0.009 (DeSoto County, FL)	0.067 (Halifax County, VA)
Rate Ratio‡	372	1.339	0.216	0.890 (Lamar County, MS)	2.565 (District of Columbia)

†RD=Risk Difference= Estimated county-level prevalence rate $_{Black}$ – Estimated county-level prevalence rate $_{White}$
‡RR=Rate Ratio=Estimated county-level prevalence rate $_{Black}$ ÷ Estimated county-level prevalence rate $_{White}$

Table 3. States* Ranked by Counties with the Largest Risk Difference and Rate Ratio

	County	Estimated Disparity
Top 5 Counties in terms of Risk Difference†:		
Virginia	Halifax County	**0.067**
	Pittsylvania County	**0.063**
North Carolina	Hoke County	**0.060**
Maryland	Caroline County	**0.054**
South Carolina	Abbeville County	**0.052**
Top 5 Counties in terms of Rate Ratio‡:		
New York	New York County	**2.106**
Colorado	Denver County	**1.905**
North Carolina	Orange County	**1.905**
	Onslow County	**1.892**
Maryland	Calvert County	**1.878**

*For the purposes of this study, The District of Columbia is not considered a state and therefore is not included in this analysis.
†RD=Risk Difference= Estimated county-level prevalence rate $_{Black}$ – Estimated county-level prevalence rate $_{White}$
‡RR=Rate Ratio=Estimated county-level prevalence rate $_{Black}$ ÷ Estimated county-level prevalence rate $_{White}$

Virginia (0.148) has the largest within state range of estimated county-level race-stratified diabetes prevalence among the states included in this study (Table 4). Although the overall estimated prevalence of diabetes in Virginia is 0.080 (8%), the race-stratified variability between the counties analyzed in this study (14.8%) is greater than this estimate,

reinforcing the necessity of looking at both county-level and race-specific data, as prevalence rates on the county level may be substantially different from state level rates and prevalence rates can differ substantially between Non-Hispanic Blacks and Whites. An examination of county-level prevalence estimates allows us to assess that the Virginia county with the greatest absolute (measured with risk difference) black-white disparity is Halifax County.

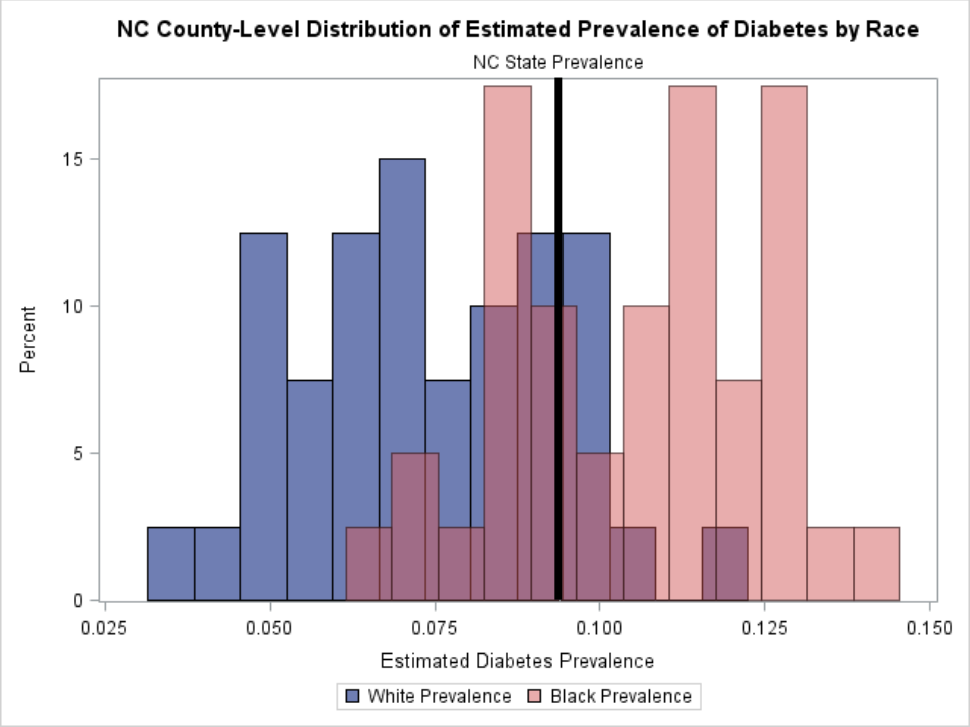
Table 4. States Ranked by the Range of Race-Stratified County-Level Diabetes Prevalence Estimates

	N*	Range§	RD† Mean	RD† SD	RD† Min	RD† Max	RR‡ Mean	RR‡ SD	RR‡ Min	RR‡ Max
Virginia	17	**0.148**	0.028	0.018	0.001	0.067	1.424	0.224	1.023	1.791
Mississippi	50	**0.122**	0.017	0.011	-0.006	0.041	1.227	0.165	0.890	1.724
Alabama	28	**0.114**	0.021	0.013	0.002	0.051	1.259	0.177	1.022	1.595
North Carolina	40	**0.106**	0.031	0.011	-0.005	0.060	1.464	0.208	0.956	1.905
Maryland	17	**0.105**	0.028	0.015	-0.005	0.054	1.430	0.210	0.947	1.878
South Carolina	39	**0.093**	0.029	0.013	-0.002	0.052	1.379	0.194	0.978	1.830
Louisiana	29	**0.086**	0.021	0.01	-0.002	0.041	1.302	0.154	0.969	1.750
Florida	37	**0.078**	0.016	0.011	-0.009	0.040	1.223	0.157	0.914	1.531
New York	9	**0.071**	0.021	0.014	-0.001	0.040	1.403	0.352	0.992	2.106
Georgia	18	**0.067**	0.016	0.012	-0.007	0.035	1.278	0.216	0.912	1.696
Ohio	7	**0.065**	0.036	0.004	0.032	0.044	1.528	0.063	1.449	1.593
New Jersey	11	**0.064**	0.028	0.008	0.008	0.043	1.441	0.149	1.080	1.612
Michigan	9	**0.063**	0.026	0.013	0.008	0.043	1.399	0.202	1.127	1.650
Texas	9	**0.060**	0.017	0.009	0.000	0.024	1.260	0.150	0.995	1.465
Delaware	3	**0.057**	0.024	0.007	0.017	0.032	1.356	0.155	1.180	1.472
District of Columbia	1	**0.054**	0.054	N/A	0.054	0.054	2.565	N/A	2.565	2.565
Arkansas	9	**0.050**	0.010	0.011	-0.006	0.025	1.142	0.143	0.934	1.293
Missouri	3	**0.045**	0.031	0.008	0.022	0.036	1.529	0.095	1.420	1.588
Tennessee	5	**0.045**	0.025	0.006	0.017	0.034	1.372	0.119	1.192	1.488
Illinois	3	**0.045**	0.026	0.011	0.016	0.038	1.463	0.169	1.316	1.648
Kentucky	2	**0.044**	0.028	0.007	0.023	0.032	1.418	0.053	1.381	1.455
Pennsylvania	4	**0.043**	0.020	0.012	0.008	0.030	1.310	0.177	1.123	1.472
Indiana	4	**0.043**	0.026	0.010	0.014	0.036	1.380	0.153	1.199	1.511
Massachusetts	1	**0.041**	0.041	N/A	0.041	0.041	1.863	N/A	1.863	1.863
Kansas	1	**0.039**	0.039	N/A	0.039	0.039	1.542	N/A	1.542	1.542
California	3	**0.039**	0.021	0.011	0.010	0.032	1.382	0.268	1.132	1.665
Oklahoma	3	**0.039**	0.023	0.004	0.020	0.028	1.360	0.040	1.326	1.404
Connect	3	**0.035**	0.020	0.002	0.018	0.023	1.363	0.100	1.281	1.474
Colorado	2	**0.034**	0.025	0.013	0.016	0.034	1.629	0.391	1.352	1.905
Nebraska	1	**0.026**	0.026	N/A	0.026	0.026	1.458	N/A	1.458	1.458
Nevada	1	**0.025**	0.025	N/A	0.025	0.025	1.358	N/A	1.358	1.358
Wisconsin	1	**0.016**	0.016	N/A	0.016	0.016	1.263	N/A	1.263	1.263
Minnesota	2	**0.014**	0.000	0.007	-0.004	0.005	1.020	0.138	0.923	1.118

*****N=the number of counties included in the analysis
§Range= the range of the race-stratified county-level prevalence estimates of diabetes within each state
†RD=Risk Difference= Estimated county-level prevalence rate $_{Black}$ – Estimated county-level prevalence rate $_{White}$
‡RR=Rate Ratio=Estimated county-level prevalence rate $_{Black}$ ÷ Estimated county-level prevalence rate $_{White}$

Looking at the variability within the states, we found that when including race-stratified estimates, the variance of the prevalence of diabetes within the state is larger than when only including the overall county-level estimates. Figure 1demonstrates this difference, showing a histogram of the Non-Hispanic White prevalence estimates next to a histogram of the Non-Hispanic Black prevalence estimates for the state of North Carolina. In Figure 1, the blue histogram depicts the county-level prevalence of diabetes for Non-Hispanic Whites, and the red histogram depicts the county-level prevalence estimates of diabetes for Non-Hispanic Blacks. The purple bars indicate the overlapping regions. The black line shows the North Carolina statewide prevalence of diabetes obtained using direct estimation of BRFSS 2006-2010 weighted data. The distribution of prevalence estimates for Blacks is shifted to the right in comparison to the distribution for Whites; the mean of the distribution of diabetes prevalence estimates for Blacks is 42.3% higher than the mean of the distribution of prevalence for Whites in North Carolina. The p-value from the two-sample test for differences in proportions is p<.0001. The variance of the distribution of diabetes prevalence estimates for Blacks is slightly larger (2.6% higher) than the variance of the distribution of prevalence for Whites in North Carolina. This difference is not statistically significant, based on the F-test on equality of variances (p=0.9379).

Figure 1. NC County-Level Distribution of Estimated Prevalence of Diabetes by Race



When looking at the overall prevalence estimates, the average variance within states is 1.75×10^{-4}, while the average variance within states stratified by race more than doubles that (3.53×10^{-4}) (Table 5). We also found that the within state variability is often larger than the variability between states. The between state variance of the states included in the above analysis (N=27) is 1.92×10^{-4}. Including all states (N=51), the between state variance is 2.18×10^{-4}. The maximum within state variance seen is 7.10×10^{-4}, more than three times larger than the between state variance when looking at all states, and more than 3.5 times larger than the between state variance of the states included in this study. Of the states included in this analysis, 11 (40.7%) have a within state variance greater than the between state variance. Further, 24 (88.9%) of the states included in this analysis have a race-stratified within state variance greater than the between state variance in terms of the estimated prevalence of diabetes.

Table 5. Variance of Within States (County-Level) Prevalence of Diabetes

	N*	Mean	SD	Minimum	Maximum
Race-Stratified	27	0.000353	0.000196	0.000035 (MN)	0.001067 (VA)
Overall	27	0.000175	0.000150	0.000004 (CO)	0.000710 (VA)

*States with only one county analyzed were not included because a range could not be calculated.

CONCLUSION

Addressing disparities based on factors such as race/ethnicity, geographic location, and socioeconomic status is a current public health priority. This study takes a first step in developing the statistical infrastructure needed to target disparities interventions and resources to the local areas with greatest disparities. Employing the methods detailed above allows investigators to apply new weights as well as incorporate multiple levels of data to estimate the prevalence of diabetes at the county level, making the targeting of interventions and resources to those areas most in need, a possibility. The PROC GLIMMIX procedure in SAS® ideally performs the necessary functions to make such estimates. This study demonstrates both the necessity to look at the county level estimates when assessing racial disparities, as well as the value of PROC GLIMMIX, alongside other tools within the SAS/STAT® software, when doing so.

REFERENCES

1. Baker, EL, White, LE, Litchtveld M. Reducing health disparities through community-based research. *Public Health Reports*. 2001;116:517–519.

2. Kim I, Keppel KG. Priority data needs: sources of national, state and local-level data and data collection systems. *Healthy People 2000 Stat Notes*. 1997;15(10; 15):1–11.

3. Jia H, Muenning P, Borawaki E. Comparison of small-area analysis techniques for estimating county level outcomes. *American Journal of Preventive Medicine*. 2004;26(5):453–460.

4. Goodman M. Comparison of Small-Area Analysis Techniques for Estimating Prevalence by Race. *Preventing Chronic Disease*. 2010;7(2).

5. NCHS. Behavioral Risk Factor Surveillance System documentation. 2001. Available at: http://www.cdc.gov/brfss/about.htm. .

6. Remington PL, Smith MY, Williamson DF, Anda RF, Gentry GC. Design, Characteristics, and usefulness of State-Based Behavioral Risk Factor Surveillance: 1981-1987. <I>*Public Health Reports*</I>. 1988;103(4):366–375.

7. Signorello LB, Schlundt DG, Cohen SS, et al. Comparing diabetes prevalence between African Americans and Whites of similar socioeconomic status. *American Journal of Public Health*. 2007;97(12):2260–7. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2089102&tool=pmcentrez&rendertype=abstract. Accessed October 3, 2011.

8. Harper S, Lynch J, Meersman SC, Breen N, Davis WW, Reichman ME. An overview of methods for monitoring social disparities in cancer with an example using trends in lung cancer incidence by area-socioeconomic position and race-ethnicity, 1992-2004. *Am J Epidemiology*. 2008;167(8):889–899.

9. Harper S LJ. *Selected Comparisons of Measures of Health Disparities: A Review Using Databases Relevant to Healthy People 2010 Cancer-Related Objectives*.; 2007.

10. Harper S, Lynch J. *Methods for Measuring Cancer Disparities: Using Data Relevant to Healthy People 2010 Cancer-Related Objectives*.; 2010.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Lucy D'Agostino
Division of Public Health Sciences
Department of Surgery
Washington University in St. Louis School of Medicine
660 S. Euclid Avenue, Campus Box 8100
St. Louis, MO 63110
Phone: 314-286-2372
E-mail: dagostinol@wudosis.wustl.edu

Melody S. Goodman, MS, PhD
Division of Public Health Sciences
Department of Surgery
Washington University in St. Louis School of Medicine
660 S. Euclid Avenue, Campus Box 8100
St. Louis, MO 63110
Phone: 314-362-1183
E-mail: goodmanm@wustl.edu