© Copyright 2018

Sungchul Park

Understanding Unintended Consequences of Risk Adjustment and Proposing Alternative Risk Adjustment

Sungchul Park

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Anirban Basu, Chair

Norma B Coe

Fahad Khalil

Program Authorized to Offer Degree:

Public Health – Health Services

University of Washington

Abstract

Understanding Unintended Consequences of Risk Adjustment
and Proposing Alternative Risk Adjustment

Sungchul Park

Chair of the Supervisory Committee:
Anirban Basu, Professor
Department of Health Services, Pharmacy, and Economics

To achieve two goals of improving quality of care and containing costs in the Medicare program, the Centers for Medicare and Medicaid Services (CMS) has reimbursed Medicare Advantage (MA) plans with a capitated amount per beneficiary. CMS uses the Hierarchical Condition Categories (HCC) risk adjustment to reflect each beneficiary's health status and more accurately estimate capitated payments to Medicare Advantage (MA) plans. However, it is debated whether the HCC model has been effective in reducing risk selection or if it has led to strategic evolutions of risk selection. This research intends to understand the intended and unintended consequences of the HCC model on the MA plans' risk selection behaviors and propose alternative risk adjustment models from the perspectives of statistics and economics, respectively.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I would first like to express my sincere gratitude to my dissertation committee members, Anirban Basu, Norma Coe, and Fahad Khalil. Thank you for patiently guiding me through this process, sharing your invaluable insights, posing thoughtful questions, encouraging me through hardships, and being kind and wonderful people to work with.

I would also like to thank additional mentors who supported me and gave me opportunities throughout my training: Emily Williams, Bianca Frogner, Doug Conrad, Dave Grembowski, Paul Fishman, and Donald Chi.

I am also thankful to my friends and peers in the Health Services PhD Program, especially the 2013 Health Services PhD cohort (Ann Nguyen, Alex Woersching, Debbie Passey, and Kara Bensely).

I would also like to acknowledge the financial support I receive for this dissertation from the Department of Health Services at the University of Washington and the National Institute of Health.

And last but not least, I would like to thank my family, especially my parents, for providing support to enable me to achieve this goal in graduate school in the United States; I could not have done it without them.

DEDICATION

This dissertation is dedicated to the memory of my beloved grandfather,

Mangdeok Baek (Kentaro Shirakawa) (1926 - 2018).

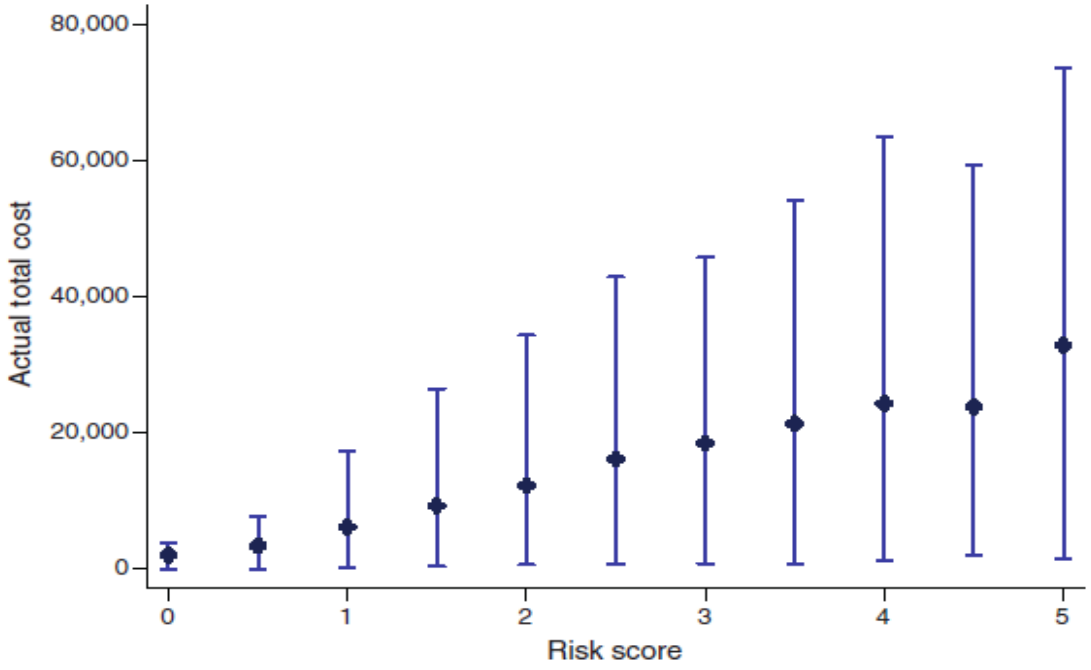
Chapter 1.

INTRODUCTION

Recent health care reforms have attempted to move away from the Fee-For-Service (FFS) payment models to the capitated payment models. Under the FFS system, health plans are paid for each test and procedure, and thus theoretically the plans do not have an incentive to selectively enroll beneficiaries with specific characteristics. However, this could potentially encourage unnecessary services, leading to high costs without additional benefits to beneficiaries. Therefore, the capitated payment system has been introduced to provide high-value care. On the other hand, under the capitated payment system, health plans receive a capitated payment per beneficiary for total care in a year, and thus have the incentive to manage care, with most of the savings in health care costs enabling them to add benefits and/or lower premiums for beneficiaries. As this approach creates incentives to encourage preventive care and better care coordination, policymakers have encouraged health plans to provide managed care as a way to achieve cost control and enhance the quality of care. However, the capitated payment structure creates an incentive for the plan to selectively enroll healthier people and avoid sicker ones. The success of such reforms hinges on correctly aligning capitated payments with a beneficiary' expected health care expenditures.

In this study, we focused on the Medicare program—the largest public health insurance program in the United States, providing the major source of insurance for the acute medical care needs of elderly and disabled persons. Private health plans contract with the Medicare program to provide elderly people Medicare Part A and Part B benefits. Medicare pays these health plans, known as Medicare Advantage (MA) plans, a capitated (per enrollee) amount to offer all Parts A and B benefits. This capitated payment structure creates incentives for MA plans to selectively

dis-enroll the high-cost beneficiaries so that, on average, they receive overpayments from the Medicare program. Empirical evidence demonstrated such selective behavior by MA plans existed (McGuire et al., 2011; Newhouse and McGuire. 2014), and as of 2006, caused an extra $30 billion in Medicare spending (Brown et al., 2014). To address this phenomenon, in 2000, the Centers for Medicare and Medicaid Services (CMS) adopted risk adjustment to their capitated payments—a process by which CMS adjusts payments to MA plans, using risk scores estimated based on each



beneficiary's demographics and diagnoses in a prior year (Pope et al., 2000). To more precisely incorporate health status differences in the calculation of payments to MA plans, in 2004, CMS introduced a new risk-adjustment model—the CMS-Hierarchical Condition Categories (HCC) model—which uses extensive inpatient and outpatient diagnostic information from the prior year to generate risk scores (Pope et al., 2004).

Figure 1. Variability of health care expenditures around risk-adjusted mean predictions

Source: Brown et al. (2014)

Some proponents of the existing risk-adjustment model argue that the CMS-HCC model has mitigated MA plans' selection behavior (McWilliams et al., 2012; Morrisey et al., 2013; Newhouse et al., 2012; Newhouse et al., 2015), but we suspect that MA plans might change their risk selection strategies in response to the CMS-HCC model (Brown et al., 2014). First, the CMS-HCC model merely focuses on predicting mean costs of beneficiaries (Pope et al., 2004). Figure 1 shows how actual health care expenditures vary among beneficiaries with the same risk score. Each dot indicates risk-adjusted mean expenditures for each risk score group. Each bar indicates the range of actual expenditures of those in each risk score group. For all beneficiaries with a given risk score, the CMS-HCC model pays the same rate to MA plans. However, the severity of condition and thus the cost of treating it can vary within a given risk score. Indeed, as shown in Figure 1, there is considerable variability of costs around the risk-adjusted mean predictions (Medicare Payment Advisory Commission, 2012). Figure 1 also shows that the variability of the within-risk-score costs is greatest for beneficiaries with the most severe health status (Manning et al., 2005). These findings suggest that the HCC model does not estimate payments as sufficiently close to the actual expenditures in a *statistical* sense. More importantly, the HCC model does not generate sufficient *economic* incentives to prevent health plans from engaging in risk selection. As such, there is a potential for MA plans to benefit financially if they selectively dis-enroll beneficiaries with higher costs than what the CMS-HCC model predicts (i.e., unprofitable beneficiaries), especially for those with high-risk scores.

To achieve such patient selection, MA plans are likely to seek screening approaches with no or little screening costs. One such approach that generates relatively low screening costs is *service-level selection*, in which, MA plans may provide different coverage levels for different services. The idea behind this selection mechanism is that MA plans would reduce coverage for

services that appeal to beneficiaries with higher costs than their risk-adjusted mean payments so that they would self-select themselves out of enrolling in the MA plan. The key feature of such behavior is that MA plans forecast services that are likely used by those who would have higher costs than those whose costs align more with the risk-adjusted mean payments. Although there is a very large literature on investigating service-level selection as a patient selection strategy (Cao and McGuire, 2003; Eggleston and Bir, 2009; Ellis et al., 2013; Ellis and McGuire, 2007; Frank et al., 2000; McGuire et al., 2014; Newhouse et al., 2013), to the best of our knowledge, there are few studies explaining mechanisms through which MA plans could engage in service-level selection against the federal government's attempt to risk adjust. Furthermore, there are very few studies proposing better risk-adjustment models.

This research intends to understand the intended and unintended consequences of the HCC model on the MA plans' risk selection behaviors and propose alternative risk adjustment models from the perspectives of statistics and economics, respectively. The specific aims of this study are:

In Chapter 2, we examined two main questions to fill the gap in evidence on the effectiveness of the CMS-HCC model. First, we examined whether MA plans strategically respond to the CMS-HCC model. Second, we examined the mechanism in which MA plans could effectuate the unintended consequence of the CMS-HCC model. Here, we first theoretically examined whether MA plans have incentive to increase copayments disproportionately more for services that appeal to unprofitable beneficiaries than other services. Then, we empirically tested two of our theoretical predictions. We examined whether MA enrollees who used more services with a large increase in copayment were more likely disenroll from MA plans. Finally, we examined whether MA disenrollment was attributable to increased burdens on out-of-pocket costs.

In Chapter 3, we proposed an alternative risk adjustment model from the perspective of *statistics*. Specifically, we developed a new evaluation metric for measuring how close predicted expenditures align with actual expenditures at the individual level. Then, we performed a comprehensive comparison of prediction accuracies at the group level, at the tail distributions, and at the individual level across 19 estimators used in the risk adjustment literature: nine parametric regression, seven machine learning, and three distributional estimators.

In Chapter 4, we proposed an alternative risk adjustment model with machine learning techniques from the perspective of *economics*. Specifically, we developed an alternative risk-adjustment method that accounts for accounts for each individual's future service-level use, using machine learning techniques. Using the same estimators and evaluation metrics used in Chapter 3, we performed a comprehensive pair comparison of our proposed method with the currently used HCC model.

Chapter 5 concludes, recapitulating and summarizing the argument.

Chapter 2.

SERVICE-LEVEL SELECTION: STRATEGIC RISK SELECTION IN MEDICARE ADVANTAGE IN RISK ADJUSTMENT

Abstract

The Centers for Medicare and Medicaid Services (CMS) uses the Hierarchical Condition Categories (HCC) risk adjustment to reflect each beneficiary's health status and more accurately estimate capitated payments to Medicare Advantage (MA) plans. However, it is debated whether the CMS-HCC model led to strategic risk selection. We examine the competing claims and analyze the risk selection behavior of MA plans in response to the CMS-HCC model. We find that the CMS-HCC model reduced MA plans' avoidance of high-cost beneficiaries at enrollment, but led to increased disenrollment of high-cost beneficiaries, conditional on illness severity, from MA plans. We explain this unintended consequence via service-level selection. First, we theoretically show that MA plans have incentives to effectuate risk selection via service-level selection, by lowering coverage levels for services that are more likely to be used by beneficiaries who could be unprofitable under the CMS-HCC model. Then, we empirically test the theoretical prediction that MA plans would raise copayments disproportionately more for services needed by unprofitable beneficiaries than for other services after the CMS-HCC implementation. This induced unprofitable beneficiaries to voluntarily disenroll from their MA plans. Further evidence supporting this selection mechanism is that those dissatisfied with out-of-pocket costs were more likely to disenroll from MA plans. We estimate that such strategic behavior led MA plans to save $5.9 billion by transferring the costs to Traditional Medicare.

2.1 BACKGROUND

Since the 1980's, Medicare beneficiaries have been able to enroll in two types of health plans: the government-run traditional Medicare (TM) program or privately-run plans. Private health plans, known as Medicare Advantage (MA) plans, contract with Medicare to provide elderly people Medicare Part A and Part B benefits. MA plans receive a capitated (per enrollee) amount from the Centers for Medicare and Medicaid Services (CMS) to offer all Parts A and B benefits. However, this capitated payment structure creates incentives for MA plans to selectively dis-enroll sicker beneficiaries so that, on average, they receive overpayments from CMS. Empirical evidence demonstrated such selective behavior by MA plans existed (McGuire et al., 2011; Newhouse and McGuire. 2014). To reduce risk selection, CMS has adjusted payments to MA plans to reflect the health status of their enrollees, a process known as risk adjustment (Pope et al., 2000). To more precisely incorporate health status differences in the calculation of capitated payments to MA plans, in 2004, CMS introduced a new risk adjustment model—the CMS-Hierarchical Condition Categories (HCC) model—which uses extensive inpatient and outpatient diagnostic information from the prior year to generate risk scores (Pope et al., 2004).

It is debated whether the CMS-HCC model has been effective in reducing risk selection (Newhouse et al., 2015) or whether it has led to a strategic evolution of risk selection (Brown et al., 2014). On one hand, it has been shown that the CMS-HCC model considerably reduced the phenomenon of avoiding sicker beneficiaries (i.e., those with high-risk scores) in TM (McWilliams et al., 2012; Morrisey et al., 2013; Newhouse et al., 2012; Newhouse et al., 2015). As risk adjustment leads to neutral payments for beneficiaries with conditions included in the risk adjustment formula, MA plans no longer have incentive to avoid those with high-risk scores if their conditions are included in the CMS-HCC model.

On the other hand, there is suggestive evidence showing that MA plans could strategically respond to the CMS-HCC model. Brown et al. (2014) argue that the CMS-HCC model merely shifted the profitable population from healthy people (i.e., those with low-risk scores) to sick ones who are over-compensated, within their risk-score. This can be achieved because, first, there is considerable variability in actual expenditures of beneficiaries around their risk-adjusted payments. For all beneficiaries with a given health condition, the CMS-HCC model is designed to adjust payments to MA plans by the same rate. However, the severity of the condition and thus the cost of treating it can vary within a given condition (Medicare Payment Advisory Commission, 2012).[1] Second, the variability of the within-risk-score expenditures is larger for those with higher risk scores (Manning et al., 2005). Because the CMS-HCC model only accounts for about 100 major conditions, this generates underpayments for those whose conditions are not accurately measured by the model, who tend to have multiple chronic conditions (Frogner et al., 2011).[2]

In this study, we examine two main questions to fill the gap in evidence on the effectiveness of the CMS-HCC model. First, we examine whether MA plans strategically respond to the CMS-HCC model. The previous literature, including Brown et al. (2014) and Newhouse et al., (2015), mainly focused on whether TM beneficiaries with high-risk scores were more likely to enroll in MA plans. However, there is another important avenue in which MA plans could achieve risk selection: getting unprofitable beneficiaries to disenroll from MA plans. We examine both enrollment and disenrollment patterns for evidence of risk selection. We hypothesize that while

[1] For example, the coefficient for breast, prostate, colorectal and other cancers (HCC10) was estimated to be 0.187 by the CMC-HCC model (Pope et al., 2004). This means that CMS increases payments by 1.187 % of the overall mean for Medicare beneficiaries with breast, prostate, colorectal and other cancers relative to equally sick beneficiaries without the condition. Depending on cancer stage, however, their actual expenditures would vary. Specifically, those with stage 4 cancer are more likely to incur higher expenditures than the rate set by CMS, whereas those with stage 1 cancer are more likely to incur lower expenditures than the rate set by CMS.

[2] Frogner et al. (2011) showed that the estimate for breast, prostate, colorectal, and other cancers (HCC 10) was only $1,835 despite its seriousness of the illness. Moreover, it was found that the interaction between chronic kidney failure and congestive heart failure had a statistically significant negative estimate, thereby reducing reimbursements for beneficiaries with these two diseases by $614. However, this is unlikely because having multiple chronic diseases requires more complex care.

the CMS-HCC model encouraged MA plans to accept TM beneficiaries with high-risk scores, MA plans strategically behave to avoid beneficiaries who could be unprofitable under the CMS-HCC model (i.e., those with higher expenditures than their risk-adjusted payments). We find that the CMS-HCC model achieved the goal of reducing selection for TM beneficiaries with high-risk scores as intended; however, we also find that it led to increased disenrollment of unprofitable beneficiaries from MA plans, leading MA plans to save costs of $5.9 billion in 2007-2009.

Second, we examine the mechanism in which MA plans could effectuate the unintended consequence of the CMS-HCC model. We hypothesize that MA plans magnify *service-level selection*, in which they provide relatively lower coverage levels for some services to discourage enrollment of unprofitable beneficiaries, as a strategic risk selection mechanism in response to the CMS-HCC model.[3] There is a large literature on investigating service-level selection as a risk selection strategy (Cao and McGuire, 2003; Eggleston and Bir, 2009; Ellis et al., 2013; Ellis and McGuire, 2007; Frank et al., 2000; McGuire et al., 2014; Newhouse et al., 2013). To the best of our knowledge, however, there is little research explaining how MA plans engage in service-level selection.[4] In this study, we first theoretically examine whether MA plans have incentive to increase copayments disproportionately more for services that appeal to unprofitable beneficiaries than other services. Then, we empirically test two of our theoretical predictions. We examine whether MA enrollees who used more services with a large increase in copayment were more likely disenroll from MA plans. Finally, we examine whether MA disenrollment was attributable to increased burdens on out-of-pocket costs. To give a preview of our results, building upon Ellis and McGuire (2007), we theoretically show that those with higher expenditures than their risk-

[3] There is evidence showing that MA plans has already used service-level selection to enforce selective enrollment (Cao and McGuire, 2003). As such, the idea of this study on service-level selection is that MA plans magnify service-level selection in response to the CMS-HCC model to avoid beneficiaries who could be costly under the CMS-HCC model.

[4] There is a literature showing that MA plans changed their drug formularies to avoid enrolling unprofitable beneficiaries under the CMS-HCC model (Han and Lavetti, 2017).

adjusted payments are more likely to use services that health plans would ration more tightly (i.e., services with higher service-level selection index). This phenomenon is more likely to be pronounced for those with higher risk scores. Then, we find evidence supporting our theoretical prediction that MA plans have actually raised copayments disproportionately more for services with higher service-level selection index (i.e., durable medical equipment, home health, inpatient hospital service) than services with lower service-level selection index (i.e., mental health specialty service, psychiatric service, and primary care physician service). Such disproportionate increases in copayment induced unprofitable beneficiaries to voluntarily disenroll from MA plans. In additional analyses, we find evidence supporting this selection mechanism that those with dissatisfaction with out-of-pocket costs were more likely to disenroll from MA plans. Consequently, the variation of total Medicare expenditures for MA enrollees with high-risk scores reduced over time. These findings indicate that service-level selection allowed MA plans to avoid the risk of enrolling unprofitable beneficiaries.

The rest of the paper is structured as follows. Section 2 provides an overview of the MA program. Section 3 reviews the literature on risk selection behaviors by MA plans. Section 4 presents the intended and unintended consequences of the CMS-HCC model on the MA plans' risk selection behaviors. Section 5 explains how MA plans could effectuate the unintended consequences via service-level selection. Specifically, we present a conceptual framework to illustrate the idea of service-level selection. Building on prior theoretical models, then we demonstrate our theoretical predictions on service-level selection in response to the CMS-HCC model. We also describe empirical analyses and discuss our results from empirical analyses. Section 6 discusses policy implications and potential future research.

2.2 THE MEDICARE ADVANTAGE PROGRAM

Medicare beneficiaries can choose to either enroll in TM or an MA plan. Private plans contract with CMS to provide MA plans that cover equivalent or greater benefits to Medicare Parts A and B benefits.[5] When individuals become eligible for Medicare, they are defaulted to TM, but can choose to stay in TM or switch to a MA plan, depending on their preferences and needs. Because MA plans offer more generous benefits and lower cost-sharing than TM, beneficiaries may prefer enrolling in the MA plan. In contrast, as MA plans have limited providers' networks (Jacobson et al., 2016), those with complex diseases may prefer TM's freedom of provider choice. To encourage beneficiaries to choose the plan that efficiently provides them care while accounting for their individual preferences, CMS adopted the "payment neutrality" approach, which sets MA payments equal to the average Medicare expenditures of TM beneficiaries in the MA enrollee's county (Medicare Payment Advisory Commission, 2014).

Over several decades, policy-makers have promoted managed care in Medicare as a way to improve quality of care while containing costs. As this approach creates incentives to encourage preventive care and better care coordination, it is especially helpful in caring for Medicare beneficiaries, 68.4 percent of whom had two or more chronic conditions and 36.4 percent had four or more chronic conditions (Lochner and Cox, 2013). In line with the encouragement of managed care, the benchmark levels have been increased to encourage plans to enter in the MA market (Medicare Payment Advisory Commission, 2014).[6] Consequently, CMS paid $170 billion to MA

[5] There are two rationales for privatization of managed care. First, capitated payments to private plans would incentivize them to actively manage their enrollees' care, leading to more efficient care provision. Also, privatization could lead to competition among private plans as well as TM, possibly lowering health care costs while improving quality of care.

[6] Prior to the Balanced Budget Act of 1997, MA plans were paid based on 95 percent of a county's average Medicare expenditures of TM beneficiaries. In the Balanced Budget Act, Congress increased benchmark levels to encourage plans to enter in the MA market. Consequently, as of 2015, MA plans are paid by 102 percent of TM costs.

plans on behalf of 16 million beneficiaries, reaching a historic high of MA's penetration of almost 31 percent of the Medicare population (Kaiser Family Foundation, 2015).

However, benefit designs by MA plans may affect whether beneficiaries choose TM or MA plans. TM under the fee-for-service (FFS) payment system are paid for each test and procedure, and thus theoretically they have no incentive to selectively accept beneficiaries with certain characteristics. In contrast, MA plans under the capitated payment system are paid a fixed amount per beneficiary, which would create an incentive to selectively accept healthier people and avoid sicker ones. To effectuate favorable selection, MA plans could vary benefit designs. For example, MA plans could increase cost-sharing for certain services to avoid unprofitable beneficiaries. This is plausible because while MA plans must provide the same services covered by TM (i.e., Medicare Parts A and B benefits), and the actuarial value of the total benefits package must at least be equivalent to TM's benefits, cost-sharing for any particular service can vary between MA plans and TM. Moreover, MA plans could restrict physician networks to distract unprofitable beneficiaries (Jacobson et al., 2016). Also, MA plans could offer additional services (e.g., dental care and vision care services) or change drug formulary to attract profitable beneficiaries (Han and Lavetti, 2017).

To mitigate favorable selection of MA plans, CMS has used a risk-adjusted payment methodology to estimate capitated payments to MA plans. Payment rates to MA plans are determined by enrollee's risk scores, county-level benchmarks set by CMS, and plan bids.[7] However, the fundamental goal of risk adjustment is to adjust payments to accurately reflect the

[7] Since 2006, CMS has implemented a competitive bidding system to determine payments to MA plans. The payment is based on bids submitted by MA plans, and then it is risk-adjusted by the CMS-HCC model. The plan bids for Parts A and B services are compared to the county-level benchmark. If the plan bid is less than the benchmark, the plan receives its bid. In addition, CMS retains 25 percent of the difference between its payment benchmark and bid. The remaining 75 percent of the difference must be returned to enrollees in the form of additional benefits or lower premium. If the plan' bid is higher than the benchmark, enrollees pay the difference in the form of a monthly premium in addition to the Medicare Part B premium.

health status of each enrollee. CMS adjusts payments to MA plans, using risk scores estimated based on each beneficiary's demographics and diagnoses in a prior year. However, its ultimate goal is not accuracy per se, but rather improved incentives (Glazer and McGuire, 2000; Van de Ven and Ellis, 2000). As such, risk adjustment intends to disincentivize health plans from selectively enrolling and caring for healthy beneficiaries, and furthermore incentivize the plans to compete based on providing high-value care. Up until 2000, the risk adjustment process only accounted for age, gender, Medicaid eligibility, institutional status, and county of residence. Starting in 2000, CMS began to use information on inpatient diagnoses to adjust payments to MA plans through the Principal Inpatient Diagnostic Cost Group (PIP-DCG) model (Pope et al., 2000).[8] To more precisely incorporate health status differences in the calculation of capitated payments to MA plans, in 2004, CMS introduced a new risk adjustment model—the CMS-HCC model—which uses extensive inpatient and outpatient diagnostic information from the prior year to generate risk scores (Pope et al., 2004).[9]

2.3 Previous Literature

The literature on risk selection in the MA program has mainly focused on one aspect of risk selection: whether TM beneficiaries with high-risk scores were less likely to enroll in MA plans. Some studies found that such selection behavior was greatly reduced after the full phase-in of the CMS-HCC model starting January 1, 2007. Using a 20 percent random sample of Medicare claims

[8] The PIP-DCG model accounts for 24 age/sex cells, interactions of Medicaid status and age/sex cells, interactions of originally disabled status and age/sex cells, working-aged status, and the 16 PIP-DCG diagnostic categories (Pope et al., 2000).

[9] The CMS-HCC model accounts for 24 age/sex cells, interactions of Medicaid status and sex and age/disabled entitlement status, interactions of originally disabled status and sex, 70 HCC diagnostic categories, interactions of diagnostic categories with entitlement by disability, and six disease interactions (Pope et al., 2004). For HCC diagnostic categories, about 2,500 diagnosis codes based on the *International Classification of Diseases* (ICD)-9 system are grouped into a small number of organized categories to generate a diagnostic profile of each person. Each HCC diagnostic category includes conditions that are clinically related to each other and have similar cost implications.

in 2003-2008, Newhouse et al. (2012) found that differences in predicted expenditure between TM-to-MA switchers (i.e., Medicare beneficiaries who enrolled in TM and switched from TM to MA next year) and TM stayers (i.e., those who enrolled in TM and remained in TM next year) declined between 2004 and 2008 by a factor of three. Also, differences in adjusted mortality rates between these two groups narrowed between 1998 and 2008 by a factor of two. Using the 2001-2007 Medicare Current Beneficiary Survey (MCBS), McWilliams et al. (2012) observed that differences in health care use and self-reported health between all TM and MA beneficiaries were narrowed from 2001-2003 to 2006-2007. They also found that differences between TM-to-MA switchers and TM stayers were narrowed. Using a five percent random sample of Medicare claims in 1999-2008, Morrisey et al. (2013) showed that the implementation of the CMS-HCC model led to increase the number of new MA enrollees and decrease the number of MA disenrollees (i.e., those who enrolled in MA and switched from MA to TM next year), resulting in increased MA enrollment.

Another strand of the literature has examined another aspect of risk selection: whether MA enrollees with high-risk scores were more likely to disenroll from MA plans. There is evidence showing that MA plans might change the targeted population for risk selection in response to the CMS-HCC model. McWilliams et al. (2012) found that compared to MA stayers (i.e., those who enrolled in MA and remained in MA next year) or TM stayers, MA-to-TM switchers (i.e., those who enrolled in MA and switched from MA to TM next year) self-reported poorer health and used more health care after the full phase-in of the CMS-HCC model. Morrisey et al. (2013) observed that after the full implementation, disenrollment from MA plans was more pronounced among the high-expenditure beneficiaries. Using a five percent sample of Medicare claims between 2006-2011, Jacobson et al. (2015) found that Medicaid-eligible beneficiaries and beneficiaries younger

than 65 years with disabilities were more likely to disenroll from MA plans. On the other hand, the low-expenditure beneficiaries were more likely to stay in MA plans and relatively younger beneficiaries aged 65 to 69 years were more likely to switch from TM to MA plans.

However, little is known about the mechanism in which MA plans might engage in risk selection beyond risk scores. To the best of our knowledge, there are few papers that argued that MA plans might strategically behave in response to the CMS-HCC model by risk-selecting based on expenditures conditional on risk scores (Brown et al., 2014; Han and Lavetti. 2017). Specifically, Brown et al. (2014) found that the CMS-HCC model reduced the phenomenon of avoiding high-risk score beneficiaries whose conditions are included in the CMS-HCC model because it increases payments for them by accounting for additional information from inpatient and outpatient claims. However, the CMS-HCC model still generates underpayments for some of those with high-risk scores as their complex health status cannot be accurately captured by the model (Frogner et al., 2011). As such, Brown et al. (2014) found that in response to the CMS-HCC model, MA plans have selectively enrolled beneficiaries with high-risk scores but low expenditures conditional on their risk scores. It is worthwhile to note that Han and Lavetti (2017) examined MA plans' strategic response to another risk adjustment model for MA plans—the CMS-RxHCC model—which is used separately to estimate plan payments for Medicare prescription drug coverage. They found that after the introduction of Medicare Part D in 2006,[10] MA plans changed their drug formulary design to avoid enrolling unprofitable beneficiaries under the CMS-RxHCC. This indicates that the CMS-HCC model cannot reflect the health status of beneficiaries

[10] The prescription drug benefits in Medicare are available through two primary coverage options: a stand-alone prescription drug plan (PDP) or a Medicare Advantage Prescription Drug (MA-PD) plan. PDP provides drug coverage to beneficiaries who remain in TM for inpatient and outpatient medical services. On the other hand, MA-PD plans provide all Medicare-covered services including prescription drug coverage to those who enroll in MA plans. Both plans must provide drug benefits at least equal in value to a standard benefit defined by the Medicare Prescription Drug, Improvement, and Modernization Act of 2003.

beyond health status captured through claims data (i.e., risk scores), creating an incentive for MA plans to avoid those whose health status could be worse than estimated.

Newhouse et al. (2015) re-examined the question of how MA plans have responded to the CMS-HCC model and found that the combination of the CMS-HCC model and a lock-in policy[11] largely reduced favorable selection after 2007, the year in which the CMS-HCC model was fully phased in. Since the full implementation of the CMS-HCC model coincided with the introduction of the lock-in policy, they acknowledged that it would be hard to distinguish the effect of the CMS-HCC model from the effect of the lock-in policy on reducing risk selection. It is worthwhile to note that Newhouse et al. (2015) intended to validate findings from a prior study of Brown et al. (2014). Using the 1994-2006 MCBS, Brown et al. (2014) found evidence showing that the CMS-HCC model did not reduce favorable selection due to MA plan's strategic response to the CMS-HCC model.[12] However, Newhouse et al. (2015) used a 20 percent random sample of Medicare claims between 2001-2011 and found that the combination of the CMS-HCC model and the lock-in policy reduced overpayments attributable to selection by roughly a factor of five (from $1,984 in 2001-2002 to $320 in 2007-2011), thereby rebutting Brown et al. (2014)'s claim.[13] However, it

[11] Beginning in 2006, CMS imposed a partial enrollment lock-in to prevent MA enrollees from switching from MA plans to TM plans monthly, limiting temporarily switches to TM plans when more generous coverage or freedom of provider choice is desired.

[12] Brown et al. (2014) concluded that favorable selection was not decreased on net because the decrease in selection along dimensions included in the formula of the CMS-HCC model was more than offset by the increase in selection conditional on the risk score. Specifically, they found that there were smaller differences in risk scores between TM-to-MA switchers and TM stayers after its initial phase-in starting January 1, 2004. This result is consistent with those of the studies showing the effect of the CMS-HCC model on reducing risk selection (McWilliams et al., 2012; Morrisey et al., 2013; Newhouse et al., 2012). However, they showed that compared to TM stayers, actual expenditures conditional on the risk score of TM-to-MA switchers substantially fell after the initial phase-in period. This suggests that MA plans might strategically behave to enroll those with expenditures lower than what is predicted by the CMS-HCC model.

[13] Such contradictory results are attributable to data and methodology differences. First, due to a relatively small sample in the MCBS, Brown et al. (2014) had to pool the years from 1994 to 2002 and then compared selection in those years with selection during the pooled years from 2004 to 2006. This might be problematic because MA reimbursement policy changed markedly during the period. For example, the Balanced Budget Act of 1997 established floors on reimbursement to MA plans for low-paying areas and restricted annual increases in reimbursement to MA plans for high-paying areas to two percent. In contrast, Newhouse et al. (2015) used the sample for the pre-implementation period from 2001 to 2003 and compared selection during the pre-implementation period with selection during the post-implementation period. Moreover, using a large sample size, the study estimated the degree of selection in each year, allowing them to control for various Medicare payment polices across years. Also, Brown et al. (2014) included all MA enrollees from the MCBS. However, starting in 2004, CMS allowed MA plans to create plans for enrollees with special needs (e.g., institutionalized or Medicaid-eligible enrollees), many of whom are non-elderly. As comparing those groups

remains unanswered whether MA plans have changed risk selection strategies in response to the CMS-HCC model to induce MA enrollees with high-risk scores but high expenditures conditional on their risk scores to drop out of their plans. Determining this aspect is critical to comprehensively understand the strategic risk selection behaviors of MA plans in response to the CMS-HCC model.

2.4 Strategic Risk Selection in Response to the CMS-HCC Model

The goal of this section is to examine the competing claims between Brown et al. (2014) and Newhouse et al. (2015) on the effectiveness of the CMS-HCC model and to comprehensively understand strategic risk selection behaviors of MA plans. To shed light on the competing claims, we replicate analyses from Brown et al. (2014) and Newhouse et al. (2015) to examine whether MA plans selectively accepted TM enrollees with lower Medicare expenditures conditional on their risk scores after the CMS-HCC model. To fully understand the MA plans' strategic risk selection behavior in response to the CMS-HCC model, we also examine whether MA plans selectively avoided MA enrollees with higher Medicare expenditures conditional on their risk scores after the CMS-HCC model.

2.4.1 *Data and Study Population*

The MCBS is a longitudinal survey of a nationally representative sample of the Medicare population. CMS annually surveys a nationally representative sample of roughly 11,000 Medicare beneficiaries each year, and link with Medicare claims data. In each MCBS dataset, three rounds of interviews per year are conducted to collect detailed information on access to and satisfaction

before and after 2004 is problematic, Newhouse et al. (2015) limited to MA enrollees who were elderly, who were not institutionalized, and who were not eligible for Medicaid.

of care, functional status, medical conditions, health care expenditures, health insurance, and other health-related topics through the four-years.

The MCBS is particularly well suited for studying risk selection in MA plans. First, it provides a nationally representative sample of the Medicare population with four-year follow-up. This allows us to track the switching behavior between TM and MA plans over time. Furthermore, the MCBS offers comprehensive information on health status and health care utilization for both TM and MA enrollees. While Medicare claims data offers complete information from Medicare-covered services for all TM beneficiaries in the sample, the claims data for MA enrollees is not publicly available. However, the MCBS obtains information on health status and health care utilization for all MA enrollees through survey. This enables us to capture comprehensive information for all TM and MA enrollees in the sample over time.

We construct two comparison groups. To examine whether the CMS-HCC model reduced the phenomenon that MA plans selectively avoid TM beneficiaries with high-risk scores, we compare TM stayers (those enrolled in TM during year t and remained in TM during year $t + 1$) and TM-to-MA switchers (those enrolled in TM during year t, but switched from TM to MA during year $t + 1$) (Panel A). In addition, we also examine the MA plans' strategic risk selection behaviors in response to the CMS-HCC model by comparing MA stayers (those enrolled in MA during year t and remained in MA during year $t + 1$) and MA-to-TM switchers (those enrolled in MA during year t, but switched from MA to TM during year $t + 1$) (Panel B). To construct Panel A and B, respectively, we first identify a sample of Medicare beneficiaries who were eligible for both Medicare Parts A and B coverage in the two consecutive years (t and $t + 1$ years) during the study period. We exclude the following types of beneficiaries from the sample: beneficiaries whose original eligibility was attributable to disability or end-stage renal disease, newly eligible

beneficiaries (since no prior claims information is available), those who died, dual-eligible beneficiaries, those who switched into Special Needs Plans, those who did not have 12 months of continuous enrollment in Medicare (both Parts A and B benefits) in year t, and those not enrolled in Medicare in January of year $t + 1$. Medicare beneficiaries are classified as TM enrollees if enrolled in TM for all 12 months of the calendar year, and classified as MA enrollees if enrolled in an MA plan for at least six months of the year and enrolled in any Medicare plan in every month of the year.

2.4.2 *Empirical Strategy*

First, we replicate analyses from Brown et al. (2014) and Newhouse et al. (2015), which examined selection patterns at different implementation times (i.e., after the initial and full phase-in of the CMS-HCC model, respectively). We perform this analysis for two purposes. The first is to examine whether a relatively small sample from the MCBS provides consistent results with a larger sample from Medicare claims. Following Newhouse et al. (2015), we compare selection during the pre-implementation period with selection during the post-implementation period. If our findings are consistent with those from Newhouse et al. (2015), then this indicates that our analysis with the MCBS provides generalizable results and insights. The second purpose is to examine whether the effectiveness of the CMS-HCC model was larger after the full phase-in of the CMS-HCC model than the initial phase-in.

We first examine whether the CMS-HCC model reduced the phenomenon of selectively avoiding TM enrollees with high-risk scores. For those who enrolled in TM during year t, risk

scores are estimated based on the risk adjustment methodology.[14] To test the hypothesis, we conduct the following difference-in-difference analysis via ordinary least squares (OLS).

$$Risk\ score_{it} = \alpha_0 + \alpha_1 Share\ of\ Year\ in\ MA_{i,t+1} + \alpha_2 Share\ of\ Year\ in\ MA_{i,t+1} \times After\ 2002_t + \alpha_3 Year\ Dummies + \epsilon_{it} \quad (1)$$

where $Risk\ score_{it}$ is beneficiary i's risk score at year t, $Share\ of\ Year\ in\ MA_{i,t+1}$ measures the share of the beneficiary's Medicare-eligible months that she stayed in MA plans in year $t+1$. $After\ 2002_t$ takes the value one for the years after 2003 and takes zero otherwise. We conduct the regression on those enrolled in TM all 12 months of the baseline years 2001-2005. Also, we perform the same regression with $After\ 2005_t$ on those enrolled in TM all months of the baseline years 2001-2002 and 2006-2008. We include year fixed effects, and use sample weights provided by the MCBS. Since we use repeated observations on individuals, standard errors are clustered at the individual level.

In the above equation, the key coefficients are those for $Share\ of\ Year\ in\ MA_{i,t+1}$ and $Share\ of\ Year\ in\ MA_{i,t+1} \times After\ 2002_t$ (or $After\ 2005_t$). Following Brown et al. (2014) and Newhouse et al. (2015), we interpret the results in a way that one simply assumes that TM-to-MA switchers spent the entire next year in MA plans so that the share of the next year spent in MA plans is one for TM-to-MA switchers and zero for TM stayers. Then, the predicted risk score for TM-to-MA switchers in 2002 is $\alpha_0 + \alpha_1$, whereas the predicted risk score for TM stayers in that year is just α_0. For subsequent years, one simply adds the coefficient of the interaction term. For

[14] The way of estimating their risk scores changed over the time period. Risk scores for the pre-implementation period (2001-2003) are estimated based on the PIP-DCG model (Pope et al., 2000). The coefficients estimated from Pope et al. (2000) are used. Risk scores for the post-implementation period (2007-2009) are estimated based on the CMS-HCC model (the HCC 2007 version 12 model). The coefficients estimated from Pope et al. (2004) are used, which is also available at the National Bureau of Economic Research website (http://www.nber.org/data/cms-risk-adjustment.html). Risk scores for the implementation period (2004-2006) are estimated by putting varying weights between the PIP-DCG and CMS-HCC models across years. In 2004, the CMS-HCC model had 30 percent weight in determining payment, in 2005, 50 percent weight, and in 2006, 75 percent weight. From 2004 to 2006, the remaining weight was on the PIP-DCG model.

those who switched beginning in 2004 or 2007, their predicted risk scores are $\alpha_0 + \alpha_1 + \alpha_2$. As such, we interpret the values of $\alpha_1 + \alpha_2$ as how much risk selection decreased or increased after adopting the CMS-HCC model. To check the robustness of our results, we perform additional specifications. We conduct the analysis excluding outliers (i.e., those with risk scores above the 95th percentile in each year), and estimate with median quantile regressions instead of OLS.

We next examine whether MA plans selectively accepted TM enrollees with lower Medicare expenditures conditional on their risk scores after the CMS-HCC model. For those who enrolled in TM during year t, total Medicare expenditures are calculated by summing any Parts A and B expenditures reported in claims data. Total Medicare expenditures were adjusted to 2009 dollars using the Consumer Price Index for All Urban Consumers (CPI-U).[15] We evaluate selection pattern after each of the two implementation points. To test the hypothesis, we conduct the following difference-in-difference analysis via OLS:

$$Expenditure_{it} = \beta_0 + \beta_1 Share\ of\ Year\ in\ MA_{i,t+1} + \beta_2 Share\ of\ Year\ in\ MA_{i,t+1} \times After\ 2002_t + \beta_3 Year\ Dummies + \beta_4 Risk\ score_{it} + \epsilon_{it} \quad (2)$$

where $Expenditure_{it}$ is beneficiary i's total Medicare expenditure at year t, and all other notation are the same as in the previous analysis. As with the above regression, we perform the regression on those who enrolled in TM all 12 months of the baseline years 2001-2005. Also, we perform the same regression with $After\ 2005_t$ on those who enrolled in TM all months of the baseline years 2001-2002 and 2006-2008.

To examine strategic risk selection behaviors of MA plans, we also perform the same analyses for those who enrolled in MA plans in year t (i.e., MA-to-TM switchers and MA stayers). Previous studies, including Brown et al. (2014) and Newhouse et al. (2015), have focused on

[15] Unless otherwise stated, all dollars amount reported in this paper are adjusted to 2009 dollars using the CPI-U.

examining whether the CMS-HCC model reduced the phenomenon of avoiding TM beneficiaries with high-risk scores, mainly due to lack of data for MA enrollees. Since the MCBS provides the data for MA enrollees, we can examine whether MA plans responded strategically to the CMS-HCC model to induce voluntary disenrollment of unprofitable MA enrollees. As with the above regression, we estimate two regressions on those who enrolled in MA plans for at least six months of the baseline years 2001-2005 as well as 2001-2002 and 2006-2008, respectively. For those who enrolled in MA plans during year t, since the claims data for MA enrollees is not publicly available, we follow the risk score estimation method from McWilliams et al. (2012).[16] Their total Medicare expenditures are estimated by summing any Parts A and B expenditures reported in claims data (if enrolled in TM) and the self-reported MA expenditures from the survey. However, the self-reported MA expenditures from the MCBS are underreported.[17] Since the self-reported expenditures are more likely to be underreported in certain populations, we additionally perform the analyses by adjusting for self-reported health status and demographic variables such as age, race, and female.[18] Using these results, we estimate cost savings attributable to such MA plans' strategic behavior in 2007-2009. Assuming that the switching rate of MA-to-TM switchers is generalizable to the entire MA population, we estimate the number of MA-to-TM switchers in the entire MA population in 2007-2009 (Kaiser Family Foundation, 2015), and then multiply it by the average excess expenditures of MA-to-TM switchers beyond their risk-adjusted payments (i.e., the values of $\beta_1 + \beta_2$).

[16] Enrollee-specific capitated payments to MA plans are calculated by multiplying county-specific benchmark rates by enrollee's demographic factors and individual HCC risk scores, modified somewhat by plan bids relative to benchmark rates. To obtain risk scores for MA enrollee each year, we divide capitated payments by county benchmark rates available from CMS.

[17] Curto et al. (2017) estimated health care spending for MA enrollees using claims data from three MA insurers (Aetna, Humana, and UnitedHealthcare). They found that MA spending per enrollee-year totaled $7704, of which $7080 was paid by MA insurers and the rest by enrollees out-of-pocket. We acknowledge that the self-reported expenditures from the MCBS are underreported. However, it is unlikely that such reporting errors have systematically changed over the study period (McWilliams et al., 2012).

[18] We adjust for self-reported health status because individuals with poor health are likely to report negative feelings toward one's health care. We also control for age, race, and female, because different demographic groups are likely to report their health and health care utilization differently.

2.4.3 *Empirical Results*

Table 1 shows baseline characteristics for the MCBS population by the three implementation periods. We find that the number and proportion of TM-to-MA switchers increased [40 (0.26 percent) and 388 (2.69 percent) for the pre- and post-implementation periods, respectively], whereas the number and proportion of MA-to-TM switchers decreased after the full phase-in period [282 (1.85 percent) and 153 (1.06 percent) for the pre- and post-implementation periods, respectively]. Moreover, we find that the difference in total Medicare expenditures between TM stayers and TM-to-MA switchers decreased after the full implementation period [$5486 (=$7915–$2429) and $2269 (=$9806–$7537) for the pre- and post-implementation periods, respectively]. However, the difference in total Medicare expenditures between MA-to-TM switchers and MA stayers increased after the period [$697 (=$4815–$4118) and $4708 (=$8257–$3549) for the pre- and post-implementation periods, respectively]. Furthermore, we find that average risk scores for TM stayers, TM-to-MA switchers, and MA stayers were relatively constant between the pre- and post-implementation periods, whereas average risk scores for MA-to-TM switchers substantially increased from 1.04 in the pre-implementation period to 1.23 in the post-implementation period.

Table 2 displays the results from re-estimating equations from Newhouse et al. (2015) (Panel A) and our own analyses (Panel B). We find that the phenomenon of avoiding TM beneficiaries with high-risk scores reduced after adopting the CMS-HCC model. For those who enrolled in TM during year t, column (1) shows that while TM-to-MA switchers had average risk scores 0.14 points lower than TM stayers in the pre-implementation period, the difference decreased by 0.04 (=–0.14+0.1) after the initial phase-in, assuming they spent full year in MA

plans. As shown in column (2), the risk score difference was almost identical in magnitude after the full phase-in. When the outliers were excluded (columns 2 and 5) or the equation was estimated via median quantile regression (columns 3 and 6), we find similar results. On the other hand, as shown in column (7)-(8), in the pre-implementation period, TM-to-MA switchers had baseline expenditures $4210.92 to $4401.86 lower than TM stayers, assuming they spent full year in MA plans. The amount of favorable selection slightly decreased during the initial phase-in period [by $4203.48 (=–$4210.92+$7.44)], and dramatically decreased after the full phase-in period [by $1803.94 (=–$4401.86+$2597.92)].

However, we find evidence of strategic risk selection of MA plans in response to the CMS-HCC model. For those who enrolled in MA plans during year t, column (1) shows that while MA-to-TM switchers had average risk scores 0.01 points lower than MA stayers in the pre-implementation period, the difference increased by 0.15 [=–0.01+(–0.14)] after the initial phase-in, assuming the stayers spent full year in MA plans and the switchers spent full year in TM. After the full phase-in, as shown in column (2), the risk score difference slightly increased by 0.22 [=–0.01+(–0.21)]. We also find robust results when the outliers were excluded or the equation was estimated via median quantile regression. Moreover, as shown in column (7)-(8), in the pre-implementation period MA stayers had baseline expenditures $917.98 to $926.49 lower than MA-to-TM switchers. However, the amount of selection increased during the initial phase-in period [by $1700.19 (=–$926.49+(–$773.70)) and even more after the full phase-in period [by $5573.89 (=–$917.98+(–$4655.91))]. We also find robust results when we adjusted for self-reported health status and demographic variables (Appendix Table A1). Based on these results, it is estimated that such strategic behavior led MA plans to save costs of $5.9 billion (=$5573.89×3.7%×28.6 million) in 2007-2009.

2.5 SERVICE-LEVEL SELECTION

The goal of this section is to examine the mechanism in which MA plans could effectuate strategic risk selection in response to the CMS-HCC model. We focus service-level selection and analyze in four steps. First, we theoretically examine whether MA plans have incentives to increase copayments disproportionately more for services needed unprofitable people. Then, we empirically test our theoretical prediction. Next, we examine whether MA enrollees used more services with a large increase in copayment were more likely disenroll from MA plans. Finally, we examine whether MA disenrollment was attributable to increased burdens on out-of-pocket costs.

2.5.1 *Conceptual Framework*

By law, MA plans are not allowed to limit enrollment based on beneficiaries' health status. However, MA plans might practice risk selection in subtle ways so that unprofitable beneficiaries voluntarily disenroll from MA plans. To achieve such risk selection, MA plans could risk-select through collecting additional data or advertising. However, because such selection mechanisms would lead to substantial screening costs, MA plans are likely to seek screening approaches with lowest costs. One such approach that generates relatively low screening costs is service-level selection because MA plans do not need to predict each beneficiary's expenditures but rather only need to predict services more likely used by beneficiaries who could be unprofitable under the CMS-HCC model.

Service-level selection is one type of risk selection, which is based on the phenomenon that unprofitable individuals are more likely to use services that are expensive to health plans subject

to capitated payments and are thus more vulnerable to under-provision by MA plans. As with risk selection, service-level selection occurs due to asymmetric information between two parties, in which health plans do not know individuals' private information about health status and preferences for health care.[19] The health plan only knows the probability of using the service at the population level, while the individual knows her need, or probability of need, for each health care service and chooses the best health plan that can satisfy her need (Ellis and McGuire, 2007; Frank et al., 2000). Since rational individuals respond to health plan design when selecting plans, reducing coverage levels for services related to financial losses (i.e., services more likely used by unprofitable individuals) would induce unprofitable individuals to voluntarily disenroll from the plan. In this way, service-level selection would allow health plans to reduce the scope of enrolling those who could be costly to them. Although unprofitable individuals enroll in the plan, service-level selection would also enable health plans to reduce their financial loss as they shift the costs to the individual.

There is suggestive evidence showing that service-level selection occurs in the MA program. Newhouse et al. (2013) estimated margins (i.e., the ratio of average revenue to average cost) across 48 HCCs and unique combinations of HCCs from data on the cost of care from two MA-health maintenance organization (HMO) plans. Despite no evidence of selection across HCCs, they showed that margins in the two plans varied greatly across HCCs. Two additional studies examine switching behavior for those with need for costly services such as nursing home

[19] Even with symmetric information, health plans can have incentives for risk selection if they are not allowed to use the private information to set premiums or benefit features (Van de Ven and Ellis, 2000). In the MA program, for example, CMS reimburses MA plans based on the costs predicted by the CMS-HCC model that only partially accounts for clinically significant medical conditions with significant costs. Because the model does not incorporate all diagnoses, payments to MA plans are too low for sicker beneficiaries and too high for healthier beneficiaries. Consequently, the imperfect risk-adjustment model can create incentives for MA plans to engage in inefficient sorting of individuals across health plans and distortion of plan benefits through service-level selection.

and home health care.[20] Rahman et al. (2015) found that a high proportion of MA enrollees with need for nursing home or home health care disenrolled from MA plans the next year. Similarly, Goldberg et al. (2016) showed that the switching rate for MA and TM beneficiaries without a nursing home stay was the same, but those who required nursing home services in the prior year were more likely to disenroll from MA plans. This phenomenon was more prominent for those with the most costly, longest nursing home stays.

However, there are few studies of the mechanisms through which MA plans could engage in service-level selection against risk adjustment. This study focuses on cost-sharing as a way to effectuate service-level selection. While cost-sharing is designed to protect people against financial risk, it also affects incentives to use more or less health care services. In the presence of low cost-sharing, an individual may use health care services more because she pays less for care than it costs. Health plans may exploit this mechanism to engage in risk selection. Implementing service-level selection through cost-sharing would likely be effective, as quantitative studies found that cost was an important consideration most MA enrollees switching to TM.[21] In this section, we first demonstrate that Medicare beneficiaries who are expected to incur higher expenditures than their risk-adjusted payments estimated by the CMS-HCC model are likely to use services that are expensive, which are more vulnerable to under-provision by MA plans. If this pattern is widespread across CMS-HCC-levels, then MA plans have strong incentives to engage in service-level selection. To avoid those with significantly higher expenditures than their risk-adjusted payments, then MA plans are likely to increase enrollees' cost-sharing more for services that

[20] Implications from Rahman et al. (2015) and Goldberg et al. (2016) may be limited due to different characteristics of the study population. Compared to Medicare beneficiaries who are eligible only for Medicare, those with need for nursing home can use Medicaid-covered services in addition to Medicare-covered services and can enroll in or exit MA plans at any time.

[21] The Government Accountability Office (2017) showed that cost-related concerns were a leading reason for disenrollment of those with poor health as well as those with better health. Moreover, McCormack et al. (2005) found that cost-related concerns were combined with other reasons, amplifying the likelihood of disenrollment from MA plans.

appeal to them than other services after the full phase-in of the CMS-HCC model. If such disproportionate increases in cost-sharing are large enough to affect individuals' plan choice, it would induce those with significantly higher expenditures than their risk-adjusted payments to voluntarily disenroll from MA plans. Consequently, the variation of total health care expenditures for MA enrollees would decrease after the full phase-in period. This effect would be more pronounced for those with high-risk scores than those with low-risk scores.

2.5.2 *Theoretical Predictions on Service-level Selection*

We build upon the two prior theoretical models on service-level selection: Frank et al. (2000) and Ellis and McGuire (2007). A health plan's profit is revenue less costs. Health plans' revenues from individual i, rev_i, typically comprise a risk-adjusted (capitated) payment. Following Ellis and McGuire (2007), we assume that a premium that the plan charges has been predetermined and thus does not influence plans' strategies to effectuate risk selection. On the other hand, the plan incurs costs for providing services. Frank et al. (2000) characterized plans' rationing as a shadow price on access to various types of care.[22] From the perspective of an individual, this can be interpreted as a threshold of clinical need or benefit that the individual must exceed to receive services. As such, a higher shadow price means tighter rationing. For service s, the plan sets a shadow price to ration the service. Let $q = \{q_s\}$ be a vector of shadow prices determined by the plan to ration services and $m_i(q) = \{m_{is}(q_s)\}$ be the vector of expenditure on service s that individual i spends as a function of the service-specific shadow price. The level of expenditure that individual i spends on service s, $m_{is}(q_s)$, is determined by the point at which the marginal benefit of expenditure for

[22] A shadow price is regarded as a device to capture various rationing strategies by a plan, which determines access to care. For example, the shadow price reflects plan decisions about capacity in various service areas as well as the makeup of networks or payment to providers.

that individual is equal to the shadow price q_s. Therefore, the plan's profit for individual i can be expressed as $rev_i - \sum_s m_{is}(q_s)$.

The plan's total profit depends on who joins. Whether an individual joins the plan is determined by her expectation of what she would receive in the plan. Let $\hat{m}_{is}(q_s)$ be the services that individual i expects to receive in a plan that rations using service-specific shadow prices q_s. From the perspective of the plan, individual i enrolls in the plan with a probability $n_i(\hat{m}_{is}(q_s))$ as a function of shadow prices. Therefore, the plan's total profit can be expressed as:

$$\pi(q) = \sum_i n_i\big(\hat{m}_{is}(q_s)\big)\left[rev_i - \sum_s m_{is}(q_s)\right] \tag{3}$$

The plan chooses each q_s to maximize expected profits in the equation (3). To find profit-maximizing values of each q_s, the equation (3) is differentiated with respect to q_s.

By differentiating the equation (3) from Frank et al. (2000), Ellis and McGuire (2007) derived the service-level selection index, which measures the plans' incentives to ration care tightly across services, I_s:

$$I_s = \sigma_\pi \phi \eta_s \left[\frac{\sigma_{\hat{m}_s}}{\bar{m}_s}\rho_{\hat{m}_s,\pi} - C\right] \tag{4}$$

where π is the plan's net profits, σ_π is the standard deviation of π, ϕ is a uniform enrollment function and is constant across service s, η_s is the demand elasticity for service s (a negative number), $\hat{m}_s$ is the individual's expected expenditure on service s, $\sigma_{\hat{m}_s}$ is the standard deviation of $\hat{m}_s$, $\bar{m}_s$ is the mean level of expected spending on service s, $\rho_{\hat{m}_s,\pi}$ is the correlation between $\hat{m}_s$ and π, and C is a numeric constant to capture terms that do not depend on service s.

The service-level selection index, I_s, measures the relative magnitude of selection incentives across services. It consists of three components: 1) the coefficient of variation of the

expected expenditure ($\widehat{m}_s$) on service s (*predictability*), $\frac{\sigma_{\widehat{m}_s}}{\bar{m}_s}$, 2) the correlation between the expected expenditure on service s ($\widehat{m}_s$) and net profits (π) (*predictiveness*), $\rho_{\widehat{m}_s,\pi}$, and 3) the demand elasticity for service s, η_s. In this study, we focus on the first two components.[23]

First, *predictability* represents how well individuals can predict service-level use. If individuals cannot predict service-level use well, service-level selection would have little to no effect on enrollment or plan profits. If individuals cannot predict service-level use at all (i.e., everyone expects themselves to be average users), predictability is zero. Selective rationing of health plans would not affect individual's plan choices, and no distortion occurs. When individuals can predict their service-level use, expected expenditures ($\widehat{m}_s$) would vary, and predictability increases. In this case, selective rationing would be effective in attracting or deterring certain types of individuals. In other words, the better the information that individuals have about their future health care use, the larger the distortion caused by the plan's selective rationing to avoid unprofitable individuals.

Second, *predictiveness* represents how use of a service is correlated with net profit per individual. This indicates whether a service is more likely to be used by those with financial gains or losses for the plan. When use of a service is negatively correlated with profits (π), the plan would want to ration the service to avoid those individuals associated with financial losses. When use of a service is positively correlated with profits, however, the plan would not want to ration the service to attract those with financial gains.

[23] Although the demand elasticity affects the magnitude of the index, it is unlikely to affect the order of the index. As shown in a literature review of Ringel et al. (2002), the price elasticity of demand for health care services is in general low. Almost services were estimated to be relatively less price-sensitive (in the range of -0.1 to -0.2 for inpatient, outpatient, and mental health services). Thus, we do not consider the demand elasticity in further analysis.

To summarize, Ellis and McGuire (2007) presented that the plan's incentives to ration at the service level are proportional to the product of *predictability* and *predictiveness*.[24] For services that are either not predictable or does not correlate with net profit, a health plan has no incentive to ration. For services with a large positive value of I_s (i.e., services that are highly predictable and negatively correlated with net profit), however, the plan has incentives to ration these services tightly. For services with a large negative value of I_s (i.e., services that are highly predictable and positively correlated with net profit), the plan has incentives to not ration these services.

Building upon Ellis and McGuire (2007), we extend the model to show that MA plans are likely to employ service-level selection as a strategic behavior in response to the CMS-HCC model.

MA plans must accept all Medicare beneficiaries who wish to join and offer at least the same benefits as TM (i.e., services covered under Parts A and B).[25] CMS pays MA plans a fixed capitated payment to cover the costs for services covered under TM. Using the CMS-HCC model, CMS calculates payments to MA plans separately for each enrollee in the plan, multiplying the plan's payment rate by the enrollee's risk score r. CMS uses the prior-year's TM data to estimate risk scores for current MA enrollees.[26] The capitation payment for an MA enrollee is based on the estimated Parts A and B payments had TM covered her directly. Following Ellis and McGuire (2007), let M_i denote total annual Medicare expenditure that individual i spends, and define $M_i =$

[24] Using Medicare claims data for 1996-1997, Ellis and McGuire (2007) measured the relative magnitude of potential selection across various types of services. For instance, hospice, home health, durable medical equipment, provider specialties of pulmonary care, oncology ambulance, and psychiatry were shown to have potential for under-coverage by managed care plans. On the other hand, eye procedures, magnetic resonance imaging (MRI), and provider specialties such as chiropractic and gynecology were found to be candidates for over-coverage. Similar patterns of health plans incentives for service-level selection were found in other studies, for example, Cao and McGuire (2003) in Medicare, Eggleston and Bir (2009) in the state employee insurance program, and Ellis et al. (2013) in commercial health plans.

[25] Since risk-adjusted payments are based on services in Parts A and B, we focus on services covered by both TM and MA plans.

[26] The data includes TM beneficiaries entitled by age or disability with continuous 12-month enrollment in TM, and thus those entitled by end stage renal disease or those without 12 months base year Medicare enrollment are excluded. For each of them, a separate risk adjustment model is used to predict their next year expenditures.

$\sum_s m_{is}(q_s)$. Let $C(r_i)$ denote the risk-adjusted (capitated) payment that an MA plan receives from CMS for individual i with a risk score r.[27] The MA plan's profit for individual i is expressed as:

$$\pi_i = C(r_i) - M_i \quad (5)$$

In an ideal risk-adjusted payment system, MA plans will have no incentive to select Medicare beneficiaries. Under an imperfect risk adjustment model, however, MA plans have incentives to discourage enrollment of those with predictably higher expenditures than their risk-adjusted payments, $\pi_i < 0$. The incentives would be stronger for those with high-risk scores than low-risk scores, as the CMS-HCC model underpredicts expenditures for those with the most severe health status (Medicare Payment Advisory Commission, 2012). This can be expressed as:

$$var(\pi_{i_h}) > var(\pi_{i_l}) \quad (6)$$

where $var(\,)$ indicates the variance of a variable. i_h and i_l indicate those with high-risk scores and low-risk scores, respectively.

Given varying selection incentives across services, as shown in Ellis and McGuire (2007), MA plans would be interested in figuring out the relationship between the magnitude of the incentive to ration services used by beneficiaries with a given risk score and the degree to how far their total expenditures are from the mean of their conditional expenditure distribution. If a service that is more likely to be used by those with substantially higher expenditures than their risk-adjusted payment is the one that MA plans want to ration more tightly, then MA plans would ration care by the order of the service-level selection index estimated from Ellis and McGuire (2007). We assume that such rationing behaviors occur across MA plans competing for beneficiaries. This suggests that switching to another MA plan is unlikely, since in a competitive market MA plans behave similarly.

[27] Since 2006, CMS has implemented a competitive bidding system to determine payments to MA plans. However, as a rebate must be returned to enrollees as a reduction in premiums or additional benefits, the bidding system is unlikely to affect $C(r_i)$.

We demonstrate how the probability of using a service by beneficiaries with expenditures higher than their risk-adjusted payment is related to the Ellis and McGuire (2007)'s service-level selection index. Define u_{is} as individual i's actual use of services s. u_{is} takes the value 1 if individual i used service s and zero otherwise. Let $\hat{P}(u_{is}) \in \{0,1\}$ denote the probability that individual i expects to use service s. To effectuate service-level selection, MA plans do not need to forecast service use at the individual level but rather focus on forecasting at the population level. Thus, define the population-level expected probabilities of using service s given that an individual i's total actual expenditure is higher than her risk-adjusted payment as follows:

$$\frac{\sum_i \left(\hat{P}(u_{is}, \pi_i < 0)\right)}{\sum_i \hat{P}(\pi_i < 0)} = \sum_i \left(\frac{\hat{P}(u_{is})\hat{P}(\pi_i < 0|u_{is})}{\hat{P}(\pi_i < 0)}\right) \quad (7)$$

where the last equality of the equation (7) is drawn from Bayes' theorem.

Then, we re-express as follows:

$$\sum_i \hat{P}(u_{is}) \sum_i \left(\frac{\hat{P}(\pi_i < 0|u_{is})}{\hat{P}(\pi_i < 0)}\right) - \sum_i \sum_{j,j\neq i} \left(\frac{\hat{P}(u_{is})\hat{P}(\pi_i < 0|u_{is})}{\hat{P}(\pi_i < 0)}\right) \quad (8)$$

The equation (6) can be written as:

$$NN\left[\frac{1}{N}\sum_i \hat{P}(u_{is})\right]\left[\frac{1}{N}\sum_i \left(\frac{\hat{P}(\pi_i < 0|u_{is})}{\hat{P}(\pi_i < 0)}\right)\right] - \frac{1}{N}\frac{1}{N}\sum_i \sum_{j,j\neq i} \left(\frac{\hat{P}(u_{is})\hat{P}(\pi_i < 0|u_{is})}{\hat{P}(\pi_i < 0)}\right) \quad (9)$$

The equation (9) indicates that the probability of using a service by unprofitable individuals is directly proportional to the second component of the first term, $\left[\frac{1}{N}\sum_i \left(\frac{\hat{P}(\pi_i<0|u_{is})}{\hat{P}(\pi_i<0)}\right)\right]$, which represents the average ratio of the probability that individual i is expected to incur a net loss given use of service s to the probability of a net loss for the individual i. This is another way to measure *predictiveness* of the service-level selection index, ($\rho_{\hat{m}_s,\pi}$), which measures the correlation of use of service with profitability to the plan. If $\left[\frac{1}{N}\sum_i \left(\frac{\hat{P}(\pi_i<0|u_{is})}{\hat{P}(\pi_i<0)}\right)\right]$ is high, it implies that service s is

more likely to be used by those with financial losses. If $\left[\frac{1}{N}\sum_i\left(\frac{\hat{P}(\pi_i<0|u_{is})}{\hat{P}(\pi_i<0)}\right)\right]$ is low, on the other hand, it implies that service s is more likely to be used by those with financial profits. Theoretically, *predictiveness* can have either negative or positive value. However, Ellis and McGuire (2007) showed that all services (except for chiropractic service) were estimated to have positive values.

To sum up, where Ellis and McGuire (2007) stopped at showing the incentives for health plans to ration care tightly across services, we go further and demonstrate the relationship between the plans' incentive to ration a service tightly and the probability of using the service by unprofitable individuals across services.

$$\frac{\sum_i\left(\hat{P}(u_{is},\pi_i<0)\right)}{\sum_i\hat{P}(\pi_i<0)} \propto \rho_{\hat{m}_s,\pi} \propto I_s \qquad (10)$$

The equation (10) presents that MA plans want to ration some services more tightly because they are more used by unprofitable beneficiaries under the CMS-HCC model. This suggests that MA plans have incentives to effectuate risk selection via service-level selection. This phenomenon is more pronounced for those with high-risk scores, as the CMS-HCC model systematically underpredicts expenditures for those with high-risk scores (Medicare Payment Advisory Commission, 2012), which indicates that net losses increase with an increase in risk score. Hence, as risk scores increase, $\left[\frac{1}{N}\sum_i\left(\frac{\hat{P}(\pi_i<0|u_{is})}{\hat{P}(\pi_i<0)}\right)\right]$ is likely to increase.

2.5.3 *Data and Study Population*

To examine service-level selection, we use two data sets: the 2001-2009 Plan Benefit Package (PBP) and the 2001-2009 MCBS.

The PBP provides information on the set of benefits that an MA plan offers (e.g., premiums, cost-sharing, and additional benefits by service). The data are submitted to CMS for benefit analysis, marketing, and beneficiary communication purposes. Recently, CMS has used the data to review and approve all benefits yearly to ensure that MA plans do not discriminate against beneficiaries with poor health or those who incur financial losses.

We identify MA plans with complete information on cost-sharing for all services covered under Parts A and B. We exclude private fee-for-service (PFFS) plans because their characteristics are similar to TM despite that PFFS plans are classified as and paid like an MA plan. We also exclude cost-based, demonstration, special needs, Medicare Savings Account, and employer-sponsored plans because they are available only to small numbers of Medicare beneficiaries. Thus, we limit analysis to HMO and preferred provider organization (PPO) plans.

We use the same MCBS population identified in Section 4.

2.5.4 *Empirical Strategy*

To test whether MA plans employed service-level selection in response to the CMS-HCC model, we compare percent changes in weighted average service-specific copayments between the pre- and post-implementation periods. We calculate service-specific copayments of 33 services covered under Medicare Parts A and B (Table 3). Most MA plans use copayments, but others use coinsurance rates. Assuming that there are marginal variations in service prices between TM and MA plans[28] and across MA plans, we convert coinsurance rates to copayments based on mean allowed charges or charges per TM beneficiaries for each service and year, which was estimated

[28] Trish et al. (2017) found that physician reimbursement in MA plans was similar to or slightly less than TM rates. For a standard mid-level office visit with an established patient, the mean MA price was 96.9 percent of TM. For physician services, mean MA reimbursement ranged from 91.3 percent of TM for cataract removal in an ambulatory surgery center to 100.2 percent of TM for the professional fee for interpretation of a computed tomographic scan in an emergency department.

from the MCBS. To map each claim or line item into the PBP services, we use the service categories used in the Medicare Options Compare Out-of-Pocket Cost (OOPC) Estimates Methodology (Centers for Medicare and Medicaid Services, 2008).[29] For inpatient hospital, skilled nursing facility, mental health specialty service, psychiatric service, and outpatient substance abuse service, MA plans can set up varying cost-sharing by a length of stay or number of visit. For these services, we calculate a copayment based on a typical length of stay or number of visit (Government Accountability Office, 2010), and then calculate a copay per day. For other services, a copayment is calculated per visit. All service-specific copayments are adjusted to 2009 US dollars by the equivalent service-specific price index (Agency for Healthcare Research and Quality). To account for varying numbers of MA plans across years, we adjust by weighting the number of MA plans in each year. Then, we plot the percent changes with respect to the service-level selection index estimated from Ellis and McGuire (2007),[30] and compare the plotted relations between the pre-and post-implementation periods.

To test whether service-level selection affected the individuals' plan switching behavior after the full phase-in period of the CMS-HCC model, we compare service-specific use during year t between switchers and stayers in the pre-implementation period with that in the post-implementation period. We measure health care utilization by type of service. For those who enrolled in TM during year t, we use claims to create a total of 29 types of services categories (Table 4). Part A claims are classified into the following five types of service (hospital inpatient visit, hospital outpatient visit, home health care, hospice, and other facility services). Part B claims

[29] The mapping identification for each service is conducted based on the Berenson-Eggers Type of Services (BETOS) codes, physician specialty codes, service type, place of service, bill type code, and revenue center code. The service-specific mapping identification is described in Centers for Medicare and Medicaid Services (2008). We use the 2009 OOPC Methodology, which is the model close to the year in which the CMS-HCC model was fully phased-in. Although the way of identifying each PBP service differs across years, dramatic changes are unlikely.

[30] We assume that the magnitude of the incentives to ration care tightly at the service level is consistent across time. Thus, we use the service-level selection index estimated using the data prior to the implementation of the CMS-HCC model.

are classified into 24 categories by the Berenson-Eggers Type of Services (BETOS) code, which is used to create clinically meaningful groupings of procedures and services to analyze Medicare expenditures by type of service. For those who enrolled in MA plans during year t, we estimate service-level use through the survey. In the survey, participants reported their use of health care by the following 7 types of service: hospital inpatient visit, home health, hospice, medical provider visit, hospital outpatient visit, prescribed medicine, and facility service. Thus, we create a total of 7 types of services categories (Table 4). To examine whether TM enrollees who used services with lower service-level selection index in year t were more likely to switch to MA plans in year $t+1$ than those who used services with higher service-level selection index in year t after the full phase-in ("intensive-margin selection"), we estimate the ratio of the proportion of TM-to-MA switchers with use of a particular type of service to TM stayers with use of the service in the pre- and post-implementation periods, respectively. To examine whether TM enrollees who used more services with lower service-level selection index in year t were more likely to switch to MA plans in year $t+1$ than those who used more services with higher service-level selection index in year t after the full phase-in ("extensive-margin selection"), we also estimate the ratio of average number of services per enrollee of TM-to-MA switchers to TM stayers in the pre- and post-implementation periods, respectively. Then, we plot the ratios with respect to the service-level selection index estimated from Ellis and McGuire (2007), and compare the plotted relations between the pre-and post-implementation periods. We perform the same analysis for those who enrolled in MA plans in year t.

To test whether service-level selection allowed MA plans to reduce the scope of enrolling those with higher expenditures than their risk-adjusted payments, especially for those with high-risk scores, we estimate the coefficients of variance of total Medicare expenditures for TM and

MA enrollees, respectively, over time by risk score. The coefficient of variance of total Medicare expenditures is estimated as the ratio of the standard deviation of total Medicare expenditures to its mean. We divide the study population into enrollees with high-risk scores and other risk scores. High-risk scores indicate above the 90th percentile of the risk score distribution in each year.

To examine the reasons of disenrollment from MA plans, we examine whether disenrollment from MA plans was related to lower satisfaction on care costs, quality of care, or access to care in year t. Satisfaction is measured by four levels: very dissatisfied, dissatisfied, satisfied, and very satisfied. To test this hypothesis, we conduct the following difference-in-difference analysis via OLS:

$$Satisfaction_{it} = \gamma_0 + \gamma_1 Share\ of\ Year\ in\ MA_{i,t+1} + \gamma_2 Share\ of\ Year\ in\ MA_{i,t+1} \times After\ 2005_t + \gamma_3 Year\ Dummies + \gamma_4 Risk\ score_{it} + \gamma_5 Health_{it} + \gamma_6 Demograhics_{it} + \epsilon_{it} \quad (11)$$

where $Satisfaction_{it}$ is beneficiary i's reported satisfaction in year t. $Health$ measures the five-category self-reported health variable (one "poor" up to five "excellent") and $Demographics$ includes age, race, and female. All other notations are the same as in the previous analysis. We perform the regression on those who enrolled in MA plans for at least six months of the baseline years 2001-2002 and 2006-2008.

2.5.5 *Empirical Results*

Table 5 shows summary statistics on MA plans by the three implementation periods of the CMS-HCC model. The mean numbers of MA plans were 481 (SD = 72) and 2,843 (SD = 348) in the pre- and post-implementation periods, respectively. The shares of HMO plans were 98.75 percent and 71.26 percent in the pre- and post-implementation periods, respectively.

Fig. 1 presents evidence of service-level selection in the MA program after the CMS-HCC model was fully phased in. Both of the fitted lines for the pre- and post- implementation periods show an upward trend with respect to the service-level selection index, but the fitted line for the post-implementation period is tilted upward more than the pre-implementation period. This indicates that MA plans increased enrollees' copayments disproportionately more for services with higher service-level selection index [e.g., durable medical equipment (the percent change of weighted average copayments between the post-implementation period and the pre-implementation period: 88.47), home health (71.93), diabetes monitoring supply (42.78), inpatient hospital psychiatric service (39.47), and inpatient hospital acute service (38.44)] than services with lower service-level selection index [e.g., mental health specialty service—individual session (5.62), psychiatric service—individual session (1.63), and primary care physician service (0.83).

Fig. 2 presents changes in service use of TM-to-MA switchers to TM stayers between the pre- and post-implementation periods. In both figures showing results for intensive and extensive margin selection, respectively, we observe that for most services, the ratios of service use of TM-to-MA switchers to TM stayers are lower than one. Also, the fitted lines for the pre- and post-implementation periods show a downward trend with respect to the service-level selection index, with an almost same slope and are below the ratio of one. However, the intercept of the post-implementation period is higher than the intercept of the post-implementation period. We also find that after the full phase-in period, TM-to-MA switchers systematically used more services across all services compared to TM switchers (not shown). However, even after the full implementation period, TM enrollees who used services with higher service-level selection index in year t were more likely to switch to MA plans in year $t + 1$ than those who used services with lower service-level selection index in year t. When the outlier service with the highest value of the service-level

selection index (hospice) was excluded, we find similar findings (Appendix Fig. 1). However, we observe that the fitted line for the post-implementation period is tilted upward more than the pre-implementation period. This indicates that TM enrollees who used services with higher service-level selection index in the full phase-in period were more likely to switch to MA plans than the equivalent population in the pre-implementation period.

On the other hand, the below two figures show changes in service use of MA-to-TM switchers to MA stayers between the pre- and post-implementation periods. In the both figures showing results for intensive and extensive margin selection, respectively, the fitted line for the post-implementation period shows a steeper slope the fitted line for the pre-implementation period. This indicates that MA enrollees who used services with higher service-level selection index after the full implementation period were more likely to disenroll from MA plans than the equivalent population in the pre-implementation period.

The left and right figures in Fig. 3 present the coefficients of variance of total Medicare expenditures for TM enrollees and MA enrollees by risk scores, respectively. For TM enrollees with both high-risk scores and other risk scores, the coefficients of variance of total Medicare expenditures decreased over time. However, we observe different patterns for MA enrollees by risk scores. The coefficients of variance of total Medicare expenditures for MA enrollees with high-risk scores show a downward trend over time, whereas those for MA enrollees with other risk scores show an upward trend over time.

Table 6 shows the results from examining the relation of MA disenrollment and satisfaction of care after the fully phase-in period. Column (1) shows that after the full phase-in period, the relation of MA disenrollment and satisfaction on care costs was the most pronounced among satisfaction measures considered in this study. Specifically, it was shown that relative to MA-to-

TM switchers, MA stayers were less satisfied with out-of-pocket costs by 0.018 points in the pre-implementation period, assuming the stayers spent full year in MA plans and the switchers spent full year in TM. However, MA stayers were more satisfied with out-of-pocket costs by 0.175 (=–0.018+0.193) points than MA-to-TM switchers after the full implementation of the CMS-HCC model. As shown in columns (2)-(8), we observe that MA stayers were more satisfied with care quality and access to care than MA-to-TM switchers in the full phase-in period. However, the magnitude of the change was lower than that for satisfaction on care costs.

2.6 DISCUSSION

The goal of this paper is to shed light on the competing claims on the effectiveness of the CMS-HCC model and to understand strategic risk selection behaviors of MA plans. We find that the CMS-HCC model achieved one goal of reducing the avoidance of high-risk score beneficiaries in MA plans. However, the CMS-HCC model also led to increased disenrollment of high-cost beneficiaries, conditional on risk score, in MA plans. We explain this unintended consequence through service-level selection. Through theoretical and empirical analysis, we show that after the full phase-in period of the CMS-HCC model, MA plans have the incentive to and did increase copayments disproportionately more for services that appeal to beneficiaries who could be unprofitable under the CMS-HCC model than other services. The disproportionate changes in copayments led to voluntary disenrollment of beneficiaries with need for these services, who tend to incur higher expenditures than their risk-adjusted payments. We also find evidence supporting our hypothesis that those who were less satisfied with out-of-pocket costs were more likely to disenroll from MA plans. Such strategic behavior led to MA plans to save $5.9 billion in 2007-

2009 by simply transferring the costs to the federal government, thereby placing significant financial burdens on the federal government.

Our study shows evidence of the intended effect of the CMS-HCC model on reducing risk selection for TM beneficiaries with high-risk scores, consistent with findings from Newhouse et al. (2015). Specifically, the differences in risk scores between TM-to-MA switchers and TM stayers reduced by more than a factor of two. Also, the differences in total Medicare expenditures between them decreased after the full phase-in period to $1803.94. This shows the intended consequence of the implementation of the CMS-HCC model. As risk adjustment leads to neutral payments for those with conditions included in the risk adjustment formula, MA plans no longer have incentive to avoid them. These findings add to earlier studies that found that a more clinically detailed risk adjustment model strengthens the incentives for MA plans to retain sick enrollees. However, the risk selection did not disappear, but rather changed in form.

The main contribution of this study is to show that MA plans magnified service-level selection in response to the CMS-HCC model to avoid beneficiaries who could be costly under the CMS-HCC model. Our theoretical analysis demonstrates that MA plans have incentives to effectuate risk selection via service-level selection, as unprofitable beneficiaries under the CMS-HCC model are more likely to use services that are expensive and are thus more vulnerable to under-provision by MA plans. It informs why MA plans more engaged in service-level selection after adopting the CMS-HCC model. Then, our empirical analysis provides supporting evidence showing that MA plans increased copayments disproportionately more for services that are more likely to be used by them. This validates our theoretical analysis and contributes to providing evidence-based policy implications. With the theoretical foundation and empirical evidence based closely on the theoretical foundation, this study adds to the body of literature regarding MA plans'

strategic response to the policy-induced change in financial incentives as an important contributor to service-level selection.

It is worthwhile to note that relatively large increases in copayments were found in two types of services. First, increases in copayments for home health service, durable medical equipment, inpatient psychiatric hospital service were likely targeted for disenrollment of beneficiaries with multiple chronic conditions. This is because their expenditures were systematically underpredicted by the CMS-HCC model. For example, the model, on average, underestimated expenditures for those with more than six chronic conditions by $608 (Government Accountability Office, 2011). Also, an increase in copayment for acute inpatient hospital service was likely intended to encourage disenrollment of those with potentially high risks because of poor health behaviors. Consequently, such service-level selection would likely lead to voluntary disenrollment of those who currently need these services as well as those who potentially have the need for these services.

Our study also shows the pronounced effect of service-level selection on MA enrollees after the full phase-in of the CMS-HCC model. Specifically, we find that MA enrollees who used services with high service-level selection index in the previous year were more likely to disenroll from MA plans in the following years than those who used services with low service-level selection index in the previous year. We also show that MA enrollees who were less satisfied with their out-of-pocket costs were more likely to disenroll from MA plans. These findings suggest that MA enrollees were more likely to disenroll from MA plans due to increased burdens on out-of-pocket costs as a result of service-level selection. Service-level selection could result in poor health status, because it is likely to lead to delayed care during the MA enrollment period, and inefficient or uncoordinated cares following enrollment switching to TM. This is especially true for those

with multiple chronic conditions, which requires more integrated and coordinated care due to complex conditions and treatment.

Although the size of MA-to-TM switchers was only about one percent of the entire Medicare population, risk selection in this population cannot be considered trivial with the following three reasons. First, the size of the population increases over time. From our data, 3.7 percent of MA enrollees left their MA plans between 2007 and 2009. However, in 2014, nearly 12 percent left their MA plans (Government Accountability Office, 2017). Also, the cost implications for this population would be significant given that the top five percent of the US population accounts for about 50 percent of total health care expenditures (Cohen, 2014). If MA-to-TM switchers keep experiencing delayed care or receiving fragmented care in non-managed care settings, this would incur even higher treatment costs. Moreover, as MA payments are partly determined by the average expenditures of TM beneficiaries at the county level, switching of high-cost MA enrollees to TM could lead to MA payment increases, possibly placing significant financial burdens on the federal government.

Our findings provide key implications for CMS in developing a better risk adjustment model. To ensure that MA plans' benefit package designs do not discriminate against beneficiaries in poor health with high health care expenditures, since 2010, CMS has reviewed all benefit packages yearly (Government Accountability Office, 2010). In addition to the review process for MA plans' benefit structures, developing a better risk adjustment model is inevitable as MA plans would continue to engage in risk selection if a risk adjustment model does not estimate capitation payments as sufficiently close to the actual expenditures. The new risk adjustment model needs to be designed to generate economic forces to prevent MA plans' strategic behaviors in engaging in service-level selection. Specifically, developing a risk adjustment model that not only conditions

on each beneficiary's risk scores but also reflect each beneficiary's potential service-level use may contribute to reducing service-level selection. This approach enables to provide overpayments for services that are more likely used by unprofitable beneficiaries and underpayments for services that are more likely used by profitable beneficiaries, thereby equalizing incentives in rationing all services (Glazer and McGuire, 2000). By regarding risk adjustment as a tax/subsidy scheme, overpaying for services in high demand by unprofitable beneficiaries and underpaying for services in high demand by profitable beneficiaries would redistribute health care costs away from profitable beneficiaries and toward unprofitable beneficiaries. As the payment for profitable services is much smaller than the payment for unprofitable services, this would penalize MA plans for attracting only those in need of profitable services. Thus, the new risk adjustment model could reduce the potential for MA plans to use service-level selection.

This study has several limitations. First, we assumed that the magnitude of the incentives to ration care tightly at the service level is consistent across time. However, it might not be true because reimbursement policies and rates change over time, possibly affecting the magnitude of the incentive across years. Following Ellis and McGuire (2007), we also assumed that all individuals share the same elasticity of demand for a certain service. However, the demand elasticity might differ across services (Manning et al., 1987) as well as individuals. Moreover, Ellis et al. (2017) further developed the service-level selection index by accounting for variation in cost-sharing, risk-adjusted profits, and demand elasticities across services. However, estimates of the new service-level selection index were empirically calculated based on commercial claims data. Considered differences in demographic profiles and health care utilization patterns between the Medicare population and the commercial insured, the estimates are unlikely to be applicable to our study. Furthermore, we assumed small variations in service prices and service utilization between

TM and MA plans and across MA plans. Thus, we converted service-specific coinsurance rates to copayments using mean allowed charges per TM beneficiaries for each year and year. However, prices in MA plans are not equivalent to those in TM (Baker et al., 2016; Trish et al., 2017). Also, there is substantial heterogeneity in cost and market power across MA plans (Glazer and McGuire, 2017). Moreover, we used self-reported data to estimate utilization and Medicare expenditures for MA enrollees, which is significantly underreported (Eppig and Chulis, 1997). However, it is unlikely that such reporting errors have systematically changed over the study period (McWilliams et al., 2012). Finally, individuals generally consider various aspects of plan benefits in changing a health plan (Government Accountability Office, 2017; McCormack et al., 2005). However, this study focuses only on cost-sharing structures. Therefore, there is the possibility that MA plans lessened other strategies of distorting service offerings such as access to specialist or provision of additional benefits in order to accept sicker TM beneficiaries who were no longer unprofitable under the CMS-HCC model, which is beyond the scope of this study.

Findings from this study indicate that the CMS-HCC model reduced the MA plans' risk selection of avoiding TM beneficiaries with high-risk scores, whereas it induced MA plans to strategically behave in response to the CMS-HCC model via service-level selection. MA plans have raised copayments disproportionately more for services needed by high-need beneficiaries than for other services, thereby inducing unprofitable beneficiaries to voluntarily disenroll from their MA plans**,** mainly due to increased out-of-pocket costs. This allows MA plans to avoid the risk of enrolling unprofitable beneficiaries. Our results provide key policy implications for CMS in moving towards a better risk adjustment model that accounts for the enrollees' predicted risk scores while generating economic incentives for MA plans that discourage service-level selection.

RERERENCES

Agency for Healthcare Research and Quality. Using appropriate price indices for analyses of health care expenditures or income across multiple years. Agency for Healthcare Research and Quality: Baltimore, MD.

Baker L.C., Bundorf M.K., Devlin A.M., Kessler D.P. 2016. Medicare Advantage plans pay hospitals less than traditional Medicare pays. Health Aff. 35 (8), 1444-1451.

Brown J., Duggan M., Kuziemko I., Woolston W. 2014. How does risk selection respond to risk adjustment? New evidence from the Medicare Advantage program. Am. Econ. Rev. 104 (10), 3335-3364.

Cao Z., McGuire T.G. 2003. Service-level selection by HMOs in Medicare. J Health Econ. 22 (6), 915-931.

Centers for Medicare and Medicaid Services. 2008. CY 2009 Medicare Options Compare Out-of-Pocket Cost (OOPC) Estimates Methodology Centers for Medicare and Medicaid Services: Baltimore, MD.

Cohen S.B. 2014. The concentration of health care expenditures and related expenses for costly medical conditions, 2012. Statistical Brief #455. Agency for Healthcare Research and Quality: Rockville, MD; 2014.

Eggleston K., Bir A. 2009. Measuring Selection Incentives in Managed Care: Evidence from the Massachusetts State Employee Insurance Program. J Risk Insur. 76 (1), 159-175.

Curto V., Einav L., Finkelstein A., Levin J.D., Bhattacharya J. 2017. Healthcare spending and utilization in public and private Medicare. NBER Working Paper No. 23090.

Ellis R.P., Jiang S., Kuo T.-C. 2013. Does service-level spending show evidence of selection across health plan types? Appl. Econ. 45 (13), 1701-1712.

Ellis R.P., Martins B., Zhu W. 2017. Demand elasticities and service selection incentives among competing private health plans. J Health Econ. 56 352-367.

Ellis R.P., McGuire T.G. 2007. Predictability and predictiveness in health care spending. J Health Econ. 26 (1), 25-48.

Eppig F.J., Chulis G.S. 1997. Matching MCBS (Medicare Current Beneficiary Survey) and Medicare data: the best of both worlds. Health Care Financ. Rev. 18 (3), 211-229.

Frank R.G., Glazer J., McGuire T.G. 2000. Measuring adverse selection in managed health care. J Health Econ. 19 (6), 829-854.

Frogner B.K., Anderson G.F., Cohen R.A., Abrams C. 2011. Incorporating new research into Medicare risk adjustment. Med. Care. 49 (3), 295-300.

Glazer J., McGuire T.G. 2000. Optimal risk adjustment in markets with adverse selection: an application to managed care. Am. Econ. Rev. 90 (4), 1055-1071.

Glazer J., McGuire T.G. 2017. Paying medicare advantage plans: To level or tilt the playing field. J Health Econ. 56 281-291.

Goldberg E.M., Trivedi A.N., Mor V., Jung H.-Y., Rahman M. 2016. Favorable risk selection in Medicare Advantage: trends in mortality and plan exits among nursing home beneficiaries. Med. Med. Res. Rev. 1 (14).

Government Accountability Office. 2010. Relationship between benefit package designs and plans' average beneficiary health status. Government Accountability Office: Washington, DC.

Government Accountability Office. 2011. Medicare Advantage: Changes improved accuracy of risk adjustment for certain beneficiaries. Government Accountability Office: Washington, DC.

Government Accountability Office. 2017. CMS should use data on disenrollment and beneficiary health status to strengthen oversight. Government Accountability Office: Washington, DC.

Han T., Lavetti K. 2017. Does Part D abet advantageous selection in Medicare Advantage? J Health Econ. 56 368-382.

Jacobson G.A., Neuman P., Damico A. 2015. At least half of new Medicare advantage enrollees had switched from traditional Medicare during 2006-11. Health Aff. 34 (1), 48-55.

Jacobson G.A., Trilling A., Neuman T., Damico A., Gold M. 2016. Medicare Advantage hospital networks: How much do they vary? Kaiser Family Foundation:

Kaiser Family Foundation. 2015. Fact Sheet on Medicare Advantage. Kaiser Family Foundation:

Lochner K.A., Cox C.S. 2013. Prevalence of multiple chronic conditions among Medicare beneficiaries, United States, 2010. Preventing Chronic Disease. 10 E61.

Manning W.G., Basu A., Mullahy J. 2005. Generalized modeling approaches to risk adjustment of skewed outcomes data. J Health Econ. 24 (3), 465-488.

Manning W.G., Newhouse J.P., Duan N., Keeler E.B., Leibowitz A., Marquis M.S. 1987. Health insurance and the demand for medical care - evidence from a randomized experiment. Am. Econ. Rev. 77 (3), 251-277.

McCormack L., Squire C., Morton J., Lynch J., Mobley L., Salib P. 2005. Disenrollment from Medicare Advantage health plans: A qualitative assessment. RTI International: Research Triangle Park, NC.

McGuire T.G., Newhouse J.P., Sinaiko A.D. 2011. An economic history of Medicare Part C. Milbank Q. 89 (2), 289-332.

McGuire T.G., Newhouse J.P., Normand S.-L., Shi J., Zuvekase S. 2014. Assessing incentives for service-level selection in private health insurance exchanges. J Health Econ. 35 (1), 47-63.

McWilliams J.M., Hsu J., Newhouse J.P. 2012. New risk-adjustment system was associated with reduced favorable selection in Medicare Advantage. Health Aff. 31 (12), 2630-2640.

Medicare Payment Advisory Commission. 2012. Report to the Congress: Medicare and the Health Care Delivery System. Medicare Payment Advisory Commission: Washington, DC.

Medicare Payment Advisory Commission. 2014. Report to the Congress: Medicare and the Health Care Delivery System. Medicare Payment Advisory Commission: Washington, DC.

Morrisey M.A., Kilgore M.L., Becker D.J., Smith W., Delzell E. 2013. Favorable selection, risk adjustment, and the Medicare Advantage program. Health Serv Res. 48 (3), 1039-1056.

Newhouse J.P, McGuire T.G. 2014. How successful is Medicare Advantage? Milbank Q. 92 (2), 351-394.

Newhouse J.P., McWilliams J.M., Price M., Huang J., Fireman B., Hsu J. 2013. Do Medicare Advantage plans select enrollees in higher margin clinical categories? J Health Econ. 32 (6), 1278-1288.

Newhouse J.P., Price M., Huang J., McWilliams J.M., Hsu J. 2012. Steps to reduce favorable risk selection in Medicare Advantage largely succeeded, boding well for Health Insurance Exchanges. Health Aff. 31 (12), 2618-2628.

Newhouse J.P., Price M., McWilliams J.M., Hsu J., McGuire T.G. 2015. How much favorable selection is left in Medicare Advantage? Am. J. Health Econ. 1 (1), 1-26.

Pope G.C., Ellis R.P., Ash A.S., Liu C.-F., Ayanian J.Z., Bates D.W., Burstin H., Iezzoni L.I., Ingber M.J. 2000. Principal inpatient diagnostic cost group model for Medicare risk adjustment. Health Care Financ. Rev. 21 (3), 93-118.

Pope G.C., Kautter J., Ellis R.P., Ash A.S., Ayanian J.Z., Iezzoni L.I., Ingber M.J., Levy J.M., Robst J. 2004. Risk adjustment of medicare capitation payments using the CMS-HCC model. Health Care Financ. Rev. 25 (4), 119-141.

Rahman M., Laura K., Trivedi A.N., Mor V. 2015. High-cost patients had substantial rates of leaving Medicare Advantage and joining Traditional Medicare. Health Aff. 34 (10), 1675-1681.

Ringel J., Hosek S.D., Vollaard B.A., Mahnovski S. 2002. The elasticity of demand for health care : A review of the literature and its application to the Military Health System. RAND: Santa Monica, CA.

Trish E., Ginsburg P., Gascue L., Joyce G. 2017. Physician Reimbursement in Medicare Advantage Compared With Traditional Medicare and Commercial Health Insurance. JAMA Intern Med. 177 (9), 1287-1295.

Van de Ven W.P.M.M., Ellis R.P. 2000. Risk adjustment in competitive health plan markets. In: Culyer J.A., Newhouse P.J. (Eds), Handbook of Health Economics, vol. I. Elsevier: Amsterdam. 755-845.

Table 1. Baseline characteristics for the MCBS sample by the implementation periods of the CMS-HCC model

	Implementation periods of the CMS-HCC model (baseline year t equals)		
	Pre-implementation period (2001-2002)	Implementation period (2003-2005)	Post-implementation period (2006-2008)
Transition frequencies, N (percent)			
TM (year t) → TM (year $t+1$)	13348 (87.59)	12533 (82.87)	10670 (74.03)
TM (year t) → MA (year $t+1$)	40 (0.26)	259 (1.71)	388 (2.69)
MA (year t) → TM (year $t+1$)	282 (1.85)	128 (0.85)	153 (1.06)
MA (year t) → MA (year $t+1$)	1569 (10.3)	2203 (14.57)	3203 (22.22)
Total	15239 (100)	15123 (100)	14414 (100)
Total Medicare expenditures at year t, Mean (SD)			
TM (year t) → TM (year $t+1$)	7915 (20289)	9235 (26869)	9806 (27280)
TM (year t) → MA (year $t+1$)	2429 (4192)	5483 (10430)	7537 (14471)
MA (year t) → TM (year $t+1$)	4815 (10677)	5208 (9872)	8257 (17283)
MA (year t) → MA (year $t+1$)	4118 (9285)	3603 (10131)	3549 (9823)
Weighted average for all beneficiaries	7453	8316	8338
Risk scores at year t, Mean (SD)			
TM (year t) → TM (year $t+1$)	1.00 (0.32)	1.00 (0.32)	1.00 (0.41)
TM (year t) → MA (year $t+1$)	0.88 (0.35)	0.96 (0.31)	0.96 (0.36)
MA (year t) → TM (year $t+1$)	1.04 (0.30)	1.19 (0.51)	1.23 (1.10)
MA (year t) → MA (year $t+1$)	1.04 (0.33)	1.05 (0.34)	1.08 (0.75)
Weighted average for all beneficiaries	1.00	1.01	1.02

Notes: Medicare enrollees were classified as TM enrollees if enrolled in TM plans for all 12 months of the calendar year, and classified as MA enrollees if enrolled in an MA plan for at least six months of the year and enrolled in any Medicare plan in every month of the year. Total Medicare expenditures for TM enrollees were estimated by summing any Part A and Part B expenditures reported in claims data, and total Medicare expenditures for MA enrollees were estimated by summing any Part A and Part B expenditures reported in claims data (if enrolled in TM) and the self-reported MA expenditures from the survey. All expenditures were adjusted to 2009 dollars using the CPI-U. Risk scores for TM enrollees were estimated from Medicare claims and risk scores for MA enrollees were estimated by dividing the reported capitation payments by county-level benchmark rates. The way of estimating the risk scores for TM enrollees varied by the implementation periods, and thus risk scores cannot be directly comparable across the three periods. Sample weights provided by the MCBS were used.

Table 2. Changes in risk selection patterns after adopting the CMS-HCC model

	Dependent variable: risk score at *year t* or total Medicare expenditure at *year t*							
	(1) Risk score	(2) Risk score	(3) Risk score	(4) Risk score	(5) Risk score	(6) Risk score	(7) Expenditure	(8) Expenditure
Panel A (Those enrolled in TM at year t)								
Share of year in MA	-0.14*	-0.20***	-0.20*	-0.14*	-0.20***	-0.20*	-4210.92***	-4401.86***
	(0.07)	(0.04)	(0.09)	(0.07)	(0.04)	(0.09)	(1086.78)	(1029.22)
Share of year in MA × after 2002	0.10	0.18***	0.18				7.44	
	(0.07)	(0.04)	(0.09)				(1428.38)	
Share of year in MA × after 2005				0.09	0.18***	0.17		2597.92
				(0.07)	(0.04)	(0.09)		(1339.25)
Risk score							17221.18***	15882.29***
							(1078.3)	(759.56)
Mean outcome variable	1.00	0.94	1.00	1.00	0.93	1.00	8537.66	8787.63
Estimated method	OLS	OLS	Median quantile	OLS	OLS	Median quantile	OLS	OLS
Evaluation period	After 2003	After 2003	After 2003	After 2006	After 2006	After 2006	After 2003	After 2006
Outliers trimmed	No	Yes	No	No	Yes	No	No	No
Observations	26180	23973	26180	24446	22380	24446	26180	24446

	(1) Risk score	(2) Risk score	(3) Risk score	(4) Risk score	(5) Risk score	(6) Risk score	(7) Expenditure	(8) Expenditure
Panel B (Those enrolled in MA at year t)								
Share of year in MA	-0.01	-0.01	-0.03	-0.01	-0.01	-0.03	-926.49	-917.98
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(788.10)	(787.54)
Share of year in MA × after 2002	-0.14**	-0.04	-0.08				-773.70	
	(0.05)	(0.04)	(0.05)				(1282.58)	
Share of year in MA × after 2005				-0.21	0.01	-0.05		-4655.91**
				(0.13)	(0.06)	(0.11)		(1733.50)
Risk score							174.09	1554.85***
							(371.69)	(368.12)
Mean outcome variable	1.04	1.01	1.05	1.07	0.98	1.07	3917.72	3901.54

Estimated method	OLS	OLS	Median quantile	OLS	OLS	Median quantile	OLS	OLS
Evaluation period	After 2003	After 2003	After 2003	After 2006	After 2006	After 2006	After 2003	After 2006
Outliers trimmed	No	Yes	No	No	Yes	No	No	No
Observations	4181	3969	4181	5207	4945	5207	4178	5206

otes: Total Medicare expenditures for TM enrollees were estimated by summing any Part A and Part B expenditures reported in claims data, and total Medicare expenditures for MA enrollees re estimated by summing any Part A and Part B expenditures reported in claims data (if enrolled in TM) and and the self-reported MA expenditures from the survey. All expenditures were justed to 2009 dollars using the CPI-U. Risk scores for TM enrollees were estimated from Medicare claims and risk scores for MA enrollees were estimated by dividing the reported capitation yments by county-level benchmark rates. "Outliers trimmed" means exclusion of individuals with risk scores above the 95th percentile in each year. Year fixed effects were included in all gressions. Sample weights provided by the MCBS were used. Standard errors, in parentheses, were clustered by the individual.

$p < 0.05$.

$p < 0.01$

$p < 0.001$**Error! Reference source not found.**

Table 3. Type of services from the PBP data

Type of service
Inpatient hospital—acute
Inpatient hospital—psychiatric
Skilled nursing facility
Comprehensive outpatient rehabilitation
Emergency care
Urgent care
Partial hospitalization
Home health service
Primary care physician service
Chiropractic service
Occupational therapy service
Physician specialist service
Mental health specialty service—individual session
Mental health specialty service—group session
Podiatrist service
Other health care professional service
Psychiatric service—individual session
Psychiatric service—group session
Physical therapy and speech/language pathology service
Diagnostic service
Radiation therapy service
Outpatient X-Ray
Outpatient hospital service
Ambulatory Surgical Center (ASC) service
Outpatient substance abuse service—individual session
Outpatient substance abuse service—group session
Cardiac rehabilitation service
Ambulance
Durable medical equipment
Medical supply
Prosthetic
Diabetes monitoring supply
Drug prescription

Notes: MA plans can set up varying copayments by a length of stay or number of visit, for example, inpatient hospital, skilled nursing facility, mental health specialty service, psychiatric service, and outpatient substance abuse service. For these services, a copayment was estimated on a basis of a typical length of stay or number of visit, according to Government Accountability Office (2010). Then, we calculated a copayment per day.

Table 4. Type of services used to examine service-level selection in the MCBS sample

Type of service	Service-level selection index	TM enrollees	MA enrollees
Hospice	2.578	Yes	Yes
Home health	0.875	Yes	Yes
Durable medical equipment	0.703	Yes	No
Hospital inpatient visit	0.592	Yes	Yes
Other	0.495	Yes	Yes
Hospital visit	0.356	Yes	No
Home visit	0.348	Yes	No
ER visit	0.265	Yes	No
Consultation	0.219	Yes	No
Other facility services	0.172	Yes	Yes
Hospital outpatient visit	0.170	Yes	Yes
Advanced imaging—CAT	0.169	Yes	No
Oncology	0.159	Yes	No
Lab test	0.144	Yes	No
Other tests	0.134	Yes	No
Standard imaging	0.119	Yes	No
Specialist	0.114	Yes	Yes
Echography	0.113	Yes	No
Ambulatory procedure	0.105	Yes	No
Imaging procedure	0.102	Yes	No
Office visit	0.096	Yes	No
Major procedure—cardiovascular	0.096	Yes	No
Minor procedure	0.095	Yes	No
Anesthesia	0.092	Yes	No
Endoscopy	0.087	Yes	No
Major procedure	0.087	Yes	No
Major procedure—orthopedic	0.083	Yes	No
Advanced imaging—MRI	0.083	Yes	No
Eye procedure	0.045	Yes	No

Notes: Services covered under Medicare Parts A and B were classified into 29 services. Specifically, Part A claims were classified into the following five types of service (hospital inpatient visit, hospital outpatient visit, home health, hospice, and other facility services). Part B claims were classified into 24 categories by the Berenson-Eggers Type of Services (BETOS) codes. Service-level selection index estimated from Ellis and McGuire (2007) was used and type of service was presented by the order of the service-level selection index. Service-level use for TM beneficiaries was estimated from MCBS claims data. Service-level use for MA enrollees was estimated from self-reported data. Prescribed medicine was categorized as "Other" and medical provider was categorized as "Specialist".

Table 5. Summary statistics on MA plans by the implementation periods of the CMS-HCC model

	Implementation periods of the CMS-HCC model		
Types of MA plans, Weighted Mean (SD)	Pre-implementation period (2001-2003)	Implementation period (2004-2006)	Post-implementation period (2007-2009)
Health Maintenance Organization (HMO)	475 (73)	1,136 (502)	2,026 (175)
Preferred Provider Organization (PPO)	6 (5)	344 (311)	817 (178)
Total	481 (72)	1,480 (813)	2,843 (348)

Notes: Plans with complete information on cost-sharing for all services covered under Medicare Parts A and B were included. Other types of MA plans such as Medicare Savings Account, and Private Fee-For-Service plans were excluded from our analysis. To account for varying numbers of MA plans across years, we adjust by weighting the number of MA plans in each year.

Table 6. Relation of MA disenrollment and satisfaction after adopting the CMS-HCC model

	Dependent variable: satisfaction rating at *year t*				
	(1) Out-of-pocket costs	(2) Quality of care	(3) Access to specialist	(4) Ease of access to care from residence	(5) Care provided in the same location
Panel B (Those enrolled in MA at year t)					
Share of year in MA	-0.018	0.019	0.015	0.005	-0.018
	(0.046)	(0.039)	(0.038)	(0.053)	(0.038)
Share of year in MA	0.193*	0.084	0.017	0.035	0.088
× after 2005	(0.087)	(0.067)	(0.065)	(0.092)	(0.078)
Risk score	-0.032*	-0.035**	-0.011	-0.018	-0.026*
	(0.015)	(0.012)	(0.012)	(0.016)	(0.011)
Mean outcome variable	2.99	3.29	3.17	3.11	3.12
Observations	4928	4746	4522	2445	4373
	(6) Availability of care nights and weekends	(7) Follow-up care	(8) Questions answered over phone	(9) Doctor's concern for your health	(10) Information about your medical condition
Panel B (Those enrolled in MA at year t)					
Share of year in MA	0.042	-0.031	0.034	0.000	0.065
	(0.036)	(0.038)	(0.049)	(0.039)	(0.035)
Share of year in MA	0.044	0.055	0.011	-0.017	-0.023
× after 2005	(0.063)	(0.065)	(0.097)	(0.062)	(0.071)
Risk score	-0.040***	-0.009	-0.012	-0.022*	-0.02
	(0.012)	(0.011)	(0.017)	(0.011)	(0.011)
Mean outcome variable	3.19	3.18	3.04	3.16	3.13
Observations	4973	4326	3587	4837	4896

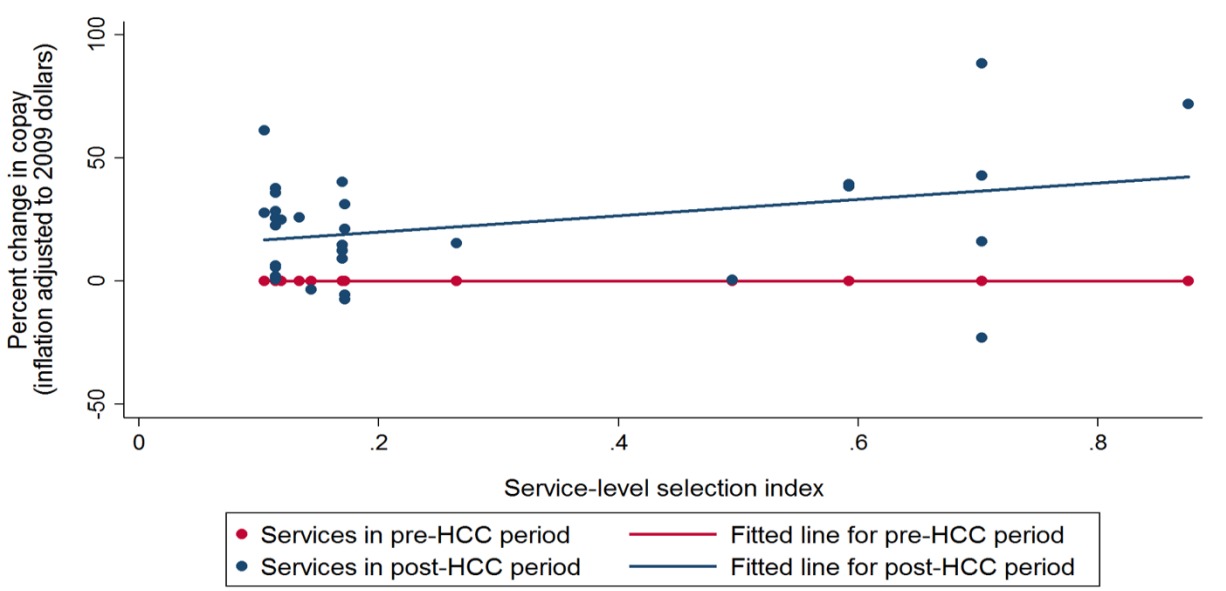
Notes: Each satisfaction measure took values from one to four ("very dissatisfied", "dissatisfied", "satisfied", "very satisfied"). Self-reported health, age, race, and female were adjusted. Risk scores for MA enrollees were estimated by dividing the reported capitation payments by county-level benchmark rates. Year fixed effects were included in all regressions. Sample weights provided by the MCBS were used. Standard errors, in parentheses, were clustered by the individual.

* $p < 0.05$.

** $p < 0.01$

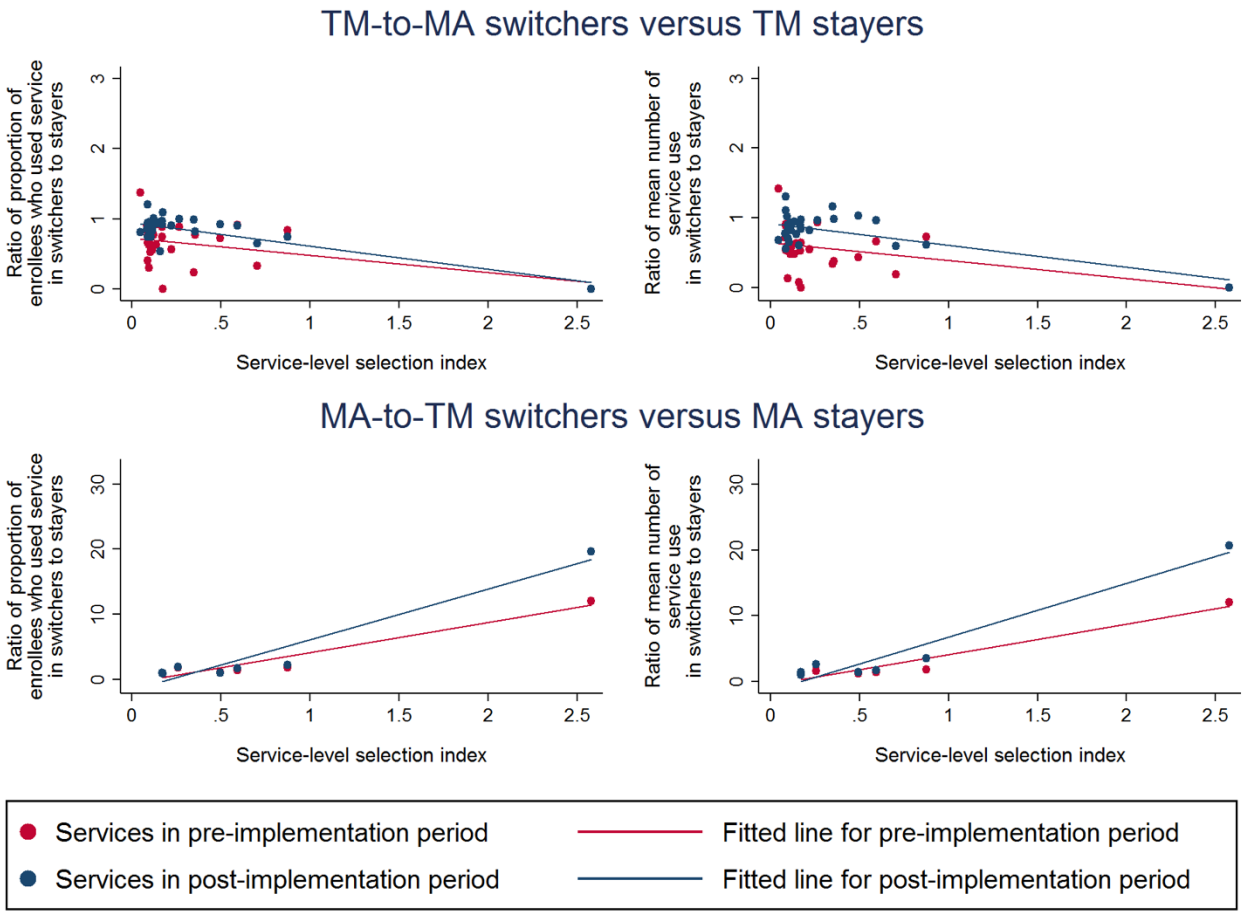
*** $p < 0.00$

Figure 1. Disproportionate percent changes in service-specific copayments between the pre-and post-implementation periods of the CMS-HCC model



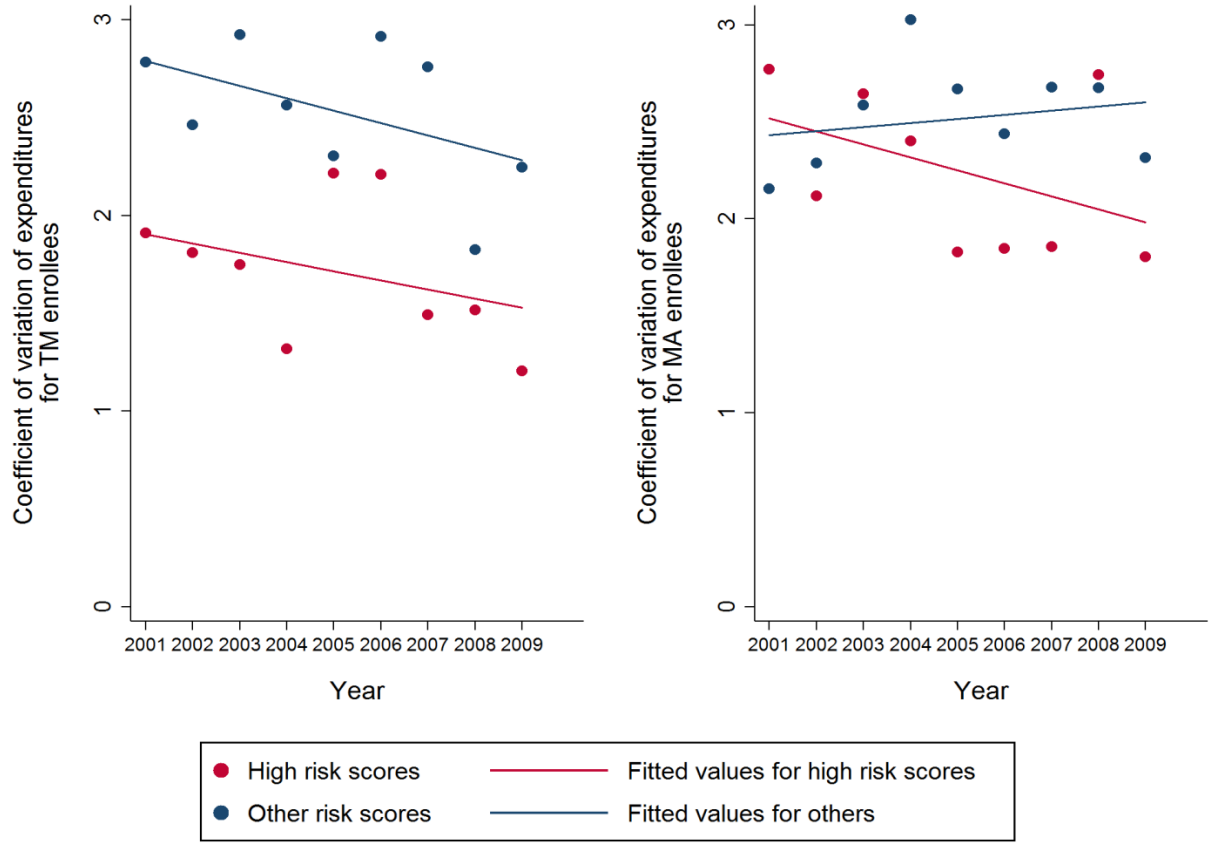
Notes: Service-level selection index estimated from Ellis and McGuire (2007) was used.

Figure 2. Changes in service-specific use of switchers to stayers between the pre- and post-implementation periods of the CMS-HCC model



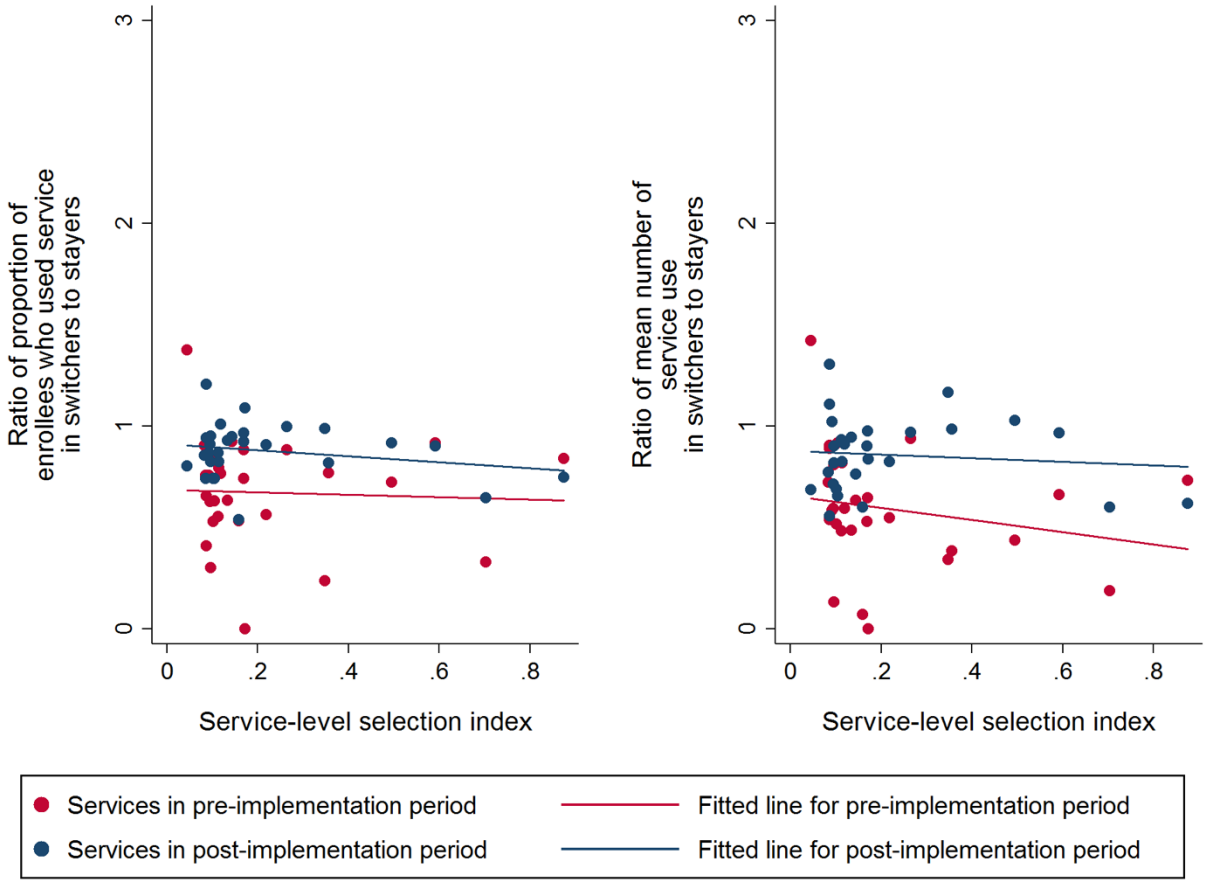
Notes: Service-level selection index estimated from Ellis and McGuire (2007) was used. The left side figures estimate the ratio of the proportion of switchers with use of a particular type of service to stayers with use of the service in the pre- and post-implementation periods, respectively ("intensive-margin selection"). The left side figures estimate the ratio of average number of services per enrollee of switchers to stayers in the pre- and post-implementation periods, respectively ("extensive-margin selection").

Figure 3. Coefficient of variance of total Medicare Expenditures for TM enrollees and MA enrollees by risk scores



Notes: The coefficient of variance was estimated as the ratio of the standard deviation to the mean. High-risk scores indicate above the 90th percentile of the risk score distribution.

Appendix Figure A1. Changes in service-specific use of switchers to stayers between the pre- and post-implementation periods of the CMS-HCC model (without hospice with the highest service-level selection index)



Notes: Service-level selection index estimated from Ellis and McGuire (2007) was used.

Appendix Table A1. Changes in risk selection patterns in the post-implementation periods of the CMS-HCC model after adjusting for demographic variables and self-reported health status

	Dependent variable: total Medicare expenditure at *year t*	
	(1)	(2)
Panel B (Those enrolled in MA at year t)		
Share of year in MA	-600.19	-700.2
	(778.57)	(775.97)
Share of year in MA × after 2002	-730.72	
	(1306.36)	
Share of year in MA × after 2005		-4685.38**
		(1774.63)
Risk score	622.29	1477.42***
	(608.73)	(410.86)
Female	-297.98	196.81
	(348.26)	(294.06)
Hispanic	-386.02	-139.86
	(888.28)	(838.51)
Asian	-102.97	-38.59
	(1201.58)	(1879.87)
Black	220.15	53.83
	(513.04)	(542.55)
Others	-1850.32**	-1135.86
	(622.5)	(1067.64)
Health status at *year t* (reference: excellent)		
Very good	-5881.94**	-2420.36
	(2073.82)	(1819.89)
Good	-8116.23***	-5971.95***
	(2046.56)	(1760.42)
Fair	-5249.31*	-3911.26*
	(2242.56)	(1804.43)
Poor	-5713.38*	-1820.31
	(2462.32)	(2384.69)
Mean outcome variable	3901.34	3900.29
Estimated method	OLS	OLS
Evaluation period	After 2003	After 2006
Observations	4071	5083

Notes: Total Medicare expenditures for MA enrollees were estimated by summing any Part A and Part B expenditures reported in claims data (if enrolled in TM) and the self-reported MA expenditures from the survey. All expenditures were adjusted to 2009 dollars using the CPI-U. Risk scores for MA enrollees were estimated by dividing the reported capitation payments by county-level benchmark rates. Year fixed effects were included in all regressions. Sample weights provided by the MCBS were used. Standard errors, in parentheses, were clustered by the individual.

* $p < 0.05$.

** $p < 0.01$

*** $p < 0.001$

Chapter 3.

ALTERNATIVE EVALUATION METRICS FOR RISK ADJUSTMENT

Abstract

Risk adjustment is instituted to counter risk selection by accurately equating payments with expected expenditures. Traditional risk-adjustment methods are designed to estimate accurate payments at the group level. However, this generates residual risks at the individual level, especially for high-expenditure individuals, thereby inducing health plans to avoid those with high residual risks. To identify an optimal risk-adjustment method, we perform a comprehensive comparison of prediction accuracies at the group level, at the tail distributions, and at the individual level across 19 estimators: nine parametric regression, seven machine learning, and three distributional estimators. Using the 2013-2014 MarketScan database, we find that no one estimator performs best in all prediction accuracies. Generally, machine learning and distribution-based estimators achieve higher group-level prediction accuracy than parametric regression estimators. However, parametric regression estimators show higher tail distribution prediction accuracy and individual-level prediction accuracy, especially at the tails of the distribution. This suggests that there is a trade-off in selecting an appropriate risk-adjustment method between estimating accurate payments at the group level and lower residual risks at the individual level. Our results indicate that an optimal method cannot be determined solely based on statistical metrics but rather needs to account for simulating plans' risk selective behaviors.

3.1 INTRODUCTION

Risk selection is a common phenomenon in the health insurance markets, where a health plan has a financial interest in encouraging low-risk individuals to enroll in the plan and discouraging high-risk individuals from enrolling (Van de Ven and Ellis, 2000). From the plan's perspective, the rationale for risk selection is based on the asymmetry of information between two parties, in which health plans do not know individuals' private information about health status and preferences for health care, and thus they cannot price their risks appropriately. Even with symmetric information, health plans can have incentives for risk selection if they are not allowed to use the private information to set premiums or benefit features.

Risk selection has been consistently observed in major public health insurance programs in the United States (US) (Riley *et al.*, 2009; McWilliams *et al.*, 2012; Newhouse *et al.*, 2012; Morrisey *et al.*, 2013; Newhouse *et al.*, 2013; Brown *et al.*, 2014; Newhouse *et al.*, 2015). For example, the Centers for Medicare and Medicaid Services (CMS) contract with private health plans, known as Medicare Advantage (MA) plans, to provide Medicare Parts A and B benefits to the elderly. CMS pay MA plans a capitated (per enrollee) amount to offer all Parts A and B benefits. Due to the rate restrictions imposed by CMS, however, this capitated payment structure creates incentives for MA plans to selectively disenroll the high-cost enrollees in order to receive overpayments from CMS. Brown *et al.* (2014) estimated that overpayments to MA plans were $30 billion in 2006, which is equivalent to eight percent of total Medicare expenditures that year. Strong incentives to select are likely to persist in state-specific Health Insurance Exchanges under the Affordable Care Act (ACA) (Weiner *et al.*, 2012; McGuire *et al.*, 2014; Montz *et al.*, 2016). Under the ACA, health plans can no longer deny coverage based on pre-existing health status. Also, they cannot differentiate premiums across individuals by any factor other than age, tobacco

use, family size, and geography. These restrictions can induce incentives for the plans to engage in risk selection through second-degree price discrimination. For example, to avoid people with multiple chronic conditions, health plans may make their benefit designs unattractive to them.

Risk adjustment plays a critical role in countering risk selection (Van de Ven and Ellis, 2000). The goal of risk adjustment is to adjust payments to health plans to accurately reflect the health status of enrollees. Due to differences in health status and health care needs, health care expenditures can vary across individuals. Risk adjustment is used to predict health care expenditures to correctly align plan payments with an individual's expected health care expenditure. With risk adjustment, health plans receive higher payments for sicker people and receive lower payments for healthier people. However, its ultimate goal is not accuracy per se, but rather improved incentives (Glazer and McGuire, 2000; Van de Ven and Ellis, 2000; Einav *et al.*, 2016). In other words, risk adjustment is implemented to disincentivize health plans from selectively enrolling and caring for healthy people, and furthermore incentivize the plans to compete based on providing high-value care. The extent to which risk selection can be reduced, therefore, depends on how well risk adjustment estimates the risk-adjusted predictions as close to the actual expenditures.

In this study, we focus on a traditional risk-adjustment methodology used in Medicare and the Exchanges. Since 2004, CMS have used the CMS-Hierarchical Condition Categories (HCC) model to adjust payments for each MA enrollee's expected expenditure (Pope *et al.*, 2004). Following the CMS-HCC model, the Department of Health and Human Services (HHS) developed a federally-certified risk-adjustment methodology to be used by states or by HHS on behalf of states, known as the HHS-HCC model. Because eligible populations for the Exchanges are fundamentally different from the Medicare population, the CMS-HCC model had to be adapted

into the HHS-HCC model.[31] The process of risk-adjusting payments based on demographics and diagnoses is as follows. First, the HCC model is used to generate each enrollee's risk score based on demographic and diagnostic information. Using the risk score, linear regression (i.e., weighted least squares) is performed to estimate the conditional mean of expenditures, $E(y_i|r_i)$, where y_i indicates an individual i's actual expenditure and r indicates her risk score. As the model is designed to be accurate at the group level, its prediction accuracy is evaluated based on group-level performance metrics. Health plans are reimbursed based on the estimated $E(y_i|r_i)$.

While these traditional risk-adjustment models have considerably reduced risk selection, especially in the MA program (McWilliams *et al.*, 2012; Newhouse *et al.*, 2012; Morrisey *et al.*, 2013; Newhouse *et al.*, 2015), they cannot fully eliminate the incentives for health plans to engage in risk selection. Because there is considerable variability in actual expenditures of enrollees around their risk-adjusted mean predictions (Brown *et al.*, 2014), health plans would receive lower payments for those with higher expenditures than their risk-adjusted payments (i.e., residual risks). As the variability of the within-risk-score expenditures is larger for those with higher expenditures, this effect would be more pronounced for those with higher expenditures. These would incentivize health plans to avoid those with residual risks, especially those with high expenditures. There is empirical evidence showing that health plans strategically behave in response to such nature of the HCC model. For instance, Morrisey *et al.* (2013) showed that MA enrollees in the highest expenditure group were more likely to leave MA plans than those in other expenditure groups. Park *et al.* (2017) explain this phenomenon in relation to service-level selection, in which health

[31] There are several differences between the CMS-HCC and HHS-HCC models. First, the CMS-HCC model uses prior year's demographic and diagnostic information to predict next year's health care expenditures (i.e. prospective model), whereas the HHS-HCC model uses current year's information to predict the same year's health care expenditures (i.e., concurrent model). Second, separate CMS-HCC models are designed for the aged (those aged 65 years old and over) and the disabled (those aged 18 to 64 years), whereas the HHS-HCC model is only designed for those aged 0 to 64 years in the individual and small group markets. Finally, the CMS-HCC model is designed to predict non-drug medical expenditures, whereas the HHS-HCC model is designed for the sum of medical and drug expenditures.

plans selectively design their mix of health care services, as those who incur higher expenditures than their risk-adjusted payments use a different mix of health care services compared to those who incur expenditures just as much their risk-adjusted payments. They found that MA plans increased copays disproportionately more for services that appeal to them than other services, thereby inducing some of them to voluntarily disenroll from MA plans. Similarly, Rahman *et al.* (2015) found that those who used high-cost services (i.e., nursing home care and home health care) were more likely to disenroll from MA plans than those who did not use.

This suggests that solely relying on group-level performance metrics be deficient in reducing the plan's incentives for risk selection. Traditional risk-adjustment models are typically evaluated based on group-level prediction metrics such as predictive ratio, which measures the accuracy of these models in predicting the average expenditure of a population, and R^2, which measures the extent to which the models can explain individual variations in expenditures (Pope *et al.*, 2011; Kautter *et al.*, 2014). In addition, other metrics such as mean prediction error (MPE), mean absolute prediction error (MAPE), and root mean square error (RMSE) have been used to evaluate group-level prediction accuracy. Because these metrics measure prediction accuracy at the aggregate level, they cannot capture true individual-level prediction inaccuracy, thereby generating the plans' incentives to avoid those with residual risks. However, this does not necessarily suggest that group-level performance accuracy should not be considered in evaluating performance of a risk-adjustment model. As health insurance is designed to pool the financial risk of a high-cost medical event across a large group of people, the traditional risk-adjustment models are intentionally designed to be accurate at the group level (Centers for Medicaid and Medicare Services, 2016).

To assess prediction accuracy across the full distribution of health care expenditures beyond evaluating prediction accuracy at the group level, Jones *et al.* (2015) adopted performance based on predicting tail probabilities, $P(\hat{Y}_i > k)$, where $\hat{Y}_i$ indicates a forecast of an individual i's expenditure[32] and k indicates an expenditure threshold. This is motivated by the fact that the traditional risk-adjustment models systematically underpredict expenditures for those with the most severe health status (Pope *et al.*, 2004; Medicare Payment Advisory Commission, 2012; Kautter *et al.*, 2014; Medicare Payment Advisory Commission, 2014) because the actual distribution of health care expenditures is heavily right skewed and long tailed. This indicates that risk selection occurs most prevalently at the tail distributions. Thus, predicting the tail probabilities accurately might reduce the incentives for health plans to avoid those with residual risks, at least to some extent. However, it might not fully eliminate the incentives, because the extent to which risk selection can be reduced is inherently not tied to how close the risk-adjusted predictions align with the actual expenditures at the tail distributions. Rather it relies on how close the risk-adjusted predictions align with the actual expenditures across all individuals.

These indicate that it is necessary to develop a complementary performance metric to the existing performance metrics, which could generate economic incentives for health plans not to avoid those with residual risks. In Figure 1 we provide a hypothetical example for illustrating how a new metric could complement the group-level performance metrics. Each dot indicates risk-adjusted mean expenditures for each risk score group. Each bar indicates the range of actual expenditures of those in each risk score group. Suppose that Estimators 1 and 2 generate the same level of MPE. However, compared to Estimator 2, Estimator 1 generates lower residual risks for intermediate risk score groups (i.e., those with risk scores of 1 and 2), but generates higher residual

[32] $\hat{Y}_i$ is distinct from the conditional mean of expenditure. It is estimated from the conditional probability density function as a stochastic quantity. A more detailed explanation is given in the Methodology and Data section.

risks for high-risk score groups (i.e., those with risk scores of 3 and 4). This means that Estimator 2 might be better suited for risk adjustment purposes of plan payment, because while achieving a similar level of group-level prediction accuracy, the estimator also produces the risk-adjusted predictions as close to the actual expenditures at the individual level, thereby reducing the scope of generating those with high residual risks. As the impact is likely to be more pronounced for those with high expenditures, assessing prediction accuracy at the tails of the expenditure distribution is also needed in selecting an appropriate and accurate risk-adjustment method. This suggests that an optimal risk-adjustment method should achieve high prediction accuracies at the group level, at the tail distributions, and at the individual level. To identify an optimal risk-adjustment method, we perform a comprehensive comparison of prediction accuracies across 19 estimators: nine parametric regression estimators, seven machine learning estimators, and three distribution-based estimators (Table 1). We compare these estimators using the existing metrics for measuring prediction accuracies at the group level and at the tail distributions. To measure prediction performance not captured by those metrics, we develop a new performance metric for measuring individual-level prediction accuracy and compare individual-level prediction accuracy of the 19 estimators.

Using the 2013-2014 MarketScan database (N=12,882,983), we find that no one estimator perform best in all prediction metrics. In other words, estimators with higher prediction accuracy at the group level do not necessarily achieve higher prediction accuracy at the individual level and vice versa. Specifically, machine learning and distribution-based estimators achieve higher group-level prediction accuracy than parametric regression estimators. Compared to machine learning and distribution-based estimators, however, parametric regression estimators show higher tail

distribution prediction accuracy and individual-level prediction accuracy, especially at the tail distributions.

We begin with describing our proposed performance metric for measuring individual-level prediction accuracy. We then describe the data, the study design, the estimators compared, and the methods of computing the performance metrics. Next, we discuss our results and conclude with a discussion.

3.2 METHODOLOGY AND DATA

3.2.1 *Primary Evaluation Metrics*

We propose two individual-level prediction performance metrics for measuring 1) overall individual-level prediction accuracy and 2) tail-specific individual-level prediction accuracy, respectively. First, the overall individual-level prediction performance metric measures what fraction of a population has underpredicted or overpredicted expenditures by greater than l percent of their actual expenditures:

$$P\left(\frac{\left|y_i - \hat{Y}_i\right|}{y_i} \times 100 > l\right) \quad (1)$$

where y_i indicates an individual i's actual expenditure and $\hat{Y}_i$ indicates a forecast of her expenditure [not her predicted conditional mean ($E(y_i|x_i) = \widehat{y_i}$)], which is estimated from the conditional probability density function as a stochastic quantity.[33] Note that plan payment is estimated based on $\widehat{y_i}$, not $\hat{Y}_i$. To distinguish $\hat{Y}_i$ from $\widehat{y_i}$, we refer to $\hat{Y}_i$ as a "forecast" of her expenditure, and $\widehat{y_i}$ as a "prediction" of her expenditure. When we refer to evaluating their

[33] Note that these quantities were also used in Jones *et al.* (2015) to estimate tail probabilities $P(\hat{Y}_i > k)/P(y_i > k)$.

performance, however, we use the terminology "prediction", for example, prediction accuracy or prediction performance.

The proposed metric reflects a different aspect of prediction performance, which is not captured in the existing prediction performance metrics. This metric quantifies the degree of misprediction to which a deviation of an individual's forecasted expenditure from her actual expenditure (i.e., forecast error) is greater than a threshold that is proportionally set as a certain percent of her actual expenditure. However, it does not measure how likely each individual is to have such forecast error. It measures what proportion of the population has the forecast error larger than the allowed threshold. For example, suppose $P\left(\frac{|y_i-\hat{Y}_i|}{y_i} \times 100 > 50\right) = 0.25$. This means that 25 percent of the population has forecast error greater than 50 percent of their actual expenditures. Thus, the metric measures how well a risk-adjustment model estimates the risk-adjusted (forecasted) expenditures as close the actual expenditures.

Equation (1) can be expressed in terms on the conditional cumulative distribution functions of $\hat{Y}_i$ as:

$$= P\left(\left(y_i - \hat{Y}_i\right) \geq \frac{ly_i}{100}\right) \times P\left(\left(y_i - \hat{Y}_i\right) \geq 0\right) + P\left(\left(y_i - \hat{Y}_i\right) < -\frac{ly_i}{100}\right) \times P\left(\left(y_i - \hat{Y}_i\right) < 0\right) \quad (2)$$

$$= P\left(\hat{Y}_i \leq y_i - \frac{ly_i}{100}\right) \times P\left(y_i \geq \hat{Y}_i\right) + P\left(\hat{Y}_i > y_i + \frac{ly_i}{100}\right) \times P\left(\hat{Y}_i > y_i\right) \quad (3)$$

$$= F_{\hat{Y}_i}\left(y_i - \frac{ly_i}{100}\right) \times F_{\hat{Y}_i}(y_i) + \left(1 - F_{\hat{Y}_i}\left(y_i + \frac{ly_i}{100}\right)\right) \times \left(1 - F_{\hat{Y}_i}(y_i)\right) \quad (4)$$

As traditional risk-adjustment models systematically underpredict expenditures for those with very high expenditures (Pope *et al.*, 2004; Medicare Payment Advisory Commission, 2012; Kautter *et al.*, 2014; Medicare Payment Advisory Commission, 2014), we also assess the tail-specific individual-level prediction accuracy:

$$P\left(\frac{|y_i - \hat{Y}_i|}{y_i} \times 100 > l | y_i > k\right) \quad (5)$$

where $y_i > k$ indicates a population in a tail distribution, whose actual expenditures exceed an expenditure threshold k.

These metrics are compared with the existing prediction performance metrics. To assess group-level prediction accuracy, we use MPE, MAPE, RMSE, R^2, and predictive ratio. To evaluate prediction accuracy at the tail distributions, we estimate a ratio of the estimated $P(\hat{Y} > k)$ to the actual proportion of expenditures in a sample observed to exceed a certain expenditure threshold k, developed by Jones *et al.* (2015). If there is a single estimator that dominates the others for all metrics, then the estimator would be an optimal risk-adjustment method. This is not only because it enables to estimate accurate plan payments at the group level but also because it diminishes the likelihood of having the individual-level residual risk, thereby reducing the incentives for health plans to avoid those with high residuals risk.

3.2.2 *Data*

We use the Truven MarketScan Commercial Claims and Encounter database between 2013-2014. We use this database because it was used to develop and validate the HCC model for the individual and small group markets under the ACA (Kautter *et al.*, 2014). The database contains inpatient, outpatient, and prescription drug claims for employees, retirees, and their dependents of over 250 medium and large employers and health plans. As enrollment and claims data are linked to detailed information on diagnosis and procedure codes across sites and over time, it allows us to identify all medical conditions diagnosed during a year. Using the database, we identify working-age adults (aged 21 to 64) who were continuously enrolled in the same type of plan [Comprehensive, Health

Maintenance Organization (HMO), Preferred Provider Organization (PPO), Point of Service (POS), or POS with capitation[34]] during 2013-2014.[35]

The outcome variable is total annual health care expenditures in 2014, including expenditures for inpatient, outpatient, and prescriptions drug services. Expenditures represent the actual paid amounts, and thus include patient payments (deductible, coinsurance, and copay), payments made by the patient's insurance plan, and any payments by other insurance providers.

To predict the second year total health care expenditures, we use the risk-adjusters of the HHS-HCC model developed for the adult population of the 2014 benefit year (Centers for Medicaid and Medicare Services, 2014). Although we use the HHS-HCC model, our goal is not to evaluate the prediction performance of the HHS-HCC model. The goal of this paper is to assess prediction performance of the 19 estimators to identify the best predictive estimator for health care expenditures when only accounted for the prior year's health status while adjusting for other factors. Thus, there are two differences between the original HHS-HCC model and the HCC model used in this study. First, the HHS-HCC model uses current year information to predict current year expenditures,[36] whereas we treat the HHS-HCC model as a prospective model, which means we use each enrollee's prior year health status to predict next year expenditures. However, most risk-adjustment models used for payment purposes are prospective, for example, MA, Medicare Part D, and the Netherlands' risk-adjustment models. Thus, evaluating prediction performance in the

[34] Because the MarketScan database does not distinguish between partially and fully capitated POS plans, individuals in capitated POS plans are traditionally excluded from analysis. Thus, we performed a sensitivity analysis after excluding those in capitated POS plans (N=80,825). However, we found consistent results with our original findings.
[35] We limited the study population to individuals who were enrolled for the two consecutive years. Since each individual's actual expenditure is used to set a relative range of allowable deviation, knowing her actual expenditure exactly is critical to accurately assess prediction performance. However, it is worthwhile noting that the HCC models account for partial year enrollees. Thus, practical implications from our study might be limited.
[36] The HHS-HCC model adopts a concurrent approach, as no previous year information on health status existed for the first year in which the model was phased-in and eligible enrollees can move in and out of enrollment between markets and move across insurers

prospective way would provide more general implications for risk adjustment of plan payment. Thus, we include 18 age-sex categories, 114 HCCs, and 16 interactions between pairs of disease groups from the prior year (2013).[37] Also, the HHS-HCC model does not account for plan type and state fixed effects, whereas we include those effects to control for different population characteristics across types of plan and states.[38] For more details on the risk-adjusters, see Appendix Table A.

3.2.3 *Quasi Monte Carlo Design*

We employ a quasi Monte Carlo design where two 1 percent random samples are selected from the full study population. Each of these samples spans data from 2013 to 2014. One of them represents the estimation set, while the other represents the validation set. Each estimator is fitted to the estimation set where the outcome is the second year total health care expenditures and the covariates are the prior year's health status. In the estimation model, a total of 148 variables (18 age-sex categories, 114 HCCs, and 16 interactions between disease groups) are used as covariates. In addition, two variables (five types of plan and 53 geographical categories) are included as fixed effects. Prediction performance for each estimator is assessed on the validation set. Using the fitted model, we predict the second year health care expenditures for those in the validation set based on their prior year's health status. This process is repeated 100 times (sampling with replacement) and the average for each performance metric is calculated over these 100 iterations.

---

[37] The HCC model is developed to obtain a clinically meaningful and statistically stable system. Specifically, tens of thousands of the *International Classification of Diseases* (ICD)-9 codes are grouped into a small number of organized categories to generate a diagnostic profile of each person. Thus, each condition category includes conditions that are clinically related and have similar cost implications.

[38] The inclusion of plan type fixed effects is to control for differences in benefit structures and care management between the types of plans, which is likely to affect access to care and use of care (e.g., coverage level and unit price). Similarly, the inclusion of state fixed effects is to control for differences in local supply side factors between states, potentially affecting access to care and use of care (e.g., number of hospitals and facilities).

The quasi Monte Carlo design enables valid comparison of prediction performance across estimators (Jones *et al.*, 2015). As the design uses an out-of-sample prediction technique, it allows us to assess prediction performance of the estimators in a rigorous and consistent manner while ensuring that results are not driven either by overfitting or traditional Monte Carlo assumptions. The benefit is maximized for big data analytics, because it reduces the number of computation steps needed to obtain a given accuracy in a traditional Monte Carlo integration. Moreover, as all estimators are fitted to the same set of covariates for the same observations, findings can inform some indications of the relative prediction performance of the estimators.

3.3 Estimation of Conditional Mean Predictions and Individual-level Forecasts

3.3.1 *Parametric Regression Estimators*

We estimate nine parametric regression estimators including linear regression and Finite Mixture Model (FMM).[39] Each estimator explicitly models the functional form for the mean of the outcome variable using an underlying distribution. Conditional mean predictions, $\widehat{y_i}$, are estimated from the conditional probability density function of each estimator. For the conditional mean for each estimator, see Jones *et al.* (2016). On the other hand, individual-level forecasts, $\widehat{Y_i}$, are estimated based on the conditional survival function of each estimator. These estimators are estimated by specifying the full conditional distribution of health care expenditures using between one and five parameters. In Table 2, we show the conditional probability density function and the conditional survival function for each estimator.

[39] Mixture models are considered to be semi-parametric as the number of components can vary. However, we use the fixed number of components, and thus FMM is essentially parametric.

We use linear regression (denoted as NORMAL) as traditional risk-adjustment models rely on linear regression. This is mainly because it is fast and easy to implement but also because it presents results on the scale of interest, which eases the interpretation of its coefficient. Also, there is empirical evidence showing that linear regression produces better predictions than nonlinear models for payment purposes (Dixon *et al.*, 2011; Ellis *et al.*, 2013). On the other hand, it has been criticized as a poor estimator for health care expenditure data that are renowned for its idiosyncrasies (e.g. kurtosis and long right tail) (Basu *et al.*, 2006; Hill and Miller, 2010; Jones, 2011).[40, 41]

We use three different estimation methods for each of three different performance metrics. First, we estimate predicted expenditures in the validation set as $E(y_i|x_i) = x_i\beta$, where x_i is a matrix of first-year covariates of an individual i in the validation set and β is a column vector of coefficients estimated from the estimation set. Second, we use the survival function (Table 2) to produce the estimate of $P(\hat{Y}_i > k|x_i)$ in the validation set. Because these estimated conditional tail probabilities vary across each possible combination of covariates for a given individual, we take the average in order to integrate out over values of x_i to provide the estimate of $P(\hat{Y}_i > k)$. Then, we estimate the ratio of $P(\hat{Y}_i > k)$ and the observed empirical proportion of expenditures in the

[40] It is worthwhile to note several limitations of linear regression in health care expenditure analysis. For example, Jones (2011) describes that "linear regression applied to the level of costs may perform poorly, due to the high degree of skewness and excess kurtosis; [ordinary least squares (OLS)] minimise the sum of squared residuals on the cost scale and may be sensitive to extreme observations. As a result, in applied work costs are often transformed prior to estimation." Similarly, Hill and Miller (2010) describe that "since linear OLS does not accommodate the skewness of health expenditure distributions, results may be dominated by outliers and may overfit the data, thus potentially reducing reliability for out-of-sample predictions. In addition, predictions for individuals are not constrained to be positive."

[41] It is worthwhile to note what Jones (2011) summarizes: "It is notable that the simple linear model, estimated by OLS, performs quite well across all of the criteria, a finding that has been reinforced for larger datasets than the one used here." Thus, we performed a sensitivity analysis to see if our results change when a very large sample is used. Specifically, we conducted the same analysis with a 50 percent random sample. We did see some relative improvements in the group-level prediction accuracy of NORMAL in terms of MAPE, RMSE, and R^2, but not MPE. However, it continued to perform poorly in terms of tail distribution prediction accuracy and individual-level prediction accuracy.

data that exceed the expenditure threshold k, $\frac{P(\hat{Y}_i>k)}{P(y_i>k)}$. Finally, we use the survival function to estimate our proposed metric in the validation set. As actual expenditures vary across individuals, the expenditure threshold for each of the four survival functions in the equation (4) varies across individuals. Using the survival function, we separately estimate four probabilities for each individual and combine them into a single estimate. Then, we carry out the same procedure as with the estimate of $P(\hat{Y}_i > k|x_i)$.

We use the generalized beta distribution of the second kind (GB2) and its special or limiting cases distributions, introduced by McDonald (1984). As GB2 specifies a very flexible four-parameter distribution, it well captures the characteristics of health care expenditures distribution, including skewness and long right-hand tail (Jones *et al.*, 2014). In this study, we use two GB2 estimators: one with a log-link function (denoted as GB2 LOG) and the other with a squared-root link function (denoted as GB2 SQRT). Also, GB2 accounts for other distributions as special or limiting cases, which allow less flexibility in skewness and kurtosis than GB2 (McDonald and Xu, 1995). As such, we use the three-parameter generalized gamma (denoted as GG), two-parameter gamma (denoted as GAMMA), two-parameter log-normal (denoted as LOGNORM), two-parameter Weibull (denoted as WEIB), and one-parameter exponential (denoted as EXP) distributions. For these special or limiting cases, we specify a log-link function.

We follow the three estimation methods described in NORMAL, but there is a difference. As these estimators cannot account for individuals with zero expenditure, we use a two-part model. To estimate $E(y_i|x_i)$, in the first part, we run logistic regression to estimate the probability of having any positive expenditures for all individuals in the validation set, $P(y_i > 0)$. In the second part, using the covariate coefficients and other parameters estimated from the estimation set, we estimate predicted expenditures for those with positive expenditures, $E(y_i|y_i > 0)$. Then,

predicted expenditures for all individuals in the validation set are estimated by multiplying the probability of having any positive expenditures and predicted expenditures for those with positive expenditures $E(y_i|x_i) = P(y_i > 0) \times E(y_i|y_i > 0)$. This process is also applied to estimations of $P(\hat{Y}_i > k)$ and our proposed performance metric.

FMM is a probabilistic model that accounts for characteristics of various subpopulations within an overall population without requiring to know detailed information about the distributions of each subpopulation (Deb and Trivedi, 1997; Deb and Burgess, 2003). Specifically, FMM allows for heterogeneity in populations either based on observed covariates and unobserved latent classes. When individual's health status is imperfectly observed, FMM enables to estimate health care expenditures by splitting a population with heterogeneity in certain characteristics by the latent health status of individuals (Cameron and Trivedi, 2005). The unknown population distribution may be empirically approximated by a mixture of distributions with finite, but a small number of mixture components (Heckman, 2001).[42] Thus, we use two gamma-distributed components of FMM with a squared-root link function in both components (denoted as FMM SQRT).[43] We use the same estimation methods as with GB2.

3.3.2 *Machine Learning Estimators*

We use seven machine learning estimators. Machine learning has been extensively used due to its various capabilities, but we focus on one aspect of machine learning. Machine learning is based on a prediction algorithm that iteratively learns from the data. Thus, it does not rely on explicit

[42] In theory, increasing the number of mixture components may enable to fit well with any distribution. However, in practice, researchers tend to use a small number of mixture components (two or three) to avoid non-convergence and time-consuming issues occurring due to the complexity of a model.

[43] Although we also used two gamma-distributed components of FMM with a log-link function in both components, the model showed very poor convergence performance (4 of 100 trials). Thus, we excluded the model from analysis.

models or distributions, thereby allowing a greater flexibility to determine how to make predictions. In other words, a predictive model is gradually improved by testing its predictions and correcting when wrong. Using this capability of machine learning, we model the relationship between health care expenditures and covariates to predict health care expenditures. Despite its popularity in computer science and statistics, it has yet to be popular in health economics, especially for health care expenditure analysis (Bertsimas *et al.*, 2008; Robinson, 2008; Buchner *et al.*, 2015; Rose, 2016).

Machine learning estimators often use a three-way cross-validation to avoid overfitting a model.[44] To be consistent with the other estimators, however, we use a one-fold cross-validation and repeat the process 100 times with sample replacement to produce the average estimates across all replications.[45] Although machine learning has the potential to improve model flexibility and predictability under certain constraints, we do not put particular constraints in the estimation procedure. This is because there is no consensus on specific criteria for selecting constraints in model selection and it is not worthwhile to make a model too complex.

Regularized linear regression is a technique used to prevent overfitting while generating a parsimonious model with high prediction accuracy (Friedman *et al.*, 2009). In the presence of collinearity, estimates from ordinary least squares can be biased and highly variable. Such an issue can be addressed by shrinking regression coefficients of covariates with weak or no contribution to an outcome variable toward zero or exactly zero. This technique is relevant to risk adjustment, because many covariates are highly correlated, possibly causing multicollinearity and unstable

[44] Typically, an entire sample is split into three disjoint sets. The first part (known as a training set) is used to fit the model. The second part (known as a validation set) is used to estimate prediction error for model selection. The last part (known as a test set) is used for assessment of the generalization error of the final chosen model.

[45] This approach might cause overfitting, especially when a sample size is small. Thus, we performed a sensitivity analysis to see if prediction performance varies depending on sample size. However, our sensitivity analysis showed no differences between findings from analyses using 1 percent random samples and those from analyses using 50 percent random samples.

estimates. We use four estimators that differ in the way of shrinking regression coefficients. Least absolute shrinkage and selection operator regression (denoted as LASSO) uses the algorithm to shrink some regression coefficients to exactly zero, thereby eliminating their contributions to the predicted outcome (Tibshirani, 1996). However, ridge regression (denoted as RIDGE) shrinks some regression coefficients toward zero, but not exactly zero (Marquardt and Snee, 1975). Elastic net regression (denoted as ENET) is a hybrid approach that combines both regularizations of the Lasso and ridge regressions (Zou and Hastie, 2005). Finally, least angle regression (LARS) (denoted as LARS) is nearly identical to Lasso, but with a simple modification the LARS algorithm provides a more efficient way to calculate all possible Lasso estimators than original Lasso (Efron *et al.*, 2004).

To obtain the estimates of $E(y_i|x_i)$, $P(\hat{Y}_i > k)$, and our performance metric, we carry out the similar estimation methods described in NORMAL. For each estimator, we choose the model with the smallest cross-validated mean squared error and use the nonzero coefficients of β estimated from the estimation set. To estimate $P(\hat{Y}_i > k)$ and our metric, it is needed to have a distributional assumption, as shown in the parametric regression estimators. However, machine learning estimators do not rely on any distributional assumptions. Thus, we assume a local normality and allow for heteroscedasticity in variances of health care expenditures. Specifically, we form percentiles of predicted expenditures in the estimation and validation sets, respectively, and compute variances of actual health care expenditures for each percentile in the estimation set. Then, we apply each of those variances from the estimation set to the equivalent percentile in the validation set.

Artificial neural network (denoted as NNET) is an information processing algorithm to explain an outcome variable by extracting complex relationships of highly interconnected

covariates within multiple layers (i.e., network) (Zhang *et al.*, 1998). The important feature of neural network is its adaptive nature, in which a self-learning algorithm replaces an explicitly programmed algorithm. Here, self-learning refers to the automatic adjustment of parameters so that the algorithm can estimate the correct outcome variable for given covariates. As neural network does not rely on a prescribed relation, the algorithm acquires information by implicitly detecting complex nonlinear relationships between the outcome variable and covariates and detecting all possible interactions between covariates. We use the similar estimation methods described in LASSO. However, the drawback is that the process of estimating $E(y_i|x_i)$ is not explicitly known. We allow the number of units in the hidden layer to be 10, as a cross-validated mean squared error was minimal when the number of hidden neurons was 10, on average.

A decision tree is a commonly used data mining method for developing a prediction algorithm by learning decision rules inferred from the data (Breiman *et al.*, 1984). Decision tree builds classification or regression models in the form of a tree structure. Although there are many different methods for growing a tree, we use a single decision tree model where every decision is based on a comparison of two covariates (denoted as SINGLE).[46] Based on a set of covariates, this method creates rules to classify a population into groups with most homogeneous characteristics (i.e., node). If the set of covariates is sufficiently homogenous among observations in each node, the tree is no longer grown. A decision tree model typically has an issue of overfitting, especially when the tree has a large number of terminal nodes (Rose, 2016). To avoid overfitting, we select a tree size that minimizes the cross-validated error. To obtain the estimates of $E(y_i|x_i)$, $P(\hat{Y}_i > k)$, and our performance metric, we follow the same estimation methods described in LASSO.

[46] In addition to SINGLE, Rose (2016) used a random forest model where decisions are made based on multiple deep decision trees. Due to its high computational burden, however, we excluded the model from analysis.

The super learner (denoted as SUPER) is an ensemble method that blends predictions from multiple candidate estimators with optimal weights (Van Der Laan *et al.*, 2007). The method itself is a prediction algorithm, which fits a set of candidate estimators to the data and then estimates the optimal weight for each of these estimators based on cross-validated risk. The optimal weights are determined by minimizing the cross-validated risk over a set of candidate estimators. Theoretically, it performs asymptotically and better than any of the candidate estimators (Van Der Laan *et al.*, 2006). Thus, the super learner enables to create a highly predictive model by objectively combining results from different prediction algorithms rather than relying on a single algorithm subjectively selected by a researcher. For candidate estimators, we use six machine learning estimators described above. We use a cross-validated mean squared error as a loss function. Thus, the super learner chosen in this study has the smallest cross-validated mean squared error. We perform the same estimation methods described in LASSO.

3.3.3 *Distributional Estimators*

We use three conditional density approximation estimators to generate counterfactual distributions.[47]

To estimate $E(y_i|x_i)$ in the validation set, we follow the estimation method outlined in Gilleskie Mroz (2004). First, the outcome variable is divided into Q discrete intervals. Then, we estimate the probabilities that each observation lies in the j interval, $p_{ij}(x_i)$, and the mean value of the outcome variable for the j interval, $\bar{y}_j$. Following Jones *et al.* (2016), we assume that the probability of lying within an interval only relies on covariates and that the mean value of the

[47] Jones *et al.* (2015) used two other distributional estimators based on quantile regression. However, we excluded them from analysis because their conditional mean values cannot be estimated.

outcome variable for a given interval does not change with covariates. By multiplying these two values for each interval and summing them up across all intervals, we estimate $E(y_i|x_i)$:

$$E(y_i|x_i) = \sum_{j=1}^{Q} p_{ij}(x_i)\bar{y}_j \quad (6)$$

In this study, we use three different ways of estimating the probabilities $p_{ij}(x_i)$. Following the method of Han Hausman (1990), we construct a categorical variable for each observation, indicating the interval into which the value of the outcome variable falls. Then, we run ordered logit regression to estimate $p_{ij}(x_i)$. We use 33 intervals because the number of intervals resulted in good convergence performance in preliminary work (denoted as HH). Following the method of Foresi Peracchi (1995), we divide the data into a set of discrete intervals, but use a series of logit regressions to estimate $p_{ij}(x_i)$. For each upper boundary of the intervals, we generate an indicator that is equal to one if the outcome variable is less than or equal to the upper boundary and zero otherwise. Based on our preliminary work, we choose 18 intervals: 0th, 5th, . . . , 85th, 90th, and 95th percentiles as boundaries (denoted as FP). In contrast, Chernozhukov *et al.* (2013) proposed more flexible distribution-based regression by using logit regression for each unique value of the outcome variable. However, as this method requires very expensive computational demands for a large sample, we adopt the linear probability model, used by de Meijer *et al.* (2013), to estimate $p_{ij}(x_i)$, (denoted as CFM).[48]

To obtain the estimate of $P(\hat{Y}_i > k)$ and our performance metric in the validation set, we follow Jones *et al.* (2015). We produce the estimate of $P(\hat{Y}_i > k^*|x_i)$, where k^* represents one of the boundaries of the intervals generated in each of these estimators. As these estimators are

[48] Jones *et al.* (2015) and de Meijer *et al.* (2013) showed that the results from the linear probability model and the method of Chernozhukov *et al.* (2013) were virtually identical.

performed without knowing the threshold k, it is not always the case where $k^* = k$. When $k^* \neq k$, we use the following simple linear interpolation formula to estimate a weighted average of $P(\hat{Y}_i > k^*|x_i)$ for the nearest two values of k^* to k.

$$P(\hat{Y}_i > k^*|x_i) = P\left(\hat{Y}_i > k_a^*|x_i\right) + \left(\frac{k - k_a^*}{k_b^* - k_a^*}\right)\left(P\left(\hat{Y}_i > k_b^*|x_i\right) - P\left(\hat{Y}_i > k_a^*|x_i\right)\right) \quad (7)$$

where k_a^* and k_b^* represent the thresholds analyzed in estimation closest below and closest above k, respectively.

To compute our proposed metric, we carry out the same procedure as with NORMAL. We follow the same process, but use three different thresholds that vary across each individual, instead of the constant threshold k across all individuals.

3.4 Evaluation of Prediction Performance Metrics

To evaluate group-level prediction accuracy in the validation sets, we use MPE, MAPE, RMSE, R^2,[49] and, predictive ratio. A perfect prediction represents zero for MPE, MAPE, and RMSE, and one for R^2 and predictive ratio. Using the second year's actual and predicted expenditures, $\hat{y}_i = E(y_i|x_i)$, we calculate the following group-level prediction performance:

$$MPE = \frac{\sum(y_i - \hat{y}_i)}{N} \quad (8)$$

$$MAPE = \frac{\sum|y_i - \hat{y}_i|}{N} \quad (9)$$

$$RMSE = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{N}} \quad (10)$$

[49] R^2 is estimated by an auxiliary regression of actual expenditures on predicted expenditures. Coefficients from the auxiliary regression are denoted by an AUX subscription and they are estimated from the estimation sets.

$$R^2 = 1 - \frac{\sum(y_i - (\alpha_{AUX} + \beta_{AUX}\hat{y}_i))^2}{\sum(y_i - \hat{y}_i)^2} \quad (11)$$

$$Predictive\ ratio = \frac{\frac{1}{N}\sum \hat{y}_i}{\frac{1}{N}\sum y_i} \quad (12)$$

To evaluate prediction accuracy at the tail distributions in the validation sets, we estimate $\frac{P(\hat{Y}_i>k)}{P(y_i>k)}$. A perfect prediction represents a ratio of one. We use the following four expenditure thresholds: $5,000, $15,000, $25,000, and $80,000. Each of them is approximately at the 75th, 90th, 95th, and 99th percentiles of the health care expenditure distribution in our data.

To evaluate individual-level prediction accuracy in the validation sets, we use two individual-level prediction performance metrics: $P(\frac{|y_i-\hat{Y}_i|}{y_i} \times 100 > l)$ and $P(\frac{|y_i-\hat{Y}_i|}{y_i} \times 100 > l\,|y_i > k)$. A perfect prediction represents zero probability. We use four different percent thresholds l: 10, 50, 100, and 250. Also, we use the same four expenditure thresholds described in above.

We report these prediction performance metrics in two ways. For each performance metric, first, we graphically present prediction accuracy of the 19 estimators. For each performance metric, we also report rankings of the 19 estimators in terms of bias. Bias indicates the averaged deviations from a perfect prediction point. Higher ranking (lower number) represents lower deviations.

3.5 RESULTS

Table 3 and Appendix Table A summarize descriptive statistics on our total study population. Our study population consisted of 12,882,983 working-aged adults who were continuously enrolled in the same plan type between 2013-2014. For a single iteration, the estimation and validation sets included 128,829 adults, respectively. Table 3 summarizes descriptive statistics for the second year

(2014) total health care expenditures. The mean and standard deviation of health care expenditures were $6,891 and $22,963, respectively. The 75th, 90th, 95th, and 99th percentiles of the expenditure distribution were $5,736, $15,384, $27,080, and $80,116, respectively. Appendix Table A summarizes descriptive statistics for the first year (2013) information on demographics, diagnoses, state, and type of plan.

Table 4 shows each estimator's performance on convergence and average computing time for a single iteration (in seconds).[50] Of the 100 iterations, convergence was achieved 97 times on average. However, FMM SQRT showed the lowest convergence performance (68 times). On the other hand, there were substantial variances in the average computing time. CFM was the most computationally demanding estimator (76,494 seconds), whereas LASSO, RIDGE, and ENET were the least computationally demanding estimators (32-33 seconds).

In Figure 2 we present the group-level prediction accuracy of the 19 estimators. The points indicate the averaged values and the capped spikes indicate the range of the values across iterations. Overall, NORMAL, machine learning, and distributional estimators outperformed the other eight parametric regression estimators. In other words, compared to the latter, the former displayed MPE closer to zero, lower MAPE, lower RMSE, higher R^2, and predictive ratio closer to one. For the other four metrics except for R^2, the former had narrower ranges than the latter.

Figure 3 shows the performance of the 19 estimators in predicting the probability of an expenditure exceeding $5,000, $15,000, $25,000, and $80,000, respectively. In forecasting the probabilities of expenditures exceeding $5,000, $15,000, and $25,000, parametric regression

[50] Average computing time represents the time spent on estimating conditional mean expenditures and measuring prediction accuracies at the group level, at the tail distribution, and at the individual level. However, it cannot be directly comparable across estimators, as all estimators were not performed with the same statistical software package on the same server. Because we used two different servers, the average computing time for each estimator might be different depending on which server is used.

estimators dominated the other estimators except for CFM. In other words, parametric regression estimators produced the least biased and the most precise estimates. For the probability of an expenditure exceeding $80,000, however, machine learning estimators yielded slightly less biased estimates than parametric regression estimators but with low precision. As a single estimator, CFM showed the least biased performance across almost all expenditure thresholds. However, the other two distributional estimators showed poorer prediction performance for higher expenditure thresholds. Finally, NORMAL showed overprediction for those whose expenditure exceeded $5,000 and $15,000, respectively, while showing underprediction for those whose expenditures exceeded $80,000.

In Figure 4 we present the overall individual-level prediction accuracy of the 19 estimators. When we defined the prediction accuracy as a forecast error less than 10 percent of each individual's actual expenditure, there were marginal differences among parametric regressions, machine learning, and distributional estimators except for FMM SQRT. When we allowed the forecast error by 50 percent, parametric regression and distributional estimators showed slightly better performance than machine learning estimators. When we allowed the forecast error by 100 percent, however, parametric regression and distributional estimators outperformed machine learning estimators. Also, FMM SQRT showed the worst individual-level prediction performance when the strictest criterion was used (10 percent forecast error), whereas it showed the best performance when the least strict criterion was applied (100 percent forecast error). In other words, FMM SQRT produced many small forecast errors, but few large forecast errors. Finally, NORMAL generated many large forecast errors, resulting in poor individual-level prediction accuracy.

Figure 5 shows the tail-specific individual-level prediction accuracy of the 19 estimators. When less strict criteria were used (10 or 50 percent forecast errors), parametric regression estimators showed slightly better performance than machine learning and distributional estimators, especially for those with expenditures exceeding $80,000. When the least strict criterion was used (100 percent forecast error), on the other hand, parametric regression and distributional estimators dominated machine learning estimators for all tail distributions considered in this study. Finally, NORMAL and machine learning estimators consistently showed poorer individual-level prediction accuracy, especially at the tail distributions.

In summary, there was no single estimator that outperformed the others for all metrics. As shown in Appendix Table B, LARS showed the highest group-level prediction accuracy, whereas the estimator's performance on other metrics was below average overall. Similarly, CFM was the best estimator for predicting tail probabilities, but the estimator's group-level prediction accuracy was around the average level. On the other hand, although GG, EXP, LOGNORM, and FMM SQRT showed the worst prediction accuracy at the group level, their overall and tail-specific individual-level prediction accuracies were highly ranked.

3.6 DISCUSSION

There are two main contributions in this paper. First, to the best of our knowledge, this is the first paper that directly compares three board classes of estimators—parametric regression estimators, machine learning estimators, and distribution-based estimators, in total 19 estimators—in predicting health care expenditures. The second contribution is the development of novel metrics used to compare individual-level prediction accuracy of these estimators. Previous literature has mainly looked at performance metrics that examine moments and features of the whole

distribution, such as R^2 and performance based on predicting the tail probabilities. We argue that while it is certainly important to look at these metrics, it is also important to examine additional metrics that measure prediction accuracies at the individual level. This is because certain risk-selection behaviors in the private insurance market, such as service-level selection, are operationalized to select out specific (but not all) individuals with high expenditures, irrespective of what their predictions from a risk-adjustment model are. Any attempt to counteract such risk selection must ensure that a risk-adjustment model can accurately predict health care expenditures at the individual level.

In line with previous literature, we found that no one estimator was clearly dominant for all metrics considered in this study. However, more importantly, we found nuances of prediction performance across estimators. In general, machine learning and distribution-based estimators achieved better group-level prediction accuracy than parametric regression estimators. NORMAL seems to perform better than other parametric estimators in terms of group-level prediction accuracy, which was also pointed out by Jones *et al*. (2015). However, parametric regression estimators showed higher tail distribution prediction accuracy and individual-level prediction accuracy, especially at the tail distributions. However, NORMAL and machine learning estimators generated poor predictability at the tail distributions and at the individual level.

This suggests that there is a trade-off in selecting an appropriate risk-adjustment method between estimating accurate payments at the group level and estimating low residual risks at the individual level. We believe that this is a new insight in the risk-adjustment literature. It indicates that the determination of an optimal risk-adjustment method would vary depending on a specific statistical metric used to evaluate prediction accuracy. Therefore, in order to select an optimal risk-adjustment estimator, future research should simulate health plans' behaviors in response to both

group-level and individual-level residual risks conditional on risk-adjusted payments. Such simulation helps to better understand the trade-off inherent in selecting an appropriate risk-adjustment method and devise an optimal risk-adjustment strategy.

Our study also found that group-level prediction accuracies tend to be more variable (across iterations) than individual-level prediction accuracies. This was expected as group-level performance metrics simply measure the magnitude of the average of the deviation between predicted and actual values across all individuals. As such, they were more likely to be disproportionately affected by the variability of less accurate prediction for those with extremely high expenditures.

There are nuances in prediction accuracies at the level of specific estimators. For example, parametric regression estimators consistently achieved high prediction accuracy at the individual level, especially at the tail distributions. Compared to the other estimators, they generated fewer large prediction errors along the expenditure distribution and generated fewer small prediction errors at the tail distributions. This implies that these flexible parametric distributions fit the health care expenditure data well, as they impose fewer restrictions on skewness and kurtosis, which allows for a greater range of estimated effects of covariates (Jones *et al.*, 2015). Within the tail-specific performance metrics, however, there was no single estimator that was superior to the others. This suggests that each estimator achieves the best fit at different points in the distribution. There are two findings that are worthwhile to note. First, the parsimonious one-parameter EXP performed better than the other parametric regression estimators with more flexible distributions. This may indicate that more flexibility does not necessarily lead to a better prediction fit. Second, FMM SQRT produced fewer large prediction errors than any other estimators, but generated much more small prediction errors than the other estimators, especially at the tail distributions. This

might be because FMM SQRT model a separate probability density function for those with relatively low expenditures, thereby avoiding large prediction errors for them. However, it might result in greater prediction errors at the tail distributions.

We found that NORMAL, the linear regression-based estimator that is typically used for the purpose of risk-adjusting plan payments, achieved a high level of prediction accuracy at the group level, whereas the other prediction performance of the estimator was poor. Especially, it produced large individual-level prediction errors for those with very high expenditures. This can be explained by the normality assumption and linearity property. Assuming the outcome variable is normally distributed, linear regression calculates a straight line through the data that results in the smallest prediction errors. However, the actual distribution of health care expenditures is heavily right skewed and long tailed. As such, linear regression underestimates observations in the tail distributions, thereby resulting in underpredictions for them. On the other hand, as linear regression relies on the linearity property, it intends to minimize the prediction errors across the entire range of the outcome variable. To cancel out the underprediction for those in the tail distributions, linear regression produces overprediction for those in other parts of the distribution. This has been shown in our study as well as other studies (Pope *et al.*, 2011; Medicare Payment Advisory Commission, 2012; Kautter *et al.*, 2014). For instance, the HHS-HCC model, on average, underestimated expenditures for those with more than six chronic conditions by $608, whereas it overestimated expenditures for those with four and five chronic conditions by $182 and $456, respectively (Government Accountablity Office, 2011). These suggest that reliance on linear regression would likely result in higher residuals risks, thereby generating inaccurate plan payments, especially for those with the most severe health status.

Our results also show that compared to parametric regression and distributional estimators, machine learning estimators produced large individual-level prediction errors more, especially at the tail distributions. This might result from prediction algorithms in machine learning and a limited set of risk-adjusters included in the HCC model. Machine learning estimators were expected to accurately predict expenditures for the high-expenditure individuals, because they are designed to capture nonlinear complex relationships between the outcome variable and covariates or between covariates. However, a sufficiently large number of observations might be needed to detect such relationships. In our data, the sample size of those with very high expenditures was not large enough to detect these relationships. Furthermore, according to the principles for risk-adjustment model development, the HCC model only accounts for clinically significant medical conditions incurring significant expenditures (Centers for Medicaid and Medicare Services, 2016). Of about 100 major medical conditions included in the model, however, the other conditions except for top 10 conditions were found to have minor contributions to predicting health care expenditures (Rose, 2016). To fully exploit the predictive ability of machine learning, it is needed to add more detailed diagnostic information (Newhouse *et al.*, 2013).

Our study has several limitations. First, our study population is not the targeted population for the HHS-HCC model because the population from the MarketScan database is more likely to be healthy than the eligible enrollees for the Exchanges. To overcome this issue, Layton *et al.* (2015); Rose (2016) drew a matched sample of individuals from the MarketScan database, who have similar characteristics to those eligible for the Exchanges. However, our goal is not to evaluate the effect of the HHS-HCC model on risk selection among those eligible for the Exchanges. Thus, we did not limit the study population to them. Second, there are various factors determining health care expenditures, but we only included demographic and major diagnostic

information. This is because the focus of this study is to examine the existing plan payment risk-adjustment policy in the US, which relies on demographic and diagnostic information. Third, we assumed the linearity between two neighboring intervals to construct the counterfactual unconditional distribution. However, such assumption is unlikely to hold at the tail distributions, as the health care expenditure data is highly skewed and has very heavy tails, thereby suggesting a nonlinear relationship between two neighboring intervals at the tail distributions.

Our study can be applied to various payment and delivery systems. First, accurately estimating a financial benchmark for each an accountable care organization (ACO) is key for determining ACO financial performance and shared savings over time. However, because benchmarks are designed to capture population health status accurately, this could generate substantial residual risks at the individual level. If benchmarks do not reflect health status not only at the group level but also at the individual level, this could unfairly penalize ACOs serving patients with high individual-level residual risks, possibly leading them to leave the voluntary ACO program. Second, starting in 2016, CMS launched a new payment and care delivery model for the treatment of cancer patients. Under the Oncology Care model, providers will receive a per-beneficiary-per-month payment for the duration of each six-month episode. However, an issue is that if episode costs are not accurately estimated at the individual level, providers will have incentives to avoid patients with high individual-level residual risks. Accordingly, it is necessary to conduct additional studies to investigate how to improve estimation accuracies for financial benchmark and episode payment not only at the group level but also at the individual level.

References

Basu A, Arondekar BV, Rathouz PJ. 2006. Scale of interest versus scale of estimation: comparing alternative estimators for the incremental costs of a comorbidity. *Health Economics* **15**(10): 1091-1107.

Bertsimas D, Bjarnadóttir MV, Kane MA, Kryder JC, Pandey R, Vempala S, Wang G. 2008. Algorithmic prediction of health-care costs. *Operations Research* **56**(6): 1382-1392.

Breiman L, Friedman J, Stone CJ, Olshen RA. 1984. *Classification and regression trees*. New York: CRC Press.

Brown J, Duggan M, Kuziemko I, Woolston W. 2014. How does risk selection respond to risk adjustment? New evidence from the Medicare Advantage program. *American Economic Review* **104**(10): 3335-3364.

Buchner F, Wasem J, Schillo S. 2015. Regression trees identify relevant interactions: can this improve the predictive performance of risk adjustment? *Health Economics* **26**(1): 74-85.

Cameron AC, Trivedi PK. 2005. *Microeconometrics: Methods and Applications*. Cambridge, UK: Cambridge University Press.

Centers for Medicaid and Medicare Services. 2014. 2014 Benefit year risk adjustment SAS version of HHS-developed risk adjustment model algorithm software. Baltimore, MD: Centers for Medicaid and Medicare Services.

Centers for Medicaid and Medicare Services. 2016. HHS-operated risk adjustment methodology meeting: Discussion paper. Baltimore, MD: Centers for Medicaid and Medicare Services Center for Consumer Information & Insurance Oversight.

Chernozhukov V, Fernandez-Val I, Melly B. 2013. Inference on counterfactual distributions. *Econometrica* **81**(6): 2205-2268.

De Meijer C, O'donnell O, Koopmanschap M, Van Doorslaer E. 2013. Health expenditure growth: looking beyond the average through decomposition of the full distribution. *Journal of Health Economics* **32**(1): 88-105.

Deb P, Burgess J. 2003. A quasi-experimental comparison of econometric models for health care expenditures. Hunter College Department of Economics Working Papers.

Deb P, Trivedi PK. 1997. Demand for medical care by the elderly: A finite mixture approach. *Journal of Applied Econometrics* **12**(3): 313-336.

Dixon J, Smith P, Gravelle H, Martin S, Bardsley M, Rice N, Georghiou T, Dusheiko M, Billings J, Lorenzo MD, Sanderson C. 2011. A person based formula for allocating commissioning funds to general practices in England: development of a statistical model. *Bmj* **343**: d6608.

Efron B, Hastie T, Johnstone I, Tibshirani R, Ishwaran H, Knight K, Loubes JM, Massart P, Madigan D, Ridgeway G, Rosset S, Zhu JI, Stine RA, Turlach BA, Weisberg S, Hastie T, Johnstone I, Tibshirani R. 2004. Least angle regression. *Annals of Statistics* **32**(2): 407-499.

Einav L, Finkelstein A, Kluender R, Schrimpf P. 2016. Beyond statistics: the economic content of risk scores. *American Economic Journal: Applied Economics* **8**(2): 195-224.

Ellis RP, Fiebig DG, Johar M, Jones G, Savage E. 2013. Explaining health care expenditure variation: large-sample evidence using linked survey and health administrative data. *Health Economics* **22**(9): 1093-1110.

Foresi S, Peracchi F. 1995. The conditional distribution of excess returns: an empirical analysis. *Journal of the American Statistical Association* **90**(430): 451-466.

Friedman J, Hastie T, Tibshirani R. 2009. *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer.

Gilleskie DB, Mroz TA. 2004. A flexible approach for estimating the effects of covariates on health expenditures. *Journal of Health Economics* **23**(2): 391-418.

Glazer J, Mcguire TG. 2000. Optimal risk adjustment in markets with adverse selection: an application to managed care. *American Economic Review* **90**(4): 1055-1071.

Government Accountablity Office. 2011. Medicare Advantage: Changes improved accuracy of risk adjustment for certain beneficiaries. Washington, DC: Government Accountablity Office,.

Han A, Hausman JA. 1990. Flexible parametric estimation of duration and competing risk models. *Journal of Applied Econometrics* **5**(1): 1-28.

Heckman JJ. 2001. Micro data, heterogeneity, and the evaluation of public policy: Nobel lecture. *Journal of Political Economy* **109**(4): 673-748.

Hill SC, Miller GE. 2010. Health expenditure estimation and functional form: applications of the generalized gamma and extended estimating equations models. *Health Economics* **19**(5): 608-627.

Jones AM. 2011. Models for health care. In *Oxford Handbook of Economic Forecasting*, Clements MP, Hendry DF (eds). Oxford University Press; 625–654.

Jones AM, Lomas J, Moore PT, Rice N. 2016. A quasi-Monte-Carlo comparison of parametric and semiparametric regression methods for heavy-tailed and non-normal data: an application to healthcare costs. *Journal of the Royal Statistical Society Series a-Statistics in Society* **179**(4): 951-974.

Jones AM, Lomas J, Rice N. 2014. Applying beta-type size distributions to healthcare cost regressions. *Journal of Applied Econometrics* **29**(4): 649-670.

Jones AM, Lomas J, Rice N. 2015. Healthcare cost regressions: going beyond the mean to estimate the full distribution. *Health Economics* **24**(9): 1192-1212.

Kautter J, Pope GC, Ingber M, Freeman S, Patterson L. 2014. The HHS-HCC risk adjustment model for individual and small group markets under the Affordable Care Act. *Medicare and Medicaid Research Review* **4**(3).

Layton TJ, Ellis RP, Mcguire TG. 2015. Assessing Incentives for Adverse Selection in Health Plan Payment Systems. *NBER Working Paper* **21531**

Marquardt DW, Snee RD. 1975. Ridge regression in practice. *American Statistician* **29**(1): 3-20.

Mcdonald JB. 1984. Some generalized functions for the size distribution of income. *Econometrica* **52**(3): 647-663.

Mcdonald JB, Xu YJ. 1995. A generalization of the beta distribution with applications. *Journal of Econometrics* **66**(1-2): 133-152.

Mcguire TG, Newhouse JP, Normand SL, Shi J, Zuvekas S. 2014. Assessing incentives for service-level selection in private health insurance exchanges. *Journal of Health Economics* **35**(1): 47-63.

Mcwilliams JM, Hsu J, Newhouse JP. 2012. New risk-adjustment system was associated with reduced favorable selection in Medicare Advantage. *Health Affairs* **31**(12): 2630-2640.

Medicare Payment Advisory Commission. 2012. Report to the Congress: Medicare and the Health Care Delivery System. Washington, DC: Medicare Payment Advisory Commission.

Medicare Payment Advisory Commission. 2014. Report to the Congress: Medicare and the Health Care Delivery System. Washington, DC: Medicare Payment Advisory Commission.

Montz E, Layton T, Busch AB, Ellis RP, Rose S, Mcguire TG. 2016. Risk-adjustment simulation: plans may have incentives to distort mental health and substance use coverage. *Health Affairs* **35**(6): 1022-1028.

Morrisey MA, Kilgore ML, Becker DJ, Smith W, Delzell E. 2013. Favorable selection, risk adjustment, and the Medicare Advantage program. *Health Services Research* **48**(3): 1039-1056.

Newhouse JP, Mcwilliams JM, Price M, Huang J, Fireman B, Hsu J. 2013. Do Medicare Advantage plans select enrollees in higher margin clinical categories? *Journal of Health Economics* **32**(6): 1278-1288.

Newhouse JP, Price M, Huang J, Mcwilliams JM, Hsu J. 2012. Steps to reduce favorable risk selection in Medicare Advantage largely succeeded, boding well for Health Insurance Exchanges. *Health Affairs* **31**(12): 2618-2628.

Newhouse JP, Price M, Mcwilliams JM, Hsu J, Mcguire TG. 2015. How much favorable selection is left in Medicare Advantage? *American Journal of Health Economics* **1**(1): 1-26.

Park S, Basu A, Coe NB, Khalil F. 2017. Service-level selection: Strategic risk selecion in Medicare Advantage in response to risk adjustment. *NBER Working Paper* **24038**.

Pope GC, Kautter J, Ellis RP, Ash AS, Ayanian JZ, Iezzoni LI, Ingber MJ, Levy JM, Robst J. 2004. Risk adjustment of medicare capitation payments using the CMS-HCC model. *Health Care Financing Review* **25**(4): 119-141.

Pope GC, Kautter J, Ingber MJ, Freeman S, Sekar R, Newhart C. 2011. Evaluation of the CMS-HCC risk adjustment model final report. Research Triangle Park: RTI International.

Rahman M, Keohane L, Trivedi AN, Mor V. 2015. High-cost patients had substantial rates of leaving Medicare Advantage and joining Traditional Medicare. *Health Affairs* **34**(10): 1675-1681.

Riley GF, Levy JM, Montgomery MA. 2009. Adverse selection in the Medicare prescription drug program. *Health Affairs* **28**(6): 1826-1837.

Robinson JW. 2008. Regression tree boosting to adjust health care cost predictions for diagnostic mix. *Health Services Research* **43**(2): 755-772.

Rose S. 2016. A machine learning framework for plan payment risk adjustment. *Health Services Research* **51**(6): 2358-2374.

Tibshirani R. 1996. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **58**: 267–288.

Van De Ven WPMM, Ellis RP. 2000. Risk adjustment in competitive health plan markets. In *Handbook of Health Economics*, Culyer JA, Newhouse PJ (eds). Amsterdam; Elsevier; 755-845.

Van Der Laan MJ, Dudoit S, Vaart Aad WVD. 2006. The cross-validated adaptive epsilon-net estimator. *Statistics & Decisions* **24**(3): 373.

Van Der Laan MJ, Polley EC, Hubbard AE. 2007. Super learner. *Statistical Applications in Genetics and Molecular Biology* **6**(1).

Weiner JP, Trish E, Abrams C, Lemke K. 2012. Adjusting for risk selection in state health insurance exchanges will be critically important and feasible, but not easy. *Health Affairs* **31**(2): 306-315.

Zhang G, Patuwo BE, Hu MY. 1998. Forecasting with artificial neural networks: the state of the art. *International Journal of Forecasting* **14**(1): 35-62.

Zou H, Hastie T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **67**(2): 301-320.

Table 1. Key for estimator labels

Estimator	Label
Parametric regression estimators	
Linear regression with the normal distribution of error	NORMAL
Generalized beta of the second kind (log-link)	GB2 LOG
Generalized beta of the second kind (squared-root link)	GB2 SQRT
Generalized gamma (log-link)	GG
Gamma (log-link)	GAMMA
Log-normal (log-link)	LOGNORM
Weibull (log-link)	WEIB
Exponential (log-link)	EXP
Two-component finite mixture of gamma densities (squared-root link)	FMM SQRT
Machine learning estimators	
Lasso regression	LASSO
Ridge regression	RIDGE
Elastic net	ENET
Least angle regression	LARS
Neural net	NNET
Single tree	SINGLE
Super learner	SUPER
Distributional estimators	
Han and Hausman (conditional density approximation estimator using ordered logit regression)	HH
Foresi and Peracchi (conditional density approximation estimator using multinomial logit regression)	FP
Chernozhukov, Fernández-Val and Melly (linear probability model)	CFM

Table 2. Forms of density functions and survival functions for parametric distributions

Model	$f(y\|X)$	$P(y > k\|X)$
NORMAL	$\frac{1}{2}\left[1 + \text{erf}(\frac{y-\mu}{\sqrt{2\sigma^2}})\right]$ where $\text{erf(y)} = \frac{1}{\sqrt{\pi}}\int_{-y}^{y} e^{-t^2} dt$	$1 - \int_{-\infty}^{y} \frac{e^{-(y-\mu)^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}}$
GB2 LOG	$\frac{ay^{ap-1}}{\exp(x\beta)^{ap} B(p,q)\left[1 + \left(\frac{y}{\exp(x\beta)}\right)^a\right]^{(p+q)}}$	$1 - I_Z(p,q)^*$ where $z = \left(\frac{k}{\exp(x\beta)}\right)^a$
GB2 SQRT	$\frac{ay^{ap-1}}{(x\beta)^{2ap} B(p,q)\left[1 + \left(\frac{y}{(x\beta)^2}\right)^a\right]^{(p+q)}}$	$1 - I_Z(p,q)^*$ where $z = \left(\frac{k}{(x\beta)^2}\right)^a$
GG	$\frac{\kappa}{\sigma y \Gamma(\kappa^{-2})}\left(\kappa^{-2}\left(\frac{y}{\exp(x\beta)}\right)^{\kappa/\sigma}\right)^{\kappa^{-2}} \exp\left(-\kappa^{-2}\left(\frac{y}{\exp(x\beta)}\right)^{\kappa/\sigma}\right)$	if $\kappa > 0$: $1 - \Gamma(z;\kappa^{-2})^{**}$ if $\kappa < 0$: $\Gamma(z;\kappa^{-2})^{**}$ where $z = \kappa^{-2}\left(\frac{k}{\exp(x\beta)}\right)^{\kappa/\sigma}$
GAMMA	$\frac{\kappa}{y\Gamma(\kappa^{-2})}\left(\kappa^{-2}\left(\frac{y}{\exp(x\beta)}\right)\right)^{\kappa^{-2}} \exp\left(-\kappa^{-2}\left(\frac{y}{\exp(x\beta)}\right)\right)$	if $\kappa > 0$: $1 - \Gamma(z;\kappa^{-2})^{**}$ if $\kappa < 0$: $\Gamma(z;\kappa^{-2})^{**}$ where $z = \kappa^{-2}\left(\frac{k}{\exp(x\beta)}\right)$
LOGNORM	$\frac{1}{\sigma y\sqrt{2\pi}} exp\left(\frac{-(\ln y - x\beta)^2}{2\sigma^2}\right)$	$1 - \Phi\left(\frac{\ln k - x\beta}{\sigma}\right)$
WEIB	$\frac{1}{\sigma y}\left(\frac{y}{\exp(x\beta)}\right)^{\frac{1}{\sigma}} \exp\left(-\left(\frac{y}{\exp(x\beta)}\right)^{\frac{1}{\sigma}}\right)$	$\exp\left(-\left(\frac{k}{\exp(x\beta)}\right)^{\frac{1}{\sigma}}\right)$
EXP	$\frac{1}{\exp(X\beta)} \exp\left(\frac{-y}{\exp(x\beta)}\right)$	$\exp\left(\frac{-k}{\exp(x\beta)}\right)$
FMM SQRT	$\sum_j^2 \pi_i \frac{y^{\alpha_j}}{y\Gamma(\alpha_j)(x\beta_j)^{2\alpha_j}} \exp\left(-\left(\frac{y}{(x\beta_j)^2}\right)\right)$	$\sum_j^2 \pi_i \left(1 - \Gamma(z;\alpha_j)\right)^{***}$ where $z = \frac{k}{(x\beta_j)^2}$

*where $I_Z(p,q) = \frac{1}{B(p,q)}\int_0^Z \frac{t^{p-1}}{(1+t)^{p+q}} dt$ is the incomplete beta function ratio.

**where $\Gamma(z;\kappa^{-2}) = \frac{1}{\Gamma(\kappa^{-2})}\int_0^Z t^{(\kappa^{-2}-1)} \exp(-t) dt$.

***where $\Gamma(z;\alpha_j) = \frac{1}{\Gamma(\alpha_j)}\int_0^Z t^{(\alpha_j-1)} \exp(-t) dt$.

Table 3. Descriptive statistics for total health care expenditures in the second year (2014)

N	12,882,983
Mean	$6,891
SD	$22,963
Skewness	28
Kurtosis	3,072
Maximum	$8,199,457
99th percentile	$80,116
95th percentile	$27,080
90th percentile	$15,384
75th percentile	$5,736
25th percentile	$416
10th percentile	$0
1th percentile	$0
Minimum	$0

Table 4. Performance on convergence and average computing time

Estimator	Number of converged models	Average computing time (second)
NORMAL	100	345
GB2 LOG	99	1,257
GB2 SQRT	99	1,324
GG	99	1,007
GAMMA	99	978
LOGNORM	99	974
WEIB	99	976
EXP	99	764
FMM SQRT	68	19,019
LASSO	100	33
RIDGE	100	33
ENET	100	32
LARS	100	45
NNET	100	501
SINGLE	100	158
SUPER	100	2,067
HH	98	318
FP	97	3,884
CFM	100	76,494

Note: Average computing time represents the time spent on estimating conditional mean expenditures and measuring at the group level, at the tail distribution, and at the individual level.

Figure 1. Example for illustrating how our proposed metric can be complementary to the existing group-level performance metrics

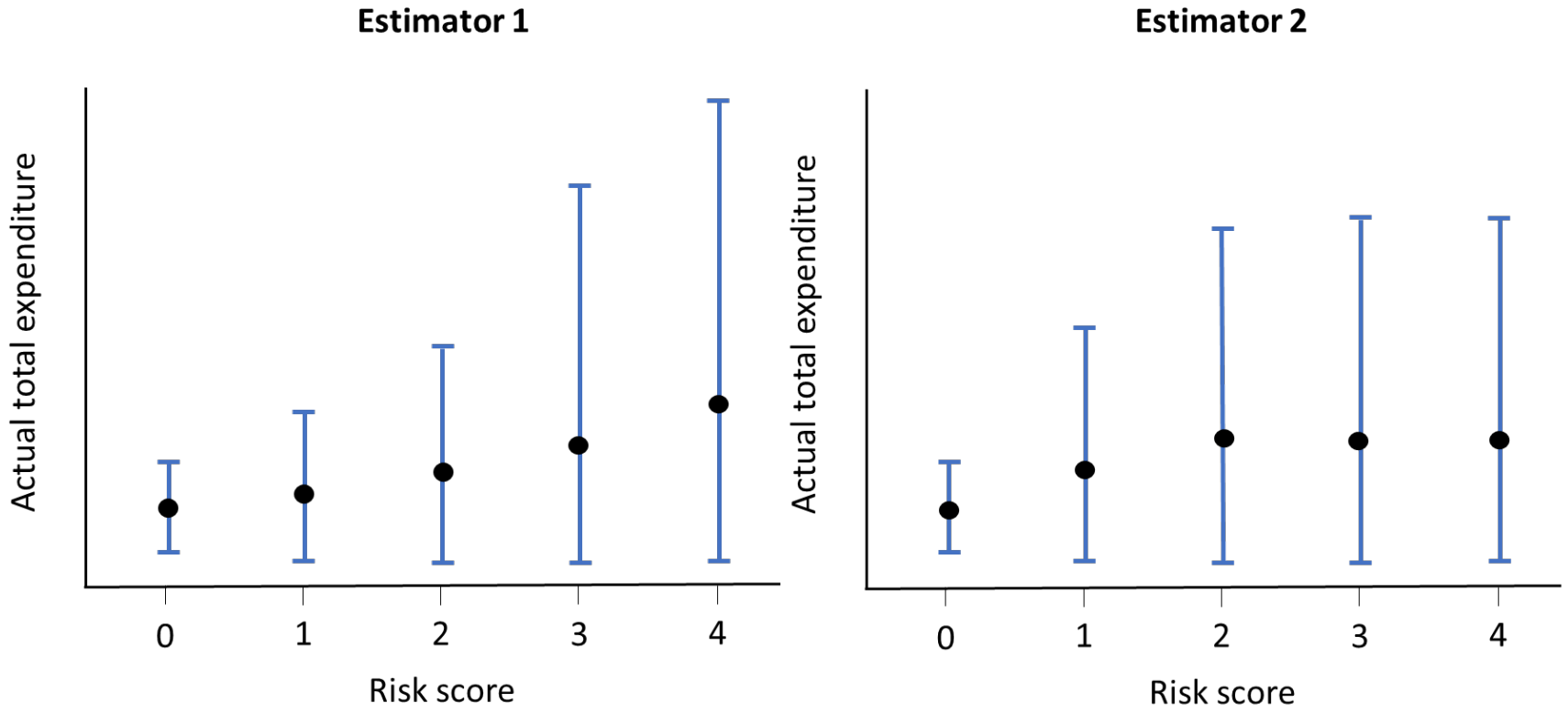
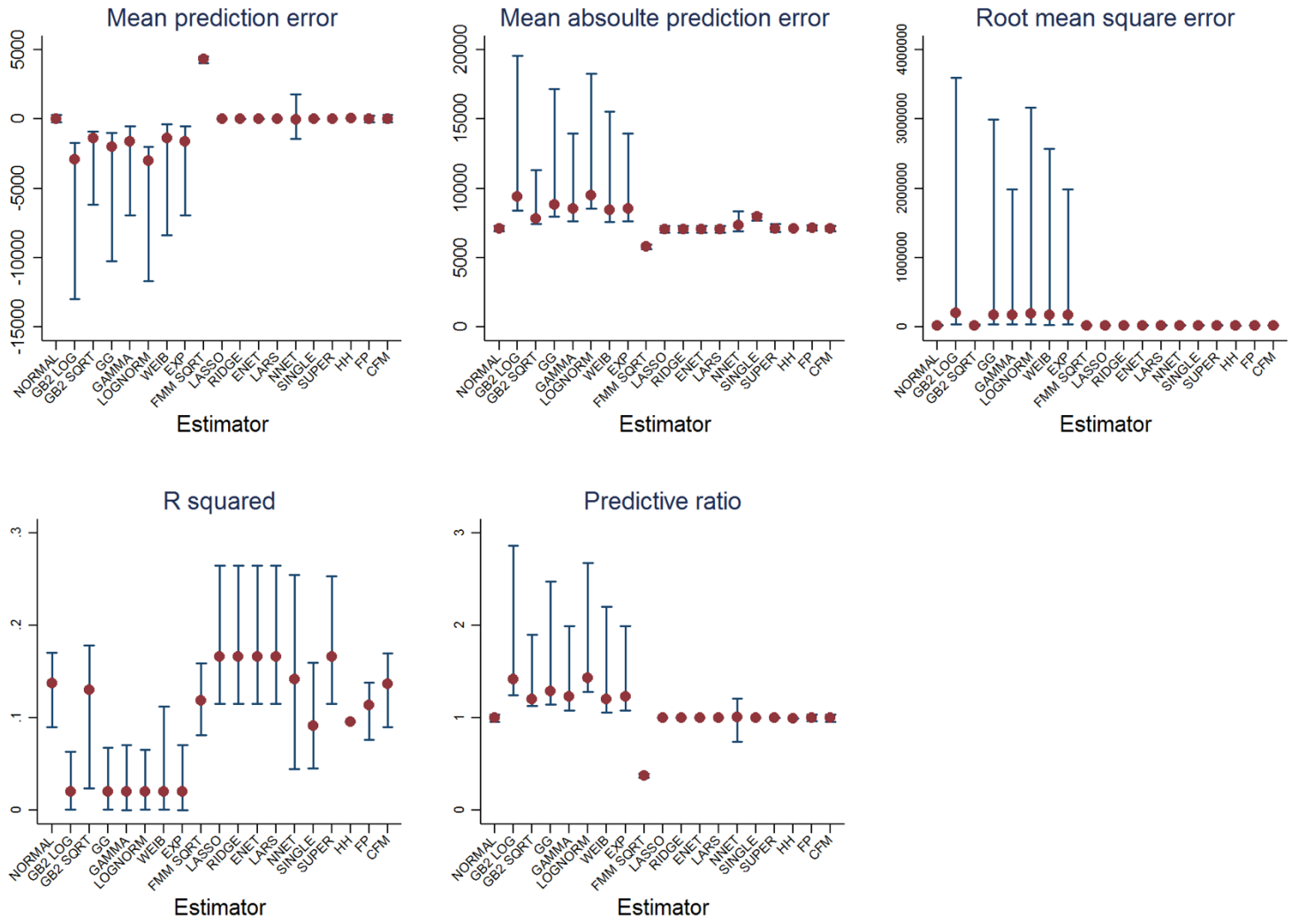
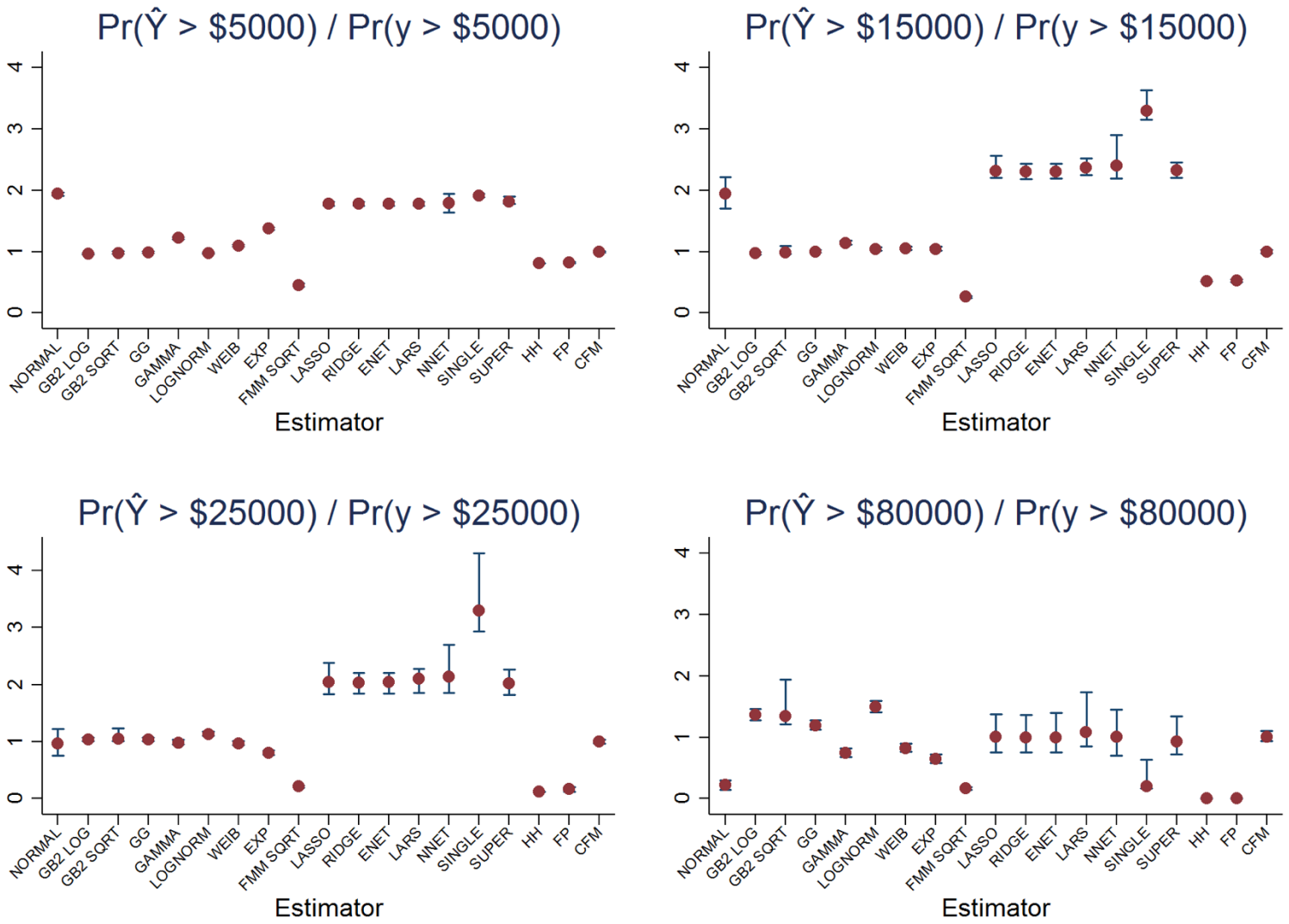


Figure 2. Evaluation of performance based on assessing group-level prediction accuracy of health care expenditures



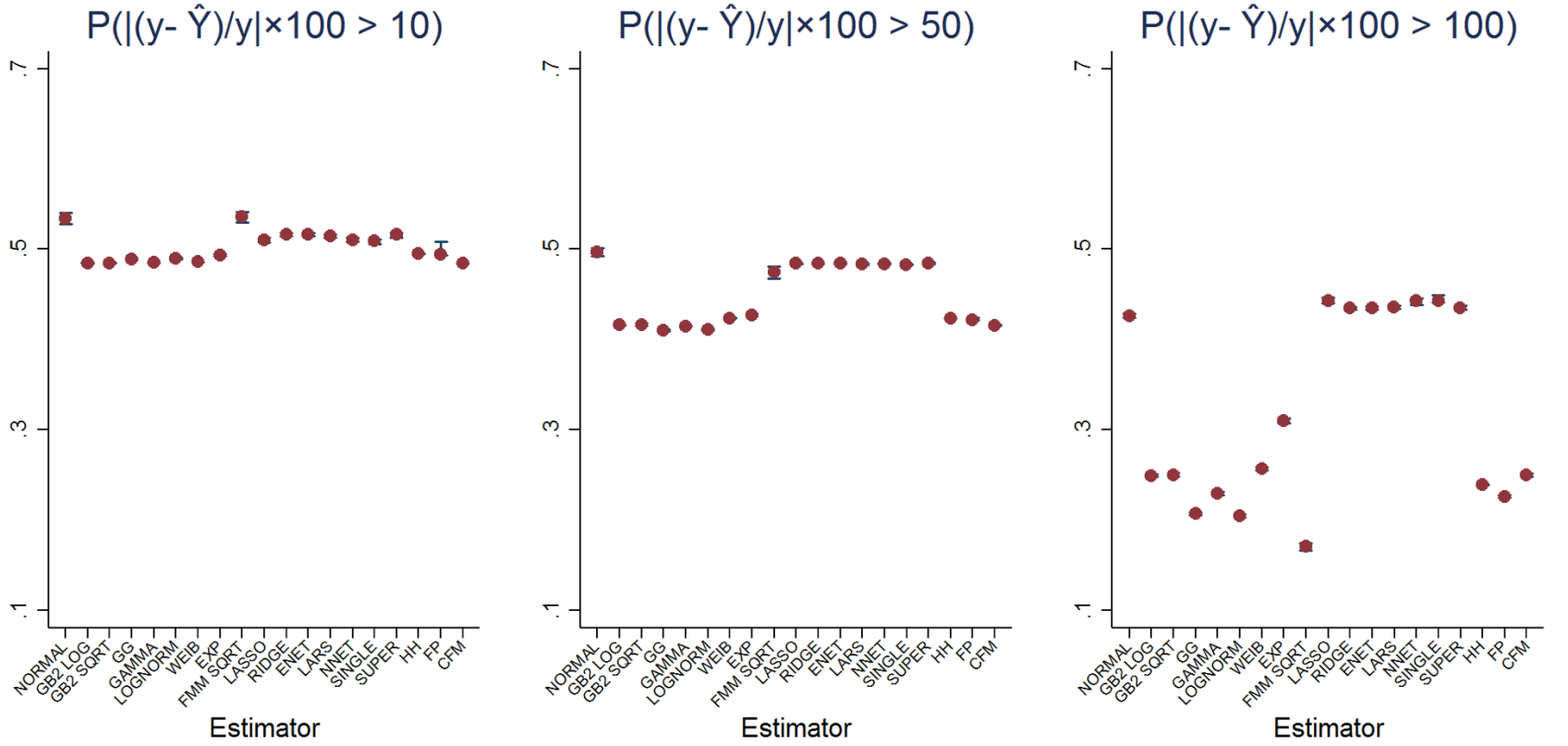
Note: A perfect prediction represents zero for mean prediction error, mean absolute prediction error, and root mean square error and one for R^2 and predictive ratio.

Figure 3. Evaluation of performance based on assessing prediction accuracy for a tail distribution of health care expenditures



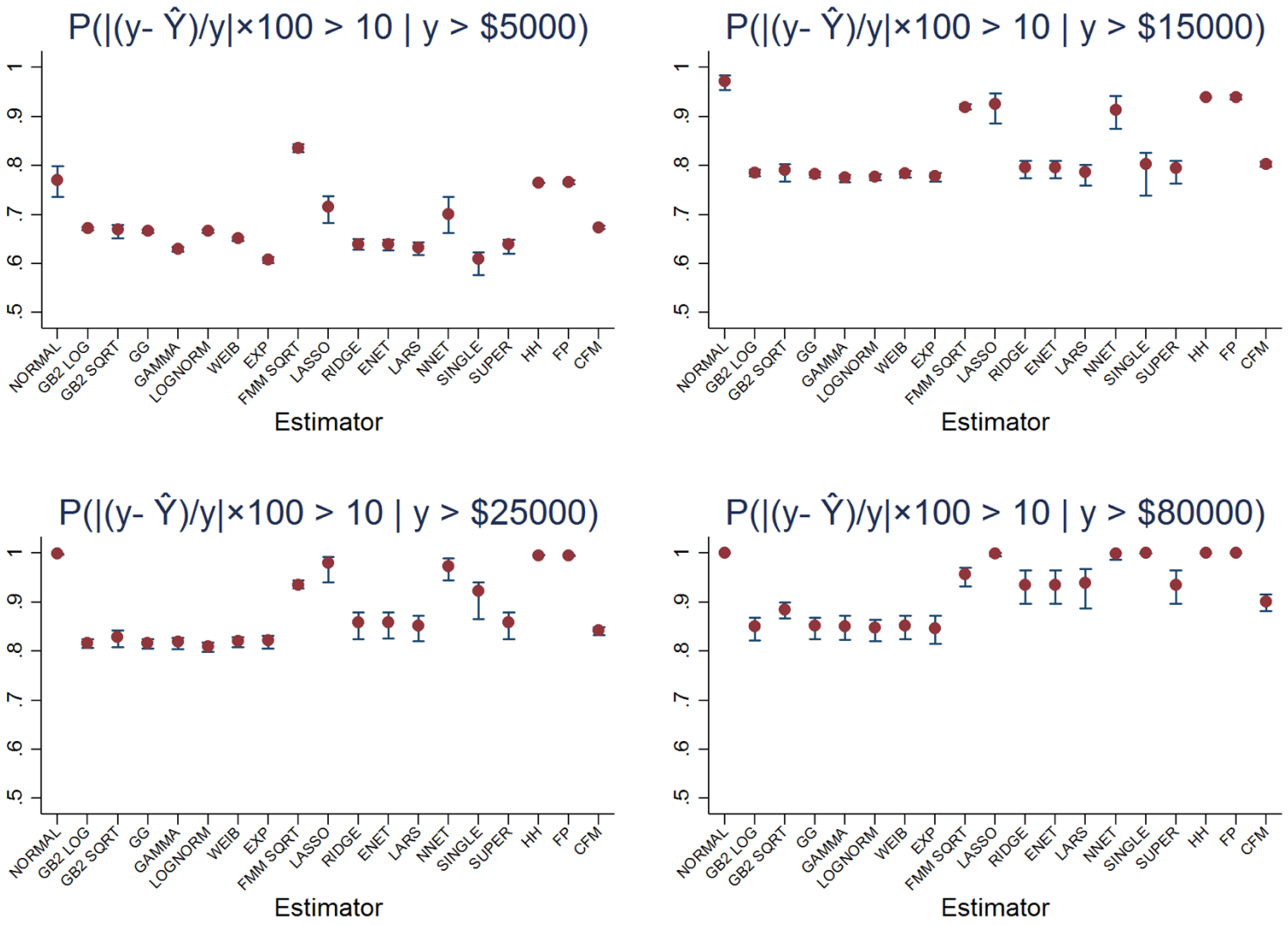
Note: A perfect prediction represents a ratio of one.

Figure 4. Evaluation of performance based on assessing overall individual-level prediction accuracy of health care expenditures



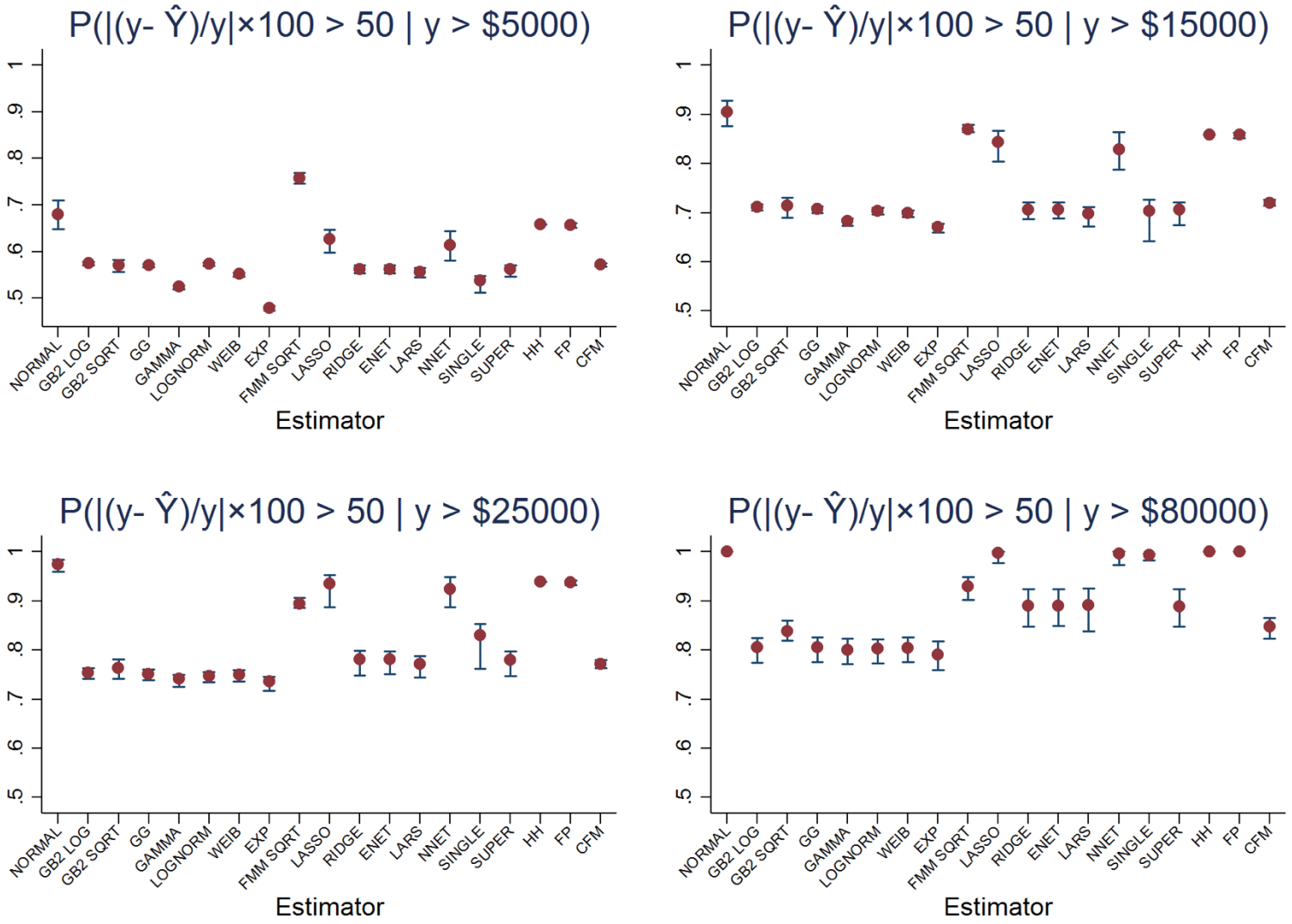
Note: A perfect prediction represents zero probability.

Figure 5. Evaluation of performance based on assessing tail specific individual-level prediction accuracy of health care expenditures (continued)



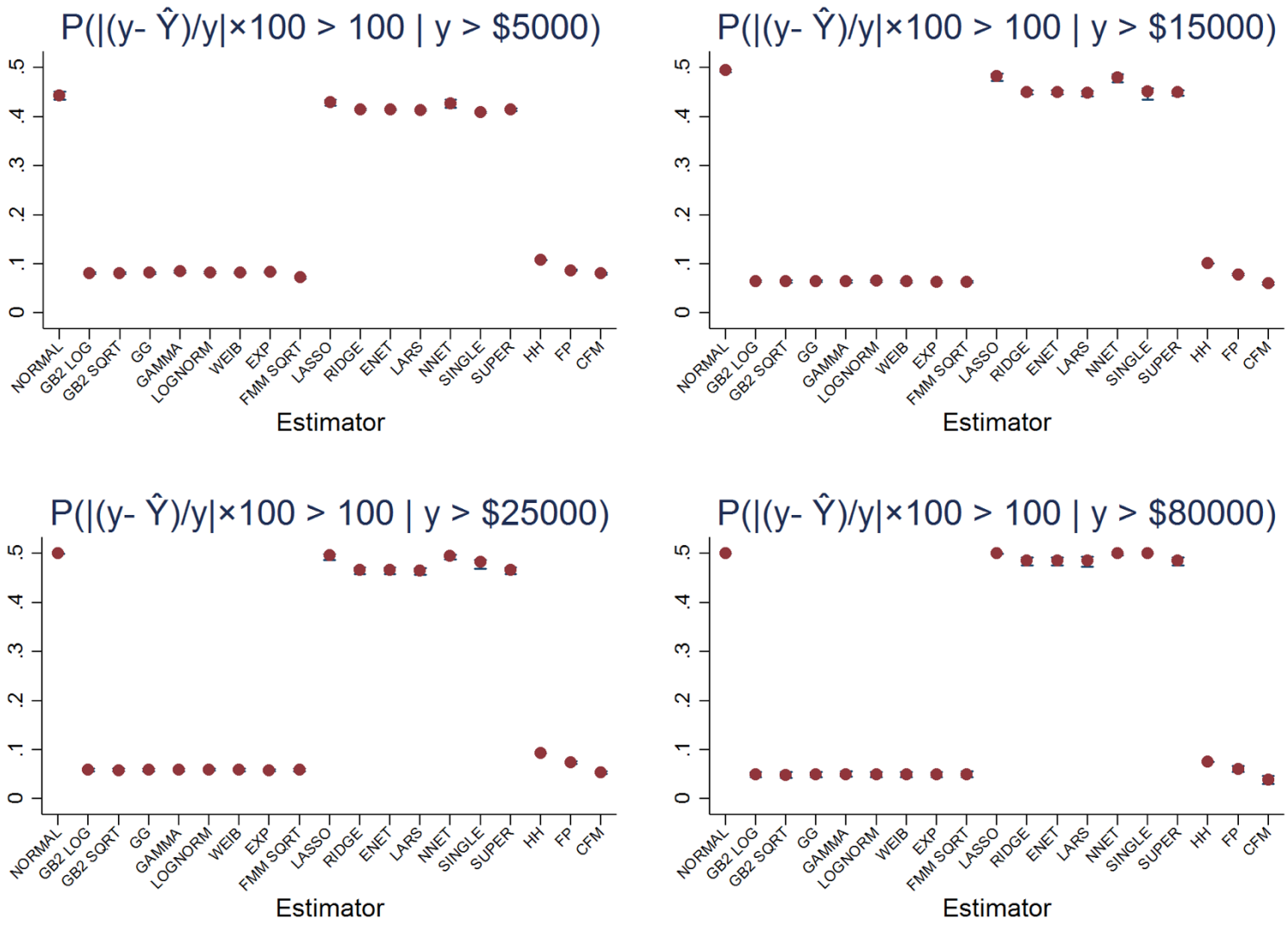
Note: A perfect prediction represents zero probability.

Figure 5. Evaluation of performance based on assessing tail specific individual-level prediction accuracy of health care expenditures (continued)



Note: A perfect prediction represents zero probability.

Figure 5. Evaluation of performance based on assessing tail specific individual-level prediction accuracy of health care expenditures (continued)



Note: A perfect prediction represents zero probability.

Appendix Table A. Descriptive statistics for information on demographics, diagnoses, geographics, and type of plan in the first year (2013) (continued)

Variable	N	%
Demographic information		
Age range 21–24 Male	574,813	4.46
Age range 25–29 Male	356,538	2.77
Age range 30–34 Male	521,852	4.05
Age range 35–39 Male	605,299	4.70
Age range 40–44 Male	725,680	5.63
Age range 45–49 Male	780,948	6.06
Age range 50–54 Male	875,991	6.80
Age range 55–59 Male	881,291	6.84
Age range 60-64 Male	642,682	4.99
Age range 21–24 Female	562,248	4.36
Age range 25–29 Female	393,790	3.06
Age range 30–34 Female	595,064	4.62
Age range 35–39 Female	676,653	5.25
Age range 40–44 Female	795,648	6.18
Age range 45–49 Female	863,752	6.70
Age range 50–54 Female	990,895	7.69
Age range 55–59 Female	1,002,496	7.78
Age range 60-64 Female	734,813	5.70
Diagnostics information		
HCC1: HIV/AIDS	8,406	0.07
HCC2: Septicemia, Sepsis, Systemic Inflammatory Response Syndrome/Shock	26,603	0.21
HCC3: Central Nervous System Infections, Except Viral Meningitis	100,313	0.78
HCC4: Viral or Unspecified Meningitis	310,806	2.41
HCC6: Opportunistic Infections	5,359	0.04
HCC8: Metastatic Cancer	32,642	0.25
HCC9: Lung, Brain, and Other Severe Cancers, Including Pediatric Acute Lymphoid Leukemia	25,538	0.20
HCC10: Non-Hodgkin`s Lymphomas and Other Cancers and Tumors	29,993	0.23
HCC11: Colorectal, Breast (Age < 50), Kidney, and Other Cancers	41,215	0.32
HCC12: Breast (Age 50+) and Prostate Cancer, Benign/Uncertain Brain Tumors, and Other Cancers and Tumors	187,043	1.45
HCC13: Thyroid Cancer, Melanoma, Neurofibromatosis, and Other Cancers and Tumors	50,155	0.39
HCC18: Pancreas Transplant Status/Complications	904	0.01
HCC19: Diabetes with Acute Complications	12,366	0.10
HCC20: Diabetes with Chronic Complications	211,066	1.64
HCC21: Diabetes without Complication	787,087	6.11
HCC23: Protein-Calorie Malnutrition	16,873	0.13
HCC26: Mucopolysaccharidosis	89	0.00
HCC27: Lipidoses and Glycogenosis	4,723	0.04
HCC29: Amyloidosis, Porphyria, and Other Metabolic Disorders	6,229	0.05

Appendix Table A. Descriptive statistics for information on demographics, diagnoses, geographics, and type of plan in the first year (2013) (continued)

Variable	N	%
HCC30: Adrenal, Pituitary, and Other Significant Endocrine Disorders	97,862	0.76
HCC34: Liver Transplant Status/Complications	2,912	0.02
HCC35: End-Stage Liver Disease	8,057	0.06
HCC36: Cirrhosis of Liver	13,052	0.10
HCC37: Chronic Hepatitis	5,850	0.05
HCC38: Acute Liver Failure/Disease, Including Neonatal Hepatitis	379,797	2.95
HCC41: Intestine Transplant Status/Complications	81	0.00
HCC42: Peritonitis/Gastrointestinal Perforation/Necrotizing Enterocolitis	12,180	0.09
HCC45: Intestinal Obstruction	33,328	0.26
HCC46: Chronic Pancreatitis	6,861	0.05
HCC47: Acute Pancreatitis/Other Pancreatic Disorders and Intestinal Malabsorption	70,369	0.55
HCC48: Inflammatory Bowel Disease	78,879	0.61
HCC54: Necrotizing Fasciitis	736	0.01
HCC55: Bone/Joint/Muscle Infections/Necrosis	38,359	0.30
HCC56: Rheumatoid Arthritis and Specified Autoimmune Disorders	129,261	1.00
HCC57: Systemic Lupus Erythematosus and Other Autoimmune Disorders	68,878	0.53
HCC61: Osteogenesis Imperfecta and Other Osteodystrophies	1,453	0.01
HCC62: Congenital/Developmental Skeletal and Connective Tissue Disorders	6,934	0.05
HCC63: Cleft Lip/Cleft Palate	621	0.00
HCC64: Major Congenital Anomalies of Diaphragm, Abdominal Wall, and Esophagus, Age < 2	1,465	0.01
HCC66: Hemophilia	2,341	0.02
HCC67: Myelodysplastic Syndromes and Myelofibrosis	1,659	0.01
HCC68: Aplastic Anemia	3,377	0.03
HCC69: Acquired Hemolytic Anemia, Including Hemolytic Disease of Newborn	1,230	0.01
HCC70: Sickle Cell Anemia (Hb-SS)	815	0.01
HCC71: Thalassemia Major	1,232	0.01
HCC73: Combined and Other Severe Immunodeficiencies	28,505	0.22
HCC74: Disorders of the Immune Mechanism	77,681	0.60
HCC75: Coagulation Defects and Other Specified Hematological Disorders	12,674	0.10
HCC81: Drug Psychosis	51,187	0.40
HCC82: Drug Dependence	12,616	0.10
HCC87: Schizophrenia	408,794	3.17
HCC88: Major Depressive and Bipolar Disorders	12,218	0.09
HCC89: Reactive and Unspecified Psychosis, Delusional Disorders	6,087	0.05
HCC90: Personality Disorders	5,171	0.04
HCC94: Anorexia/Bulimia Nervosa	777	0.01
HCC96: Prader-Willi, Patau, Edwards, and Autosomal Deletion Syndromes	6,469	0.05
HCC97: Down Syndrome, Fragile X, Other Chromosomal Anomalies, and Congenital Malformation Syndromes	3,181	0.02

Appendix Table A. Descriptive statistics for information on demographics, diagnoses, geographics, and type of plan in the first year (2013) (continued)

Variable	N	%
HCC102: Autistic Disorder	2,212	0.02
HCC103: Pervasive Developmental Disorders, Except Autistic Disorder	8,057	0.06
HCC106: Traumatic Complete Lesion Cervical Spinal Cord	99	0.00
HCC107: Quadriplegia	2,921	0.02
HCC108: Traumatic Complete Lesion Dorsal Spinal Cord	107	0.00
HCC109: Paraplegia	2,982	0.02
HCC110: Spinal Cord Disorders/Injuries	14,097	0.11
HCC111: Amyotrophic Lateral Sclerosis and Other Anterior Horn Cell Disease	1,956	0.02
HCC112: Quadriplegic Cerebral Palsy	888	0.01
HCC113: Cerebral Palsy, Except Quadriplegic	4,620	0.04
HCC114: Spina Bifida and Other Brain/Spinal/Nervous System Congenital Anomalies	5,535	0.04
HCC115: Myasthenia Gravis/Myoneural Disorders and Guillain-Barre Syndrome/Inflammatory and Toxic Neuropathy	19,580	0.15
HCC117: Muscular Dystrophy	2,124	0.02
HCC118: Multiple Sclerosis	37,192	0.29
HCC119: Parkinson`s, Huntington`s, and Spinocerebellar Disease, and Other Neurodegenerative Disorders	9,857	0.08
HCC120: Seizure Disorders and Convulsions	91,604	0.71
HCC121: Hydrocephalus	5,098	0.04
HCC122: Non-Traumatic Coma, and Brain Compression/Anoxic Damage	7,905	0.06
HCC125: Respirator Dependence/Tracheostomy Status	2,598	0.02
HCC126: Respiratory Arrest	778	0.01
HCC127: Cardio-Respiratory Failure and Shock, Including Respiratory Distress Syndromes	39,846	0.31
HCC128: Heart Assistive Device/Artificial Heart	281	0.00
HCC129: Heart Transplant	1,038	0.01
HCC130: Congestive Heart Failure	116,539	0.90
HCC131: Acute Myocardial Infarction	17,788	0.14
HCC132: Unstable Angina and Other Acute Ischemic Heart Disease	36,421	0.28
HCC135: Heart Infection/Inflammation, Except Rheumatic	31,422	0.24
HCC142: Specified Heart Arrhythmias	157,683	1.22
HCC145: Intracranial Hemorrhage	8,614	0.07
HCC146: Ischemic or Unspecified Stroke	28,897	0.22
HCC149: Cerebral Aneurysm and Arteriovenous Malformation	8,380	0.07
HCC150: Hemiplegia/Hemiparesis	10,364	0.08
HCC151: Monoplegia, Other Paralytic Syndromes	2,810	0.02
HCC153: Atherosclerosis of the Extremities with Ulceration or Gangrene	4,400	0.03
HCC154: Vascular Disease with Complications	15,929	0.12
HCC156: Pulmonary Embolism and Deep Vein Thrombosis	66,726	0.52
HCC158: Lung Transplant Status/Complications	640	0.00
HCC159: Cystic Fibrosis	2,032	0.02

Appendix Table A. Descriptive statistics for information on demographics, diagnoses, geographics, and type of plan in the first year (2013) (continued)

Variable	N	%
HCC160: Chronic Obstructive Pulmonary Disease, Including Bronchiectasis	181,447	1.41
HCC161: Asthma	490,596	3.81
HCC162: Fibrosis of Lung and Other Lung Disorders	43,164	0.34
HCC163: Aspiration and Specified Bacterial Pneumonias and Other Severe Lung Infections	11,800	0.09
HCC183: Kidney Transplant Status	10,421	0.08
HCC184: End Stage Renal Disease	8,614	0.07
HCC187: Chronic Kidney Disease, Stage 5	3,730	0.03
HCC188: Chronic Kidney Disease, Severe (Stage 4)	5,967	0.05
HCC203: Ectopic and Molar Pregnancy, Except with Renal Failure, Shock, or Embolism	7,102	0.06
HCC204: Miscarriage with Complications	2,417	0.02
HCC205: Miscarriage with No or Minor Complications	34,611	0.27
HCC207: Completed Pregnancy With Major Complications	8,531	0.07
HCC208: Completed Pregnancy With Complications	97,301	0.76
HCC209: Completed Pregnancy with No or Minor Complications	103,090	0.80
HCC217: Chronic Ulcer of Skin, Except Pressure	41,028	0.32
HCC226: Hip Fractures and Pathological Vertebral or Humerus Fractures	11,438	0.09
HCC227: Pathological Fractures, Except of Vertebrae, Hip, or Humerus	4,491	0.03
HCC251: Stem Cell, Including Bone Marrow, Transplant Status/Complications	2,863	0.02
HCC253: Artificial Openings for Feeding or Elimination	17,054	0.13
HCC254: Amputation Status, Lower Limb/Amputation Complications	3,115	0.02
Severe illness indicator × HCC6	975	0.01
Severe illness indicator × HCC8	4,664	0.04
Severe illness indicator × HCC9	6,201	0.05
Severe illness indicator × HCC10	3,189	0.02
Severe illness indicator × HCC35	1,636	0.01
Severe illness indicator × HCC38	9,998	0.08
Severe illness indicator × HCC115	2,166	0.02
Severe illness indicator × HCC135	3,500	0.03
Severe illness indicator × HCC145	3,326	0.03
Severe illness indicator × HCC153	1,111	0.01
Severe illness indicator × HCC154	4,147	0.03
Severe illness indicator × HCC163	4,969	0.04
Severe illness indicator × HCC253	5,145	0.04
Severe illness indicator × aggregate HCC grouping G3	5,111	0.04
Severe illness indicator × aggregate HCC grouping G6	551	0.00
Severe illness indicator × aggregate HCC grouping G8	2,945	0.02
Geography		
Unknown region	401,994	3.12
Connecticut	278,352	2.16

Appendix Table A. Descriptive statistics for information on demographics, diagnoses, geographics, and type of plan in the first year (2013) (continued)

Variable	N	%
Maine	83,696	0.65
Massachusetts	208,692	1.62
New Hampshire	91,048	0.71
Rhode Island	28,859	0.22
Vermont	7,572	0.06
New Jersey	300,922	2.34
New York	1,301,026	10.1
Pennsylvania	420,697	3.27
Illinois	352,035	2.73
Indiana	448,742	3.48
Michigan	524,102	4.07
Ohio	702,797	5.46
Wisconsin	217,493	1.69
Iowa	69,598	0.54
Kansas	56,911	0.44
Minnesota	91,151	0.71
Missouri	224,273	1.74
Nebraska	32,248	0.25
North Dakoda	5,572	0.04
South Dakoda	9,201	0.07
Washington DC	6,535	0.05
Delaware	74,046	0.57
Florida	614,139	4.77
Georgia	432,233	3.36
Maryland	103,524	0.80
North Carolina	271,103	2.10
South Carolina	549,128	4.26
Virginia	245,991	1.91
West Virginia	45,557	0.35
Alabama	176,444	1.37
Kentucky	205,148	1.59
Mississippi	168,472	1.31
Tennessee	322,453	2.50
Arkansas	69,930	0.54
Louisiana	490,358	3.81
Oklahoma	80,544	0.63
Texas	713,662	5.54
Arizona	131,802	1.02
Colorado	154,451	1.20
Idaho	132,555	1.03
Montana	10,126	0.08
Montana	71,751	0.56
New Mexico	114,215	0.89

Appendix Table A. Descriptive statistics for information on demographics, diagnoses, geographics, and type of plan in the first year (2013) (continued)

Variable	N	%
Utah	77,082	0.60
Wyoming	8,163	0.06
Alaska	10,637	0.08
California	1,335,884	10.37
Hawaii	1,968	0.02
Oregon	134,353	1.04
Washington	267,975	2.08
Puerto Rico	5,773	0.04
Type of plan		
Comprehensive	287,861	2.23
Health maintenance organization	1,720,222	13.35
Point of service	995,744	7.73
Preferred provider organization	9,798,331	76.06
Point of service with capitation	80,825	0.63

Note: To predict total health care expenditures in 2014, we use the risk-adjusters of the Department of Health and Human Services-Hierarchical Condition Categories model developed for the adult population of the 2014 benefit year.

Appendix Table B. Rankings of estimators in terms of bias (continued)

	Prediction performance metrics								
	Group-level prediction accuracy					Prediction accuracy for tail distributions: $\frac{P(\hat{Y}_i>k)}{P(y_i>k)}$			
Ranking	Mean prediction error	Mean absolute prediction error	Root mean square error	R^2	Predictive ratio	k = \$5,000	k = \$15,000	k = \$25,000	k = \$80,000
1	LARS	FMM SQRT	LARS	LARS	LASSO	CFM	CFM	CFM	LASSO
2	LASSO	LARS	ENET	ENET	RIDGE	GG	GG	GAMMA	CFM
3	ENET	ENET	LASSO	LASSO	SINGLE	GB2 SQRT	GB2 SQRT	GB2 LOG	ENET
4	RIDGE	LASSO	RIDGE	RIDGE	ENET	LOGNORM	GB2 LOG	WEIB	RIDGE
5	SINGLE	RIDGE	SUPER	SUPER	LARS	GB2 LOG	LOGNORM	GG	NNET
6	SUPER	HH	NNET	NNET	SUPER	WEIB	EXP	NORMAL	SUPER
7	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL	FP	WEIB	GB2 SQRT	LARS
8	CFM	CFM	CFM	CFM	CFM	HH	GAMMA	LOGNORM	WEIB
9	FP	SUPER	FP	GB2 SQRT	FP	GAMMA	FP	EXP	GG
10	HH	FP	SINGLE	FMM SQRT	HH	EXP	HH	FMM SQRT	GAMMA
11	NNET	NNET	FMM SQRT	FP	NNET	FMM SQRT	FMM SQRT	FP	GB2 SQRT
12	WEIB	GB2 SQRT	GB2 SQRT	HH	WEIB	ENET	NORMAL	HH	EXP
13	GB2 SQRT	SINGLE	HH	SINGLE	GB2 SQRT	LASSO	ENET	SUPER	GB2 LOG
14	GAMMA	WEIB	GAMMA	GG	GAMMA	RIDGE	RIDGE	RIDGE	LOGNORM
15	EXP	GAMMA	EXP	LOGNORM	EXP	LARS	LASSO	ENET	NORMAL
16	GG	EXP	WEIB	WEIB	GG	NNET	SUPER	LASSO	SINGLE
17	GB2 LOG	GG	GG	GB2 LOG	GB2 LOG	SUPER	LARS	LARS	FMM SQRT
18	LOGNORM	GB2 LOG	LOGNORM	GAMMA	LOGNORM	SINGLE	NNET	NNET	FP
19	FMM SQRT	LOGNORM	GB2 LOG	EXP	FMM SQRT	NORMAL	SINGLE	SINGLE	HH

Note: Higher ranking (lower number) represents better performance. Estimators inside boarders indicate the same ranking.

Appendix Table B. Rankings of estimators in terms of bias (continued)

	Prediction performance metrics						
	Overall individual-level prediction accuracy: $P\left(\frac{\lvert y_i-\hat{Y}_i\rvert}{y_i}\times 100 > l\right)$			Tail-specific individual-level prediction accuracy: $P\left(\frac{\lvert y_i-\hat{Y}_i\rvert}{y_i}\times 100 > l \lvert y_i > k\right)$			
	l = 10	l = 50	l = 100	l = 10			
Ranking				k = \$5,000	k = \$15,000	k = \$25,000	k = \$80,000
1	CFM	GG	FMM SQRT	EXP	GAMMA	LOGNORM	EXP
2	GB2 SQRT	LOGNORM	LOGNORM	SINGLE	LOGNORM	GG	LOGNORM
3	GB2 LOG	GAMMA	GG	GAMMA	EXP	GB2 LOG	GAMMA
4	GAMMA	CFM	FP	LARS	GG	GAMMA	GB2 LOG
5	WEIB	GB2 SQRT	GAMMA	SUPER	WEIB	WEIB	GG
6	GG	GB2 LOG	HH	ENET	GB2 LOG	EXP	WEIB
7	LOGNORM	FP	GB2 LOG	RIDGE	LARS	GB2 SQRT	GB2 SQRT
8	EXP	WEIB	GB2 SQRT	WEIB	GB2 SQRT	CFM	CFM
9	FP	HH	CFM	LOGNORM	SUPER	LARS	SUPER
10	HH	EXP	WEIB	GG	ENET	SUPER	ENET
11	SINGLE	FMM SQRT	EXP	GB2 SQRT	RIDGE	ENET	RIDGE
12	NNET	SINGLE	NORMAL	GB2 LOG	CFM	RIDGE	LARS
13	LASSO	LARS	ENET	CFM	SINGLE	SINGLE	FMM SQRT
14	LARS	NNET	RIDGE	NNET	NNET	FMM SQRT	NNET
15	SUPER	LASSO	SUPER	LASSO	FMM SQRT	NNET	LASSO
16	RIDGE	SUPER	LARS	HH	LASSO	LASSO	SINGLE
17	ENET	RIDGE	LASSO	FP	FP	FP	NORMAL
18	NORMAL	ENET	NNET	NORMAL	HH	HH	FP
19	FMM SQRT	NORMAL	SINGLE	FMM SQRT	NORMAL	NORMAL	HH

Note: Higher ranking (lower number) represents better performance. Estimators inside boarders indicate the same ranking.

Appendix Table B. Rankings of estimators in terms of bias (continued)

	Prediction performance metrics							
	Tail-specific individual-level prediction accuracy: $P\left(\frac{\lvert y_i-\hat{Y}_i\rvert}{y_i}\times 100 > l \lvert y_i > k\right)$							
	l = 50				l = 100			
Ranking	k = \$5,000	k = \$15,000	k = \$25,000	k = \$80,000	k = \$5,000	k = \$15,000	k = \$25,000	k = \$80,000
1	EXP	EXP	EXP	EXP	FMM SQRT	CFM	CFM	CFM
2	GAMMA	GAMMA	GAMMA	GAMMA	CFM	EXP	EXP	GB2 SQRT
3	SINGLE	LARS	LOGNORM	LOGNORM	GB2 SQRT	FMM SQRT	GB2 SQRT	EXP
4	WEIB	WEIB	WEIB	WEIB	GB2 LOG	WEIB	WEIB	FMM SQRT
5	LARS	SINGLE	GG	GB2 LOG	GG	GAMMA	FMM SQRT	WEIB
6	SUPER	LOGNORM	GB2 LOG	GG	WEIB	GB2 SQRT	GAMMA	GG
7	ENET	SUPER	GB2 SQRT	GB2 SQRT	LOGNORM	GG	GG	GB2 LOG
8	RIDGE	ENET	CFM	CFM	EXP	GB2 LOG	GB2 LOG	LOGNORM
9	GG	RIDGE	LARS	SUPER	GAMMA	LOGNORM	LOGNORM	GAMMA
10	GB2 SQRT	GG	SUPER	RIDGE	FP	FP	FP	FP
11	CFM	GB2 LOG	ENET	ENET	HH	HH	HH	HH
12	LOGNORM	GB2 SQRT	RIDGE	LARS	SINGLE	LARS	LARS	SUPER
13	GB2 LOG	CFM	SINGLE	FMM SQRT	LARS	SUPER	SUPER	ENET
14	NNET	NNET	FMM SQRT	SINGLE	SUPER	ENET	ENET	RIDGE
15	LASSO	LASSO	NNET	NNET	ENET	RIDGE	RIDGE	LARS
16	FP	FP	LASSO	LASSO	RIDGE	SINGLE	SINGLE	NNET
17	HH	HH	FP	NORMAL	NNET	NNET	NNET	LASSO
18	NORMAL	FMM SQRT	HH	HH	LASSO	LASSO	LASSO	SINGLE
19	FMM SQRT	NORMAL	NORMAL	FP	NORMAL	NORMAL	NORMAL	NORMAL

Note: Higher ranking (lower number) represents better performance. Estimators inside boarders indicate the same ranking.

Chapter 4.

IMPROVING PLAN PAYMENT RISK ADJUSTMENT WITH MACHINE LEARNING: ACCOUNTING FOR SERVICE-LEVEL PROPENSITY SCORES TO REDUCE SERVICE-LEVEL SELECTION

Abstract

Objective. To propose an alternative risk-adjustment model accounting for service-level propensity scores with machine learning and then compare its prediction accuracy of estimating plan payments with that of the currently used Hierarchical Condition Category (HCC) model.

Data Sources. 2013–2014 Truven MarketScan database.

Study Design. We implemented the alternative model in a two-step process. First, we used the generalized boosting model to predict the probabilities that each enrollee would use each service based on her demographic and diagnostic characteristics (service-level propensity scores). Second, we incorporated the service-level propensity scores into the HCC model. Then, we performed a paired performance comparison of the alternative model and the HCC model across 19 estimators: nine parametric estimators, seven machine learning estimators, and three distributional estimators. Prediction accuracies were evaluated at the three level: group level, tail distribution, and individual level.

Principal Findings. The alternative model more accurately estimated plan payments when combined with machine learning estimators, especially for enrollees with high expenditures. However, negligible improvements were observed in parametric and distributional estimators.

Conclusions. Accounting for the service-level propensity scores in risk adjustment with machine learning has the potential to reduce incentives to avoid enrollees with higher expenditures than their estimated plan payments.

Key words. risk adjustment, machine learning, service-level selection, service-level propensity scores

4.1 INTRODUCTION

There is a long-standing concern about risk selection in the health insurance market. Risk selection occurs because health plans cannot vary premiums due to government regulation policies. In the regulated market, risk selection can be mitigated by adjusting payments to health plans to more accurately reflect beneficiary health status, a process known as *risk adjustment*. However, the goal of risk adjustment is not merely to adjust plan payments as close to the actual expenditures in a statistical sense. If risk adjustment is aimed at achieving purely statistical objects, it would not sufficiently generate economic incentives for health plans to discourage risk selection, potentially leading to health plan's strategic behavior to avoid enrollees who could be unprofitable (i.e., those with higher expenditures that their plan payments) (Einav et al. 2016; Glazer and McGuire 2000). Thus, optimal risk adjustment needs to account for statistical as well as economic forces to prevent health plans' strategic behaviors in a particular context and objective.

In this study, we focused on the risk-adjustment model used in the Medicare Advantage (MA) program and the Exchanges. Since 2004, the Centers for Medicare and Medicaid Services (CMS) has used the CMS-Hierarchical Condition Categories (HCC) model to adjust payments for Medicare enrollees (Pope et al. 2004). Following the CMS-HCC model, the Department of Health and Human Services (HHS) developed a federally-certified risk-adjustment model to be used by states or by HHS on behalf of states, known as the HHS-HCC model (Kautter et al. 2014). These HCC models use inpatient and outpatient diagnostic information to generate risk scores. Using the risk scores, the risk-adjusted expenditures of each enrollee are estimated. Finally, health plans receive the risk-adjusted expenditures as plan payments.

However, there are two limitations of the HCC models. First, the HCC models use linear regression to estimate health care expenditures. Despite its practical features, linear regression

does not fit well to a highly skewed distribution of health care expenditure. Also, it cannot capture complex nonlinear relationships and interactions between variables. Hence, the HCC models based on linear regression tend to insufficiently capture health conditions for enrollees with high-risk scores, who tend to have complex health conditions, thereby leading under-prediction of their plan payments (Kautter et al. 2014; Medicare Payment Advisory Commission 2012, 2014; Pope et al. 2004). Second, the HCC models are designed to estimate accurate payments at the group level. The HCC models pay the same rate for all enrollees with a given health condition, but health care expenditures can vary by the severity of the condition (Medicare Payment Advisory Commission 2012). This creates substantial variability in actual expenditures of enrollees around their risk-adjusted payments. Moreover, the variability of the within-risk-score expenditures is larger for those with higher risk scores (Manning, Basu, and Mullahy 2005). Thus, the HCC models focused on group-level accuracy generate large prediction inaccuracy at the individual level, especially for those with high-risk scores.

These suggest that health plans have incentives to strategically respond to the HCC models to avoid unprofitable enrollees, especially for those with high-risk scores. Park et al. (2017) found that MA plans effectuated *service-level selection* as a strategic behavior in response to the CMS-HCC model to avoid unprofitable enrollees. Specifically, MA plans induced unprofitable enrollees to voluntarily switch from MA plans to traditional Medicare (TM) by raising copays disproportionately more for services needed by them. Another study found that MA enrollees switching to TM used high-cost services (e.g., nursing home and home health care) more frequently than those continuously staying in MA plans (Rahman et al. 2015). Hence, the switching rates from MA plans to TM varied greatly by disease after the implementation of the CMS-HCC model (Park et al. 2018). Evidence on service-level selection was found in the Exchanges, showing

potential under-provision of care for cancer mental health and substance abuse (Barry et al. 2012; McGuire et al. 2014; Montz et al. 2016; Shrestha et al. 2017; Weiner et al. 2012).

There is an approach to identifying an alternative risk-adjustment model by moving beyond the estimation and evaluation methodology used in the HCC models. As described above, the HCC models only rely on linear regression and prediction accuracy is merely evaluated at the group level, thereby incompletely capturing health status and thus inaccurately estimating plan payments at the individual level. To identify the best model in a *statistical* sense within the HCC framework, Park and Basu (2018) applied the HCC model to 19 estimators used to analyze health care expenditures in the previous literature (Jones, Lomas, and Rice 2015; Rose 2016) and then examined their prediction accuracies not only at the group level but also at the tail distributions and individual level. The idea behind this approach is that if there is a statistical model that outperforms others in all prediction accuracies, this model would be the best risk-adjustment model from the statistical perspective. However, they found that no one estimator performed best in all prediction accuracies. This finding suggests that an optimal risk-adjustment model cannot be determined solely based on the statistical perspective, but rather it needs to account for statistical as well as economic forces to prevent service-level selection.

In this study, we proposed an alternative risk-adjustment model in both a statistical and economic sense. Using machine learning algorithms, we developed the alternative model that not only conditions on each enrollee's demographics and diagnoses but also reflects each enrollee's service-level propensity scores. To determine which estimator is the best to reduce service-level selection, next, we performed a paired comparison of the alternative model and the HCC model across 19 estimators at the three levels: group level, tail distributions, and individual level.

4.2 METHOD

4.2.1 *Data and Study Sample*

We used the Truven MarketScan Commercial Claims and Encounter database, which was used to develop and evaluate the HCC model for the individuals and small group markets established by the ACA (Kautter et al. 2014). The database contains inpatient, outpatient, and prescription drug claims for employees, spouses, and dependents covered by employer-sponsored private health plans. Using the database between 2013-2014, we identified a sample of working-age adults (aged 21 to 64) continuously enrolled in Comprehensive, Health Maintenance Organization (HMO), Preferred Provider Organization (PPO), Point of Service (POS), or POS with capitation during 2013-2014.

4.2.2 *Study Design*

Drawing a one percent random sample of the study population, we employed a quasi-Monte-Carlo design where each estimator was fitted to the first-year's individual-level characteristics of estimations sets and its prediction accuracy was assessed on the second-year's total health care expenditures of validation sets. To fit an estimator, we used two risk-adjustment models: 1) the HCC model adjusting for demographic and diagnostic characteristics and 2) the alternative model adjusting for the demographic and diagnostic characteristics as well as service-level propensity scores. Using each of the 19 estimators, we estimated two sets of $\widehat{y_i}$ and $\widehat{Y_i}$, one of which was estimated based on the HCC model and the other was based on the alternative model. $\widehat{y_i}$ indicates an enrollee i's predicted conditional mean. $\widehat{Y_i}$ indicates a forecast of individual-level expenditure, which is estimated from the predicted conditional probability density function as a stochastic

quantity. To distinguish $\hat{Y}_i$ from $\widehat{y_i}$, we referred to $\widehat{y_i}$ as a "prediction" of her expenditure, and $\hat{Y}_i$ as a "forecast" of her expenditure. We repeated the process described above 100 times while sampling with replacement. Using the estimated $\widehat{y_i}$ and $\hat{Y}_i$, we calculated each estimator's performance metrics for measuring prediction accuracies at the group level, at the tail distributions, and at the individual level. Finally, we calculated the average for each performance metric across these iterations.

4.2.3 *HCC Model and Alternative Model*

We used two risk-adjustment models. First, we used the HHS-HCC model developed for the adult population of the 2014 benefit year (Centers for Medicaid and Medicare Services 2014). We included 18 age-sex categories, 114 HCCs, and 16 interactions between disease groups. HCCs are mapped from individual diagnoses from five-digit *International Classification of Diseases* (ICD)-9 codes to a much smaller number of categories. Each HCC is selected to be clinically meaningful and have substantial cost implications. Using the HCC model described above, we estimated $\widehat{y_i}$ and $\hat{Y}_i$. However, there are several differences between the original HHS-HCC model and the HHC model used in this study. For example, to control for differences in care management between plan type and differences in health care supply between states, respectively, we also included five types of plan (Comprehensive, HMO, PPO, POS, or POS with capitation) and 53 state categories. We refer to Park and Basu (2018) for other differences and justification.

Second, we used the alternative risk-adjustment model accounting for service-level propensity scores. We followed a two-step process. We predicted the probabilities that each enrollee would use each type of service in the second year based on her first-year's demographic and diagnostic characteristics. The rationale for this approach is that the probabilities rely solely

on individual-level characteristics on demographics and diagnoses, not utilization, and thus it is unlikely to induce an incentive for health plans to provide more services regardless of clinical need, which is consistent with the principles of risk adjustment (Medicare Payment Advisory Commission 2011; Pope et al. 2011). Following Ellis, Martins, and Zhu (2017), all services were classified into mutually exclusive 33 types of services based on the MarketScan's definitions of service location and procedures. To better predict the service-level propensity scores for each type of service, we used the 148 demographic and diagnostic variables described above as well as a (standardized) risk score calculated from the HCC model and interactions between the risk score and each of the demographic and diagnostic variables. To predict the service-level propensity scores, we used the generalized boosting model (GBM). We chose GBM because it is an ensemble approach where new models are added to correct the errors made by existing models. Models are added sequentially until no further improvements can be made. To construct the best GBM, we followed the following steps: Since there is a trade-off between the number of trees and learning rate, we initially set a high number of trees (10,000) and learning rate (0.05) Then, we sequentially decreased the learning rate to find the best trade-off relationship. We used a random subset of 80 percent of the estimation set to fit a model and the remaining 20 percent to test the fit out of sample. Model performance was evaluated based on the area under the curve (AUC) of the receiver operating characteristics (ROC). Then, we performed a grid search to find the best combination of the learning rate along with other parameters. Next, we figured out the range of the maximum depth of variable interactions for the top five models, and then performed a grid search to find the best combination among parameters. The model with the highest AUC was selected as the best final model. Using the final model, we predicted the service-level propensity scores for those in the validation set. To see how well GBM predict, we ran logistic regression and compared AUC

between GBM and logistic regression across the 33 types of services. Second, we included the service-level propensity scores in the HCC model as covariates, and then estimated $\widehat{y_i}$ and $\widehat{Y_i}$.

4.2.4 *Estimation of Plan Payments*

Following Park and Basu (2018), we used 19 estimators (Table 1): nine parametric regression estimators, seven machine learning estimators, and three distributional estimators. We briefly describe each estimator and explain how to estimate $\widehat{y_i}$ and $\widehat{Y_i}$ by these three estimator groups. We refer to Park and Basu (2018) for detailed explanation of each estimator.

Parametric Regression Estimators We used nine parametric regression estimators. First, we used linear regression (NORMAL) because the HCC models rely on linear regression (Kautter et al. 2014). We used three different estimation methods for each of three different performance metrics. First, we estimated predicted expenditures in the validation set as $E(y_i|x_i) = x_i\beta$, where x_i is a matrix of first-year covariates of an enrollee i in the validation set and β is a column vector of coefficients estimated from the estimation set. Second, we used the survival function (Appendix Table A) to produce the estimate of $P(\hat{Y}_i > k|x_i)$ in the validation set. Because these estimated conditional tail probabilities vary across each possible combination of covariates for a given enrollee, we took the average in order to integrate out over values of x_i to provide the estimate of $P(\hat{Y}_i > k)$. Then, we estimated the ratio of $P\left(\hat{Y}_i > k\right)$ and the observed empirical proportion of expenditures in the data that exceed the expenditure threshold k, $\frac{P(\hat{Y}_i>k)}{P(y_i>k)}$. Finally, we used the survival function to estimate the metric for measuring individual-level prediction accuracy in the validation set. As actual expenditures vary across enrollee, the expenditure threshold for each of the four survival functions in the equation (4) varies across enrollees. Using the survival function,

we separately estimated four probabilities for each enrollee and combined them into a single estimate. Then, we carried out the same procedure as with the estimate of $P(\hat{Y}_i > k|x_i)$.

We also used the generalized beta distribution of the second kind (GB2), and its special or limiting cases distributions. Since the GB2 accounts for a flexible four-parameter distribution, it enables to reflect the characteristics of the health care expenditure data. For the GB2, we used two estimators: one with a log-link function (GB2 LOG) and the other with a squared-root link function (GB2 SQRT). For the special or limiting cases distributions, we used five estimators: the three-parameter generalized gamma distribution (GG), two-parameter gamma (GAMMA), two-parameter log-normal (LOGNORM), two-parameter Weibull distributions (WEIB), one-parameter exponential distribution (EXP). We followed the three estimation methods described in NORMAL, but there is a difference. As these estimators cannot account for enrollees with zero expenditure, we used a two-part model. To estimate $E(y_i|x_i)$, in the first part, we ran logistic regression to estimate the probability of having any positive expenditures for all enrollees in the validation set, $P(y_i > 0)$. In the second part, using the covariate coefficients and other parameters estimated from the estimation set, we estimated predicted expenditures for those with positive expenditures, $E(y_i|y_i > 0)$. Then, predicted expenditures for all enrollees in the validation set were estimated by multiplying the probability of having any positive expenditures and predicted expenditures for those with positive expenditures $E(y_i|x_i) = P(y_i > 0) \times E(y_i|y_i > 0)$. This process was also applied to estimations of $P(\hat{Y}_i > k)$ and the metric for measuring individual-level prediction accuracy.

Finally, we used finite mixture models (FMM). Since FMM specify finite mixture distributions, it accounts for heterogeneity in populations either based on observed covariates and unobserved latent classes. In this study, we used two gamma-distributed components of FMM with

a squared-root link function in both components (FMM SQRT). We used the same estimation methods as with GB2.

Machine Learning Estimators We used seven machine learning estimators. Machine learning estimators often use a three-way cross-validation to avoid overfitting a model. To be consistent with the other estimators, however, we used a one-fold cross-validation and repeated the process 100 times with sample replacement to produce the average estimates across all replications. Although machine learning has the potential to improve model flexibility and predictability under certain constraints, we did not put particular constraints in the estimation procedure. This is because there is no consensus on specific criteria for selecting constraints in model selection and it is not worthwhile to make a model too complex.

First, we used four regularized linear regressions. Regularized linear regression is similar to NORMAL, but different in the way that it shrinks regression coefficients of covariates with weak or no contribution to an outcome variable toward zero or exactly zero. We used four estimators with different shrinkage approaches: least absolute shrinkage and selection operator regression (LASSO) (Tibshirani 1996), ridge regression (RIDGE) (Marquardt and Snee 1975), elastic net regression (ENET) (Zou and Hastie 2005), and least angle regression (LARS) (Efron et al. 2004). To obtain the estimates of $E(y_i|x_i)$, $P(\hat{Y}_i > k)$, and the metric for measuring individual-level prediction accuracy, we carried out the similar estimation methods described in NORMAL. For each estimator, we chose the model with the smallest cross-validated mean squared error and use the nonzero coefficients of β estimated from the estimation set. To estimate $P(\hat{Y}_i > k)$ and the metric for measuring individual-level prediction accuracy, it was needed to have a distributional assumption, as shown in the parametric regression estimators. However, machine learning estimators do not rely on any distributional assumptions. Thus, we assumed a local normality and

allow for heteroscedasticity in variances of health care expenditures. Specifically, we formed percentiles of predicted expenditures in the estimation and validation sets, respectively, and computed variances of actual health care expenditures for each percentile in the estimation set. Then, we applied each of those variances from the estimation set to the equivalent percentile in the validation set.

Furthermore, we used artificial neural network (NNET). NNET builds an algorithm to better predict the outcome by detecting complex nonlinear relationships between the outcome variable and covariates and detecting all possible interactions between covariates (Zhang, Eddy Patuwo, and Y. Hu 1998). However, since the algorithm is not explicitly known, it leads to the difficulty of understanding and interpretation of the selected algorithm. We used the similar estimation methods described in LASSO. However, the drawback is that the process of estimating $E(y_i|x_i)$ is not explicitly known. We allowed the number of units in the hidden layer to be 10, as a cross-validated mean squared error was minimal when the number of hidden neurons was 10, on average.

Moreover, we used single decision tree (SINGLE). SINGLE builds a classification or regression model in the form of a tree structure by classifying a population into subsets (i.e., node) so as to have the most homogeneous outcome based on covariates (Breiman et al. 1984). The classification is performed by two covariates, and repeated until the outcome is sufficiently identical among observations in each node. To avoid overfitting, we selected a tree size that minimizes the cross-validated error. To obtain the estimates of $E(y_i|x_i)$, $P(\hat{Y}_i > k)$, and the metric for measuring individual-level prediction accuracy, we followed the same estimation methods described in LASSO.

Finally, we used super learner (SUPER). SUPER builds a weighted combination of estimators with optimal weights (Van Der Laan, Polley, and Hubbard 2007). Optimal weights are typically estimated based on cross-validated mean squared error. We only considered six machine learning estimators described above as a candidate estimator. The estimator with the smallest cross-validated mean squared error was selected as the final SUPER estimator. We performed the same estimation methods described in LASSO.

Distributional Estimators We used three distributional methods: conditional density approximation estimator using ordered logit regression (HH), conditional density approximation estimator using multinomial logit regression (FP), and linear probability model (CH). To estimate $\hat{y}_i$, we followed the estimation method described in Gilleskie and Mroz (2004). First, the outcome variable was grouped into Q discrete intervals. Next, we calculated the probabilities that each observation lies in the j interval, $p_{ij}(x_i)$, and the mean value of the outcome variable for the j interval, $\bar{y}_j$. Following Jones et al. (2016), we assumed that only the probability of lying within an interval relies on covariates and that the mean value of the outcome variable for a given interval does not vary with covariates. By multiplying these two values for each interval and summing them up across all intervals, we estimated $\hat{y}_i = \sum_{j=1}^{Q} p_{ij}(x_i)\bar{y}_j$. In this study, we used three different ways how to categorize the outcome variable and calculate the probabilities, $p_{ij}(x_i)$. For HH, we generated a categorical variable for each observation, indicating the interval into which the value of the outcome variable falls, and ran an ordered logit regression to calculate the probabilities $p_{ij}(x_i)$ (Han and Hausman 1990). We performed preliminary work to establish the largest number of intervals while retaining good convergence performance, and finally selected 33 intervals. For FP, we generated a categorical variable that takes value one if the outcome variable is less than or equal to the upper boundary, and zero otherwise. Then, we ran a series of logit

regressions. (Foresi and Peracchi 1995). From our preliminary work, we select 18 intervals: 0th , 5th, . . . , 85th, 90th, 95th percentiles as boundaries. For CH, we generated an indicator that is equal to one if the outcome variable is less than or equal to each unique value of the outcome variable and zero otherwise. Although the original method of Chernozhukov, Fernandez-Val, and Melly (2013) proposed to use a logit regression for each unique value of the outcome variable, it requires very expensive computational demands for large sample. Alternatively, we used the linear probability model by de Meijer et al. (2013), to estimate the probabilities, $p_{ij}(x_i)$, (CH). Previous studies that the results from the linear probability model and the method of Chernozhukov et al. (2013) were virtually identical (Jones et al. 2015). To obtain the estimate of $P(\hat{Y}_i > k)$ and the metric for measuring individual-level prediction accuracy in the validation set, we followed Jones et al. (2015). We produced the estimate of $P(\hat{Y}_i > k^*|x_i)$, where k^* represents one of the boundaries of the intervals generated in each of these estimators. As these estimators were performed without knowing the threshold k, it is not always the case where $k^* = k$. When $k^* \neq k$, we used the following simple linear interpolation formula to estimate a weighted average of $P(\hat{Y}_i > k^*|x_i)$ for the nearest two values of k^* to k : $P(\hat{Y}_i > k^*|x_i) = P\big(\hat{Y}_i > k_a^*\big|x_i\big) + \left(\frac{k-k_a^*}{k_b^*-k_a^*}\right)\Big(P\big(\hat{Y}_i > k_b^*\big|x_i\big) - P\big(\hat{Y}_i > k_a^*\big|x_i\big)\Big)$, where k_a^* and k_b^* represent the thresholds analyzed in estimation closest below and closest above k, respectively. To compute the metric for measuring individual-level prediction accuracy, we carried out the same procedure as with NORMAL. We followed the same process, but used three different thresholds that vary across each individual, instead of the constant threshold k across all individuals.

4.2.5 *Evaluation of Prediction Performance Metrics*

We used three different prediction performance metrics to identify the best estimator for reducing service-level selection. To assess the prediction accuracy at the group level, we used five metrics: mean prediction error (MPE), $\frac{\sum(y_i-\hat{y}_i)}{N}$, mean absolute prediction error (MAPE), $\frac{\sum|y_i-\hat{y}_i|}{N}$, root mean square error (RMSE), $\sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{N}}$, R^2, $1-\frac{\sum(y_i-(\alpha_{AUX}+\beta_{AUX}\hat{y}_i))^2}{\sum(y_i-\hat{y}_i)^2}$, and predictive ratio, $\frac{\frac{1}{N}\sum\hat{y}_i}{\frac{1}{N}\sum y_i}$. The highest prediction accuracy takes value one for MPE, MAPE, and RMSE, and zero for R^2 and predictive ratio, respectively.

To assess the prediction accuracy at the tail distributions, we used the evaluation metric, $\frac{P(\hat{Y}_i>k)}{P(y_i>k)}$, developed by Jones et al. (2015). The metric measures a ratio of the estimated $P(\hat{Y}>k)$ to the actual proportion of a population whose expenditures are higher than the expenditure threshold k. The highest prediction accuracy represents a ratio of one. We used four expenditure thresholds ($5,000, $15,000, $25,000, and $80,000), each of which was at the 75th, 90th, 95th, and 99th percentiles of the expenditure distribution in our data, respectively.

To assess the individual-level prediction accuracy, we used the two metrics developed by Park and Basu (2018): overall individual-level prediction accuracy, $P\left(\frac{|y_i-\hat{Y}_i|}{y_i}\times 100>l\right)$ and tail-specific individual-level prediction accuracy, $P\left(\frac{|y_i-\hat{Y}_i|}{y_i}\times 100>l|y_i>k\right)$. The first metric measures what portion of a population has forecast error greater than l percentage of their actual expenditures. For example, $P\left(\frac{|y_i-\hat{Y}_i|}{y_i}\times 100>10\right)=0.70$ means that for 70 percent of the population, a risk-adjustment model produces under-forecasted or over-forecasted expenditures by greater than than 10 percentage of their actual expenditures. Furthermore, the HCC models systematically under-predict expenditures for those with very high expenditures (Medicare

Payment Advisory Commission 2012, 2014; Pope et al. 2004). Thus, we used the second metric to assess the individual-level prediction accuracy at the tail distributions. The highest prediction accuracy represents zero probabilities. We used three percentages thresholds l (10, 50, and 100). We used the same three expenditure thresholds described above.

4.3 RESULTS

Table 2 presents descriptive statistics on our total study population. Our total study population comprised 12,882,983 working-aged adults continuously enrolled in the same plan type between 2013-2014. The estimation and validation sets of each iteration included 128,829 adults, respectively. Table 2 also provides descriptive statistics for health care use by service type. Among the 33 types of service, the most frequently used top five services were non-specialty visits (72.02 percent), laboratory (67.03 percent), pharmacy (61.81 percent), specialty visits (50.36 percent), and prevention-related services (36.53 percent). On the other hand, the least frequently used top five services were hospice (0.01 percent), dialysis (0.08 percent), radiology (therapeutic) (0.23 percent), PET scans (0.28 percent), and home visits (0.69 percent).

Table 3 shows improvements of prediction performance on AUC with GBM over logistic regression across the 33 types of service. GBM outperformed logistic regression in predicting service-level use in the second year among all service type categories. Specifically, the service-specific average AUC estimated based on GBM ranged from 0.61 to 0.94 (chiropractic and maternity, respectively), whereas the service-specific average AUC estimated based on logistic regression ranged from 0.56 to 0.93. GBM achieved approximately 10 percent improvements on AUC, on average, across the 33 types of service. Substantial improvements were in hospice (56

percent), laboratory (18 percent), surgical procedures (12 percent), surgical supplies and devices (12 percent), and pharmacy (12 percent).

Table 4 presents descriptive statistics for the second-year total health care expenditures. The mean expenditures were $6,891. The 75th, 90th, 95th, and 99th percentiles of the expenditure distribution were $5,736, $15,384, $27,080, and $80,116, respectively. The first-year information on demographics, diagnoses, state, and type of plan are presented in Appendix Table B.

Figure 1 presents comparison of prediction accuracies between the alternative model and the HCC model across the 19 estimators. The first five graphs show the group-level prediction accuracies. When compared the group-level prediction accuracies across the 19 estimators based on the HCC model, machine learning estimators had higher group-level prediction accuracies than parametric regression and distributional estimators. When the alternative model was applied, these machine learning estimators had even more accurate group-level predictions relative to the corresponding estimator based on the HCC model. Such increase in group-level prediction accuracy was more pronounced for MAPE and R^2 because MPE, RMSE, and predictive ratio already achieved a relatively high level of prediction accuracy with the HCC model. However, there were small improvements in parametric regression and distributional estimators.

The next four graphs in Figure 1 show the prediction accuracies at the tail distributions. Among the 19 estimators based on the HCC model, machine learning estimators tended to over-forecast expenditures above the 75th and 90th percentiles, respectively, of the expenditure distribution two times higher than parametric regression and distributional estimators. However, such over-forecast decreased as the expenditure threshold increased, and this over-forecast disappeared for those above the 99th percentile of the expenditure distribution. When the alternative model was applied, machine learning estimators had more accurate forecasts compared

to the HCC model with the corresponding estimators, especially for expenditures above the 75th and 90th percentiles. However, there was no substantial difference between the alternative model and the HCC model in predicting the probabilities of expenditures exceeding $80,000.

The last 16 graphs show the individual-level prediction accuracies, the first four of which shows the overall individual-level prediction accuracies and the others shows the tail-specific individual-level prediction accuracies. There was no substantial difference between the alternative model and the HCC model in overall individual-level prediction accuracy. However, the alternative model had higher individual-level prediction accuracy at certain expenditure thresholds and percentage thresholds with certain estimators. Specifically, when we defined the prediction accuracy as a forecast error less than 10 percent and 50 percent, respectively, of each individual's actual expenditures, ridge regression and elastic net based on the alternative model higher individual-level prediction accuracy for those with very high expenditures (those above the 95th and 99th percentiles of the expenditure distribution) compared to the corresponding estimators using the HCC model. However, we allowed the forecast error by 100 percent, there were marginal differences between the alternative model and the HCC model.

4.4 DISCUSSION

In this paper, we proposed an alternative risk-adjustment model using machine learning algorithms, which accounts for service-level propensity scores to prevent service-level selection. Machine learning algorithms enable to implement this alternative model in two ways. First, we found that service-level propensity scores were more accurately predicted based on GBM, thereby reflecting the residual health status not captured by the HCC model more accurately as well as generating stronger economic incentives to discourage health plans from engaging in service-level

selection. Second, we found that this alternative model improved prediction accuracies of health care expenditures when combined with machine learning estimators. Specifically, machine learning estimators based on the alternative model generally had higher prediction accuracies at the group level and at the tail distributions relative to the corresponding estimators based on the HCC model. Although improvements in overall individual-level prediction accuracy were negligible, machine learning estimators based on the alternative model has higher individual-level prediction accuracy, especially for those with very high expenditures compared to the corresponding estimators based on the HCC model.

Our findings showed that accounting for service-level propensity scores in risk adjustment enables to more accurately reflect health status beyond HCCs, thereby leading accurate plan payments, especially for those with high expenditures. From the statistical perspective, the idea behind this model is to more precisely incorporate health status differences in the calculation of plan payments by accounting for the residual health status not captured by the HCC model. Previous studies have shown that some services are more likely to be used by beneficiaries whose health status could be worse than estimated by the HCC model (Park et al. 2017; Rahman et al. 2015). This suggests that use of these services serve as a signal to capture the residual health status. Adjusting plans payments based on each beneficiary's future service-level use allows to more accurately capture her current health status. It is worthwhile noting that this alternative model is feasibly implemented with machine learning algorithms. In this study, GBM was used to estimate service-level propensity scores because performance of logistic regression was relatively low. This suggests that there may be complex nonlinear relationships and interactions in capturing the residual health status, which traditional statistical models were not able to capture.

More importantly, the alternative model is designed to reduce service-level selection by generating economic forces to discourage service-level selection. It has been shown that health plans have incentives to engage in service-level selection in the MA program (Ellis and McGuire 2007; Park et al. 2017; Park et al. 2018; Rahman et al. 2015) and the Exchanges (Barry et al. 2012; McGuire et al. 2014; Montz et al. 2016; Shrestha et al. 2017; Weiner et al. 2012). To reduce service-level selection, this alternative model seeks to equalize incentives in rationing all services by providing overpayments for services needed by unprofitable beneficiaries and underpayments for services needed by profitable beneficiaries, which originates from *optimal risk adjustment* (Glazer and McGuire 2000). This would redistribute health care costs away from services likely to be used by the low-cost individuals and toward services likely to be used by the high-cost individuals, thereby reducing health plans' incentives for service-level selection. It is worthy to note that the model is designed not to induce the incentive to offer more services regardless of clinical needs. Service-level propensity scores were estimated based on prior year's health status, not actual utilization. Otherwise, this would likely induce health plans to use specific services more in order to receive higher payments.

Our study showed that the alternative model generated more accurate plan payments when combined with machine learning estimators. When compared across estimators based on the HCC model, machine learning estimators had higher group-level prediction accuracies than parametric estimators and distributional estimators (Park and Basu 2018). In this study, we found that these estimators had even higher group-level prediction accuracy with the alternative model, especially in MAPE and R^2 relative to the corresponding estimators with the HCC model. Moreover, the machine learning estimators based on the alternative model had higher prediction accuracy in predicting the probabilities of expenditures exceeding \$5,000, \$15,000, \$25,000, respectively,

compared to the corresponding estimators based on the HCC model. Finally, the machine learning estimators based on the alternative model had higher individual-level prediction accuracy for those with very high expenditures (those above the 95th and 99th percentiles of the expenditure distribution) compared to the corresponding estimators based on the HCC model, especially for ridge regression and elastic net. This suggests that machine learning estimators have the potential to detect complex nonlinear relationships and interactions that parametric estimators and distributional estimators were not able to capture, especially for those with high expenditures, who tend to have complex health status.

Accounting for the service-level propensity scores in risk adjustment with machine learning has the potential to reduce incentives to avoid enrollees with higher expenditures than their estimated plan payments. The differential impact of machine learning estimators seems to be most pronounced in estimating plan payments for those with high expenditures. Our study provides policy implications on the design of an alternative risk-adjustment model, which enables to improve prediction accuracy for health care expenditures while reducing service-level selection.

References

Barry, C. L., J. P. Weiner, K. Lemke, and S. H. Busch. 2012. "Risk adjustment in health insurance exchanges for individuals with mental illness." *Am J Psychiatry* 169(7): 704-9.

Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen. 1984. *Classification and regression trees*. New York: CRC Press.

Centers for Medicaid and Medicare Services. 2014. "2014 Benefit year risk adjustment SAS version of HHS-developed risk adjustment model algorithm software." Baltimore, MD: Centers for Medicaid and Medicare Services.

Chernozhukov, V., I. Fernandez-Val, and B. Melly. 2013. "Inference on counterfactual distributions." *Econometrica* 81(6): 2205-68.

de Meijer, C., O. O'Donnell, M. Koopmanschap, and E. van Doorslaer. 2013. "Health expenditure growth: Looking beyond the average through decomposition of the full distribution." *Journal of Health Economics* 32(1): 88-105.

Efron, B., T. Hastie, I. Johnstone, R. Tibshirani, H. Ishwaran, K. Knight, J. M. Loubes, P. Massart, D. Madigan, G. Ridgeway, S. Rosset, J. I. Zhu, R. A. Stine, B. A. Turlach, S. Weisberg, T. Hastie, I. Johnstone, and R. Tibshirani. 2004. "Least angle regression." *Annals of Statistics* 32(2): 407-99.

Einav, L., A. Finkelstein, R. Kluender, and P. Schrimpf. 2016. "Beyond statistics: the economic content of risk scores." *American Economic Journal: Applied Economics* 8(2): 195-224.

Ellis, R. P., B. Martins, and W. J. Zhu. 2017. "Demand elasticities and service selection incentives among competing private health plans." *Journal of Health Economics* 56: 352-67.

Ellis, R. P. and T. G. McGuire. 2007. "Predictability and predictiveness in health care spending." *Journal of Health Economics* 26(1): 25-48.

Foresi, S. and F. Peracchi. 1995. "The conditional distribution of excess returns: an empirical analysis." *Journal of the American Statistical Association* 90(430): 451-66.

Gilleskie, D. B. and T. A. Mroz. 2004. "A flexible approach for estimating the effects of covariates on health expenditures." *Journal of Health Economics* 23(2): 391-418.

Glazer, J. and T. G. McGuire. 2000. "Optimal risk adjustment in markets with adverse selection: an application to managed care." *Am. Econ. Rev.* 90(4): 1055-71.

Han, A. and J. A. Hausman. 1990. "Flexible parametric estimation of duration and competing risk models." *Journal of Applied Econometrics* 5(1): 1-28.

Jones, A. M., J. Lomas, P. T. Moore, and N. Rice. 2016. "A quasi-Monte-Carlo comparison of parametric and semiparametric regression methods for heavy-tailed and non-normal data: an application to healthcare costs." *Journal of the Royal Statistical Society Series a-Statistics in Society* 179(4): 951-74.

Jones, A. M., J. Lomas, and N. Rice. 2015. "Healthcare cost regressions: going beyond the mean to estimate the full distribution." *Health Econ* 24(9): 1192-212.

Kautter, J., G. C. Pope, M. Ingber, S. Freeman, and L. Patterson. 2014. "The HHS-HCC risk adjustment model for individual and Small Group markets under the affordable care act." *Medicare and Medicaid Research Review* 4(3).

Manning, W. G., A. Basu, and J. Mullahy. 2005. "Generalized modeling approaches to risk adjustment of skewed outcomes data." *J Health Econ* 24(3): 465-88.

Marquardt, D. W. and R. D. Snee. 1975. "Ridge regression in practice." *American Statistician* 29(1): 3-20.

McGuire, T. G., J. P. Newhouse, S.-L. Normand, J. Shi, and S. Zuvekase. 2014. "Assessing incentives for service-level selection in private health insurance exchanges." *J Health Econ* 35(1): 47-63.

Medicare Payment Advisory Commission. 2011. "Report to the Congress: Medicare Payment Policy." Washington, DC: Medicare Payment Advisory Commission.

Medicare Payment Advisory Commission. 2012. "Report to the Congress: Medicare and the Health Care Delivery System." Washington, DC: Medicare Payment Advisory Commission.

Medicare Payment Advisory Commission. 2014. "Report to the Congress: Medicare and the Health Care Delivery System." Washington, DC: Medicare Payment Advisory Commission.

Montz, E., T. Layton, A. B. Busch, R. P. Ellis, S. Rose, and T. G. McGuire. 2016. "Risk-Adjustment Simulation: Plans May Have Incentives To Distort Mental Health And Substance Use Coverage." *Health Affairs* 35(6): 1022-28.

Park, S. and A. Basu. 2018. "Alternative evaluation metrics for risk adjustment methods." *Health Econ* 27(6): 984-1010.

Park, S., A. Basu, N. B. Coe, and F. Khalil. 2017. “Service-level selection: strategic risk selection in Medicare Advantage in response to risk adjustment.” *NBER Working Paper No. 24038*.

Park, S., P. Fishman, L. White, E. B. Larson, and N. B. Coe. 2018. “Plan switching between traditional Medicare and Medicare Advantage plans: variation by disease and costs.”

Pope, G. C., J. Kautter, R. P. Ellis, A. S. Ash, J. Z. Ayanian, L. I. Iezzoni, M. J. Ingber, J. M. Levy, and J. Robst. 2004. “Risk adjustment of medicare capitation payments using the CMS-HCC model.” *Health Care Financ. Rev.* 25(4): 119-41.

Pope, G. C., J. Kautter, J. M. Ingber, S. Freeman, R. Sekar, and C. Newhart. 2011. “Evaluation of the CMS-HCC Risk Adjustment Model ”. Research Triangle Park, NC: RTI International

Rahman, M., K. Laura, A. N. Trivedi, and V. Mor. 2015. “High-cost patients had substantial rates of leaving Medicare Advantage and joining Traditional Medicare.” *Health Aff. (Millwood)* 34(10): 1675-81.

Rose, S. 2016. “A Machine Learning Framework for Plan Payment Risk Adjustment.” *Health Serv Res* 51(6): 2358-74.

Shrestha, A., S. Bergquist, E. Montz, and S. Rose. 2017. “Mental Health Risk Adjustment with Clinical Categories and Machine Learning.” *Health Serv Res*.

Tibshirani, R. 1996. “Regression shrinkage and selection via the Lasso.” *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 58: 267–88.

Van Der Laan, M. J., E. C. Polley, and A. E. Hubbard. 2007. "Super learner." *Statistical Applications in Genetics and Molecular Biology* 6(1).

Weiner, J. P., E. Trish, C. Abrams, and K. Lemke. 2012. "Adjusting For Risk Selection In State Health Insurance Exchanges Will Be Critically Important And Feasible, But Not Easy." *Health Affairs* 31(2): 306-15.

Zhang, G., B. Eddy Patuwo, and M. Y. Hu. 1998. "Forecasting with artificial neural networks: The state of the art." *International Journal of Forecasting* 14(1): 35-62.

Zou, H. and T. Hastie. 2005. "Regularization and variable selection via the elastic net." *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 67(2): 301-20.

Tables

Table 1 Key for estimator labels

Estimator	Label
Parametric regression estimators	
Linear regression with the normal distribution of error	NORMAL
Generalized beta of the second kind (log-link)	GB2 LOG
Generalized beta of the second kind (squared-root link)	GB2 SQRT
Generalized gamma (log-link)	GG
Gamma (log-link)	GAMMA
Log-normal (log-link)	LOGNORM
Weibull (log-link)	WEIB
Exponential (log-link)	EXP
Two-component finite mixture of gamma densities (squared-root link)	FMM SQRT
Machine learning estimators	
Lasso regression	LASSO
Ridge regression	RIDGE
Elastic net	ENET
Least angle regression	LARS
Neural net	NNET
Single tree	SINGLE
Super learner	SUPER
Distributional estimators	
Han and Hausman (conditional density approximation estimator using ordered logit regression)	HH
Foresi and Peracchi (conditional density approximation estimator using multinomial logit regression)	FP
Chernozhukov, Fernández-Val and Melly (linear probability model)	CFM

Table 2 Descriptive statistics on health care use by type of services

Types of service	N	%
Non-specialty visits	11,861,891	72.02
Home visits	112,947	0.69
Prevention	6,017,096	36.53
Maternity	381,349	2.32
Mental health	1,828,050	11.10
Substance abuse	124,635	0.76
Surgical procedures	269,875	1.64
Surgical supplies and devices	263,669	1.60
Non-surgery inpatient procedures	505,067	3.07
Specialty visits	8,294,615	50.36
Dialysis	13,578	0.08
PT, OP, and speech therapy	2,221,379	13.49
Chiropractic	1,287,468	7.82
Hospice	2,236	0.01
Emergency room	2,289,684	13.90
Room and board (surgical)	282,498	1.72
Room and board (medical and other)	285,464	1.73
CAT scans	1,122,982	6.82
Mammograms	2,909,050	17.66
MRIs	1,216,270	7.38
PET scans	46,685	0.28
Radiology (diagnostic)	4,643,718	28.19
Radiology (therapeutic)	37,860	0.23
Ultrasounds	2,427,326	14.74
Diagnostic services	4,468,924	27.13
Laboratory	11,039,827	67.03
Pharmacy	10,180,083	61.81
Facility-based pharmacy	1,094,317	6.64
Specialty drugs, injections	2,703,674	16.42
Non-surgery supplies and devices	2,198,911	13.35
Durable medical equipment	673,987	4.09
Transportation	285,588	1.73
Other	3,400,749	20.65

Table 3. Prediction performance on service-level use between logistic and GBM regressions

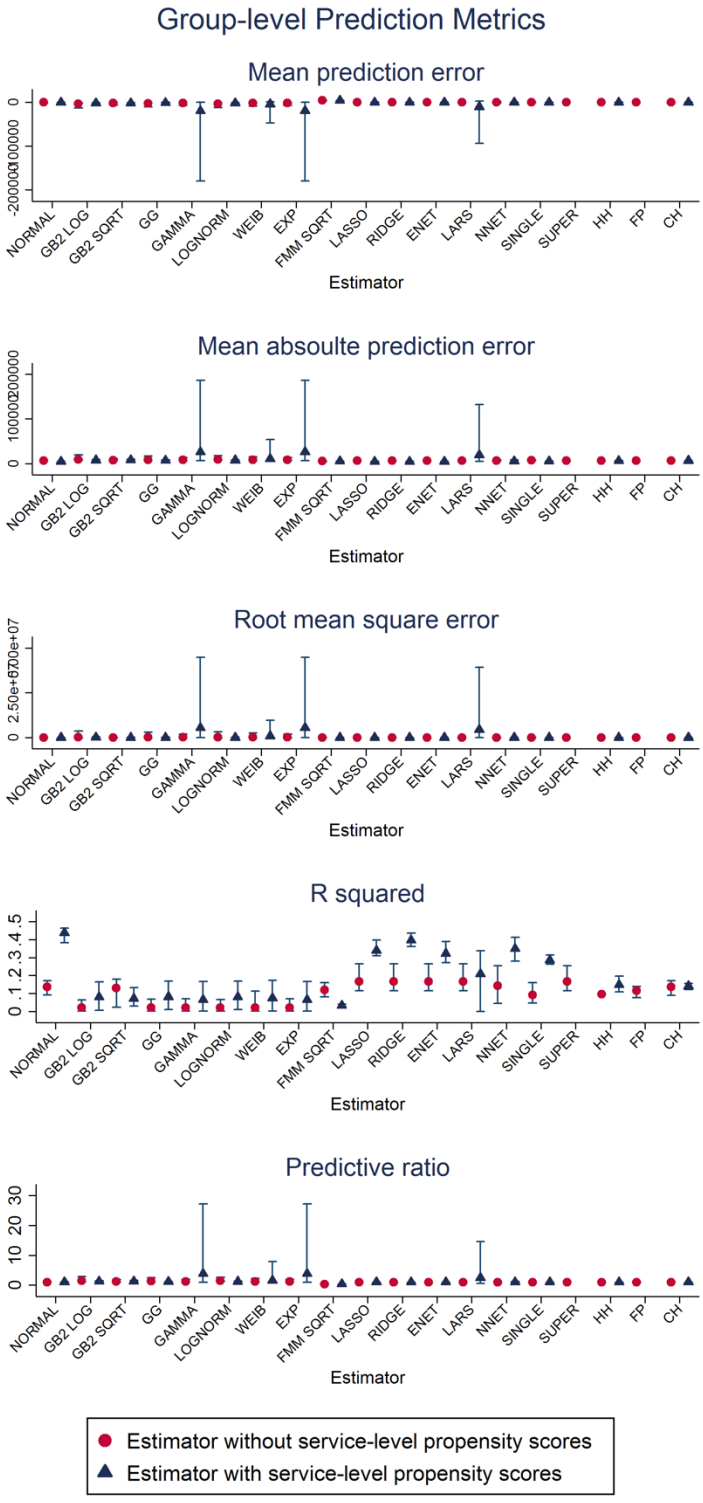
	The Area Under the ROC Curve					
	Logistic regression			GBM regression		
Types of service	Mean	Min	Max	Mean	Min	Max
Non-specialty visits	0.677	0.674	0.678	0.733	0.730	0.736
Home visits	0.777	0.766	0.789	0.826	0.808	0.835
Prevention	0.672	0.671	0.673	0.716	0.713	0.718
Maternity	0.929	0.927	0.930	0.940	0.937	0.942
Mental health	0.678	0.675	0.681	0.710	0.707	0.713
Substance abuse	0.749	0.728	0.769	0.762	0.744	0.781
Surgical procedures	0.708	0.698	0.718	0.793	0.783	0.797
Surgical supplies and devices	0.707	0.698	0.716	0.794	0.782	0.801
Non-surgery inpatient procedures	0.707	0.703	0.715	0.777	0.772	0.782
Specialty visits	0.692	0.688	0.695	0.740	0.737	0.743
Dialysis	0.888	0.845	0.939	0.927	0.874	0.960
PT, OP, and speech therapy	0.613	0.608	0.615	0.662	0.658	0.665
Chiropractic	0.570	0.565	0.573	0.606	0.602	0.612
Hospice	0.557	0.504	0.628	0.868	0.715	0.943
Emergency room	0.594	0.592	0.596	0.649	0.644	0.652
Room and board (surgical)	0.708	0.699	0.715	0.792	0.783	0.801
Room and board (medical and other)	0.735	0.725	0.745	0.805	0.797	0.812
CAT scans	0.676	0.670	0.683	0.727	0.721	0.734
Mammograms	0.898	0.896	0.899	0.939	0.938	0.940
MRIs	0.651	0.644	0.655	0.689	0.642	0.705
PET scans	0.809	0.786	0.820	0.841	0.662	0.885
Radiology (diagnostic)	0.668	0.666	0.670	0.722	0.667	0.730
Radiology (therapeutic)	0.781	0.755	0.808	0.826	0.668	0.877
Ultrasounds	0.707	0.703	0.711	0.765	0.740	0.773
Diagnostic services	0.705	0.703	0.708	0.752	0.733	0.768
Laboratory	0.721	0.719	0.724	0.852	0.819	0.896
Pharmacy	0.616	0.613	0.618	0.687	0.673	0.695
Facility-based pharmacy	0.708	0.703	0.711	0.738	0.669	0.761
Specialty drugs, injections	0.635	0.630	0.639	0.685	0.633	0.704
Non-surgery supplies and devices	0.681	0.677	0.684	0.713	0.670	0.729
Durable medical equipment	0.672	0.666	0.679	0.738	0.684	0.751
Transportation	0.651	0.642	0.659	0.701	0.626	0.724
Other	0.651	0.648	0.652	0.702	0.678	0.708

Table 4. Descriptive statistics for total health care expenditures in the second-year (2014)

N	12,882,983
Mean	$6,891
SD	$22,963
Skewness	28
Kurtosis	3,072
Maximum	$8,199,457
99th percentile	$80,116
95th percentile	$27,080
90th percentile	$15,384
75th percentile	$5,736
25th percentile	$416
10th percentile	$0
1th percentile	$0
Minimum	$0

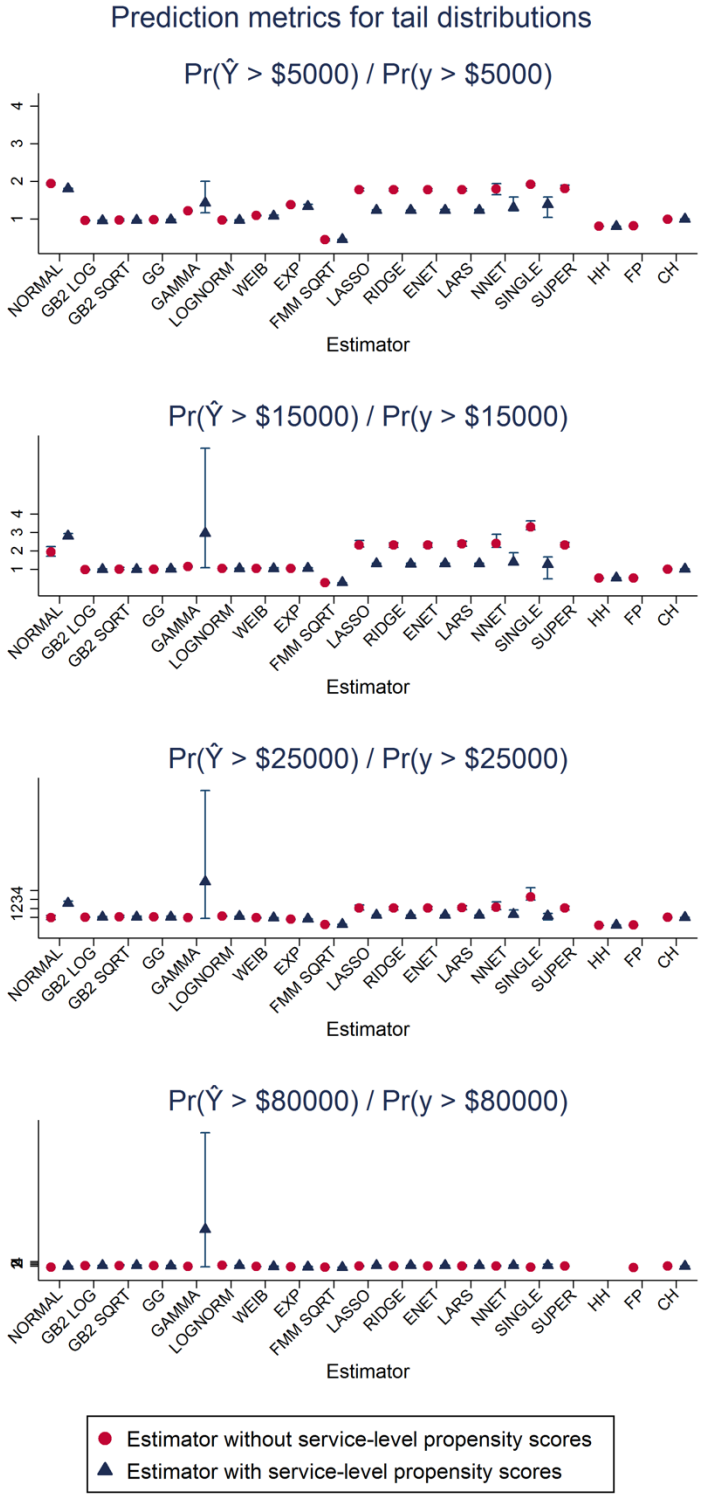
Figures

Figure 1 Comparison of prediction accuracy between the HCC model and the alternative model (continued)



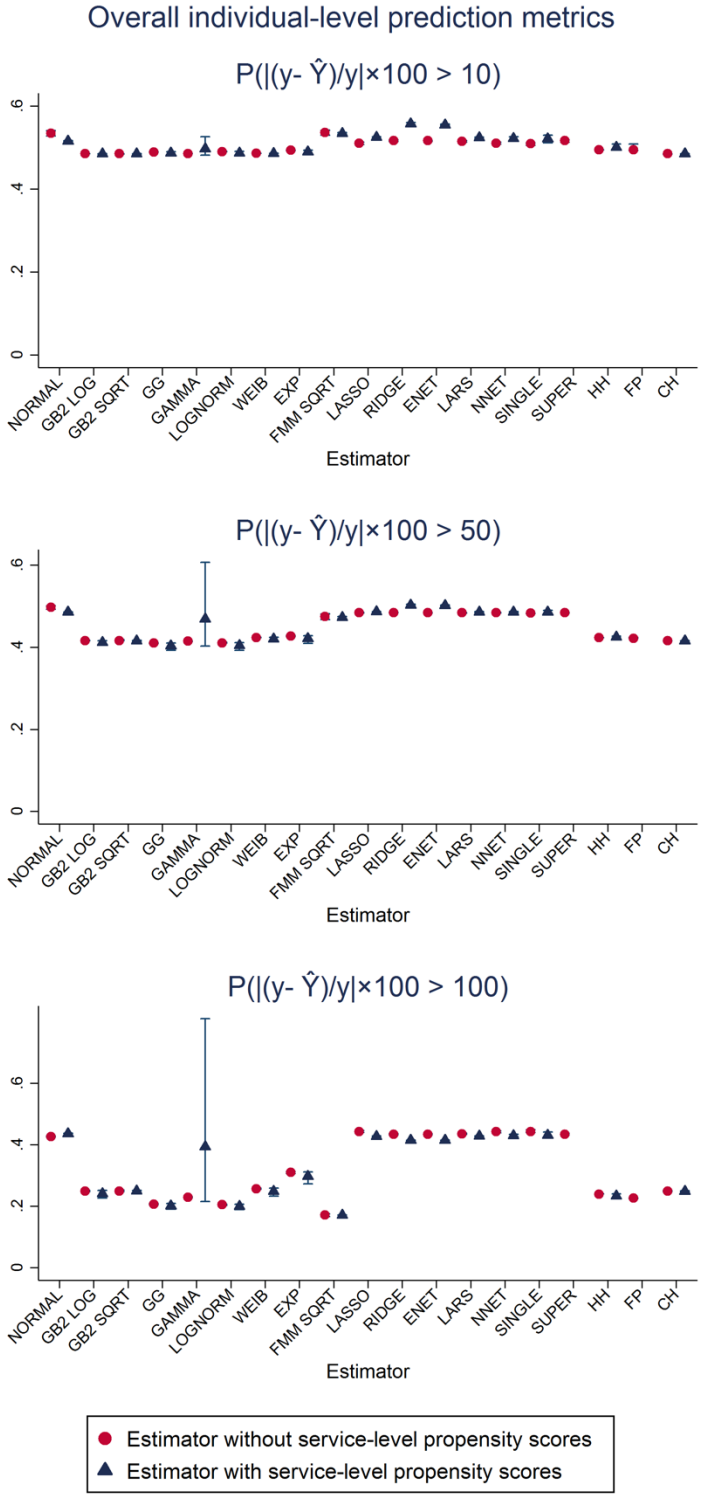
Note: A perfect prediction represents zero for mean prediction error, mean absolute prediction error, and root mean square error, and one for R^2 and predictive ratio.

Figure 1 Comparison of prediction accuracy between the HCC model and the alternative model (continued)



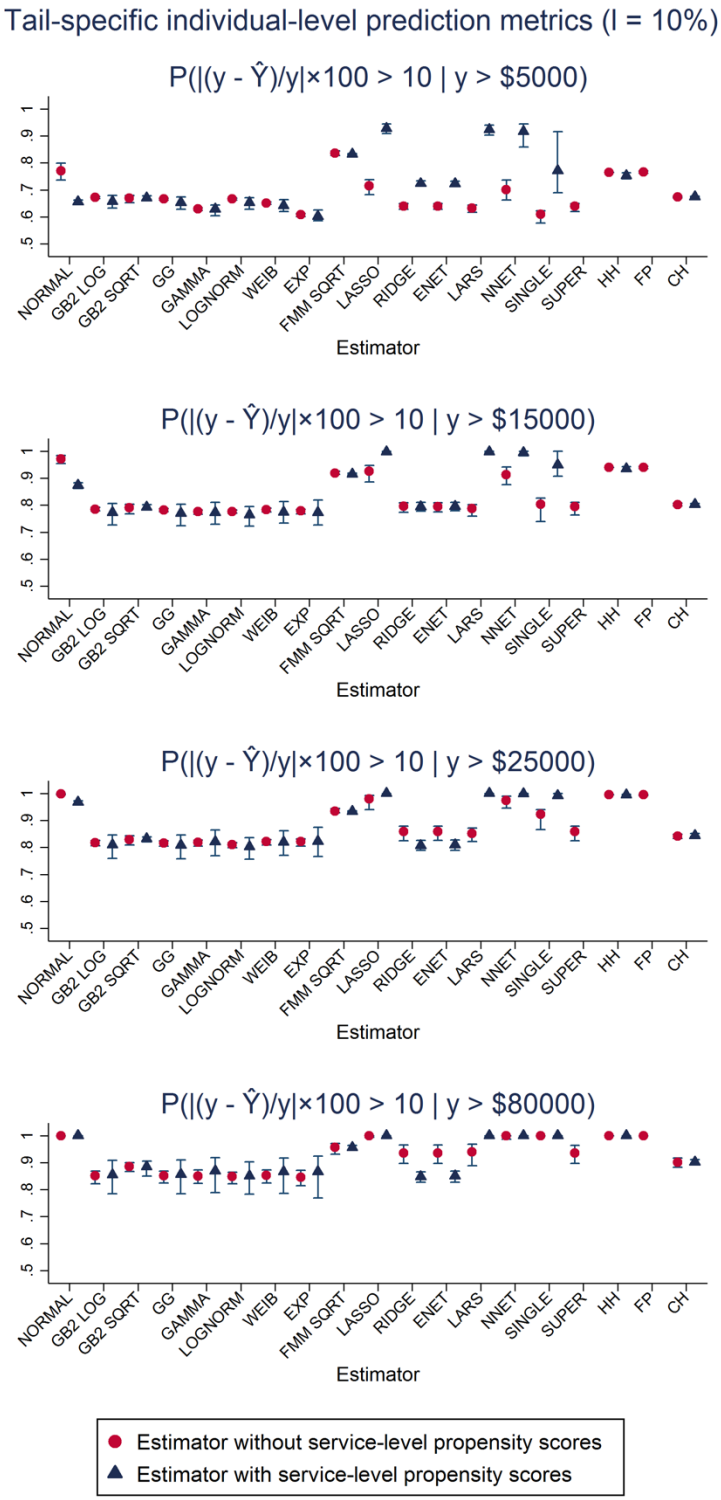
Note: A perfect prediction represents a ratio of one.

Figure 1 Comparison of prediction accuracy between the HCC model and the alternative model (continued)



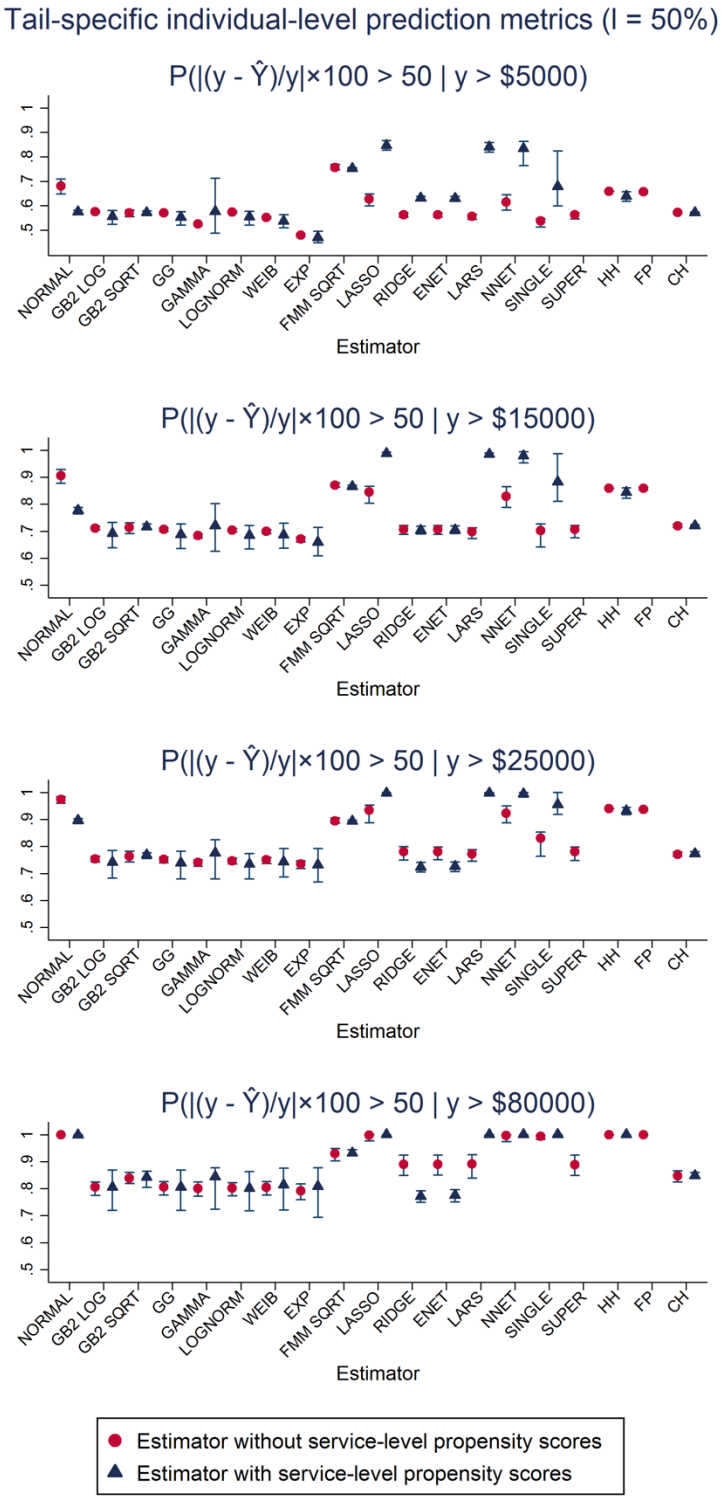
Note: A perfect prediction represents a zero probability.

Figure 1 Comparison of prediction accuracy between the HCC model and the alternative model (continued)



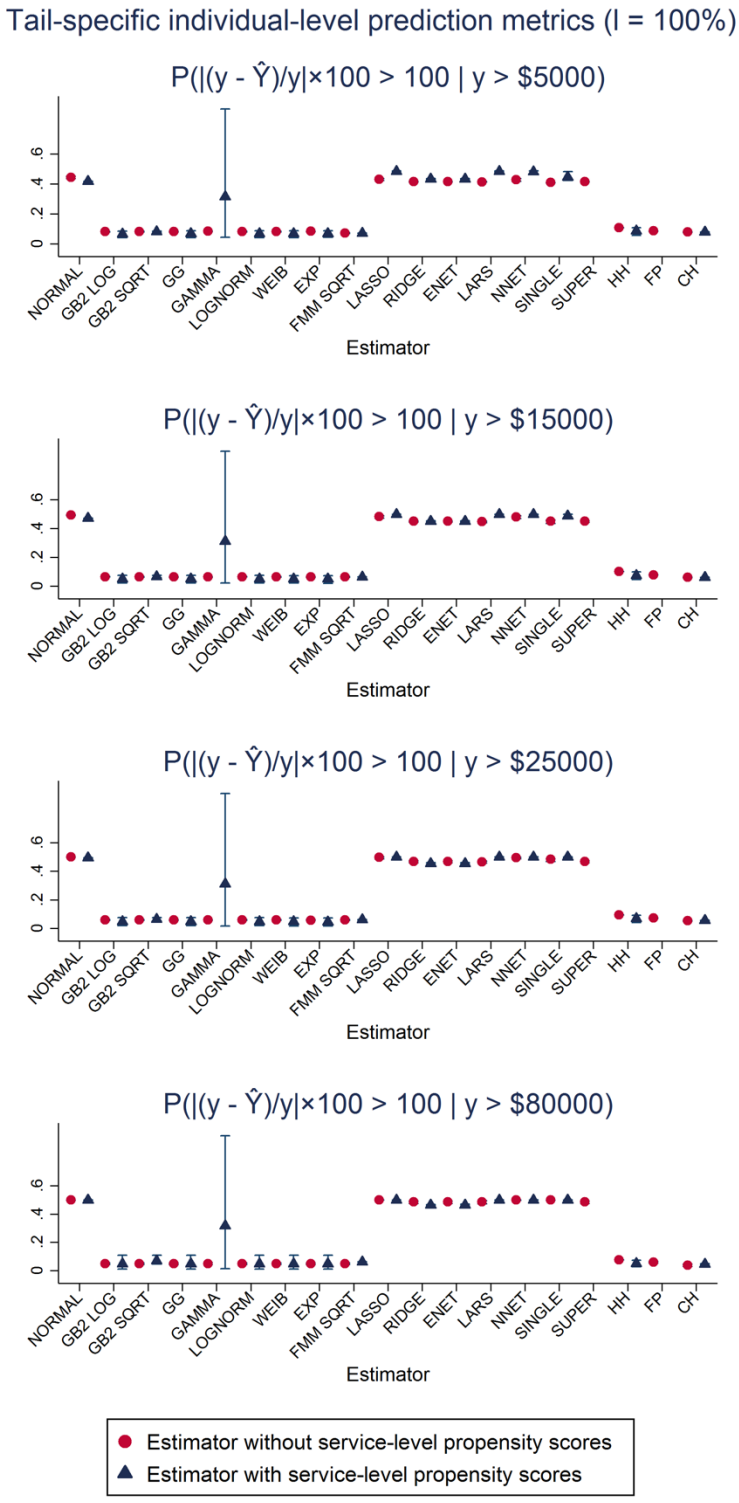
Note: A perfect prediction represents a zero probability.

Figure 1 Comparison of prediction accuracy between the HCC model and the alternative model (continued)



Note: A perfect prediction represents a zero probability.

Figure 1 Comparison of prediction accuracy between the HCC model and the alternative model (continued)



Note: A perfect prediction represents a zero probability.

Chapter 5.

CONCLUSION

There is a long-standing concern about risk selection in the health insurance market. Over several decades, policy makers have promoted managed care as a way to improve quality of care while containing costs. As this approach creates incentives to encourage preventive care and better care coordination, it is especially helpful in caring for Medicare beneficiaries, 68.4 percent of whom had two or more chronic conditions and 36.4 percent had four or more chronic conditions. To achieve the goals of improving quality of care and containing costs in the Medicare program, CMS has reimbursed MA plans, with a capitated amount per beneficiary. As an unintended consequence, however, MA plans began to selectively enroll healthier people as a means to receive overpayments.

CMS have tried to reduce risk selection by adjusting payments to MA plans to reflect the health status of their enrollees, a process known as *risk adjustment*. It is used to predict health care expenditures to correctly align plan payments with a beneficiary's expected health care expenditure. Its ultimate goal is to discourage health plans from selectively enrolling and caring for healthy beneficiaries, thereby encouraging the plans to compete based on providing high-value care. To more precisely incorporate health status differences in the calculation of payments to MA plans, in 2004, CMS introduced a new risk adjustment model—the HCC model. However, it is debated whether the HCC model has been effective in reducing risk selection or whether it has led to a strategic evolution of risk selection.

The overarching goal of this research is to understand the intended and unintended consequences of the HCC model on the MA plans' patient selection behaviors and propose alternative risk adjustment models from the perspectives of statistics and economics, respectively.

In Chapter 2, we found that the HCC model reduced MA plans' avoidance of high-cost beneficiaries at enrollment, but led to increased disenrollment of high-cost beneficiaries, conditional on illness severity (i.e., unprofitable beneficiaries), from MA plans. We explained this unintended consequence via service-level selection. We found that MA plans raised copayments disproportionately more for services needed by unprofitable beneficiaries than for other services after the HCC implementation. This induced unprofitable beneficiaries to voluntarily disenroll from their MA plans. Such strategic behavior led to reduced access to some services, such as home health care, resulting in delayed care and low care satisfaction. We estimated that this led MA plans to save $5.9 billion by transferring the costs to the federal government, thereby placing significant financial burdens on the federal government.

In Chapter 3, we proposed an alternative risk adjustment model from the perspective of statistics. The HCC model has two limitations, leading to the unintended patient selection. First, the HCC model uses a poor statistical model for health care expenditure data. Second, the HCC model is designed to be merely accurate at the group level, thereby generating mispredictions at the individual level. This induces MA plans to avoid beneficiaries with higher expenditures than their estimated payments. To identify the best predictive statistical model, we performed a comprehensive comparison of prediction accuracies across 19 statistical models. To evaluate model performance, we developed a new evaluation metric for measuring how close predicted expenditures align with actual expenditures at the individual level. Then, we compared the group-level and individual-level prediction accuracies of the 19 models, including 7 machine learning

models. However, we found that no one model performed best in all prediction accuracies. This finding suggests that an optimal risk adjustment model cannot be determined solely based on statistical metrics.

In Chapter 4, we proposed an alternative risk adjustment model from the perspective of statistics. The HCC model generates underpayments for services more used by unprofitable beneficiaries, thereby inducing MA plans to engage in service-level selection. Using machine learning techniques with big data, we developed an alternative risk adjustment method that accounts for each individual's future service-level use, thus generating economic incentives for MA plans to discourage service-level selection. Then, to determine which model is the best model to address service-level selection, we performed a paired comparison of our proposed method with the HCC method across the 19 estimators used in Aim 2. We found that our proposed model improved group-level and individual-level prediction accuracies, especially for those with high expenditures. The improvement was more pronounced in several machine learning models. This finding suggests that accounting for service-level propensity scores in risk adjustment with machine learning have the potential to more accurately estimate plan payments, potentially reducing service-level selection.

This research provides key policy implications for reform of the health care systems in the United States. First, this research sheds light on the mechanism in which MA plans could effectuate the unintended consequence of the HCC model. Second, this study offers policy implications for CMS in improving risk adjustment methodology that estimates payments to MA plans more accurately while discouraging service-level selection. If our proposed model can contribute to markedly reducing service-level selection incentives by MA plans, it would enable Medicare beneficiaries to receive the right care in the right place at the right time. It would also contribute

to reducing overpayments to MA plans, indicating the potential financial gain to the federal government. This research can be also applied to various payment and delivery systems. For example, under the Affordable Care Act (ACA), by law health plans can no longer differentiate premiums by health status, possibly inducing incentives to engage in risk selection. Constructing better Medicare's risk adjustment systems will help to inform policies to protect risk selection under ACA.

VITA

Sungchul Park, MPH, is a doctoral candidate in the Department of Health Services. He earned his bachelor's degree in both economics and financial engineering from Korea University (South Korea) in 2010, and MPH degree from Kyoto University (Japan) in 2013. He is anticipated to receive his PhD in Health Services from the Department of Health Services at the University of Washington in June 2018, with an emphasis on Health Economics. Sungchul's research lies at the intersection of health economics and health policy, specifically in the areas of health care financing and insurance; payment and delivery systems; long-term care; policy evaluation; and racial/ethnic/socioeconomic disparities affecting the above areas. He is most interested in developing new methodologies grounded in policy, through use of predictive modeling, simulation modeling, and machine learning and big data analytics. He will be continuing his research in health economics and health policy by starting in the fall as an assistant professor at the Department of Health Management and Policy at the Dana and David Dornsife School of Public Health, Drexel University.