VARIATION IN USE OF $4 GENERIC PROGRAM AND POTENTIAL SAVINGS AMONG MEDICARE BENEFICIARIES--- BIOSTATISTICS STUDENT'S INTERNSHIP EXIT REPORT

by

Lei Zhou

M.D., Tianjin Medical University, China, 2005

Submitted to the Graduate Faculty of

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2010

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Lei Zhou

It was defended on

June 14, 2010

and approved by

Committee Chair:

Advisor:
Vincent C. Arena, PhD.
Associate Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Member:

Thesis advisor:
Yuting Zhang, PhD
Assistant Professor
Department of Health Policy & Management
Graduate School of Public Health
University of Pittsburgh

Jeanine Buchanich, PhD.
Research Assistant Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

VARIATION IN USE OF $4 GENERIC PROGRAM AND POTENTIAL SAVINGS AMONG MEDICARE BENEFICIARIES---
BIOSTATISTICS STUDENT'S INTERNSHIP EXIT REPORT

Lei Zhou, M.S.

University of Pittsburgh, 2010

As an option to fulfill the MS thesis requirement at the Department of Biostatistics, I worked as an intern under the supervision of Dr. Yuting Zhang at the Department of Health Policy & Management, Graduate School of Public Health, University of Pittsburgh, from January to June 2010. During the internship, I have been fully involved in some of Dr. Zhang's projects and have made the following contributions.

First, I consolidated different pharmacy event data and medical claims data obtained from multiple sources into several analytic databases for those projects. The end products in this step included the analytic datasets, data dictionary for each corresponding dataset, and the SAS programming codes.

After completion of the dataset construction, I had opportunities to fully apply the statistical skills I have learned during my coursework on a specific project, entitled "Variation in the use of $4 generic prescription and potential savings among Medicare beneficiaries." Under the supervision of Dr. Zhang as well as collaborating with other colleagues, I played the leading role in data analysis, the interpretation of results and writing of a manuscript for publication.

Public Health Relevance: Our research on these projects focused on evaluating the strengths and weaknesses of the Medicare prescription drug program, especially its effects on

vulnerable American populations such as under-served minorities, patients with severe mental health and multiple medical conditions. Through our research, public policy might be improved to eliminate health disparities in populations. Our findings from the project have important policy implications for optimizing cost-effective use of prescription plans to the public.

Through this half-year long internship, I have had great opportunities to learn study design, data management, statistical analysis and hypothesis testing in a real world setting, to apply statistics/econometrics knowledge to large existing data, to evaluate the effects of health care policy and interventions on medical spending and health outcomes. In addition, I have practiced advanced SAS programming skills in manipulating the large datasets.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my internship supervisor, Dr. Yuting Zhang. Thank you so much for giving me the opportunity to work with and learn from you. I appreciate all of the time and resources you extended to me along the way. The experience has been invaluable.

I would like to thank Dr. Vincent Arena for taking the time to review this report. And I appreciate your informative guidance on my coursework study during the past two years as my academic advisor.

I would like to thank Dr. Jeanine Buchanich for taking time to review this report and make comments on it.

I would also like to thank Ms. Yi-Jen Lin for her great work on some related calculation by using the ArcGIS software.

Much gratitude is owed to my wife, Yali Wu and my family for their encouragement and support throughout my tenure in the program.

1.0 INTRODUCTION TO INTERNSHIP

According to departmental degree requirements for the Master of Science, MS students could be permitted with approval of their primary advisors and the Biostatistics MS/MPH Program Committee to complete an internship as an option for fulfillment of the MS thesis requirement. By complying with this requirement, the purpose of this report is to explain what I did and learned during my internship under the supervision of Dr. Zhang, Yuting, at the Department of Health Policy and Management, Graduate School of Public Health, University of Pittsburgh, from January to June, 2010.

During the internship period, I made two primary contributions. First, I assembled various analytic datasets for future analysis from the source data as well as the detailed data dictionary for each corresponding dataset, where the source data are obtained from multiple sources including pharmacy event data, medical claims data, census data, and drug database. I used PROC SQL language from SAS® 9.2 to manage and manipulate these large existing real-world data (usually several hundred thousands or millions of observations). The description about the activities during this step is in Section 2.0 of this report.

Following the completion of data construction, I moved onto the task of investigating the variation in the use of $4 generic prescriptions among Medicare beneficiaries, where I played a leading role in data analysis as a statistician, interpreting and publishing the results as a co-author. The detailed results about this project are shown in Section 3.0 of this report.

2.0 DATA MANAGEMENT AND DATASET CONSTRUTION

As an intern, I have been fully involved in several projects evaluating the strengths and weaknesses of the Medicare prescription drug program (a.k.a Medicare Part D), which was enacted as part of the Medicare Prescription Drug, Improvement, and Modernization Act MMA of 2003 (MMA, Section 101) and went into effect on January 1, 2006, to subsidize the costs of prescription drugs for Medicare beneficiaries in the United States. [1]

In this first step of the internship, I identified those variables of interest which denoted the beneficiaries' demography, socioeconomics, insurance status, medicine and medication uses, such as spending and counts, and drug information like name, strength, dosage form, and national drug code (NDC). These variables could be derived from various source data such as pharmacy event data and medical claims data obtained from the Centers for Medicare & Medicaid Services (CMS), census data obtained from the US Census 2000 database, area medical supply data obtained from Health Resources and Services Administration (HRSA), and drug information obtained from the First DataBank database

Following the identification of these variables, I linked those source data through some specific identifiers and assembled the variables of interest from these source datasets into several analytic datasets for future analyses. In addition, I created a detailed data dictionary for each analytic dataset including the variable name, data type, and description. In this way, other colleagues who would focus on some specific project in the future can easily find their needed variables by going through these dictionaries.

Subsection 2.1 through 2.4 briefly described the source data we have been using, and in subsection 2.5, I showed an example to elaborate my activities on dataset construction.

2.1 CMS-CCW SOURCE DATA

CMS has contracted with the Buccaneer Computer Systems and Services, Inc. (BCSSI) to establish the Chronic Condition Data Warehouse (CCW). The CCW contains existing CMS Medicare beneficiary claims data and are available for services beginning January 1, 1999 through the most current year of Medicare data available, for a 5% random sample of Medicare beneficiaries. [1]

The major data we used included the beneficiary annual summary file, Part D denominator file, Part D event data file and Part D plan characteristics file. The detailed guidelines for using these data and data descriptions can be found from the website of http://www.resdac.umn.edu .

2.1.1 Beneficiary Annual Summary File

The Beneficiary Annual Summary File, available and updated annually since 1999, contains demographic and enrollment information about each beneficiary enrolled in Medicare during a calendar year. For our current projects, we used the 2007 data. From this data, I identified the following variables: beneficiary unique identifier (*bene_id*), state and county codes (*state_cd, cnty_cd*), ZIP code (*bene_zip*), date of birth (*bene_dob*), date of death (*bene_dod*), gender (*sex*), months of Part A/B/both A and B enrollment (*a_mo_cnt, b_mo_cnt, ab_mo_cnt*), months of Medicaid coverage (a.k.a. state buy-in coverage, *buyin_mo*), months of managed care enrollment (*hmo_mo*) and Medicare status code (*ms_cd*) which can be used to define those who are disabled. The variables indicating different kinds of medical spending (inpatient, outpatient, nursing facility, physician visits, medical equipment, and hospice) and those 21 chronic conditions are also identified from this data. The complete definitions for these variables are shown in Appendix A.

2.1.2 Medicare Part D Denominator File

The denominator file, similar to the beneficiary annual summary file, available since 1999, also contains demographic and enrollment information about each beneficiary enrolled in Medicare during a calendar year. The variables I identified from 2007 data include the beneficiary unique identifier (*bene_id*), age (*age*), RTI (Research Triangle Institute) race (*rti_race*, which is more accurate than another race variable), months of Part D plan coverage (*plncovmo*), months of dual eligible (*dual_mo*, i.e., both Medicare and Medicaid coverage), monthly cost share group code (*cstshr<mon#>*, which can be used to determine the low-income subsidy status), monthly plan contract ID and benefit package ID (*cntrct<mon#>, pbpid<mon#>*). This file is not available after March of 2010 due to its incorporation into the beneficiary summary file since then. The data dictionary about this data is shown in Appendix B.

2.1.3 Medicare Part D Event Data File

The PDE data are person-drug level claims data, containing prescription drug claims information for each beneficiary. From this source data, I identified the following variables: beneficiary unique identifier (*bene_id*), service provider ID (*prvdr_id*), prescriber ID (*prscrbid*), product service ID (*prdsrvid*), RX service date (*srvc_dt*), plan contract ID and benefit package ID (*plncntrc, plnpbprc*), days of supply (*dayssply*), drug coverage status code (*drcvstcd*), all kinds of drug costs (*totalcst, ptpayamt, lics_amt, othtroop, cpp_amt, npp_amt*), benefit phase of the Part D event (*bnftphas*). The detailed explanation about these variables is shown in Appendix C.

2.1.4 Medicare Part D Plan Characteristics File

Each state may have its own types of prescription plans based on the standard Medicare Part D benefit design. The plan characteristics file contains all the information for each specific Part D prescription plan. The variables we used include plan contract ID and benefit package ID (*cntrctid, plan_id*), drug benefit type (*drgbentp*), type of gap coverage (*gapcovtp*), deductible amount (*ded_amt*), how initial coverage limit is defined (*icl_app*), ICL amount (*icl_amt*). The complete data dictionary is shown in Appendix D.

2.2 FIRST DATABANK SOURCE DATA

First DataBank, a subsidiary of Hearst Corporation, is the leading provider of electronic drug information to the healthcare industry. We obtained the drug information data from their most widely used drug database, First DataBank's National Drug Data File Plus (NDDF Plus), which combines a comprehensive set of drug database elements, drug pricing and clinical information with multiple types of unique drug identifiers.

After carefully reading the NDDF Plus documentation, I extracted those data from the database, which include the following drug information: NDC number (*ndc*), drug dosage form (*gcdf_desc*), drug strength (*str60*), package size (*ps*), generic name (*gnn60*), brand name (*bn*), multiple source indicators (*ndcgi1, gcnseq_gi*), generic indicator (*gni*), enhanced therapeutic classifications (*etc_id, etc_name*). And then I linked these data through those specific drug identifiers to consolidate those variables into one dataset.

The therapeutic classification is complicated to define. The enhanced therapeutic classification system was designed using a parent-to-child relationship hierarchy, i.e., for some etc_id's, they identify the therapeutic classifications that are at the most top of the hierarchy, and for some etc_id's, they are related to these parent id's and identify a low-level classification associated to their parent levels. So for our convenience, I separated these parent-to-child hierarchy etc_id's into several single-level classification id's, such as top level (tc_1), second level (tc_2), and third level (tc_3) and so on to the eighth level (tc_8).

The detailed variable description is shown in Appendix E, and the entity relationship diagrams of these datasets are shown in Figure 1.

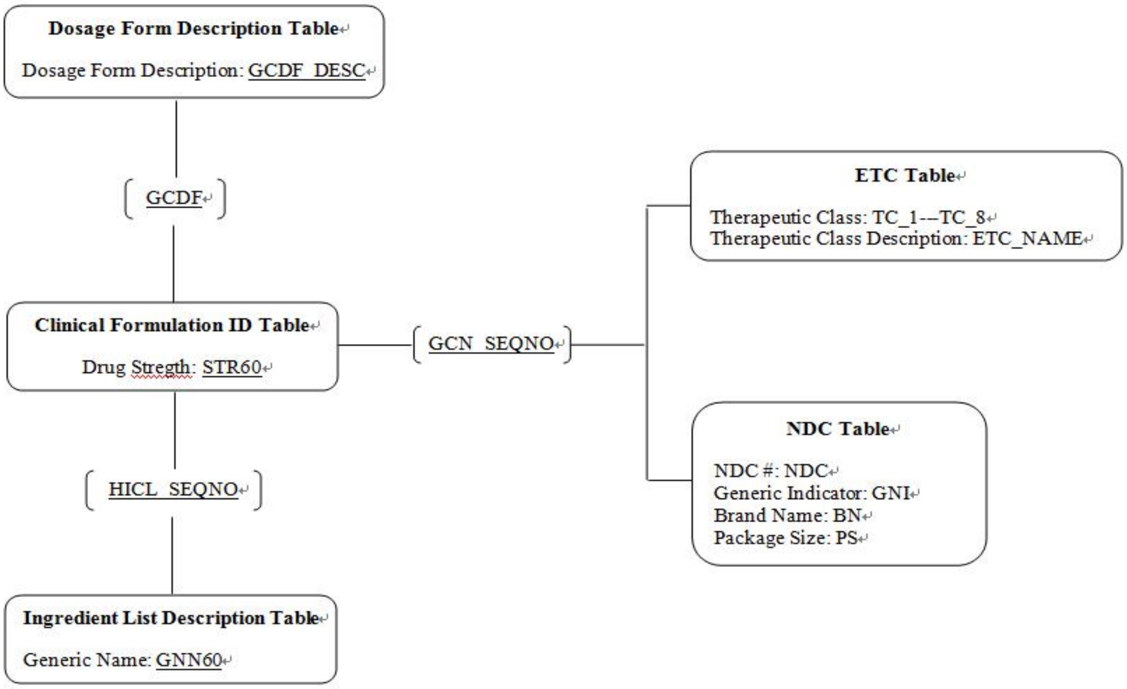

Dosage Form Description Table
Dosage Form Description: GCDF_DESC
GCDF
Clinical Formulation ID Table
Drug Stregth: STR60
GCN_SEQNO
HICL_SEQNO
Ingredient List Description Table
Generic Name: GNN60
ETC Table
Therapeutic Class: TC_1--TC_8
Therapeutic Class Description: ETC_NAME
NDC Table
NDC #: NDC
Generic Indicator: GNI
Brand Name: BN
Package Size: PS

Figure 1. Entity relationship diagram of First DataBank Source Data

2.3 U.S. CENSUS DATA, AREA SUPPLY DATA AND ZIP-HRR DATA

2.3.1 U.S. Census 2000 Data

U.S. Census 2000 Data can be downloaded from the website of U.S. Census Bureau (http://factfinder.census.gov/servlet/DownloadDatasetServlet?_lang=en). We used the data coming from the Summary File 3 (SF3) database, which consists of 813 detailed tables regarding the social, economic and housing characteristics information in Census 2000 data compiled from a sample of approximately 19 million housing units that received the Census 2000 long-form questionnaire. SF3 presents data for the United States, the 50 States, the District of Columbia and Puerto Rico in a hierarchical sequence down to the block group for many tabulations, from which we used the 860 ZIP-code level data to obtain the total population, population for each race/ethnic group, population for each gender, population for each age group, population below the poverty line, household income information, household income for different race/ethnic group, employment and education information.

I downloaded the tables including above variables from SF3 database, then imported them into SAS and linked them together through the ZIP-code.

2.3.2 Area Medical Supply Data

For the area medical supply data, the source data can be downloaded from the website of Health Resources and Services Administration (HRSA). The following datasets were mainly used:

1). Physician Characteristics 2006 data, derived from 2006 American Medical Association Physician Master file (AMA MF), which provides the characteristics of primary care physicians (family physicians, general internists, and pediatricians), specialists, obstetricians /gynecologists at ZIP-code level.

2). Health Professional Shortage Area (HPSA) and Medically Underserved Areas/Medically Underserved Populations (MUA/P) file linked to ZIP-codes, from which researchers could obtain the information about the percentage of 2006 estimated population living in 2007 Primary Care HPSA at ZIP-code level.

2.3.3 Zip-HRR Mapped Data

These data can be downloaded from the Dartmouth Atlas of Health Care. The ZIP-code to Hospital Referral Region (HRR) crosswalk file allows researchers to aggregate data at the ZIP-code level to HRR level. The future analysis on the geographic variation in medication uses would be primarily focused on HRR level.

2.3.4 Data Merging and Creating Segregation Index

After obtaining the above mentioned source data, I merged the census data, area medical supply data and ZIP-HRR mapped data through the unique ZIP-code into one consolidated dataset, which contained all the variables of interest mentioned above. These variables can be adjusted as covariates for future analysis on variation in medication uses.

Another important factor that can be adjusted as a covariate is the index of dissimilarity or segregation, i.e., the segregation of African American or Hispanic to Non Hispanic White, which was calculated based on HRR level. The following equation was employed to calculate the index,[2]

$$D = 0.5 * \sum \left| \frac{P\mathrm{i}_{\mathrm{b/h}}}{P_{b/h}} - \frac{P\mathrm{i}_{\mathrm{w}}}{P_{w}} \right|, \text{ where}$$

$P\mathrm{i}_{\mathrm{b/h}}$ is the Black or Hispanic population in census tract i (ZIP-code area)

$P_{b/h}$ is the total Black or Hispanic population in each HRR area

$P\mathrm{i}_{\mathrm{w}}$ is the non-Hispanic White population in census tract i (ZIP-code area)

P_{w} is the total non- Hispanic White population in each HRR area

2.4 CMS-HCC/RXHCC RISK SCORES

The CMS Hierarchical Condition Categories (HCC) and Prescription Drug Hierarchical Condition Categories (RxHCC) models are implemented to adjust Medicare capitation payments to health care plans for the health and prescription expenditure risk of their enrollees. These risk scores are also considered as important covariates in our future variation analysis.

The CMS website (https://www.cms.gov/) provides HCC and RxHCC software to be downloaded to calculate these risk scores. For example, the HCC software includes a SAS program that calls several SAS macros to create HCC score variables using coefficients from different regression models.

2.5 EXAMPLE OF DATASET CONSTRUCTION---DONUT HOLE STUDY

After obtaining all of above source data, I began to assemble data to create the final analytic dataset for each project. I took one of the ongoing projects, the “donut hole” study, as an example to show how these source data can be linked together to construct analytic data.

2.5.1 Introduction to “Donut Hole”

The standard benefit that Part D plans offer is defined in terms of the benefit structure, which may vary by year. In 2007, this benefit required a $265 deductible, and the beneficiary paid 25% of the cost of covered Part D prescription drugs up to an initial coverage limit of $2,400. [3] Once the initial coverage limit was reached, the beneficiary entered into the Coverage Gap period, more commonly referred to as the “Donut Hole”, in which they paid the full cost of their prescriptions up to $5,451 in total, i.e. their true out-of-pocket expenditures (TrOOP) on formulary drugs for this year reached $3,800, which is the sum of the deductible, initial copayment before “donut hole” threshold, and payment in coverage gap. After this second threshold, the beneficiary entered into the catastrophic coverage phase, in which they only paid 5% coinsurance and the plan paid the rest 95% in excess of $5,451. [3] The standard benefit design is shown below in Figure 2.

This project mainly applied the pre-post comparison methods on existing large healthcare data to investigate the inferable causality between Part D policy interventions and its health outcomes, especially the effects of the “Donut Hole” for vulnerable American populations such as under-served minorities, patients with severe mental health and multiple medical conditions, i.e., the racial/ethnic disparity or geographic variations in use of different medications and medical care services before and after the “Donut Hole” was under investigation.

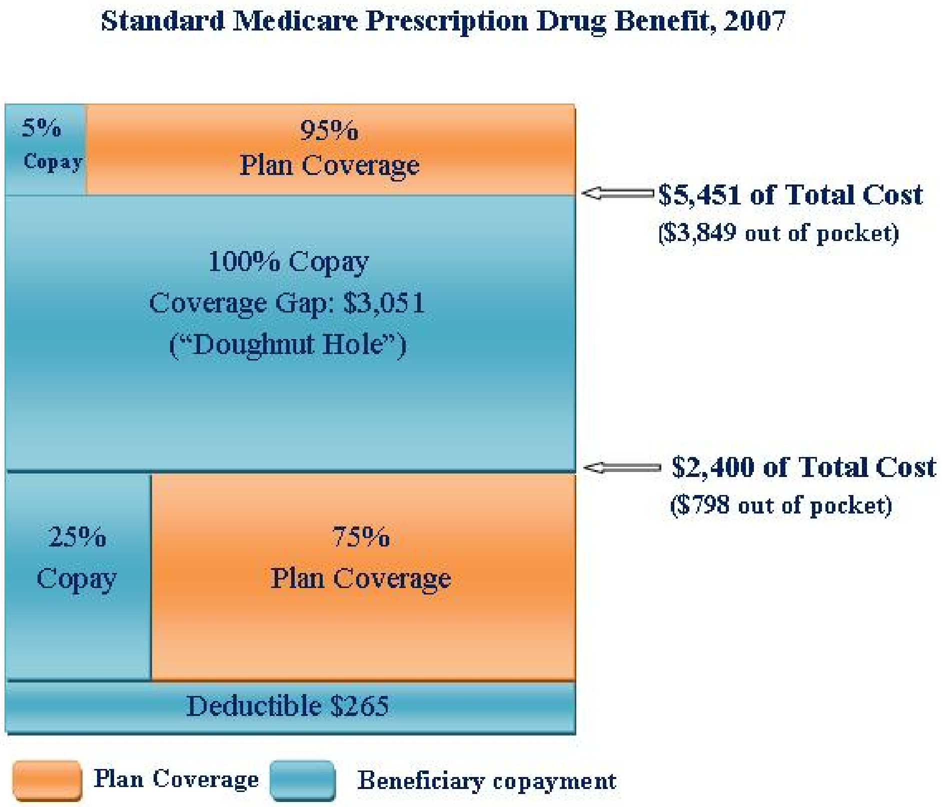


Figure 2. Standard Medicare Prescription Drug Benefit Design, 2007

2.5.2 Selection criteria

In this study, we wanted to investigate the variation on medication uses, comparing before and after donut-hole period information. The study population was identified from the beneficiary annual summary file as well as the Part D denominator file. I linked these two files together through the unique beneficiary identifier (*bene_id*) to form a consolidated file, on which the selection criteria were implemented,

1) Beneficiary still alive (bene_dod with missing value);

2) 12-month continuous enrollment in both Part A and B program (*ab_mo_cnt* = 12);

3) 12-month continuous enrollment in Part D program (*plncovmo* = 12);

4) No managed care or non-fee-for-service coverage in any month (*hmo_mo* = 0);

5) All in stand-alone Part D plans (the first letter of each month's contract ID is 'S', i.e., *CNTRCT<mon#>* = "S")

2.5.3 Defining outcome variable and other covariates variables

Following the above step, the study population was determined. The outcome variable and covariates were defined next. As we investigated the variation on medication uses as well as the pre- and post- donut hole information, the first thing was to define the pre- and post- donut hole period, and next summarize the drug spending and counts for each beneficiary during these two periods. To fulfill this approach, I linked the Part D event data with above defined study sample through the beneficiary identifier (*bene_id*) to get all drug claims for each beneficiary in the study sample. Then I identified those drug events in the phases that covered the "donut hole" (*bnftphas* = 'DI', 'DC', 'PI', 'PC', 'II', 'IC', 'CC') or catastrophic period (*bnftphas* = 'DC', 'PC', 'IC', 'CC'), and then created two indicators to describe whether or not each drug event was in the "Donut Hole" or catastrophic phase, where if the drug event's phase was in those phases, the indicator would be assigned a value of '1'. After that, I defined the first date triggering the "Donut Hole" or catastrophic period, which was the same with the earliest prescription date (*srvc_dt*) in these periods for each beneficiary. Once obtaining these information, I moved onto summarizing the person-level total and monthly averaged drug spending and counts for pre- and post- donut hole period respectively.

I created the outcome variables, drug spending and counts, for each beneficiary in the study sample. Next, I linked other source data with the study sample to add some other covariates, e.g., linking with Part D plan characteristics file through contract ID (*cntrctid*) and benefit package ID (*plan_id*) to add deductible amount (*ded_amt*), initial coverage limit type (*icl_app*), initial coverage limit amount (*icl_amt*), drug benefit type (*drgbentp*), gap coverage type (*gapcovtp*); linking with CMS-HCC/RxHCC risk score files through beneficiary identifier (*bene_id*) to add these two types of risk score; linking with census-area supply data to add HRR number, segregation indices through beneficiary's ZIP-code.

The entity relationship diagrams of these datasets are shown in Figure 3.

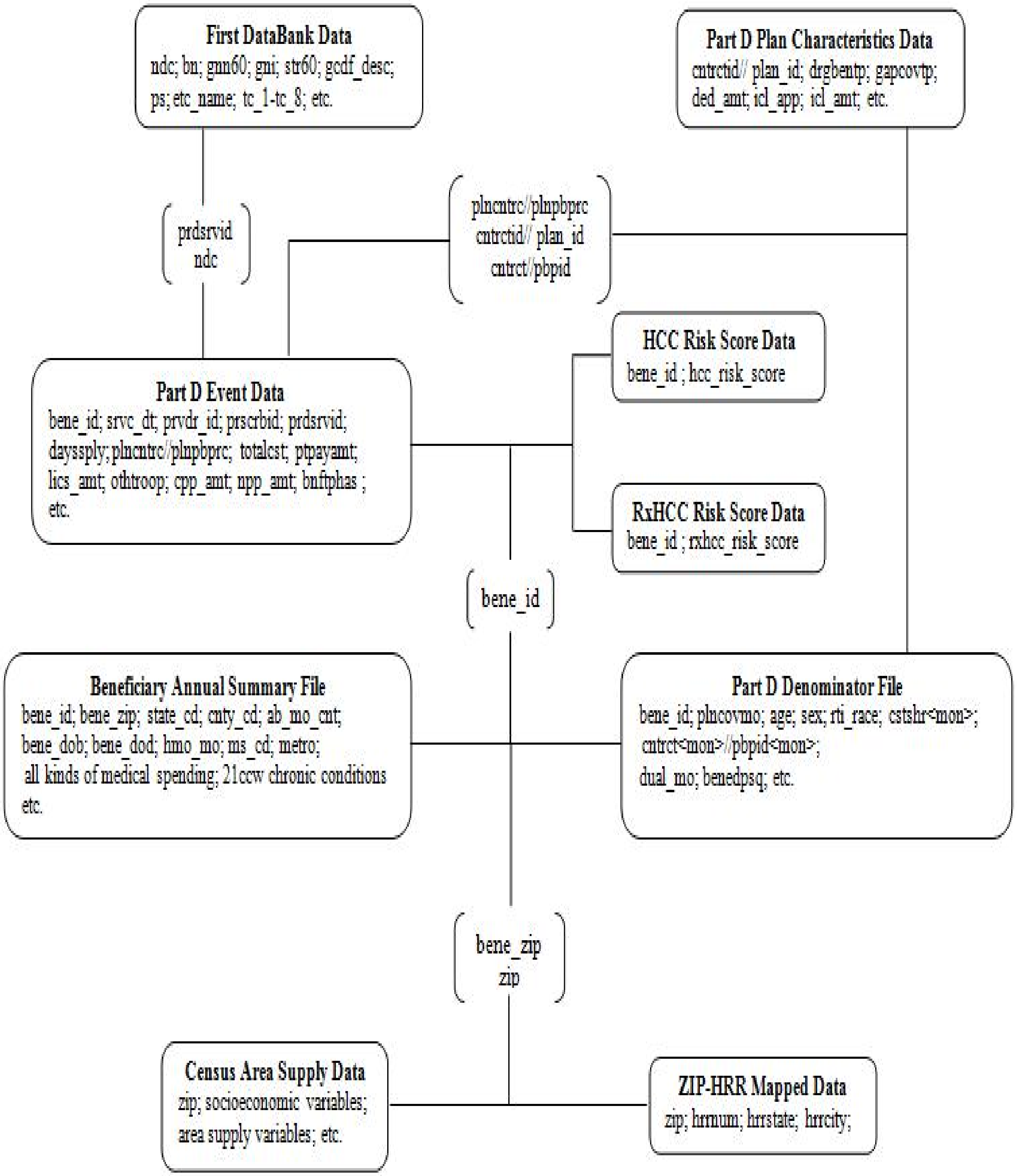

First DataBank Data
ndc; bn; gnn60; gni; str60; gcdf_desc;
ps; etc_name; tc_1-tc_8; etc.
Part D Plan Characteristics Data
cntrctid// plan_id; drgbentp; gapcovtp;
ded_amt; icl_app; icl_amt; etc.
prdsrvid
ndc
plncntrc//plnpbprc
cntrctid// plan_id
cntrct//pbpid
HCC Risk Score Data
bene_id ; hcc_risk_score
Part D Event Data
bene_id; srvc_dt; prvdr_id; prscrbid; prdsrvid;
dayssply; plncntrc//plnpbprc; totalcst; ptpayamt;
lics_amt; othtroop; cpp_amt; npp_amt; bnftphas ;
etc.
RxHCC Risk Score Data
bene_id ; rxhcc_risk_score
bene_id
Beneficiary Annual Summary File
bene_id; bene_zip; state_cd; cnty_cd; ab_mo_cnt;
bene_dob; bene_dod; hmo_mo; ms_cd; metro;
all kinds of medical spending; 21ccw chronic conditions
etc.
Part D Denominator File
bene_id; plncovmo; age; sex; rti_race; cstshr<mon>;
cntrct<mon>//pbpid<mon>;
dual_mo; benedpsq; etc.
bene_zip
zip
Census Area Supply Data
zip; socioeconomic variables;
area supply variables; etc.
ZIP-HRR Mapped Data
zip; hrrnum; hrrstate; hrrcity;

Figure 3. Entity relationship diagram of "Donut Hole" Study

2.6 CONCLUSIONS

During my internship, I served as Data Manager, SAS programmer and Statistician. As data manager and SAS programmer, I was responsible for maintaining and manipulating the source data, creating analytic datasets from the source data, and generating detailed readable data dictionaries for each dataset I have created. I have not only practiced my skills in advanced SAS programming language, but also improved my ability to keep things in an organized way. For example, when I wrote the SAS code, I always kept in mind to put comments as much as possible. In this way, people including me, could easily go back to the codes to check the problems or repeat the procedures. Another example is that I generated an Excel log file to record the information for each dataset I have created, including the updated date, source data I have used, the description of the new dataset, the purpose of this dataset and the SAS codes to creating it. In this way, people can easily find the dataset they want and understand how the dataset was constructed.

When calculating the segregation index and generating the HCC and RxHCC risk scores, I worked as a statistician. I conducted online literature searches to find references on how to calculate the segregation index and figure out what methods those references have used. I made decisions on what method I should use to calculate the index. For the risk scores, I read carefully the software documentation to understand what regression models they are using to calculate the scores and what the purpose of each step is when the software calls those SAS macros. In this way, I was able to function independently.

3.0 VARIATION IN USE OF $4 GENERIC PROGRAM AND POTENTIAL SAVINGS AMONG MEDICARE PART D BENEFICIARIES

Spending for prescription drugs in US is rising. It was $234.1 billion in 2008 and $216 .7 billion in 2006, nearly 6 times the $40 billion spent in 1990. [4] Generic drugs are typically less expensive than brand-name drugs, and prices for generics have historically increased less than those for brand-name drugs. [4-6] The U.S. Food and Drug Administration (FDA) examines the generic formulations and approves them as bioequivalent to brand-name drugs in safety and quality.[7] Therefore, use of generic formulation instead of a brand-name for multisource drug (i.e., those with more than one generic available) could be one potential strategy for limiting drug expenditures.

Wal-Mart first launched a highly-discounted drug program in 2006, which is called *$4 Prescription Program* and covers a 30-day generic prescription at $4 or a 90-day generic prescription at $10.[8] Then during the following years several retail stores, like Target, Giant Eagle, Walgreen, and CVS, launched their own low-cost generic programs which are quite similar to Wal-Mart’s $4 program and the number of covered generic drugs keeps increasing to above 400 (can be seen from the $4 prescriptions lists from the websites of those retailers). Nowadays, due to the low cost and easy access to these low-cost programs, we could assume a huge amount of potential savings on the medication spending for the nation. However, we could find no existing investigation about the exact potential savings as these low-cost programs are relatively new to the society.

We used the 2007 Medicare Part D data and First DataBank data to identify those who ever used at least one drug that were commonly available from the Wal-Mart *$4* and then estimated potential savings by switching prescriptions to $4 programs among those who paid more than $4 per 30 days for these drugs. Also we investigated the variation in use of these $4 drugs to see if the beneficiaries' insurance status, demography, living area (i.e., urban or rural, the distance to the closest Wal-Mart pharmacy store) and other factors could affect the likelihood of use of these drugs.

3.1 METHODS

3.1.1 Data

Wal-Mart first launched this low-cost generic program to the nation in 2006, and the number of $4 generics in the lists from 2006 to 2009 didn't change dramatically. Only a slight increase can be seen. And we compared the $4 generic lists from Wal-Mart with other pharmacy retailers as well, and found the similarities among them. Thus we used the latest updated 2009 list to determine which drugs are available through the $4 program. For the current analysis, we focused on those in tablet and capsule forms, which is easy to calculate the drug use and spending.

First DataBank data contain complete drug information, such as NDC number, brand name, equivalent generic name, strength, dosage form, package size, generic or brand indicator, and the therapeutic class. We matched this data with Wal-Mart's $4 list by drug name to identify all NDC's that are available from Wal-Mart's $4 program list.

For drug spending and counts, we used 2007 Medicare Part D event data and matched them with above matched NDC data through drugs' NDC number. In this way, we obtained the drug expenditure information for all NDC's that are available in Wal-Mart's $4 program list.

In the end, we merged this drug information data with the "donut hole" study sample through the beneficiaries identifier and thus, identified all those 2007 Medicare Part D beneficiaries who ever used the Wal-Mart's $4 program drugs. The final data was based on drug claim level.

3.1.2 Study population

3.1.2.1 Drug Claims

For all the drug claims, we separated them into three sub-groups,

1) All claims that were probably filled through $4 program: beneficiaries paid $4 for 30-day or $10 for 90-day and they have no other plan payments or federal low-income-subsidy (LIS) or other public assistance subsidy, i.e., the total drug cost was just the $4 program cost;

2) All claims that were considered as potential savings by switching to $4 program: beneficiaries paid more than $4 program cost, thus they could save by switching their prescriptions to $4 program, and Medicare plans, LIS, and public subsidy would also save by such switching because they would not pay anymore;

3) All claims that were not considered as potential savings: beneficiaries paid less than $4 program cost because Medicare plans, LIS, and public subsidy covered most part of the drug costs, thus they have no incentives to switch to $4 program.

Among each sub-group, we identified those brand-name claims as well as generic claims.

3.1.2.2 Subjects

We assigned the beneficiaries to two sub-groups,

1) Those filled at least one claim through $4 program, no matter if they paid greater or less than $4 for their other claims. We considered them as current users.

2) Those filled at least one potential saving claim, but excluding those who also filled at least one claim through $4 program or those subsidized by LIS. We considered the second sub-group as non-users.

We compared the demography between current users and non-users to see if the use of $4 program varies significantly.

3.1.3 Potential savings

We calculated three parts of potential savings among the second sub-group of claims.

First, we calculated the potential savings by switching from regular generic to the $4 program. In this way, the beneficiary would pay $4 for 30-day or $10 for 90-day, and the plan and public subsidy would not pay anymore.

Second, we calculated the potential savings by switching from brand name to the $4 program, which was similar to the savings in first part. The calculations for these two parts are as following,

Total saving = $\sum_{all\ claims}(Total\ drug\ cost - \$4\ program\ cost)$

Beneficiaries saving = $\sum_{all\ claims}(Beneficiary's\ copayment - \$4\ program\ cost)$

Plan saving = $\sum_{all\ claims}(\ Plan\ payment)$

Public Subsidy saving = $\sum_{all\ claims}(\ Public\ subsidies\ payment)$

Last part, we calculated the potential savings by switching from brand name to regular generic. We calculated this part of savings because it might not be feasible and available to the society if everyone switches to $4 program. This calculation was based on drug level. First we calculated the per 30-day average drug cost for each generic drug. Then using these per 30-day average costs and the days of supply of the brand version drugs, we estimated their total costs assuming they were switched to their corresponding regular generic versions. In this way, we estimated the savings from brand name to regular generics. However, this estimation was based on drug level; thus, we could not estimate the person level average saving.

Total saving = $\sum_{all\ drugs}\ \{Total\ drug\ cost\ for\ brand\ name\ drug - \ (per\ 30 - day\ cost\ for\ alternative\ generic\ drug * (days\ of\ supply\ for\ brand\ name)/30\}$

For those claims in the third sub-group, the beneficiaries received subsidies and therefore paid less than $4 for these drugs, they would not have the incentives to use this $4 program. We only calculated the potential savings from the perspectives of Medicare plan and other public subsidies.

3.1.4 Variation in use of $4 generic program

By applying the chi-square test and the t-test, we compared the demographic factors between the current users and the non-users to check what kind of population could be more or less likely to use this $4 program.

As Wal-Mart is the first national store that offered this $4 program since 2006, we assumed that majority of the users in our data (year of 2007) obtained the access to $4 program from Wal-Mart. In this case, we calculated the distance for each beneficiary to the closest Wal-

Mart based on their ZIP-code of residence. This calculation was accomplished by using the ArcGIS software. We obtained the longitude and latitude information for each Wal-Mart store around the United States as well as for the center of each ZIP-code area. Then by using the ArcGIS software, we calculated the distance from the center of each ZIP-code area to its closest Wal-Mart store. Those beneficiaries who lived in the same ZIP-code area have the same distance. Then by fitting a logistic regression model, we checked the likelihood of using $4 program among different levels of multiple factors including this distance. Then we checked the residuals, outliers and goodness-of-fit for the model.

3.2 RESULTS

3.2.1 Summary of Prescription Claims among Medicare Part D Beneficiaries

We found that (shown in Table 1), in 2007 there were totally 9,918,962 claims for 106 kinds of drugs that were commonly available from Wal-Mart's $4 program, filled by 595,693 Medicare beneficiaries.

Among these claims, only 9.8% of them were filled through 76 kinds of brand-name drugs by 22.1% of the sampled beneficiaries, and the rest were filled through 103 kinds of generic drugs by 98.6% of the beneficiaries. The beneficiaries overlapped between these two kinds of claims because one beneficiary could fill multiple prescription claims.

Among these brand-name drug claims, about 48% of them were filled with beneficiary's copayment greater than $4 program cost (i.e., $4 for 30-day, $10 for 90-day), thus they were considered as potential saving claims. And 51% of claims were filled with beneficiary's copayment less than $4 program cost due to the subsidies by LIS or other public assistance programs. We excluded this part from the calculation of potential savings because of lack of incentives to switch to $4 program.

Among the generic drug claims, only 24.7% could be considered as potential saving claims, and the rest were excluded from the calculation of potential savings because they were subsidized by LIS (73%) or they were filled through the $4 program (2.3%).

Table 1. Summary of Prescription Claims among Medicare Part D Beneficiaries

	# of Claims	# of Beneficiary*	% of total claims	% of Beneficiary	% of brand name claims	% of generic claims	# of drugs
Total Claims	9,918,962	595,693	-	-	-	-	106
Brand-name Claims	972,103	131,812	9.8	22.1	-	-	76
Not using $4, could save	465,218	73,776	4.7	12.4	47.9	-	75
Copayment ≤ $4	496,363	74,650	5.0	12.5	51.1	-	74
Generic Claims	8,946,859	587,357	90.2	98.6	-	-	103
Not using $4, could save	2,212,048	244,550	22.3	41.1	-	24.7	102
Copayment ≤ $4	6,528,304	478,510	65.8	80.3	-	73.0	103
Currently using $4	206,507	33,840	2.1	5.7	-	2.3	101

**The numbers don't add up because one beneficiary could fill multiple claims.*

3.2.2 Demography of study population

Table 2 shows that 33,840 beneficiaries were defined as current users and 270,918 were defined as non-users. Compared to non-users, current users were more likely to be white (92.5% vs. 91.7%, *p<0.0001*), younger (72.7±8.6 vs. 74.1±9.0*, p<0.0001*), younger than 74 (59.8% vs. 53%, *p<0.0001*), live outside of the urban area (59.5% vs. 67.1%, , *p<0.0001*) and having 4 or more chronic conditions (21.9% vs 21.1%, *p=0.0005*). The significant p-values are most likely due to the large sample size of the two populations.

Table 2. Demography of study population

	Non-users (270,918)	Current users (33,840)	*p-value**
Female (%)	62.6	63.0	*0.147*
Race (%)			*<0.0001*
White	91.7	92.5	*<0.0001*
Black	3.9	3.8	*0.380*
Hispanic	2.5	2.4	*0.265*
Asian	1.1	0.7	*<0.0001*
Native	0.2	0.2	*0.240*
Metropolitan Area (%)	67.1	59.5	*<0.0001*
Age Group (%)			*<0.0001*
<65	6.8	8.9	*<0.0001*
65-74	46.2	50.9	*<0.0001*
75-84	34.3	31.9	*<0.0001*
85-99	12.7	8.3	*<0.0001*
Chronic Conditions (%)			*<0.0001*
0	13.4	12.6	*<0.0001*
1-3	65.5	65.5	*0.982*
4 or more	21.1	21.9	*0.0005*
Age (Mean±STD)	74.1±9.0	72.7±8.6	*<0.0001*
Risk Score (Mean±STD)	0.89±0.29	0.92±0.29	*<0.0001*

3.2.3 Potential Savings

Table 3 and Table 4 show that if all generic potential saving claims switch to $4 generic claims, the total annual saving and beneficiaries' saving would be $17,591,736 and be $6,949,582, respectively by 244,550 beneficiaries. And corresponding per person savings would be $71.94 (95% CI, $71.51-$72.36) and $28.42 (95% CI, $28.25-$28.58). Only part of these people had Medicare plan payment and public subsidy for their prescriptions, the total plan savings would be $10,466,366 and the public subsidy saving would be $150,074 by 164,403 beneficiaries; per person saving would be $63.66 (95% CI, $63.20-$64.12) and $31.34 (95% CI, $29.84-$32.84), respectively.

For brand-name potential saving claims, if they switch to $4 generic claims, the total annual saving would be $19,780,945 by 73,776 beneficiaries, and per person saving would be $268.12 (95% CI, $265.52-$270.73). And the total beneficiaries' saving would be $7,793,562; per person saving would be $105.64 (95% CI, $104.50-$106.78). Only part of these people had Medicare plan payment and public subsidy for their prescriptions, the total plan savings would be $9,804,742 by 43,989 beneficiaries; per person saving would be $222.90 (95% CI, $220.43-$225.36), and the public subsidy saving would be $2,151,831 by 10,548 beneficiaries; per person saving would be $204 (95% CI, $199.77-$208.24).

Totally, if all potential saving claims switch to $4 generic program regardless of brand-name claims or generic claims, the total and beneficiaries' annual savings would be $37,372,680 and $14,743,144 by 267,285 beneficiaries, and per person saving would be $139.82 (95% CI, $138.88-140.77) and $55.16 (95% CI, $54.75-$55.56), respectively. And the plan saving would be $20,271,109 by 185,017 beneficiaries; per person saving would be $109.56 (95% CI, $108.74-$110.39). For public subsidy, the total savings would be $2,301,905 by 14,471 beneficiaries; per person saving would be $159.07 (95% CI, $155.67-$162.47).

In addition, we estimated the potential savings for switching the brand-name claims to regular generic claims, because sometimes it's not quite reasonable and feasible for all brand-name claims to be switched to the $4 generic claims. In this case, the total and beneficiaries' annual savings would be $14,414,229 and $5,506,894, and the plan saving would be $6,849,071, and the public subsidy saving would be $2,043,624. These savings were estimated based on drug levels instead of person levels which has been discussed in method section, thus per person savings were not estimated.

Table 3. Summary of Potential Savings

	Total Saving	Beneficiary Saving	Plan Saving	Public Subsidies Saving	# of Beneficiaries
From regular to $4 generic[1]	$17,591,736	$6,949,582	$10,466,366	$150,074	244,550
From brand name to $4 generic[2]	$19,780,945	$7,793,562	$9,804,742	$2,151,831	73,776
Total saving to $4 generic*	$37,372,680	$14,743,144	$20,271,109	$2,301,905	267,285
From brand name to regular generic[3]	$14,414,229	$5,506,894	$6,849,071	$2,043,624	73,776

*** = 1 + 2**

Table 4. Summary of Potential Savings (Per person)

	Total Saving	Beneficiary Saving	Plan Saving	Public Subsidies Saving
From regular generic to $4 generic	$71.94 ($71.51-$72.36)	$28.42 ($28.25-$28.58)	$63.66 ($63.20-$64.12) (n=164,403)	$31.34 ($29.84-$32.84) (n=4,789)
From brand name to $4 generic	$268.12 ($265.52-$270.73)	$105.64 ($104.50-$106.78)	$222.90 ($220.43-$225.36) (n=43,989)	$204.00 ($199.77-$208.24) (n=10,548)
Total saving to $4 generic	$139.82 ($138.88-140.77)	$55.16 ($54.75-$55.56)	$109.56 ($108.74-$110.39) (n=185,017)	$159.07 ($155.67-$162.47) (n=14,471)

For those beneficiaries who received federal LIS or other public subsidies, they paid less than $4 for most of their claims; they would not have incentives to switch themselves. We only calculated the total cost of these claims as well as their copayments, payment by the plans and public subsidies, which is shown in Table 5. For those generic claims, the plans paid $34,410,710 in a year for 478,510 beneficiaries, and public subsidies paid $22,443,909. And for those brand name claims, the plans paid $8,794,458 in a year for 74,650 beneficiaries, and public subsidies paid $8,212,482. However, if these brand name claims switch to the regular generic claims, the plans could still save $5,085,988 in total, and public subsidies would save $5,684,736 in total.

Table 5. Summary of Spending for those with federal or public assistance

	Total Cost	Beneficiary Copayment	Plan Payment	Public Subsidies payment	# of Beneficiaries
Generic claims	$64,678,548	$7,751,557	$34,410,710	$22,443,909	478,510
Brand name claims	$17,827,548	$796,694	$8,794,458	$8,212,482	74,650
Total	$82,506,096	$8,548,251	$43,205,168	$30,656,391	483,541
Saving from brand name to regular generic	$11,049,040	$263,699	$5,085,988	$5,684,736	74,650

3.2.4 Variation in use of $4 program

To investigate the variation in use of $4 program, we fitted logistic regression models, modeling on the probability of using $4 program. For the response variable, we created an indicator for the current users with value of '1' versus the non-users with value of '0'. And the potential explanatory variables identified from our data included demography factors (age, gender, and race), geographic residence information (metropolitan area indicator, distance to the closest Wal-Mart), socioeconomic information (percentage of poor within each ZIP-code area, percentage of who ever finished high school within each ZIP-code area), medical conditions (disabled indicator, number of chronic conditions, risk scores). The model is expressed as

Logit(p) = log(p/1-p) = $\alpha + \beta' x$

where p denotes the response probability to use $4 program, and α is the intercept parameter and β is the vector of slope parameters.

We applied the stepwise selection method (Hosmer and Lemeshow, 2000) to choose the important variables. The p-value for variable entry was set at 0.15 and that for variable removal was set at 0.2 to obtain a continued “significant “contribution.[9] The Hosmer-Lemeshow test for goodness-of-fit model check was applied after selection of variables.

The results are shown in Table 6, all the variables were important and included in the model. Table 7 shows that for per 5 years older, the beneficiary is 10% less likely to use the $4 program (OR = 0.9, 95% CI, 0.89-0.91). A woman is 4% more likely to use the $4 program than a man, holding other variables the same. (OR = 1.04, 95% CI, 1.01-1.06). For those who live in rural area, per 5 miles further to Wal-Mart store, they are 14% less likely to use the $4 program (OR = 0.86, 95% CI, 0.85-0.87); and per 10 miles further, 26% less likely to use (OR = 0.74, 95% CI, 0.72-0.75). For those live in metropolitan area, per 5 miles further to Wal-Mart store, they are 23% less likely to use the $4 program (OR = 0.77, 95% CI, 0.75-0.79); and per 10 miles further, 41% less likely to use (OR = 0.59, 95% CI, 0.57-0.62). Asian (OR = 0.71, 95% CI, 0.62-0.82) and black (OR = 0.83, 95% CI, 0.78-0.88) people are less likely to use than white people, while Hispanic (OR = 1.13, 95% CI, 1.04-1.22) are more likely to use. Other results are shown in Table 7 as well.

Table 6. Estimated coefficient of logistic model

Effect	Coefficients	Std	*p*
Age	-0.0207	0.001	<*.0001*
Female (vs. Male)	0.0182	0.006	*0.0041*
Distance	-0.0303	0.001	<*.0001*
Metropolitan area (vs. Rural area)	-0.1192	0.009	<*.0001*
Distance*Metropolitan	-0.0217	0.002	<*.0001*
Race (vs. White)			
Black	-0.08	0.044	*0.0684*
Hispanic	0.226	0.049	<*.0001*
Asian	-0.235	0.070	*0.0008*
American Native	0.102	0.118	*0.3880*
Percentage of poor	0.019	0.001	<*.0001*
Percentage of who finished high school	0.029	0.001	<*.0001*
Number of chronic conditions	0.017	0.004	<*.0001*
Disabled (vs. Non Disabled)	-0.097	0.014	<*.0001*
Risk score	0.312	0.025	<*.0001*

Table 7. Estimated odds ratio (OR) for each variable in the logistic model

Effect	Estimated OR	95% CI Lower limit	95% CI Upper limit
Age (per 5 years older)	0.902	0.894	0.909
Female (vs. Male)	1.037	1.012	1.063
Metropolitan area (vs. Rural area)	0.788	0.761	0.816
Distance (per 5 miles further, rural area)	0.860	0.851	0.868
Distance (per 10 miles further, rural area)	0.739	0.724	0.754
Distance (per 5 miles further, metropolitan area)	0.771	0.754	0.788
Distance (per 10 miles further, metropolitan area)	0.594	0.569	0.621
Race (vs. Wihte)			
Black vs. White	0.828	0.777	0.883
Hispanic vs. White	1.125	1.038	1.219
Asian vs. White	0.709	0.616	0.816
American Native vs. White	0.993	0.764	1.292
Percentage of poor	1.020	1.018	1.021
Percentage of who finished high school	1.030	1.028	1.031
Number of chronic conditions	1.018	1.009	1.026
Disabled (vs. Non disabled)	0.824	0.779	0.872
Risk score	1.366	1.301	1.434

The Hosmer-Lemeshow test (Table 8) shows that this model did not fit adequately (p<0.05). However, from the residual plots, we could find only few outliers, which indicated that the model fitted well. The plots are shown in Appendix F.

Table 8. Hosmer-Lemeshow Goodness-of-Fit Test

Group	Total	Observed Events	Expected Events	Observed Non Events	Expected Non Events
1	28865	1065	1163.618236	27800	27701.38176
2	28866	1891	1902.842473	26975	26963.15753
3	28865	2336	2342.766745	26529	26522.23326
4	28865	2716	2697.854664	26149	26167.14534
5	28865	2977	3019.489509	25888	25845.51049
6	28866	3330	3328.505112	25536	25537.49489
7	28868	3763	3654.818949	25105	25213.18105
8	28865	4160	4030.171422	24705	24834.82858
9	28865	4515	4518.410484	24350	24346.58952
10	28857	5411	5505.506005	23446	23351.49399

Chi-square	DF	*p-value*
20.1475	8	*0.0098*

Distance and percentage of poor and percentage of who finished high school were calculated based on ZIP-code level, i.e., for those beneficiaries lived in the same ZIP-code area, they have the same information of these three variables. Thus there is a correlated data issue. We applied the GEE model with repeated measures to treat ZIP-code area as a cluster. The estimated coefficients for each variable are shown in Table 9. We found that most of the estimated coefficients are similar to those in the above model, which is probably because of the large number of clusters (28,956 ZIP-codes) with small number of observations (0 to 258) in each cluster, as well as the very small working correlation which is 0.00002. The GEE fit criteria (QIC) is about 195831, which is huge and not supporting the adequate model fit.

Table 9. Estimated Coefficients of GEE Model

Effect	Coefficients	Std	*p*
Age	-0.0207	0.001	<*.0001*
Female (vs. Male)	0.0359	0.0125	*0.0042*
Distance	-0.0303	0.0015	<*.0001*
Metropolitan area (vs. Rural area)	-0.240	0.0272	<*.0001*
Distance*Metropolitan	-0.0216	0.0038	<*.0001*
Race (vs. White)			
Black	-0.189	0.0364	<*.0001*
Hispanic	0.117	0.0442	*0.0079*
Asian	-0.344	0.0880	<*.0001*
American Native	-0.395	0.1059	*0.0002*
Percentage of poor	0.019	0.0013	<*.0001*
Percentage of who finished high school	0.029	0.0010	<*.0001*
Number of chronic conditions	0.017	0.004	<*.0001*
Disabled (vs. Non Diabled)	-0.195	0.029	<*.0001*
Risk score	0.313	0.024	<*.0001*

3.3 DISCUSSIONS

We found that in 2007, among those beneficiaries taking the drugs available from $4 generic program, only 5.7% of them (33,840 out of 595,693) actually filled their prescriptions through this program, which accounted for only 2% of the total prescription claims in the year. And 45% of the beneficiaries (267,285 out of 595,693) could potentially save $37,372,680 in total and $139.82 per person (95%CI, $138.88-$140.77) by switching their prescriptions to $4 generic program. As the data we used was only 5% sample of total Medicare beneficiaries, the total potential savings would probably be much more than this amount.

For our current analysis, we assumed these beneficiaries filled their prescriptions through Wal-Mart because Wal-Mart first offered this program since 2006 and our analysis was based on 2007 data. In fact, it is not quite reasonable or feasible for all the beneficiaries to go to Wal-Mart to fill their prescriptions. We can verify this later when the new pharmacy data are available in near future, which would indicate where they filled their prescriptions. In that case, we need to recalculate the distance from their residence area to the stores where they actually filled the prescriptions.

For the variation in use of $4 generic program, we found younger people and women are more likely to use this program. Further distance to Wal-Mart can reduce the likelihood of use. Black and Asian people are less likely to use compared with white people, while Hispanic are more likely to use. In addition, in area with higher percentage of poor or higher percentage of who finished high school, people are more likely to use this program. Our findings reinforce the importance of understanding the drivers of variation in use of $4 program. Both areal-level variation and patient characteristics could potentially affect the use of $4 program. These findings may offer us an opportunity to gain insight into the potential for public policy actions to improve the value of the health care delivered in the United States. In future analysis, we need to investigate more factors, like beneficiaries' insurance status, access to other retail stores.

In conclusion, at present, the *$4 program* is still new to both patients and healthcare providers who know little about it. Patients, especially those without insurance or those with low income, could get great benefit from this highly discounted program to lower their high costs on medications. In addition, our research intended to remind the healthcare providers of these low-cost prescription programs. It is quite helpful for physicians to get familiar with these programs and provide these options to their patients to help ease their financial burden from taking medications.

4.0 SUMMARY OF INTERNSHIP

During my internship, I acted as data manager, programmer and statistician. I played a leading role on a small piece of project, which has provided me great opportunity to learn how to be fully involved into a research project, how to think and resolve problem independently, how to write up a research paper and how to deal with multiple tasks to meet the deadline. I also obtained great opportunity to apply my statistical knowledge in a real-world setting.

I hope my experience in this student internship is helpful for those MS student who wish to fulfill their degree requirements through this option.

APPENDIX A:

DATA DICTIONARY OF BENEFICIARY ANNUAL SUMMARY FILE

Variable	Data Type	Description
BENE_ID	Char	Encrypted 723 Beneficiary ID
STATE_CD	Char	State code (SSA)
CNTY_CD	Char	County code (SSA)
BENE_ZIP	Char	Zip code of residence
METRO	Char	Metro Status
SEX	Char	Sex
BENE_DOB	Num	Date of birth (Date)
BENE_DOD	Num	Date of death (Date)
MS_CD	Char	Medicare status code
A_MO_CNT	Num	Number of Months enrolled in Part A
B_MO_CNT	Num	Number of Months enrolled in Part B
AB_MO_CNT	Num	Number of Months enrolled in both Part A and B
HMO_MO	Num	Number of non Fee-for-Service Months
BUYIN_MO	Num	Number of Months Medicaid Coverage
MEDREIMB_IP	Num	Inpatient annual Medicare reimbursement amount
BENRES_IP	Num	Inpatient annual beneficiary responsibility amount
PPPYMT_IP	Num	Inpatient annual primary payer reimbursement amount
MEDREIMB_SNF	Num	Skill Nursing Facility annual Medicare reimbursement amount
BENRES_SNF	Num	Skill Nursing Facility annual beneficiary responsibility amount
PPPYMT_SNF	Num	Skill Nursing Facility annual primary payer reimbursement amount
MEDREIMB_OP	Num	Outpatient Institutional annual Medicare reimbursement amount
BENRES_OP	Num	Outpatient Institutional annual beneficiary responsibility amount
PPPYMT_OP	Num	Outpatient Institutional annual primary payer reimbursement amount
MEDREIMB_CAR	Num	Carrier annual Medicare reimbursement amount
BENRES_CAR	Num	Carrier annual beneficiary responsibility amount
PPPYMT_CAR	Num	Carrier annual primary payer reimbursement amount
MEDREIMB_DME	Num	Durable Medical Equipment annual Medicare reimbursement amount
BENRES_DME	Num	Durable Medical Equipment annual beneficiary responsibility amount
PPPYMT_DME	Num	Durable Medical Equipment annual primary payer reimbursement amount
MEDREIMB_HH	Num	Home Health Agency annual Medicare reimbursement amount
PPPYMT_HH	Num	Home Health Agency annual primary payer reimbursement amount
MEDREIMB_HS	Num	Hospice annual Medicare reimbursement amount
BENRES_HS	Num	Hospice annual beneficiary responsibility amount

Variable	Data Type	Description
PPPYMT_HS	Num	Hospice annual primary payer reimbursement amount
IPSTY	Num	Annual number of Inpatient admissions in calendar year
OPVST	Num	Annual number of Outpatient Institutional visits in calendar year
SNF_COVDYS	Num	Annual number of Skill Nursing Facility covered days in calendar year
PHSVST	Num	Annual number of physician office visits in calendar year
AMI	Num	Chronic Condition Warehouse: Acute Myocardial Infarction
ALZH	Num	Chronic Condition Warehouse: Alzheimer`s Disease
ALZHDMTA	Num	Chronic Condition Warehouse: Alzheimer`s Disease and Related Disorders or Senile
ATRIALFB	Num	Chronic Condition Warehouse: Atrial Fibrillation
CATARACT	Num	Chronic Condition Warehouse: Cataract
CHRNKIDN	Num	Chronic Condition Warehouse: Chronic Kidney Disease
COPD	Num	Chronic Condition Warehouse: Chronic Obstructive Pulmonary Disease
CHF	Num	Chronic Condition Warehouse: Heart Failure
DIABETES	Num	Chronic Condition Warehouse: Diabetes
GLAUCOMA	Num	Chronic Condition Warehouse: Glaucoma
HIPFRAC	Num	Chronic Condition Warehouse: Hip/Pelvic Fracture
ISCHMCHT	Num	Chronic Condition Warehouse: Ischemic Heart Disease
DEPRESSN	Num	Chronic Condition Warehouse: Depression
OSTEOPRS	Num	Chronic Condition Warehouse: Osteoporosis
RA_OA	Num	Chronic Condition Warehouse: RA/OA
STRKETIA	Num	Chronic Condition Warehouse: Stroke / Transient Ischemic Attack
CNCRBRST	Num	Chronic Condition Warehouse: Female Breast Cancer
CNCRCLRC	Num	Chronic Condition Warehouse: Colorectal Cancer
CNCRPRST	Num	Chronic Condition Warehouse: Prostate Cancer
CNCRLUNG	Num	Chronic Condition Warehouse: Lung Cancer
CNCRENDM	Num	Chronic Condition Warehouse: Endometrial Cancer
AMIE	Num	Earliest indication of Acute Myocardial Infarction (Date)
ALZHE	Num	Earliest indication of Alzheimer`s Disease (Date)
ALZHDMTE	Num	Earliest indication of Alzheimer`s Disease and Related Disorders
ATRIALFE	Num	Earliest indication of Atrial Fibrillation (Date)
CATARCTE	Num	Earliest indication of Cataract (Date)
CHRNKDNE	Num	Earliest indication of Chronic Kidney Disease (Date)
COPDE	Num	Earliest indication of Chronic Obstructive Pulmonary Disease (Date)
CHFME	Num	Earliest indication of Heart Failure (Date)
DIABTESE	Num	Earliest indication of Diabetes (Date)
GLAUCMAE	Num	Earliest indication of Glaucoma (Date)
HIPFRACE	Num	Earliest indication of Hip/Pelvic Fracture (Date)
ISCHMCHE	Num	Earliest indication of Ischemic Heart Disease (Date)
DEPRSSNE	Num	Earliest indication of Depression (Date)
OSTEOPRE	Num	Earliest indication of Osteoporosis (Date)
RA_OA_E	Num	Earliest indication of RA/OA (Date)
STRKTIAE	Num	Earliest indication of Stroke / Transient Ischemic Attack (Date)
CNCRBRSE	Num	Earliest indication of Female Breast Cancer (Date)
CNCRCLRE	Num	Earliest indication of Colorectal Cancer (Date)
CNCRPRSE	Num	Earliest indication of Prostate Cancer (Date)
CNCRLNGE	Num	Earliest indication of Lung Cancer (Date)
CNCENDME	Num	Earliest indication of Endometrial Cancer (Date)
BASF_YR_NUM	Char	BASF Year

APPENDIX B:

DATA DICTIONARY OF PART D DENOMINATOR FILE

Variable	Data Type	Description
BENE_ID	Char	Encrypted 723 Beneficiary ID
STATE_CD	Char	SSA State Code
CNTY_CD	Char	SSA County Code
BENE_ZIP	Char	Zip Code of Residence
BENE_DOB	Num	Date of Birth
SEX	Char	Sex
AGE	Num	Age at Beginning of Bene Enrollment
MS_CD	Char	Medicare Status Code
BUYIN01	Char	Jan. Medicare Entitlement/Buy-In Indicator
BUYIN02	Char	Feb. Medicare Entitlement/Buy-In Indicator
BUYIN03	Char	Mar. Medicare Entitlement/Buy-In Indicator
BUYIN04	Char	Apr. Medicare Entitlement/Buy-In Indicator
BUYIN05	Char	May Medicare Entitlement/Buy-In Indicator
BUYIN06	Char	Jun. Medicare Entitlement/Buy-In Indicator
BUYIN07	Char	Jul. Medicare Entitlement/Buy-In Indicator
BUYIN08	Char	Aug. Medicare Entitlement/Buy-In Indicator
BUYIN09	Char	Sep. Medicare Entitlement/Buy-In Indicator
BUYIN10	Char	Oct. Medicare Entitlement/Buy-In Indicator
BUYIN11	Char	Nov. Medicare Entitlement/Buy-In Indicator
BUYIN12	Char	Dec. Medicare Entitlement/Buy-In Indicator
HMOIND01	Char	Jan. HMO Indicator
HMOIND02	Char	Feb. HMO Indicator
HMOIND03	Char	Mar. HMO Indicator
HMOIND04	Char	Apr. HMO Indicator
HMOIND05	Char	May HMO Indicator
HMOIND06	Char	Jun. HMO Indicator
HMOIND07	Char	Jul. HMO Indicator
HMOIND08	Char	Aug. HMO Indicator
HMOIND09	Char	Sep. HMO Indicator
HMOIND10	Char	Oct. HMO Indicator
HMOIND11	Char	Nov. HMO Indicator
HMOIND12	Char	Dec. HMO Indicator
CNTRCT01	Char	Jan. Encrypted Contract ID
CNTRCT02	Char	Feb. Encrypted Contract ID
CNTRCT03	Char	Mar. Encrypted Contract ID
CNTRCT04	Char	Apr. Encrypted Contract ID
CNTRCT05	Char	May Encrypted Contract ID

Variable	Data Type	Description
CNTRCT06	Char	Jun. Encrypted Contract ID
CNTRCT07	Char	Jul. Encrypted Contract ID
CNTRCT08	Char	Aug. Encrypted Contract ID
CNTRCT09	Char	Sep. Encrypted Contract ID
CNTRCT10	Char	Oct. Encrypted Contract ID
CNTRCT11	Char	Nov. Encrypted Contract ID
CNTRCT12	Char	Dec. Encrypted Contract ID
PBPID01	Char	Jan. Encrypted Plan Benefit Package ID
PBPID02	Char	Feb. Encrypted Plan Benefit Package ID
PBPID03	Char	Mar. Encrypted Plan Benefit Package ID
PBPID04	Char	Apr. Encrypted Plan Benefit Package ID
PBPID05	Char	May Encrypted Plan Benefit Package ID
PBPID06	Char	Jun. Encrypted Plan Benefit Package ID
PBPID07	Char	Jul. Encrypted Plan Benefit Package ID
PBPID08	Char	Aug. Encrypted Plan Benefit Package ID
PBPID09	Char	Sep. Encrypted Plan Benefit Package ID
PBPID10	Char	Oct. Encrypted Plan Benefit Package ID
PBPID11	Char	Nov. Encrypted Plan Benefit Package ID
PBPID12	Char	Dec. Encrypted Plan Benefit Package ID
CSTSHR01	Char	Jan. Cost Share Group Code
CSTSHR02	Char	Feb. Cost Share Group Code
CSTSHR03	Char	Mar. Cost Share Group Code
CSTSHR04	Char	Apr. Cost Share Group Code
CSTSHR05	Char	May Cost Share Group Code
CSTSHR06	Char	Jun. Cost Share Group Code
CSTSHR07	Char	Jul. Cost Share Group Code
CSTSHR08	Char	Aug. Cost Share Group Code
CSTSHR09	Char	Sep. Cost Share Group Code
CSTSHR10	Char	Oct. Cost Share Group Code
CSTSHR11	Char	Nov. Cost Share Group Code
CSTSHR12	Char	Dec. Cost Share Group Code
HMO_MO	Char	HMO Coverage Count
BUYIN_MO	Char	State Buy-In Coverage Count
PLNCOVMO	Char	Plan Coverage Months Number
DUAL_MO	Char	Dual Eligible Months Number
RTI_RACE	Char	RTI (Research Triangle Institute) Race C
BENEDPSQ	Num	BENE_ID w/ More than One Record

APPENDIX C:

DATA DICTIONARY OF PART D EVENT DATA

Variable	Data Type	Description
BENE_ID	Char	Encrypted 723 Beneficiary ID Number
SRVC_DT	Num	RX Service Date (DOS)
PRVDR_ID	Char	Service Provider ID
PRSCRBID	Char	Prescriber ID
PRDSRVID	Char	Product Service ID
PLNCNTRC	Char	Plan Contract Record ID
PLNPBPRC	Char	Plan PBP Record Number
QTYDSPNS	Num	Quantity Dispensed
DAYSSPLY	Num	Days of Supply
FILL_NUM	Num	Fill Number

DRCVSTCD	Char	Drug Coverage Status Code:
		C = Covered
		E = Supplemental drugs
		O = Over-the-counter drug

PTPAYAMT	Num	Patient Pay Amount
OTHTROOP	Num	Other TrOOP Amount
LICS_AMT	Num	Low Income Cost Sharing Subsidy Amount (
CPP_AMT	Num	Covered D Plan Paid Amount (CPP)
NPP_AMT	Num	Non-Covered Plan Paid Amount (NPP)
TOTALCST	Num	Gross Drug Cost

BNFTPHAS	Char	The benefit phase of the Part D Event
		Blank = not a covered drug
		DD = Deductible phase
		DP = Deductible to Pre-ICL
		DI = Deductible to ICL (Coverage Gap)
		DC = Deductible to Catastrophic
		PP = Pre-ICL phase
		PI = Pre-ICL to ICL
		PC = Pre-ICL to Catastrophic
		II = ICL (Coverage Gap) phase
		IC = ICL to Catastrophic
		CC = Catastrophic

TIER_ID	Char	Medicare Part D formulary tier identifier

APPENDIX D:

DATA DICTIONARY OF PLAN CHARACTERISTICS FILE

Variable	Data Type	Description
CNTRCTID	Char	Encrypted Contract ID
PLAN_ID	Char	Encrypted Plan ID
DRGBENTP	Char	Drug Benefit Type 1 = Defined Standard Benefit 2 = Actuarially Equivalent Standard 3 = Basic Alternative 4 =Enhanced Alternative
GAPCOVTP	Char	Type of gap coverage offered
DED_APP	Char	How Deductible is applied: 1-Medicare Defined; 2-Plan Defined; 3-No Deductible
DED_AMT	Num	Deductible Amt.
ICL_APP	Char	How ICL is applied: 1-Medicare Defined; 2-Plan Defined; 3-No ICL
ICL_AMT	Num	ICL Amt.
OOPT_AMT	Num	OOP Threshold Amt.

APPENDIX E:

DATA DICTIONARY OF FIRST DATABANK DATA

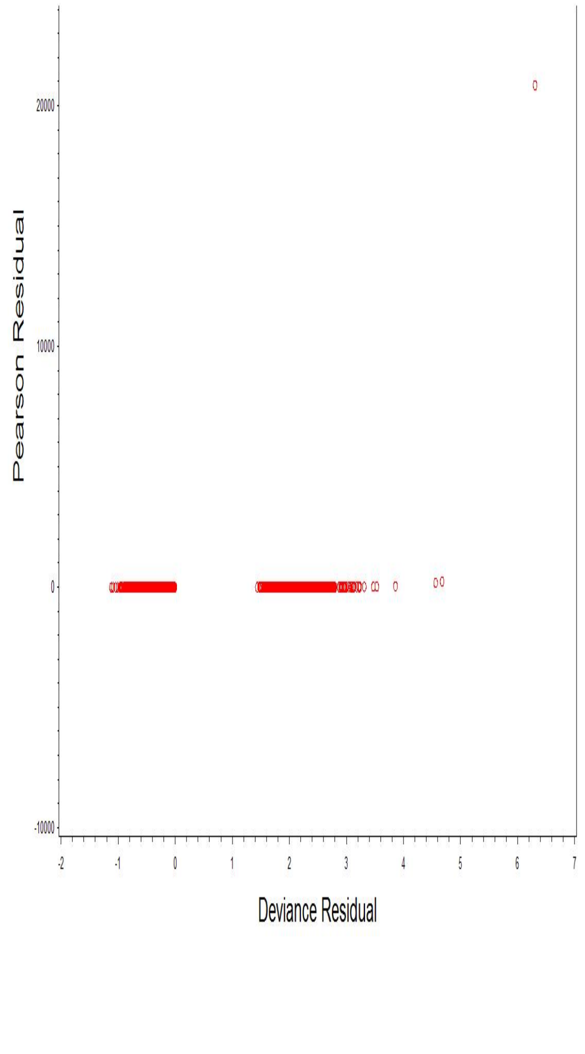
Variable	Data Type	Description
NDC	Char	NDC
BN	Char	Brand Name
GNN60	Char	Generic Name

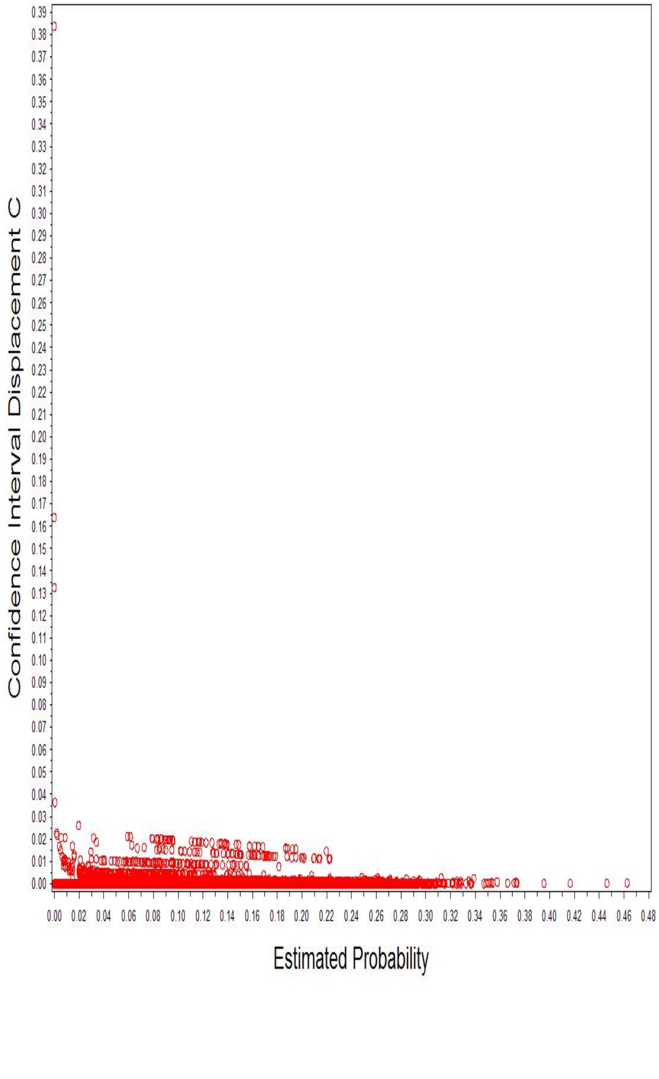
GNI	Char	Generic Name Indicator:
		0 = Non-drug Item;
		1 = Generically Named;
		2 = Brand Named

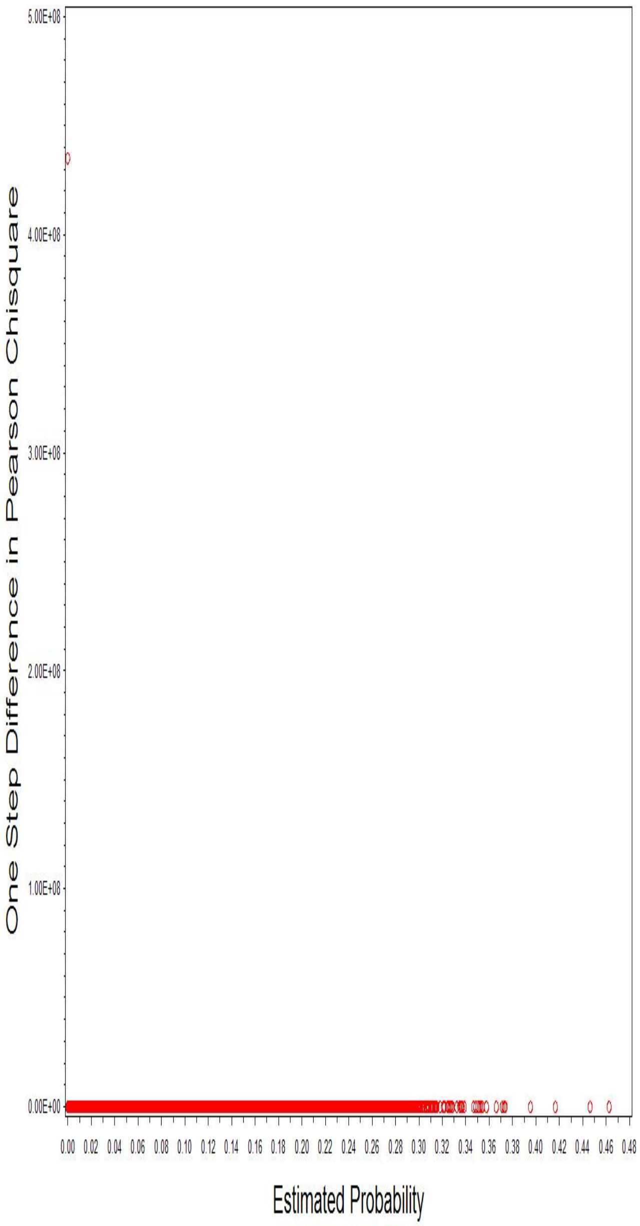
STR60	Char	Drug Strength Description
GCDF_DESC	Char	Dosage Form Code Description
PS	Num	Package Size
ETC_NAME	Char	ETC Therapeutic Class Description
TC_1	Char	Top level ETC Class
TC_2	Char	2nd level ETC Class
TC_3	Char	3rd level ETC Class
TC_4	Char	4th level ETC Class
TC_5	Char	5th level ETC Class
TC_6	Char	6th level ETC Class
TC_7	Char	7th level ETC Class
TC_8	Char	8th level ETC Class

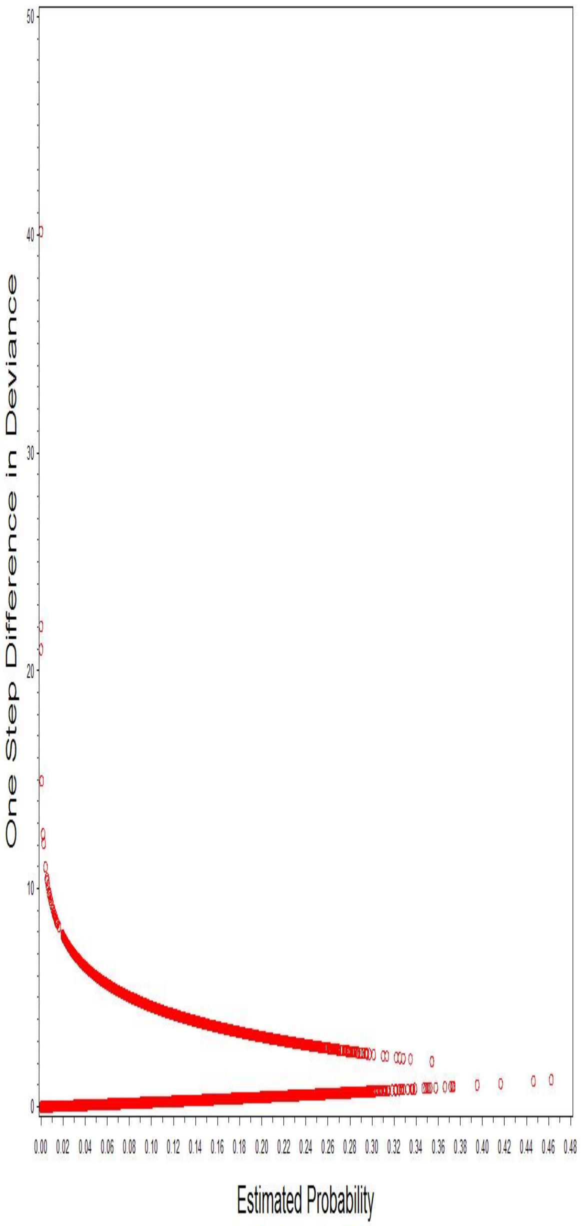
APPENDIX F:

MODEL DIAGNOSTICS PLOTS FOR LOGISTIC REGRESSION






One Step Difference in Pearson Chisquare
5.00E+08
4.00E+08
3.00E+08
2.00E+08
1.00E+08
0.00E+00
0.00 0.02 0.04 0.06 0.08 0.10 0.12 0.14 0.16 0.18 0.20 0.22 0.24 0.26 0.28 0.30 0.32 0.34 0.36 0.38 0.40 0.42 0.44 0.46 0.48
Estimated Probability


One Step Difference in Deviance
50
40
30
20
10
0
0.00 0.02 0.04 0.06 0.08 0.10 0.12 0.14 0.16 0.18 0.20 0.22 0.24 0.26 0.28 0.30 0.32 0.34 0.36 0.38 0.40 0.42 0.44 0.46 0.48
Estimated Probability

BIBLIOGRAPHY

[1] The Centers for Medicare & Medicaid Services (CMS). Chronic Condition Data Warehouse User Manual. Ver.1.6, Jan.2010. Accessed at http://ccwdata.org/downloads/CCW_UserManual.pdf.

[2] Miyares IM. Segregation: Index of Dissimilarity. Feb. 2010. Accessed at http://www.geo.hunter.cuny.edu/~imiyares/Segregation.htm

[3] The Centers for Medicare & Medicaid Services (CMS). Chronic Condition Data Warehouse Part D Data User Manual. Ver.3.0, Jan.2010. Accessed at http://www.ccwdata.org/downloads/CCW_PartD_UserManual_201006.pdf.

[4] The Henry J. Kaiser Family Foundation. Prescription Drug Trends. May 2010. Accessed at http://www.kff.org/rxdrugs/upload/3057-08.pdf

[5] The Henry J. Kaiser Family Foundation. Prescription Drug Trends. Sep. 2008. Accessed at http://www.kff.org/rxdrugs/upload/3057_07.pdf

[6] Henry J. Kaiser Family Foundation. Prescription Drug Trends. May 2007. Accessed at http://www.kff.org/rxdrugs/upload/3057_06.pdf

[7] Nightingale SL. From the Food and Drug Administration: Promotional Practices of Pharmacy Benefits Management Companies. JAMA. 1998; 279(9):645. [PMID: 9496968]

[8] Wal-Mart Stores Inc. Wal-Mart Announces Accelerated Rollout of $4 Generic Prescription Program in 14 States. Jan. 8, 2010. Accessed at http://walmartstores.com/FactsNews/NewsRoom/6021.aspx

[9] Hosmer DW, Lemeshow S. (2000). Model-Building Strategies and Methods for Logistic Regression. In: Applied Logistic Regression. 2nd ed., 116-128. New York: John Wiley & Sons, Inc.