Utilization Thresholds in Risk Adjustment Systems

Eran Politzer*

Abstract

Risk adjustment systems, that reallocate funds among competing health insurers, often use risk adjustors that are based on utilization. The level of utilization that triggers an adjustor – the utilization threshold – is frequently chosen implicitly and uniformly. I study utilization thresholds empirically in the setting of the U.S. Marketplaces. I demonstrate how an explicit choice of such thresholds, tailored to each adjustor, may improve the prediction fit of the risk adjustment system and decrease the incentives to game it. Using simulations, I find that a single alternative threshold may improve the prediction fit in some disease groups by up to 14%. A choice of multiple utilization thresholds, guided by a regression tree algorithm, may improve fit furthermore while taking into account the effect on gaming incentives.

Keywords: Health insurance, risk adjustment, utilization threshold
JEL Classification: I11, I13

*The Department of Health Care Policy, Harvard Medical School, and Federmann School of Public Policy and Governance, The Hebrew University of Jerusalem (email: eran.politzer@mail.huji.ac.il)

1 Introduction

Risk adjustment schemes are a cornerstone of a functioning managed competition market for health insurance. They reallocate funds among competing health plans based on the risk of their enrollees, and by that decrease plans' incentives to select profitable (typically healthier) enrollees and deter unprofitable ones (see Ellis, Martins, and Rose (2018) and Layton et al. (2018) for a review). Such selection incentives may hurt consumers' welfare by distorting the behavior of insurers, the design of their services and networks (Glazer and McGuire 2000), and the sorting of consumers between plans (Einav and Finkelstein 2011). To improve their predictive accuracy, risk adjustment systems have long ago advanced from relying only on age and gender adjustors that are exogenous to plans' influence, and now often use adjustors established from medical claims.[1] [2] All these adjustors depend on enrollees' utilization of services either directly, e.g. adjustors based on the utilization of prescription drugs, or indirectly, e.g. diagnoses-based adjustors that are established during provider-patient interactions (Geruso and McGuire 2016). Any adjustor based on utilization requires a decision regarding the minimum level of utilization that will trigger the adjustor – the utilization threshold. This decision is often made implicitly and uniformly for all adjustors. For example, when setting diagnoses-based adjustors, policy makers rarely debate what is the number of times a diagnosis has to appear in claims over the year, implicitly choosing a utilization threshold of a single appearance for all adjustors. This is not the only possible choice – for example, the risk adjustment system in Germany's Social Health Insurance requires that out-patient diagnoses appear twice over the year, in two separate quarters. Explicit thresholds for adjustors based on the use of prescription drugs are also used in some non-U.S. countries (see Table A1 in the Appendix), but in most cases thresholds are uniform for all drug-adjustors.

This paper studies utilization thresholds and examines how their level may affect the performance of the risk adjustment model. Explicit thresholds would be desirable when a certain level of utilization is more predictive of spending and less prone to gaming, comparing to the baseline. I show that finding optimal thresholds is an

1. The approach in this paper aligns with conventional risk adjustment, that aims to pay insurers as close as possible to the expected cost of their enrollees. An alternative approach – optimal risk adjustment – examines risk adjustment as a tool to incentivize insurers and thus achieve an economic goal, such as efficiency. See Glazer and McGuire (2012) for a review.

2. Better predictive accuracy may decrease plans' incentive for cost saving – a tradeoff long acknowledged in the literature. Moreover, using adjustors that are endogenous to plans' influence may create opportunities for manipulation.

empirical question, and they may be unique for each adjustor. A simple example, presented in section 2, demonstrates how a threshold for a drug-based adjustor may increase the fit, relative to a zero-threshold baseline, when patients' total costs increase with higher utilization of the drug. The same threshold would decrease the fit, when total costs are a decreasing function of the drug's utilization. Furthermore, a higher threshold does not necessarily reduce the incentives for gaming, as the opportunity for gaming depends on the existing utilization patterns – a higher threshold may increase the number of patients susceptible to gaming if a larger group is left just below it. Higher thresholds may also increase the potential revenues from gaming if the payment per patient above the threshold rises as the threshold increases.

While several established measures exist for the prediction fit of a risk adjustment system, there is no consensus on how to measure the ex-ante incentives to game the system by increasing utilization. I suggest new measures, based on the potential net revenue plans can gain from gaming an adjustor. The measures focus on patients that may be susceptible to gaming – those that already have some utilization (e.g. filled a prescription for insulin, included in a drug-adjustor), and those with a potential for new utilization (e.g. diagnosed with diabetes, but have no insulin prescription). The main measure limits the scope of the gaming activity so it applies only to patients on the margin of crossing the utilization threshold, making the measured incentives more actionable and realistic (e.g. it limits additional prescribing to no more than 30 days of supply). While this paper studies ex-ante gaming-incentives, it does not examine the ex-post response of insurers to these incentives. Nor does it distinguish between welfare-increasing and welfare-decreasing gaming.

In addition to setting the level of the utilization threshold for an existing adjustor, I examine also the general case in which multiple thresholds may be set, possibly adding new adjustors to the system. The choice of the number of thresholds and their levels is again an empirical question. This choice could be guided by machine learning algorithms that tackle a parallel challenge of splitting data – dividing observations into groups with the most homogeneity (See Ellis, Martins, and Rose (2018) for a review of the use of such algorithms for risk adjustment). I employ a regression tree algorithm – CART – to choose the utilization thresholds. The algorithm's loss function favors splits that increase the system's fit, while minimizing the potential revenue to insurers from gaming the system. I explore several possible weights on fit vs. gaming-incentives.

I study utilization thresholds empirically in the setting of the ACA Marketplaces,

mainly examining days' supply thresholds for prescription-drug adjustors. I also explore thresholds for diagnosis-related adjustors, based on the number of times a diagnosis appears in patients' claims.[3] To simulate thresholds, I use the IBM Truven Marketscan database, that holds claims from employers and commercial health plans, and was used to develop the Marketplaces' risk adjustment model (Kautter et al. 2014). I use the data from 2015 and 2016 for calibration and learning, and apply the risk adjustment model to enrollees in 2017.

For ten drug-adjustors (RXCs), I simulate multiple days' supply thresholds of up to 360 days, re-estimating the model's coefficients in each iteration. The results show a unique pattern of the prediction fit and the gaming incentives for each adjustor.[4] For six adjustors, an alternative utilization threshold would improve the fit for the disease group related to the drug (by up to 14% of the baseline fit) and the overall fit (by up to 0.1% of the baseline fit). The fit-maximizing threshold is 30 days' supply for anti hepatitis-C agents, 120 days for immune suppressants, 150 days for multiple sclerosis agents, 180 days for anti-HIV agents and for cystic fibrosis agents, and 210 days for non-insulin anti diabetes agents. In an important result, setting these utilization thresholds pose no trade-off between fit and gaming incentives as both are improved. Using my preferred measure – plans' potential revenue from gaming through addition of up to 30 days of supply – the gaming incentives are reduced by 25% to 96%, relative to the baseline. For four other drug adjustors – antiarrhythmics, phosphate binders, inflammatory bowel disease, and insulin – fit is maximized with the baseline zero-days threshold.

I show that utilization thresholds may be beneficial also for diagnoses-based adjustors by examining a threshold for the number of times a diagnosis appears in claims. A threshold that requires four appearances of the "Diabetes without Complications" group of diagnoses (CC21), would both increase fit and reduce the gaming incentives relative to the baseline single-appearance threshold.

Using regression trees, I find that multiple thresholds could be beneficial for some adjustors. For example, maximizing fit, the tree algorithm suggests five thresholds for RXC1 (anti-HIV agents): 0, 120, 180, 270, and 330 days' supply.[5] However,

3. The paper examines thresholds for each adjustor separately. In practice, thresholds would most likely apply to multiple adjustors at the same time. The methodology presented in the paper may be used to examine possible interactions between thresholds for several adjustors.

4. I calculate the "individual fit" – the R-square of a model that predicts individuals' costs using the adjustors included in the risk-adjustment formula.

5. I limit the algorithm to find regression trees with at most 3 levels, i.e. limit the number of thresholds to 7. I also restrict the thresholds to be multiples of 30 days, between 0 and 360.

when the loss function assigns some weight also to reducing gaming incentives, the algorithm's results change. With a weight of 0.2 on gaming incentives (vs. 0.8 on fit), only a single threshold of 150 days' supply is recommended. In some cases, multiple thresholds may improve fit, but violate common principles of risk adjustment systems. For example, the regression tree algorithm recommends four thresholds for RXC3 (0, 60, 180, 270), but requires adjustors with negative coefficients for all but the first threshold. Such coefficients would violate the monotonicity principle stated in Pope et al. (2004), requiring that insurers are not penalized for additional reporting.

The rest of the paper proceeds as follows. Section 2 demonstrates that finding a fit-maximizing threshold is an empirical question. Section 3 introduces measures for the incentives to game the system. Section 4 describes the risk adjustment model in the Marketplaces, and section 5 describes the data. Section 6 presents the simulations and the regression tree algorithm used to select thresholds. Section 7 examines the impact of thresholds for drug-adjustors on fit and incentives for gaming, and Section 8 explores these effects for morbidity-based adjustors. I discuss the results in section 9. Section 10 concludes.

2 The Choice of Utilization Thresholds

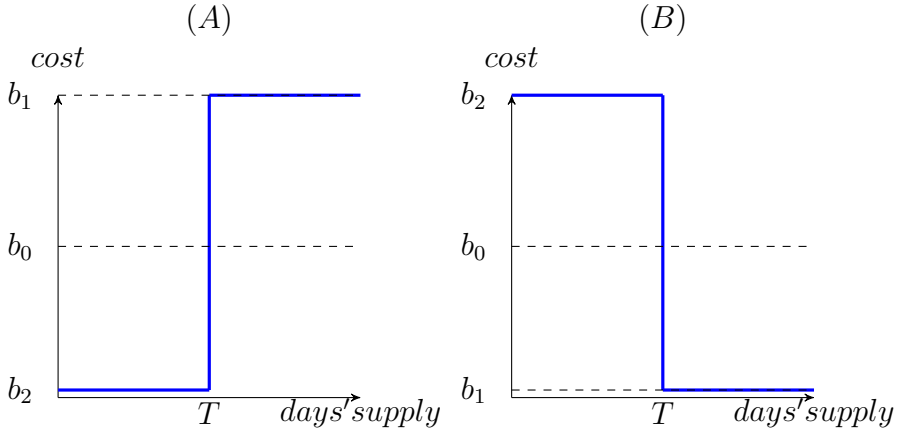
2.1 A single utilization threshold

Intuition may suggest that a higher utilization threshold always harms the fit, as information about some utilizers seems to be ignored. However, this intuition is wrong as a rule. Consider a simple risk adjustment system with only one adjustor that indicates the use of drug X. For each patient using the drug, a plan receives a risk-adjustment payment that equals the average of the additional costs for all drug-X users.[6] With a non-zero utilization threshold, payment is the average additional cost of patients with utilization above the threshold. Suppose that the cost of drug X is negligible and the number of users is small relative to the number of non-users. Figure 1 presents two possible distributions of the additional cost of drug users, ordered by the number of days' supply in their prescriptions.

Costs may be higher for high utilizers of the drug (panel A) if a higher use indicates a severe chronic condition with additional co-morbidities. Alternatively, costs may

6. This will be the payment if the coefficient for the drug-X adjustor comes from an OLS estimation of enrollees' costs on a constant and the single adjustor. This kind of estimation is the typical way to set coefficients for adjustors in risk adjustment formulas.

Figure 1. Possible distributions of enrollees' costs, by days' supply of drug X



The figure presents two possible distributions of the additional costs for users of a certain drug ("drug X") by the number of days' supply in their filled prescriptions over the year. In panel A, high-cost patients have high utilization of the drug, while in panel B low-cost patients have higher utilization. b_0 is the average additional cost over all users of drug X. Hence, b_0 equals also the payment to the plan for each such patient. Setting a utilization threshold of T days' supply changes the payment due to patients below the threshold to zero. Patients above the T-days threshold have an average cost of b_1, and hence the plans receive a payment of b_1 for them. The threshold increases the individual fit in panel A as each type of patient receives the correct payment. However, the threshold decreases the fit in panel B, as all patients receive zero payment.

be lower for high utilizers (panel B) if higher and continuous use indicates a patient with good drug adherence and a controlled disease. With a zero-days threshold, the average additional cost for all patients with the drug-adjustor turned on is b_0, and hence the payment to the plan is also b_0. Setting a utilization threshold that requires prescriptions of at least T days' supply has a very different impact on the individual fit in these two cases. In panel A, setting a threshold T would improve the fit, as the adjustor would pay zero for low-cost patients and the correct cost b_1 for high cost patients. In panel B, setting an identical threshold of T days' supply, would decrease the fit as both high-cost patients and low-cost patients would receive zero payment. This is equivalent to eliminating the adjustor altogether and fit must be better with the adjustor than without it.

2.2 Multiple utilization thresholds

The case in the simple example above restricts the drug-adjustor to have a single utilization threshold. This section relaxes the restriction and discusses the choice of multiple thresholds, practically allowing to add new adjustors to the risk adjustment scheme. Any additional adjustor would weakly increase the fit of the scheme, as

long as overfitting does not become an issue due to small sample sizes or inflation of adjustors. In Panel A of Figure 1, when $b_2 > 0$, it is easy to see that adopting two thresholds – $0, T$ – paying b_2, b_1 accordingly, would lead to a higher fit than using any single threshold.

However, some additional adjustors may not be acceptable from policymakers' point of view. Returning to panel B of Figure 1, consider the choice of two thresholds – $0, T$. Patients with some utilization of the drug below T days, would receive a high compensation of b_2, while patients with utilization higher than T days would receive a lower amount of b_1 (or zero). It is easy to see that fit would be better with two thresholds – plans receive exactly the additional cost of each patient. However, a lower compensation when utilization increases would violate monotonicity – one of the common principles guiding the development of risk adjustment systems (Pope et al. 2004). This principle states that insurers should not be penalized for additional recording of diagnoses, or in our case – recording of additional days of supply provided to a patient. Violation of this principle may incentivize insurers to game the system by skimping, or by under-reporting the number of days' supply.

3 Measuring the Incentives for Gaming

A major concern when choosing adjustors for risk adjustment systems is that plans and providers should not be able to readily manipulate them to increase plan payments (Ellis, Martins, and Rose 2018). Unlike age or gender, adjustors based on utilization are susceptible to gaming. When The Centers for Medicare & Medicaid Services (CMS), the federal agency that administers the risk adjustment system in the Marketplaces, added drug adjustors to the risk-adjustment model, it acknowledged that this "may provide an incentive to overprescribe medications" (CMS 2016b). Such concerns have led CMS to exclude some drug groups from the model because medical professionals judged that they are "particularly subject to intentional or unintentional discretionary prescribing variation or inappropriate prescribing by health plans or providers" (CMS 2017). Gaming of drug-adjustors is a concern also in other risk adjustment systems, e.g. Lamers and van Vliet (2003) discuss the potential for gaming the drug adjustors in the Dutch risk-adjustment scheme.[7]

Unlike fit, there is no consensus about how to measure the incentives for gam-

7. The authors propose several measures to decrease the opportunities for gaming, among them – setting days' supply thresholds.

ing.[8] One indirect measure appears in Lamers, van Vliet, and van de Ven (1999) that examines the proposed Pharmacy-based Cost Group (PCG) model in the Netherlands. They calculate the ratio between the capitation payment for those assigned to one of the PCGs and their pharmacy costs. They find that revenues are on average about four times as high as the cost of drugs. While this ratio may suggest that there is some room for insurers to increase revenues by prescribing more drugs, it doesn't quantify the gaming incentives directly. The approach in this paper is similar to the way Behrend, Felder, and Busse (2007) examine the gaming opportunities of the then-proposed drug adjustors in the German risk adjustment system (IPHCC+RxGroups). The authors simulate three specific cases of plans' gaming behaviors and calculate their net monetary returns per insured person: substitution to an alternative drug for hypertension that leads to a higher risk score[9]; increasing the prevalence of antidepressants use among patients already diagnosed with depression[10]; and increasing the use of diabetes drugs by diagnosing previously unidentified diabetic patients and supplying them with very short prescriptions.[11]

This paper extends the scant literature about gaming incentives and suggests new and more general measures for plans' incentives to game the system by increasing utilization. I define gaming here as *any* increase in utilization that may lead to higher profit for the plan. Before turning to defining new measures, I note, first, that I quantify the *ex-ante* incentives to increase utilization and the paper does not study the actual *ex-post* response of the plans.[12] Like ex-ante fit measures that don't directly explore ex-post selection behaviors, measuring ex-ante incentives for gaming could assist policy makers to examine and compare proposals for risk adjustment

8. The literature is wider on the incentives for cost saving, that are related to incentives for gaming as both may depend on the effect of current plan's spending on its future revenue. To quantify incentives for cost-saving, Geruso and McGuire (2016) use the "power" concept to measure the share of costs borne by the plan at the margin. However, the power measure seems insufficient to measure incentives for gaming. Some gaming activities may entail no further utilization or cost at the enrollee level. Even when gaming requires additional utilization, e.g. prescribing additional days' supply, the resulting increase in revenue could be much higher than the additional cost, leading to a negative "power" measure – below the usual 0 to 1 range of this statistic.

9. The simulation moves all the patients prescribed with ACE inhibitors to Angiotensin II receptor blockers – an alternative drug that leads to a higher risk score.

10. To simulate the change, the authors randomly assign antidepressant drugs to patients diagnosed with depression, so the prevalence of the drug use increases by 30%.

11. This simulation examines an increase of 4 percentage points in the prevalence of diabetes treatment among the relevant age groups – a 33% increase of the baseline prevalence.

12. Most research of *ex-post* gaming around utilization thresholds focuses on "upcoding", which is essentially a gaming activity that pushes patients across a single-appearance threshold for diagnosis-based adjustors (for example, see Geruso and Layton (2020)).

schemes before they come into effect. The magnitude of the elasticity of gaming to gaming incentives is, however, unclear – the reluctance of providers to increase utilization and the potential cost for patients may inhibit gaming in practice, but plans may incentivize providers, change formularies, establish pre-authorization conditions, and set non-linear price schedules to convince providers and patients to increase utilization. Second, I note that gaming is not necessarily sub-optimal. Incentives to increase utilization may improve welfare if plans would have skimped on services or drugs without them. Lastly, I note that though this paper focuses on plans' gaming by increasing utilization, other forms of gaming may be possible. This includes delaying utilization around the end of a calendar year to cross a threshold in at least one of the years (this is especially relevant to short curative treatments), or gaming by decreasing utilization (e.g. when the risk adjustment system violates the monotonicity principle). Pharmaceutical companies and Pharmacy Benefit Managers may also participate in gaming through changes in drug prices and drug packaging.

To define measures of the incentives for gaming one has to choose first the *relevant population*, for which gaming is examined. This choice creates a tradeoff – widening the population may enable a more comprehensive examination of incentives, but in most cases these incentives will be less and less actionable. For example, a plan may have a theoretical incentive to prescribe *everyone* with a drug if the resulting payment is higher than the drug's cost. However, it will be very hard for a plan to act on such an incentive – make providers prescribe unnecessary drugs to healthy individuals, overcoming their reservations due to professional ethics and intrinsic concern for their patients, and convince individuals to fill these prescriptions. In contrast to that, it will be most likely much easier for a plan to make providers and patients lengthen justified prescriptions for patients that already use a drug, especially if there is a gray area around the desired utilization. A similar tradeoff exists when choosing the *gaming activity* that the measure examines – limiting the scope of the gaming activity will most likely make the incentives more actionable. For example, an incentive for a gaming activity that includes prescribing additional 180 days of supply is likely less actionable than an incentive to game the system by prescribing one more day of supply (regardless of the cost of the drug).

With these tradeoffs in mind, the measures I define focus on the population of patients with either an existing adjustor-related utilization or at least a potential for such utilization (e.g. patients with a prescription for a drug included in a drug-adjustor, or patients diagnosed with a disease related to the drug). Regarding the

scope of the gaming activity, my main measure focuses on a marginal gaming activity that limits the additional utilization (e.g. allowing no more than 30 days of additional supply when gaming a prescription). As this measure focuses on patients at the margin of passing the threshold, gaming their utilization is arguably easier and the incentives to do so are more actionable. In the Appendix (section A.1), I also examine a case of no limits to the additional utilization (e.g. adding any number of additional days of supply) – a measure that could serve as an upper bound of the gaming incentives for this activity. While these choices of population and gaming activity are definitely not the only ones possible, they may serve as a first-step in measuring gaming incentives.

To quantify the extent of the gaming incentives I use the potential net revenue to the plan from the gaming activity, per person in the target sub-population. These per-capita measures allow to compare gaming incentives between groups, and are compatible with gaming activities performed at the group-level. As the share of patients that could be gamed in each group may also matter, Figure A4 in the Appendix presents the share of enrollees that are "gameable", i.e. the enrollees for whom the gaming activity in my main measure yields a profit to the plan.

The incentives measure I define is presented below for the case of additional days of supply in drug adjustors, but can be easily adapted to reflect other types of utilization. Section 8 examines incentives for a gaming activity that adds a single office visit to increase the number of times a diagnosis appears in claims.

My main measure focuses on the disease group, limiting the scope of the gaming activity to prescribing no more than 30 additional days of supply to push patients across the threshold. The measure calculates the average net revenue from the gaming activity, per patient in the disease group. The measure's definition:

$$\begin{aligned} &\frac{1}{N}\sum_i \Delta R_i - \min_j(\Delta sup_{ij} * cost_j) \\ &s.t.\ sup_{ij} + \Delta sup_{ij} > T \ \ \& \ \ \Delta sup_{ij} <= 30 \end{aligned} \tag{1}$$

where N is the number of patients in the disease group, ΔR_i is the additional revenue to the plan due to pushing patient i over the threshold, sup_{ij} is the annual number of days' supply in patient i's prescriptions of drug j (included in the drug-adjustor), Δsup_{ij} is the number of additional days' supply prescribed as part of the gaming – limited to 30 days, and $cost_j$ is the daily cost of drug j. The incentive measure is calculated by using the cheapest way for patient i to cross the threshold of T days'

supply.

4 Drug-adjustors in the Marketplaces' Risk Adjustment

The risk adjustment scheme for the plans in the U.S. Marketplaces includes two components: the Department of Health and Human Services' (HHS) risk adjustment model and a transfer formula (Layton, Montz, and Shepard 2018). The basic model predicts this year's plan liability for enrollees based on their age, sex and the diagnoses drawn from their claims, producing a risk score for each person.[13] The transfer formula redistributes plans' premium revenues by the average risk score in each plan and other factors.

Beginning in the 2018 benefit year, CMS started using a "hybrid drug-diagnosis" risk adjustment model in the Marketplaces, adding adjustors indicating a filled prescription for the included drugs (CMS 2016a, 2016b). For example, a patient who filled a prescription for insulin will have a higher risk score, potentially increasing the risk adjustment transfer to her plan, whether she has a diabetes diagnosis in one of her claims or not. The drug adjustors are meant to indicate health risk when a diagnosis is missing. This can happen due to a mistake, to avoid stigma, or because the patient did not visit a physician. However, the drug adjustors appear independently in the risk adjustment model, and are not used only to turn on a related diagnosis-adjustor. The drug adjustors may also provide information on the severity of a diagnosed illness. To do this, the model includes interactions of drug-adjustors and their related diagnosis-adjustors. In the model, no minimum utilization is required for a prescription to increase a patient's risk score, e.g. a prescription of insulin for a single day will suffice to increase the score, and will have the same effect as a prescription for a year's supply.

The baseline risk adjustment model in this paper is the CMS 2019 model (*HHS-HCC V0519*), that includes ten drug-adjustors (RXCs). Each RXC is a prescription drug category that may include several drugs, identified by their National Drug Code (NDC). CMS chose RXCs that are closely related to diagnoses that were already included in the model within Hierarchical Condition Categories adjustors (HCC), that group diagnoses. Each RXC appears in the model as both an independent adjustor and within an interaction with its paired HCCs. Table 1 describes the RXC-HCC

13. The prediction model produces 15 sets of risk adjustment coefficients: three age-specific models (adult, child and infant), and five models specific for each coverage level in the Marketplaces (platinum, gold, silver, bronze, catastrophic).

disease groups in the 2019 model.

Table 1. Drug-Diagnosis Pairs in the 2019 Marketplaces Risk Adjustment Model

RXC	RXC Label	Related Diagnoses (HCCs)
1	Anti-HIV Agents	HIV/AIDS
2	Anti-Hepatitis C (HCV) Agents	Chronic Hepatitis C, Cirrhosis of Liver, End-Stage Liver Disease, and Liver Transplant
3	Antiarrhythmics	Specified Heart Arrhythmias
4	Phosphate Binders	End Stage Renal Disease, Kidney Transplant, Chronic Kidney Disease – Stage 5, Chronic Kidney Disease – Severe (Stage 4)
5	Inflammatory Bowel Disease Agents	Inflammatory Bowel Disease, Intestine Transplant
6	Insulin	Diabetes, Pancreas Transplant
7	Anti-Diabetic Agents, Except Insulin and Metformin Only	Diabetes, Pancreas Transplant
8	Multiple Sclerosis Agents	Multiple Sclerosis
9	Immune Suppressants and Immunomodulators	Rheumatoid Arthritis and Specified Autoimmune Disorders, Systemic Lupus Erythematosus and Other Autoimmune Disorders, Inflammatory Bowel Disease, Intestine Transplant
10	Cystic Fibrosis Agents	Cystic Fibrosis, Lung Transplant

The coefficients in the 2019 version of the model are based on an average of the coefficients separately estimated for the years 2014, 2015 and 2016 (CMS 2018b). CMS adjusts the coefficients post-estimation for clinical reasonableness and to decrease gaming. Fearing inappropriate prescribing when an inexpensive drug treats a medically expensive condition, CMS restricted the payment for two of the drug-adjustors included in the model – RXC3 (Antiarrhythmics) and RXC4 (Phosphate Binders) – to less than the average cost of supplying the drugs.[14]

CMS considered in a 2016 White Paper (CMS 2016a) whether to require a utilization threshold to trigger a drug indication – either require multiple prescriptions for the same drug, or prescriptions totalling at least 30 or 60 days' supply. CMS' clinical consultants suggested that for some potential RXCs, a minimum days' supply utilization threshold would be useful to distinguish severely ill patients from those

14. Payment due to these RXC adjustors was a priori set to be equal to the average annual per capita cost of the drugs in the RXC (in the calibration dataset). In addition to that, the RXC-HCC interaction term was set to zero for both RXCs.

with milder conditions. However, CMS decided not to include these RXCs in the model, requesting feedback from the public.

Prescription drugs serve as adjustors in risk adjustment models in other countries as well. In most cases, some minimum utilization threshold is required to trigger an indication. In Germany, 183 days' supply are required for drug adjustors to validate most chronic diseases, 42 days are required for diseases with medication to be taken as needed, and 10 days are required for acute diseases. Switzerland, the Netherlands and the Czech Republic demand prescriptions of at least 180 days' supply for most drug groups. The Netherlands has a 90 days threshold for some specific groups, and no threshold at all for extremely high-cost drugs. See Table A1 in the Appendix for more details on the use of drug adjustors in these countries.

5 Data

This paper uses the IBM Truven MarketScan database of medical claims from the employer-sponsored insurance market to measure spending, record diagnoses, and examine the utilization of prescription drugs. Utilization of drugs is measured by the number of days' supply, i.e., the number of days for which supply will last for the patient when using the maximum dose prescribed.[15] The Truven database was used to develop the original Marketplace payment system (Kautter et al. 2014), and until recently was used exclusively in updating it.[16] I estimate the risk adjustment coefficients using the 2015 and 2016 versions of the database, and use the 2017 version to simulate payments under different utilization thresholds (two more scoring years, 2018 and 2019, are used in the Appendix). The analytic sample is composed of adults, between ages 21 and 65. It includes individuals who had coverage for both prescription drugs and mental health, were continuously enrolled for twelve months, and had fee-for-service claims data for the whole period (i.e. no encounter data from managed care plans).[17] Enrollees with a negative sum of their total spending for the year are excluded from the sample. Table 2 reports summary statistics for the 10,898,743 individuals in the analytic sample. The average annual cost of these enrollees is $7,049 (this sum includes plans' spending on in-patient and out-patient

15. This measure, appearing in U.S. pharmacy claims, is different than the number of Defined Daily Doses – a uniform standard dose defined for each drug by the World Health Organization.

16. Starting in the model for 2019, CMS is gradually shifting to using claims data from the plans in the Marketplaces themselves (EDGE data), instead of the Truven database.

17. Underpayment for partial year enrollees (Ericson, Geissler, and Lubin 2018) may be exacerbated by utilization thresholds. This issue is not examined in the paper.

services, as well as on prescription drugs, and also out-of-pocket payments by the enrollees). 6.6% of enrollees have a prescription for a drug included in one of the ten RXC drug-adjustors. The cost of treating these patients is 4 times higher than the cost of the average enrollee. Table 2 also presents the share of patients and the average cost for each RXC-HCC disease group, and within it – for patients with a prescription for the RXC drugs.

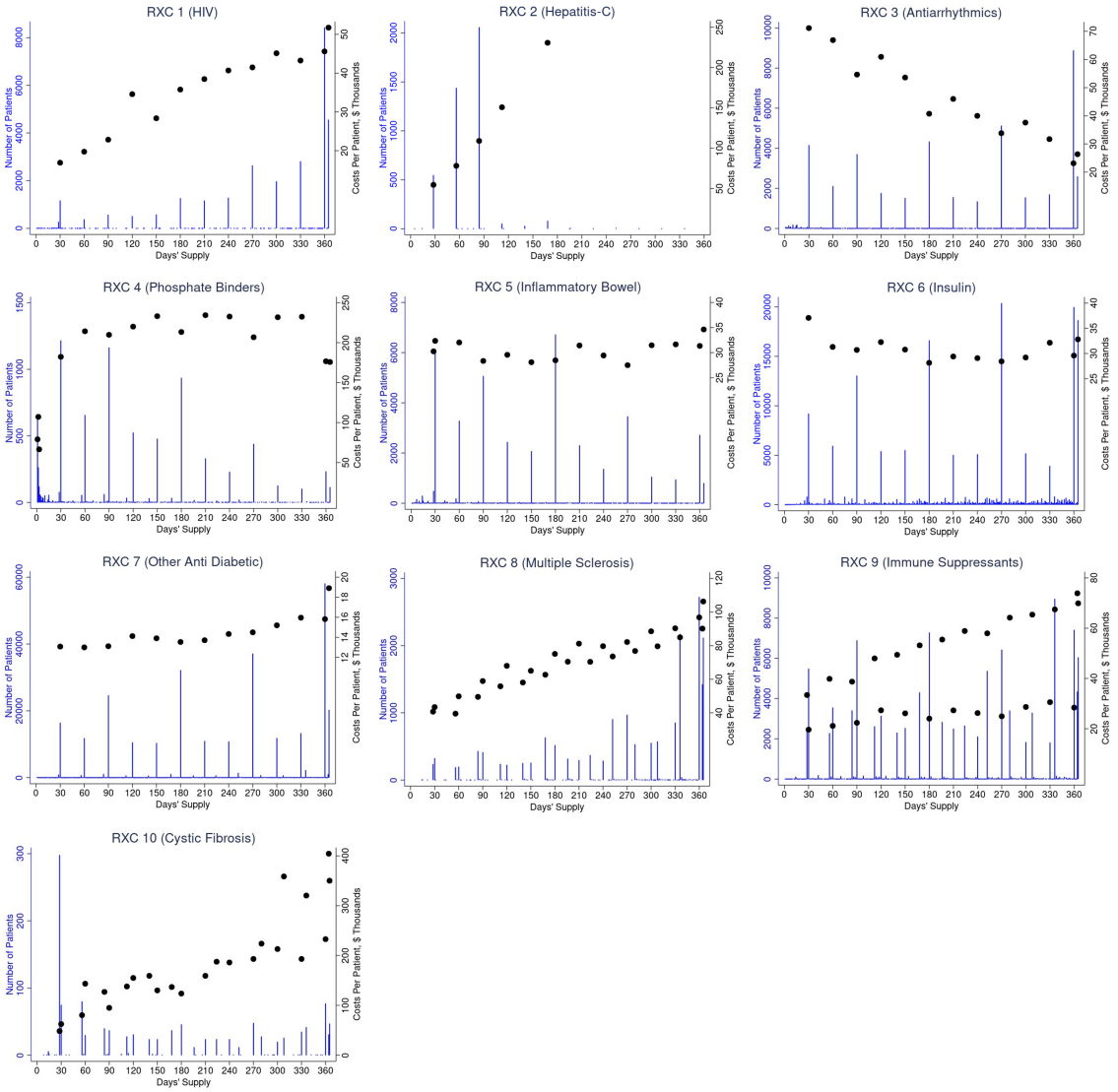
Table 2. Descriptive statistics for the 2017 sample (N=10,898,743)

Variable	Share of Enrollees (%)	Mean Cost ($)		
All	100	7,049		
Share of:				
Females	52.4	7,830		
21-29	17.7	3,850		
30-39	20.3	5,148		
40-49	23.5	6,425		
50-65	38.6	9,893		
Patients w. Any RXC	6.6	28,962		
	Diagnosed or Prescribed		Prescribed	
Disease group	Share of Enrollees (%)	Mean Cost ($)	Share of Disease Group (%)	Mean Cost ($)
RXC-HCC 1 (HIV)	0.27	40,787	94	41,873
RXC-HCC 2 (Hepatitis-C)	0.28	53,900	14	96,991
RXC-HCC 3 (Antiarrhythmics)	1.33	34,626	30	43,457
RXC-HCC 4 (Phosphate Binders)	0.24	107,608	34	185,153
RXC-HCC 5 (IBD)	0.73	33,457	53	30,238
RXC-HCC 6 (Insulin)	7.38	18,983	24	30,940
RXC-HCC 7 (Other Diabetes)	7.57	18,669	35	14,968
RXC-HCC 8 (Multiple Sclerosis)	0.27	65,417	62	82,007
RXC-HCC 9 (Immunosuppressants)	2.34	32,534	44	42,648
RXC-HCC 10 (Cystic Fibrosis)	0.02	136,067	58	152,230

This table presents summary statistics for the analytic sample. For each RXC-HCC disease group (named by the drug category), the share of enrollees and their mean cost is shown for both the whole group, and the subgroup of patients prescribed with an RXC drug (these patients may also be diagnosed with a related disease). Costs include both plans' spending and out-of-pocket payments by patients.

For each RXC drug group, Figure 2 presents the distribution of patients with a prescription, and their total annual costs, by their annual number of days' supply.

Figure 2. The distribution of prescribed patients and their average annual cost, by annual number of days' supply



For each RXC, the graph shows: 1. The number of patients by the number of annual days' supply in their prescriptions (bars, left axis); 2. The average annual costs of patients by the number of annual days' supply in their prescriptions (dots, right axis). Costs are shown only for days' supply categories with at least 1% of the prescribed patients in the RXC. Patients with no prescription are excluded. All patients with 365 days' supply or more are top-coded to the 365 days category (average costs are weighted by the number of patients). For RXC9, the apparent two series of costs are costs for patients with days' supply in multiples of 30, and costs for patients with days' supply in multiples of 28 (see text below).

The numbers of days' supply are mostly bunched in multiples of 30-days or 28-days.[18] Many patients have less than a year of supply even for chronic conditions, possibly reflecting partial adherence, a new diagnosis during the year, or a change in medication. Patients with more than a year-worth of supply are top coded in the figure and included in the 365 days' supply category.[19] Figure 2 also shows the average annual costs of patients for each days' supply category. Figure A2 in the Appendix breaks down these costs by the type of spending: RXC drugs, other drugs, outpatient, and inpatient. Costs increase as a function of the number of days' supply for RXCs 1,2,8 and 10, mostly due to increase in the costs of the RXC drugs themselves. Costs decrease with days' supply for RXC 3, where the spending on the cheap RXC drugs remains low, while inpatient costs decrease with a higher use of the RXC drugs. Costs are mostly stable for RXCs 4, 5, 6 and 7, sometimes due to decreasing inpatient costs offsetting the increase in the costs of RXC drugs. For RXC 9, costs are stable for patients with 30-days multiples of prescriptions, and are higher and increasing for patients with 28-days multiples. The diverging costs are due to inclusion of both cheap and expensive drugs in this category.[20]

I use the 2019 HHS-HCC risk adjustment methodology, implemented in CMS' *HHS-HCC V0519* software,[21] to calculate the risk scores, and thus the risk adjustment payment for each person. The V0519 methodology is used to decide which adjustors should be turned on for each enrollee. To calculate the risk scores for each enrollee the adjustors' vector is multiplied by their corresponding coefficients, which I re-estimate in each step (see details in the next section). While CMS adjusts the risk adjustment coefficients post-estimation for clinical reasonableness and to decrease gaming, I apply no restrictions on the estimated coefficients. The correlation between the predicted spending by the CMS model and its coefficients and my baseline model, estimated on my sample, is 95.3%. The individual fit of my baseline model is 40.1%, close to the 41.2% to 41.4% range of fit that CMS published for their 2019 model (CMS 2018a).

18. The bunching presumably reflects the common packaging of the RXC drugs (e.g. a month-worth of supply for a drug taken once a week will often have 28 days of supply – 4 weeks). It happens despite the growing popularity of 90-days prescriptions in both mail-orders and retail pharmacies.

19. During a single year, some patients may fill prescriptions with more than a year's worth of supply if, for example, a long prescription is filled toward the end of the year.

20. The RXC group includes both cheap drugs dispensed in daily pills – mainly Hydroxychloroquine Sulfate, and very expensive biologics dispensed in a weekly injection (e.g. Humira).

21. The 2019 HHS Risk adjustment software can be downloaded here: https://www.cms.gov/CCIIO/Resources/Regulations-and-Guidance/

6 Empirical Methods

6.1 Simulations of multiple levels for a single threshold

I use simulations to examine the effect of a single days' supply utilization threshold on the model's fit and the incentives for gaming. I use the CMS software to turn on the risk adjustors for all the data years and then, for each drug group (RXC), simulate thirteen thresholds – all the 30-days multiples between 0 and 360 days. Each simulation of a single threshold includes four steps:

1. **Adjust the drug-adjustor**: Turn off the drug adjustor for all enrollees whose annual number of days' supply is lower than the simulated threshold. This step is done for all the data years.[22]

2. **Reestimate model coefficients**: Reestimate the risk adjustment model using the enrollees' revised risk adjustors (from the previous step) for each of the years 2015 and 2016. The dependent variable in these estimations is the annual total cost of the enrollee.[23] Average the coefficients from the two estimations to get the payment coefficients for the modified 2017 risk adjustment model.

3. **Recalculate risk scores**: Calculate new risk scores for the 2017 enrollees, using the revised adjustors for this year (step 1), and the new coefficients (step 2). Calculate the payment (i.e. the predicted cost) for each enrollee.

4. **Calculate fit and gaming incentives**: Calculate the fit measures, comparing the annual 2017 costs with the updated payments (step 3). Calculate the incentives to prescribe more to cross the simulated threshold.

For each simulated threshold, individual R^2 fit measures are calculated.[24] To assess gaming incentives, I calculate for each patient the additional revenue the plan

22. The number of days' supply is summed up at the single-drug level (NDC-by-NDC) and not at the RXC level (by drug group). e.g. two 90-days prescriptions for different drugs in the same RXC do not sum up to cross a 180-days threshold.

23. I use the cost not covered by the Marketplace risk sharing scheme, that pays 60% of costs above $1 million. This cost is also used for fit calculations. "Payment system fit" is a more general fit statistic to calculate the fit including these omitted costs. See Layton et al. (2018) for details on this measure.

24. Risk adjustment models are essentially prediction models, and a fit statistic measures the accuracy of their prediction. The most common fit statistic is the R^2 individual-level fit. This is the R^2 of a regression of enrollees' cost on the adjustors included in the risk model. Two fit measures are calculated: First, the individual fit for the entire population of enrollees. Second, the individual fit for the patients in each RXC-HCC disease group.

would receive if the patient crosses the threshold.[25] I also calculate the minimum cost of the additional days of supply required for the crossing.[26] [27]

6.2 Regression trees for multiple thresholds

Regression trees may guide the choice of multiple utilization thresholds for a drug-based risk adjustor, or for any adjustor based on utilization. This choice requires partitioning the range of patients' utilization by a splitting variable, so more cost-homogeneous groups of patients are formed, thus improving the model's fit. Regression trees may help us find such a partition. The paper employs a common algorithm that identifies such trees – the Classification And Regression Tree algorithm (CART). For drug-based adjustors, the splitting variable is the number of days' supply in patients' prescriptions. I use a version of CART to choose days' supply thresholds: At the root node, the algorithm splits the entire range of days of supply, finding the first threshold. It recursively searches for an additional threshold in each subset of this range (i.e. in each child node), below and above the chosen threshold. CART is a "greedy" algorithm – it finds a *local* optimum for a partition in each iteration and not a *global* optimum, but it provides an efficient way to guide the choice of thresholds. To identify the best threshold at each tree node (i.e. for each new adjustor), I simulate multiple additional thresholds as described in the previous section. The chosen threshold at each node is the one that minimizes a loss function that may consider changes in both fit and the cumulative 30-days gaming incentives, relative to the baseline of a single zero-days threshold:

$$(1-\lambda)(1-\frac{Fit_{n,t}}{Fit_{1,0}})+\lambda\frac{GamingIncentives_{n,t}}{GamingIncentives_{1,0}} \tag{2}$$

where $Fit_{n,t}$ is the R-squared individual fit for the disease group after adding the utilization threshold t at node n, $Fit_{1,0}$ is the baseline fit with a single zero threshold. $GamingIncentives_{n,t}$ is the cumulative incentives for gaming by prescribing up to

25. This calculation uses the coefficients of the risk adjustors re-estimated in each simulation.

26. I assume that an additional supply that allows the patient to *just* cross the threshold is possible. In practice, the cost of gaming could be higher if prescriptions must be rendered in multiples of certain days' supply due to availability of specific dosages. Costs could also be higher if adding days of supply requires an additional service from a provider (e.g. an office visit). The costs of gaming could be lower if plans shift patients from one drug, not included in the RXC, to another one that is included in the RXC. I ignore all such options in the calculations.

27. For each patient with an existing RXC prescription, the cost of an additional day of supply is the average of the daily cost in her own prescriptions. For diagnosed patients with no prescription, the cost of an additional day is the average daily cost for all prescribed RXC patients.

30 more days of supply to patients in the disease group, after adding threshold t at node n. $GamingIncentives_{1,0}$ is the baseline gaming incentive. λ is a parameter that determines the relative weight the loss function applies to the change in gaming incentives versus the change in fit.

I limit each regression tree to three levels, i.e. no more than seven thresholds, restrict the possible thresholds to multiples of 30 days, between 0 and 360, and avoid thresholds with a difference of 30 days or less from the previous threshold in a parent node. These restrictions are mainly employed to limit the computational resources required to create the regression trees, and may be relaxed in alternative versions. However, increasing the number of thresholds and reducing the distance between them may increase the risk of overfitting the data, and may require other stopping criteria for the recursion (e.g. a minimal number of patients within the partition).

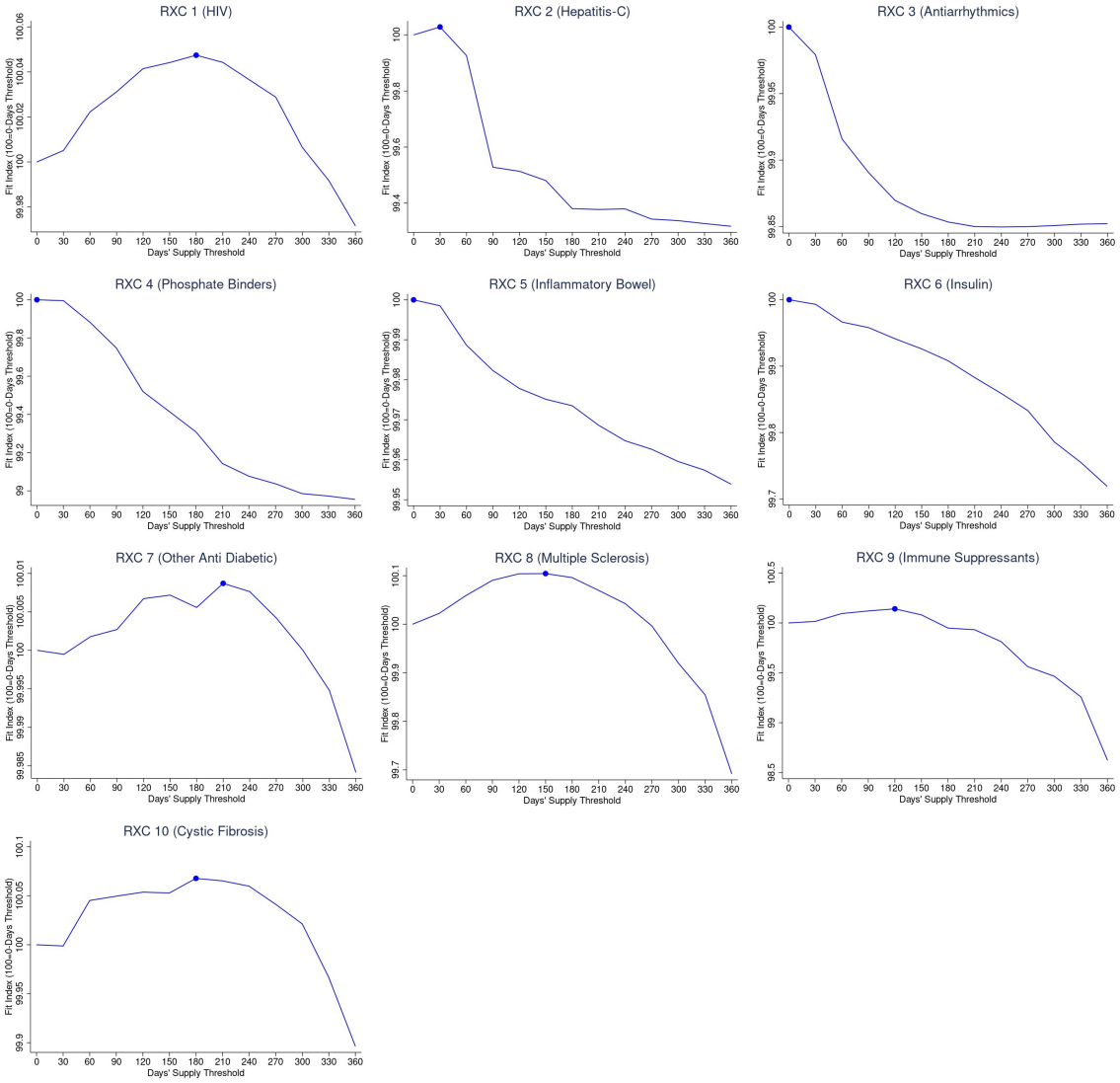
7 Results

7.1 Simulation of days' supply thresholds for a single adjustor

For each threshold, two individual R^2 fit measures are presented: First, Figure 3 presents the individual fit for the entire population of enrollees. Second, Figure 4 presents the individual fit for the patients in each RXC-HCC disease group. This group includes the patients with any prescription for a drug in the RXC and the patients that are diagnosed with one of the related diseases (described in Table 2). Both figures present the results by days' supply thresholds of each RXC. They normalize the baseline zero-threshold fit to 100, and hence show the percent change in the fit, relative to the baseline. A circle on each line denotes the fit-maximizing threshold.

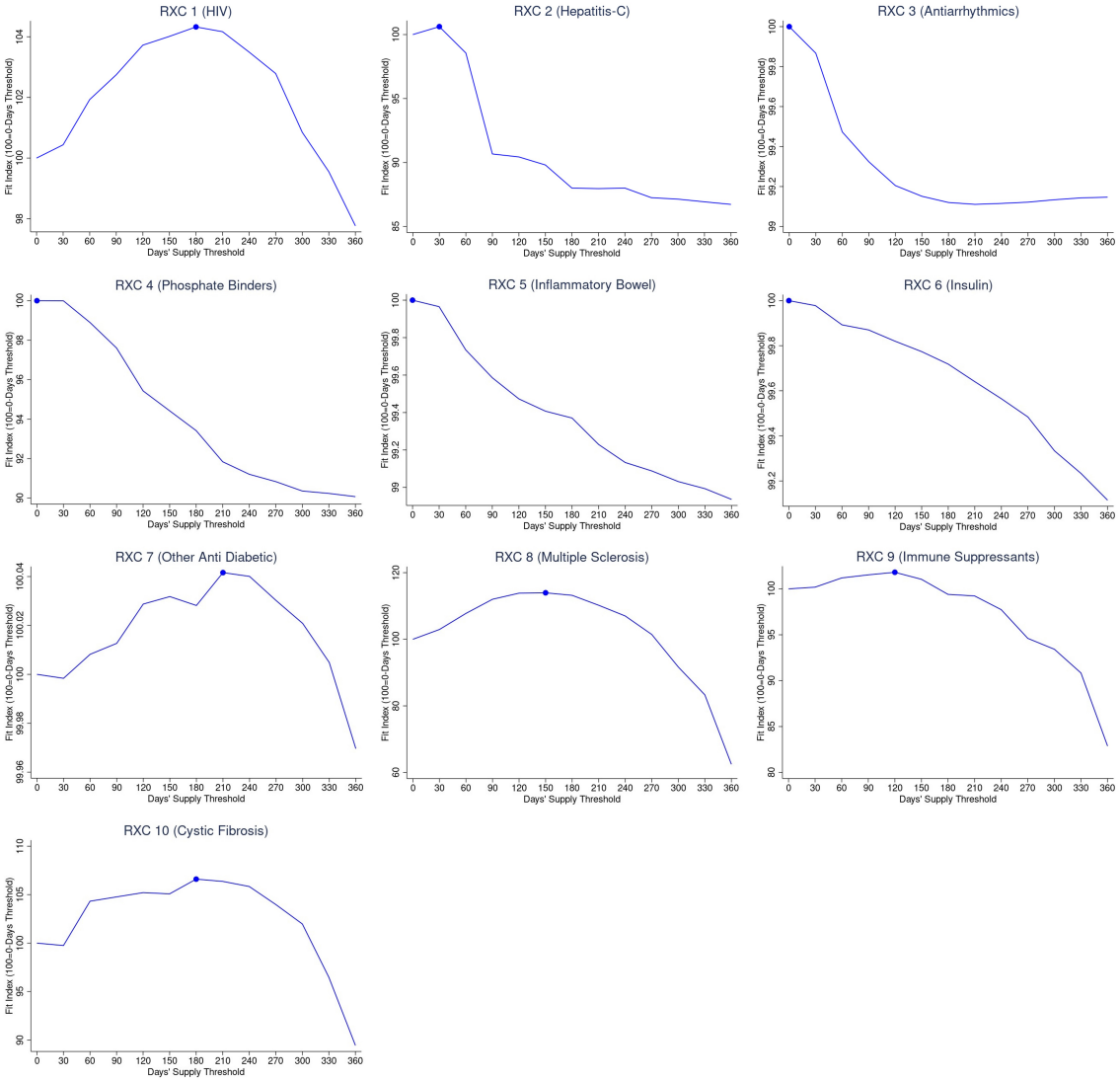
The results show a unique impact of thresholds on the fit for each RXC. While the overall fit in the population doesn't vary a lot since the number of patients in most groups is quite small (Figure 3), the fit for patients within each disease group may be affected in a significant way by the choice of the threshold (Figure 4). For four drug groups – RXC3 (Antiarrhythmics), RXC4 (Phosphate Binders), RXC5 (Inflammatory Bowel Disease), and RXC6 (Insulin) – both fit measures are maximized with the baseline zero-days utilization threshold. For six other RXCs, non-zero days' supply thresholds may improve both the overall fit and the fit for the disease group. The greatest improvement is achieved in RXC8 (MS Agents), where a threshold of 150 days maximizes the fit in the disease group, improving it by 14%. Overall fit increases by 0.1%. 180-days thresholds maximize the fit for RXC1

Figure 3. Individual fit for all the enrollees, by days' supply threshold (0-days=100)



For each RXC drug-adjustor, the figure presents an index of the individual R^2 fit statistic for all enrollees under each of the thirteen simulated days' supply thresholds between 0 and 360. The baseline zero-threshold fit is normalized to 100, and hence the graphs show the percent change in the fit, relative to the baseline. A circle on each line denotes the fit with the fit-maximizing threshold.

Figure 4. Individual fit for enrollees in the RXC-HCC disease group, by days' supply threshold (0-days=100)



For each RXC drug-adjustor, the figure presents an index of the individual R^2 fit statistic, for enrollees in the RXC-HCC disease group, under each of the thirteen simulated days' supply thresholds between 0 and 360. The baseline zero-threshold fit is normalized to 100, and hence the graphs show the percent change in the fit, relative to the baseline. A circle on each line denotes the fit with the fit-maximizing threshold. The RXC-HCC group includes patients prescribed with a drug included in the RXC or diagnosed with a related diseases (HCC).

(Anti HIV Agents), improving the fit in the disease group by 4.3%, and for RXC10 (CF Agents), where fit in the disease group increases by 6.6%. For RXC9 (Immune Suppressants), a 120-days threshold improves the fit in the disease group by 1.8%. The improvement in fit is smaller for RXC2, where a 30-days threshold increases the fit in the disease group by 0.6%, and for RXC7 (Non-Insulin Anti Diabetic Agents), where a 210-days threshold increases the fit in the disease group by only 0.04%.

Figure A3 in the Appendix presents the fit results of the simulations for two additional years – 2018 and 2019, using the same calculated coefficients. In almost all RXCs, the fit-maximizing thresholds are identical or very close in all the three years, implying a significant amount of stability of the results over time.

7.2 Determinants of the impact of thresholds on fit

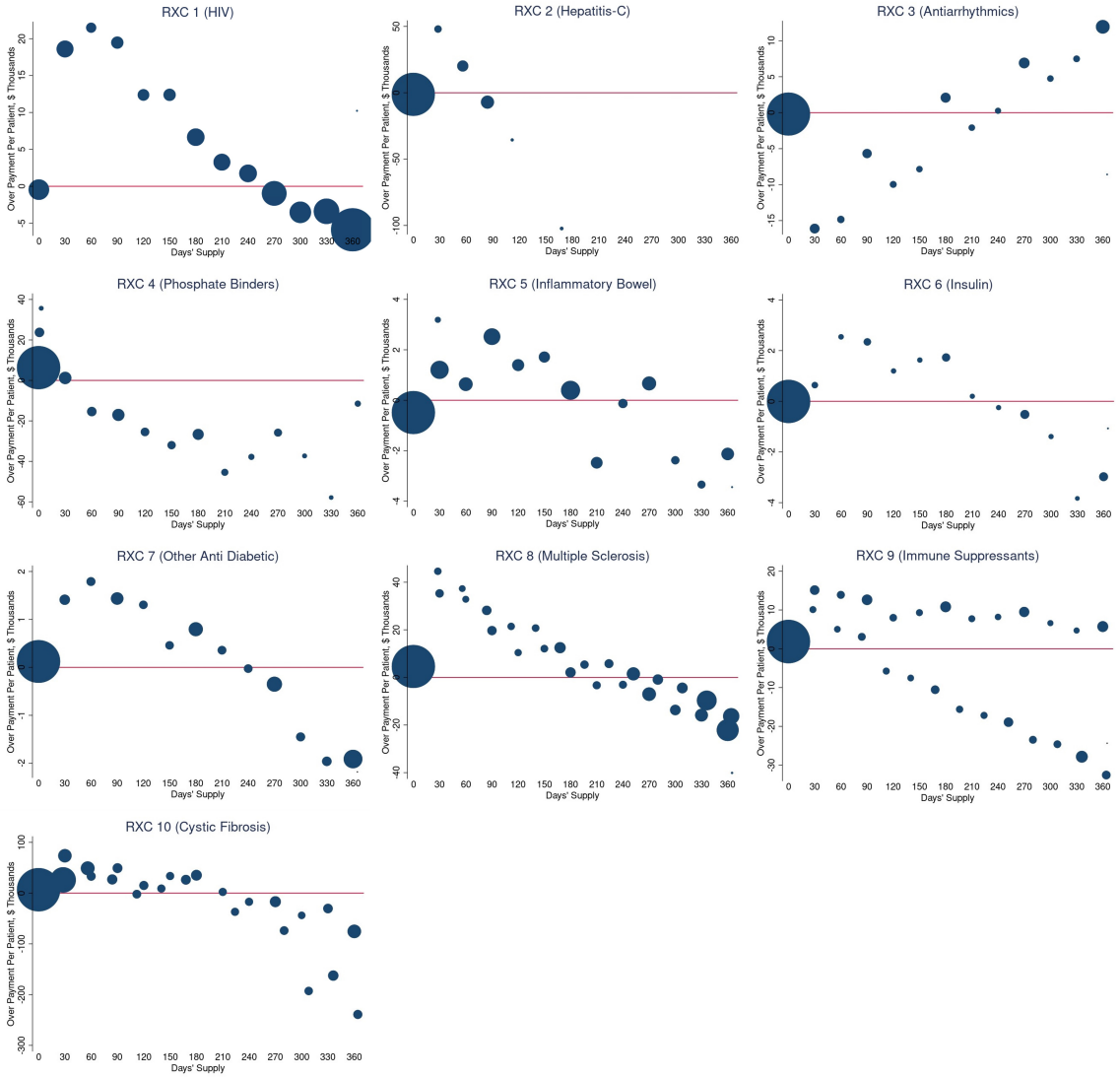
The individual fit statistic is based on the difference between each enrollee's actual costs and the costs predicted by the risk adjustment model. To better understand the effect of non-zero thresholds on the fit, Figure 5 explicitly examines these prediction errors in each disease group. By the number of days supply of RXC drugs, the figure presents patients' over- or under-compensation at the baseline with a zero-days threshold.[28] The size of the bubbles in the graph indicates the number of people in each days' supply group (normalized separately for each disease group).

In the groups related to RXCs 1,2,7,8,9, and 10, patients with shorter prescriptions are mostly over compensated while patients with longer prescriptions are under-compensated. This may explain why a non-zero threshold improves the fit for these RXCs – it decreases payments for the over-compensated below the threshold, as the drug-adjustor is turned off for them, and increases payments to the under compensated above the threshold, as the adjustor's coefficient is re-estimated to better fit their higher costs. This also confirms the intuition from the simple example presented in Figure 1 – for almost all these RXCs, cost is an increasing function of the number of days' supply (see Figure 2). In such cases, a non-zero threshold may improve the fit. However, this pattern of over- and under-compensation does not guarantee that a non-zero threshold improves fit, as RXCs 5 and 6 demonstrate – fit is maximized for these disease groups with the baseline zero-days threshold.

RXC3 presents an opposite example. The cost for patients in the disease group decreases as a function of the number of days' supply, basically similar to the dis-

28. The over- or under- compensation presented are relative to each disease group total gap in compensation.

Figure 5. Relative Over/Under-compensation in the disease group with a zero-days threshold, by patients' days' supply



For each disease group, the figure presents the average over- or under-compensation for patients, by the annual number of days' supply in their prescriptions for the RXC's drugs. It shows the compensation gap in the baseline scenario of a zero-days threshold, examining patients with 0 days of supply (i.e. diagnosed only) up to 365 days of supply. The dollar gap in compensation is calculated by the difference between the average payment to the patients in each days' supply subgroup, and the average baseline costs to the plans (which are the actual costs minus the costs covered by the reinsurance program). The gap is standardized for each group by subtracting the patients-weighted mean gap. The size of the bubble around every point indicates the number of patients in each subgroup (normalized separately for each graph).

tribution in panel B of Figure 1. As a result of the decreasing cost, patients with shorter prescriptions are under-compensated, while patients with longer prescriptions are over compensated. In such cases, a non-zero threshold will make things worse for the under-compensated below the threshold, further lowering their predicted costs. For the patients above the threshold, the payment will be reduced to better match their actual costs. The total effect on the fit depends on the relative amounts of under and over compensation in each days' supply group, weighted by a quadratic loss function. In RXC3 any non-zero threshold leads to a worse fit.

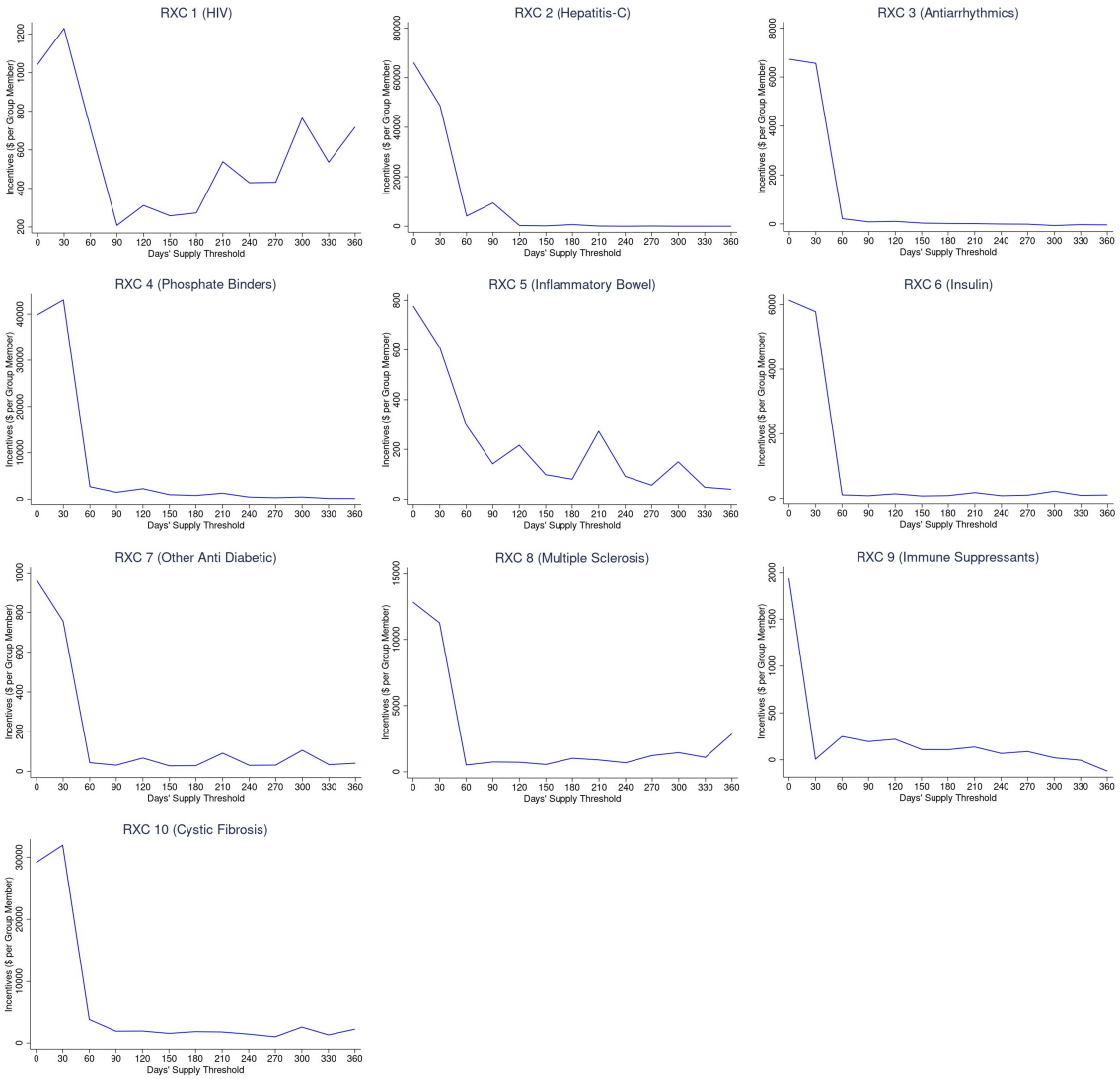
7.3 Incentives for gaming

Figure 6 presents the measure of gaming incentives – the net revenue from prescribing up to 30 more days of supply so patients with an existing prescription or a related diagnosis cross the threshold. It presents the incentives for each RXC, by the simulated days' supply thresholds. The measure is strongly affected by the distribution of patients by days' supply. For all groups, short thresholds allow the large group of diagnosed patients without prescription to be gamed, leading to high gaming incentives in thresholds up to 30 days of supply. The diagnosed-only group is excluded when thresholds are higher than 30 days, decreasing the incentives beyond this point. For example, for RXC1, a threshold of 90 days' supply creates the lowest incentive for this marginal gaming activity. With such a threshold, the gaming yields a net revenue of \$208 per patient in the HIV disease group. Incentives are mostly non-monotonous, emphasizing the need for an empirical analysis to identify the effect of each threshold.

7.4 Fit and incentives for gaming with multiple thresholds

The empirical analysis so far was restricted to a single threshold, exploring only different levels that trigger an existing drug-adjustor. This section eliminates the restriction and tries to find a better partitioning of the days' supply space, potentially using more than a single threshold. The regression tree found by the CART algorithm for RXC1 (Anti-HIV agents), when the loss function puts no weight on gaming incentives ($\lambda = 0$), is presented in Figure 7. The algorithm identifies a threshold of 180 days' supply as the fit-maximizing threshold in the root node (as identified in the simulations above). It then recursively searches for additional thresholds that improve the fit and finds: 0 and 120 days on the range below 180, and 330 and 270 days on the range above 180. The figure presents the coefficients calculated for the

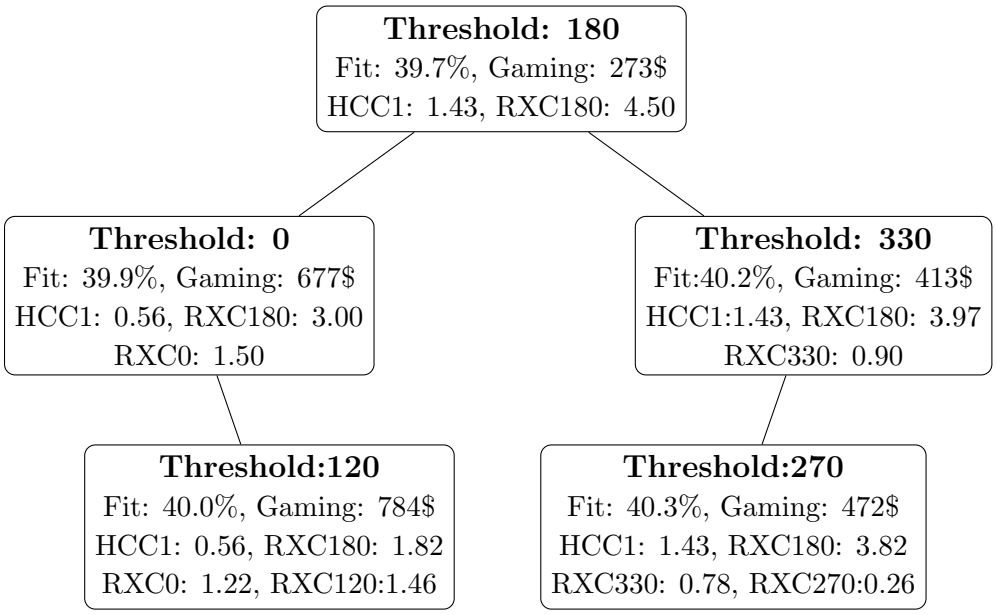
Figure 6. Net revenue from marginal gaming of prescriptions limited to 30 additional days of supply, by days' supply thresholds of each RXC



For each disease group, the figure shows the net revenue to a plan from a gaming behaviour that prescribes up to 30 additional days of supply of the RXC's drugs to push patients across a days' supply threshold. The net revenue, per member of the disease group, is shown for all thresholds that are multiples of 30-days, between 0 and 360. The net revenue of each patient is calculated by subtracting the minimal cost of the additional drugs, required to cross the threshold, from the additional revenue accrued to the plan from having the patient cross the threshold and have a higher risk score.

RXC adjustors in each node, the resulting fit in the disease group, and the cumulative marginal gaming incentives.

Figure 7. Best thresholds for RXC1, as chosen by the regression tree algorithm, with a zero weight on gaming incentives



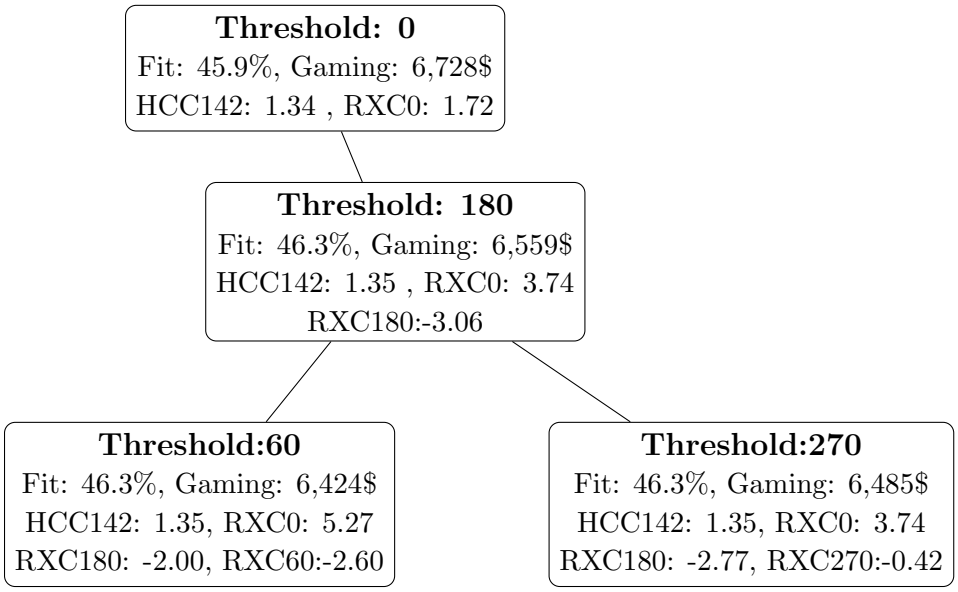
The figure presents the tree created by the CART algorithm that recursively searches for utilization thresholds to maximize the individual fit for the RXC1 disease group (i.e. patients either diagnosed with HIV or prescribed with Anti-HIV agents). Each node shows the fit-maximizing threshold, the individual R-squared fit for the group, and the cumulative incentive for gaming by prescribing up to 30 more days of supply of the RXC drugs. Each node also presents the coefficients of the relevant adjustors: HCC1 is the diagnosis-based adjustor for HIV; RXC180 is the adjustor for Anti-HIV agents, with a utilization threshold of 180, and similarly RXC0, RXC330, RXC120, and RXC270 signify RXC adjustors with different utilization thresholds.

The results change as the loss function puts more weight on the gaming incentives. With a weight of $\lambda = 0.2$ on the incentives, the regression tree includes only a single threshold of 150 days' supply.

The regression tree for RXC3 (Antiarrhythmics), when all the weight is on fit ($\lambda = 0$), is presented in Figure 8 and serves as another example. The fit-maximizing threshold in the root node is 0-days (as in the simulations above). The algorithm continues to search for thresholds above zero that improve the fit, and finds 180-days in the second level of the tree, and 60 and 270 days at the third level.

However, the coefficients estimated for the adjustors of the additional thresholds are negative. At the second level of the tree, adding one day of supply to a patient with 179 days' supply prescription, will *decrease* her risk score by 3.06, cutting the payment to the plan by more than 20 thousand dollars. Thus, an additional adjustor

Figure 8. Best thresholds for RXC3, as chosen by the regression tree algorithm, with a zero weight on gaming incentives



The figure presents the tree created by the CART algorithm that recursively searches for utilization thresholds to maximize the individual fit for the RXC3 disease group (i.e. patients either diagnosed with Specified Heart Arrhythmias or prescribed with Antiarrhythmics). Each node shows the fit-maximizing threshold, the individual R^2 fit for the group, and the cumulative incentive for gaming by prescribing up to 30 more days of supply of the RXC drugs. Each node also presents the coefficients of the relevant adjustors: HCC142 is the diagnosis-based adjustor for Specified Heart Arrhythmias; RXC0 is the adjustor for Antiarrhythmics, with a utilization threshold of 0-days, and similarly RXC180, RXC60, and RXC270 signify RXC adjustors with different utilization thresholds.

for a 180-days threshold would violate the monotonicity principle, set in Pope et al. (2004) and widely adopted by policy makers in developing risk adjustment systems. This principle states that insurers should not be penalized for additional recording of diagnoses, and in the current context – for additional recording of days of supply in patients' prescriptions. Violation of monotonicity may create an incentive for gaming the system by either skimping on the provision of the drug, under-reporting its use, or providing it in ways that are not recorded in claims (e.g. sending a coupon to cover out-of-pocket purchases). I don't examine such gaming behaviors in this paper, that focuses on incentives for gaming only by increasing utilization.

8 Utilization Thresholds for Non-drug Adjustors

Utilization thresholds exist not only in adjustors based directly on utilization, like the consumption of prescription drugs, but also in any adjustor that is related to uti-

lization indirectly, such as morbidity-based adjustors. These are based on data that appear in claims, and thus depend on the utilization of services (Geruso and McGuire 2016). In this section I examine utilization thresholds for two diagnosis-based adjustors to demonstrate how the impact of these thresholds is again an empirical question. The examined thresholds are applied to the number of times a diagnosis appears in a patient's claims.[29] When counting the number of appearances, a diagnosis is counted at most once per day, and once per hospital admission. Figure 9 examines the effect of simulated thresholds for two related adjustors: To the left, CC19 that indicates "Diabetes with acute complications", and on the right, CC21, that indicates "Diabetes without complications".[30]

The bars in Panel A in the figure present the distribution of patients for which the CC adjustor is turned on (regardless of the HCC hierarchy), by the number of times the diagnosis appear in their claims.[31] The dots in panel A present the average annual cost of patients by the number of appearances. Most patients with a CC19 diagnosis (Diabetes with acute complications) have only one claim that denotes it, and for 46% of them it is an in-patient claim from an hospital admission. Average costs are higher for patients with one appearance ($48,624) than for patients with two ($38,194), three ($38,508), or four appearances ($47,543).[32] In contrast to that, most patients with a CC21 diagnosis (a much larger group) have more than one appearance of this diagnosis, and almost all of these appearances are recorded in an out-patient setting. The annual cost of patients with one appearance is $17,320 and cost increases almost monotonously as patients have a higher number of appearances.

Panel B presents the R^2 individual fit for the group of patients with a diagnosis included in the CC group, by appearances thresholds. The fit at the baseline scenario – a single appearance threshold – is indexed to 100. The baseline scenario provides the best fit for CC19. For CC21 a 4-appearances threshold maximizes the fit and improves it by about 2% .

29. Alternative thresholds may be applied only to outpatient or inpatient diagnoses, examine whether the diagnoses were recorded in separate quarters of the years, require a minimum cost of a claim to allow the diagnosis included in it to be used for calculating risk scores, etc.

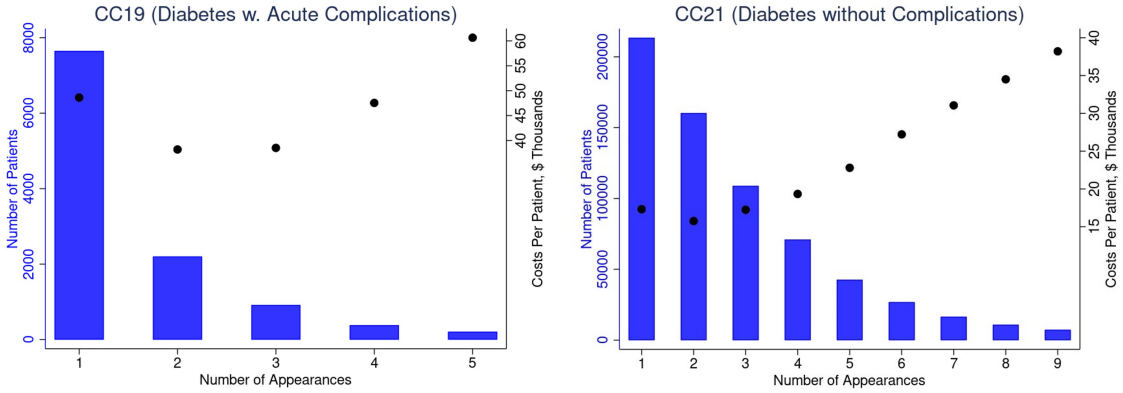
30. Each CC may group several related diagnoses codes and, when triggered on, has an effect on a patient's risk score only if a CC higher in the HCC hierarchy is not also triggered. For diabetes, CC21 (Diabetes without complications) is the lowest HCC in the hierarchy, below CCs that describe diabetes with chronic complications (CC20), and diabetes with acute complications (CC19).

31. If a patient has several diagnoses that are included in the same CC, the diagnosis with the highest number of appearances is used.

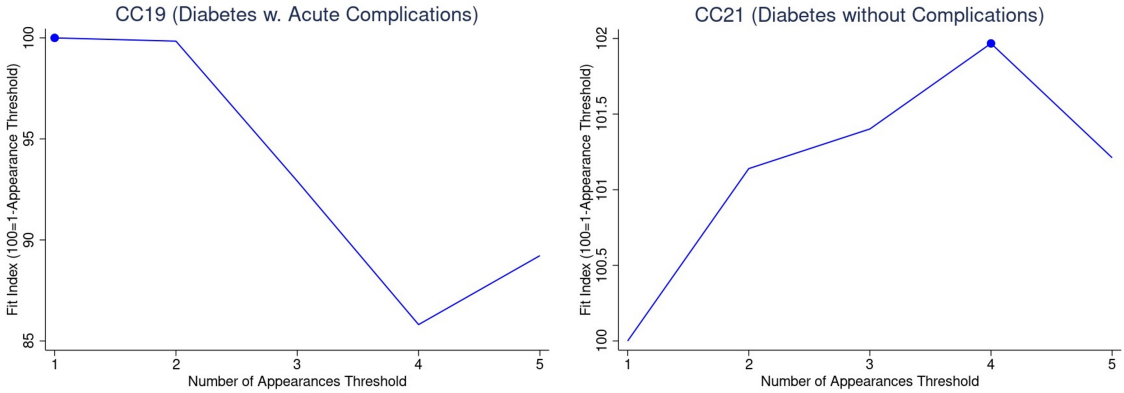
32. The share of diagnoses documented in outpatient setting increases for patients with more than one appearance: 81% with two appearances, and 88% with three appearances.

Figure 9. Appearances thresholds for diagnoses-based adjustors: Distribution of enrollees and their costs, fit, and incentives for gaming

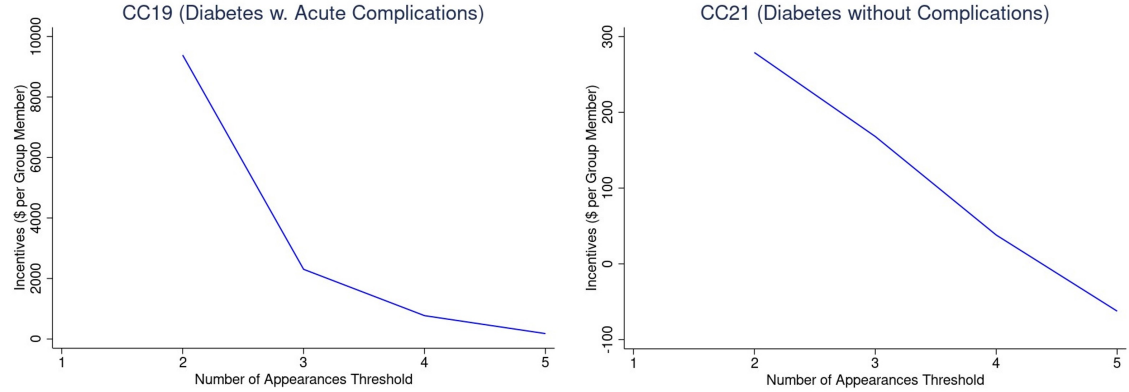
(a) The distribution of patients with a CC diagnosis and their average annual cost, by the number of annual appearances of the diagnosis



(b) Individual fit for patients with a CC diagnosis, by appearances threshold



(c) Net revenue from gaming CC adjustors by adding at most a single appearance of a diagnosis to patients with an existing CC diagnosis, by appearances threshold



Panel A presents the distribution (bars, left axis) and average annual costs (dots, right axis) of patients, by the number of times that a diagnosis included in the Condition Category (CC) appears in their annual claims. Panel B shows the R^2 individual fit for patients with a CC diagnosis, by the number-of-appearances threshold. Panel C presents the potential net revenue to plans from a gaming activity that targets patients with an existing CC diagnosis, and adds at most a single appearance of the diagnosis, by inducing an additional out-patient claim. Revenue is measured as dollars per enrollee with an existing CC diagnosis.

Panel C presents the net revenue from a gaming activity that adds at most a single additional appearance of a CC diagnosis to patients that are already diagnosed (a measure equivalent to the marginal incentive measure for drug-adjustors, but excludes the non-diagnosed). The net revenue for each gamed patient is the revenue due to crossing the threshold and turning the CC on,[33] minus the direct cost of gaming, defined here as the cost of an additional out-patient visit in which the diagnosis is coded.[34] This kind of incentive for gaming decreases monotonously with higher thresholds for both CCs.

To conclude, a utilization threshold that requires more than one appearance of a diagnosis to turn on a CC-adjustor may increase fit for some adjustors, and decrease it for others. Higher thresholds may be more beneficial for CCs in which a large number of patients have more than one appearance of the diagnosis (e.g. chronic conditions rather than acute episodes), and when patients' costs increase in the number of appearances (then patients with a higher number of appearances tend to be under-compensated). Here again, simulations (and regression trees) may serve as tools to choose utilization thresholds that will improve the model's fit, taking into account the incentives to game it.

9 Discussion

9.1 The lack of a tradeoff between fit and the incentives for gaming

When implementing the new hybrid drugs-diagnoses risk adjustment system, CMS declared that it is seeking to "strike a reasonable balance between increasing predictive accuracy and reducing incentives for overprescription" (CMS 2016b). This reflects a common belief in the existence of a tradeoff between fit and the incentives for gaming. However, this paper shows that such a tradeoff doesn't always exist. For four out of the ten RXCs, non-zero thresholds can both improve the fit and reduce the incentives for gaming the prescription behavior: A 30-days threshold for RXC2 (Anti Hepatitis-C agents) improves fit in the disease group by 0.6%. It also reduces the gaming incentives by a quarter; For RXC7 (Non-insulin anti-diabetes agents), a 210-days threshold improves fit for the disease group only by 0.01%, but decreases the

33. Turning the CC on may yield no additional revenue if a CC higher in the hierarchy of Condition Categories is already turned on.

34. If the patient has out-patient visits in which a CC-diagnosis appears, than their average cost is used. If the patient has only in-patient claims, then the average cost of an out-patient visit for the whole group is used.

marginal incentives for gaming by 90%; For RXC8 (MS Agents), a 150 days' supply threshold increases the disease group's individual fit by 14%. The net revenue from marginal gaming under such threshold is 96% lower than the baseline incentive with a zero-days threshold; Lastly, in RXC9 (Immune suppressants), a 120-days threshold increases the individual fit in the disease group by 1.8%, and decreases the marginal gaming incentive by 89%. For RXC1 (Anti-HIV), and RXC10 (Cystic Fibrosis), a non-zero threshold improves fit and decreases the marginal incentive for gaming, but increases the net incentives from unlimited gaming.

9.2 Dynamic vs. Static Incentives for Gaming

A caveat to the incentive measures defined in this paper is that the measures are static in nature, i.e. they measure the incentives for gaming within a single year (in the concurrent payment system used at the Marketplaces). The expected incentives may be lower for later years, especially when the risk adjustment model is re-estimated using claims data from the Marketplaces themselves (data that CMS began using in 2019), and when more plans, with a larger share of relevant patients, game the system. In such a case, even if the absolute return to gaming decreases, a plan that avoids gaming alone may suffer financially, as the predicted costs of gamed RXCs decreases. Behrend, Felder, and Busse (2007) show that in a prospective payment system, gaming may be lucrative to health plans as long as the share of plans gaming and the share of patients gamed are not too high.

Another dynamic aspect is the potential effect of gaming on patients' selection into plans. Gaming activity that easily provides longer prescriptions to drugs that treat a certain disease, may attract patients with the disease to the plan. Such adverse selection may change the incentives in the following years.

Lastly, alternative thresholds for drug-based adjustors may also change the incentives that pharmaceutical companies and Pharmacy Benefit Managers (PBMs) face, adding another possible dimension of gaming. These players nay change drug prices, adapt drug packaging, and tailor innovations in response to change in days' supply thresholds. In this paper, I abstract away from all the above issues.

10 Conclusions

I examine the role of utilization thresholds in risk adjustment systems and their effect on the fit of the risk adjustment model and the incentives for gaming it. The

sign and size of these impacts is an empirical question that, inter alia, depends on the cost distribution of enrollees by utilization. I study thresholds in the setting of the U.S. Marketplaces, examining both days' supply thresholds for prescription-drug adjustors and thresholds that depend on the number of times a diagnosis appears in claims for morbidity-based adjustors. I show that for some adjustors, there is no tradeoff between fit and incentives for gaming when applying a non-zero threshold. The paper demonstrates how simulations and regression trees may guide the choice of utilization thresholds and allow policy makers to explicitly balance their desire for a better fit with their concerns from gaming.

ACKNOWLEDGEMENTS

I am grateful to Tom McGuire for his guidance. For their helpful comments and explanations, I thank Konstantin Beck, John Bertko, Itamar Caspi, Randall Ellis, Keith Marzilli Ericson, Kobi Glazer, Lukas Kauer, Richard van Kleef, Tim Layton, Michael McWilliams, Joe Newhouse, Sonja Schillo, Erik Schokkaert, Wynand van de Ven, and participants of the Harvard Health Care Policy Department's Health Economics seminar, and the Risk-Adjustment Network meeting. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Behrend, C., S. Felder, and R. Busse. 2007. "Susceptibility to Strategy of the Drug Component of the IPHCC+RxGroups Classification System in a Risk-Adjusted Morbidity Compensation Scheme - A Conceptional and Data-Supported Analysis (German)." *Das Gesundheitswesen* 69 (1): 1–10. https://www.thieme-connect.com/products/ejournals/html/10.1055/s-2007-968171.

CMS. 2018a. *2019 Final HHS Risk Adjustment Model Coefficients,* April 12, 2018. https://www.hhs.gov/guidance/document/2019-final-hhs-risk-adjustment-model-coefficients.

———. 2016a. *March 31, 2016, HHS-Operated Risk Adjustment Methodology Meeting, Discussion Paper.* https://www.cms.gov/CCIIO/Resources/Forms-Reports-and-Other-Resources/Downloads/RA-March-31-White-Paper-032416.pdf.

CMS. 2016b. *Patient Protection and Affordable Care Act; HHS Notice of Benefit and Payment Parameters for 2018.* https://www.federalregister.gov/documents/2016/12/22/2016-30433/patient-protection-and-affordable-care-act-hhs-notice-of-benefit-and-payment-parameters-for-2018.

Einav, Liran, and Amy Finkelstein. 2011. "Selection in Insurance Markets: Theory and Empirics in Pictures." *Journal of Economic Perspectives* 25, no. 1 (February 1, 2011): 115–138. ISSN: 0895-3309, accessed August 27, 2022. https://doi.org/10.1257/jep.25.1.115. https://pubs.aeaweb.org/doi/10.1257/jep.25.1.115.

Ellis, Randall P., Bruno Martins, and Sherri Rose. 2018. "Risk Adjustment for Health Plan Payment." In *Risk Adjustment, Risk Sharing and Premium Regulation in Health Insurance Markets: Theory and Practice,* edited by Thomas G. McGuire and Richard C. van Kleef, 55–104. London: Academic Press (Elsevier).

Ericson, Keith M. Marzilli, Kimberley H. Geissler, and Benjamin Lubin. 2018. "The Impact of Partial-Year Enrollment on the Accuracy of Risk-Adjustment Systems: A Framework and Evidence." *American Journal of Health Economics* 4, no. 4 (November): 454–478. ISSN: 2332-3493, 2332-3507, accessed August 27, 2022. https://doi.org/10.1162/ajhe_a_00108. https://www.journals.uchicago.edu/doi/10.1162/ajhe_a_00108.

Geruso, Michael, and Timothy Layton. 2020. "Upcoding: Evidence from Medicare on Squishy Risk Adjustment." *Journal of Political Economy* 128, no. 3 (March): 984–1026. ISSN: 0022-3808, 1537-534X, accessed September 2, 2022. https://doi.org/10.1086/704756. https://www.journals.uchicago.edu/doi/10.1086/704756.

Geruso, Michael, and Thomas G. McGuire. 2016. "Tradeoffs in the Design of Health Plan Payment Systems: Fit, Power and Balance." *Journal of health economics* 47:1–19.

Glazer, Jacob, and Thomas G. McGuire. 2012. "Optimal Risk Adjustment." In *The Elgar Companion to Health Economics, Second Edition,* by Andrew Jones, 14021. Edward Elgar Publishing. ISBN: 978-0-85793-813-8, accessed August 31, 2022. https://doi.org/10.4337/9780857938138.00040. http://www.elgaronline.com/view/9781849802673.00040.xml.

Glazer, Jacob, and Thomas G McGuire. 2000. "Optimal Risk Adjustment in Markets with Adverse Selection: An Application to Managed Care." *American Economic Review* 90, no. 4 (September 1, 2000): 1055–1071. ISSN: 0002-8282, accessed August 27, 2022. https://doi.org/10.1257/aer.90.4.1055. https://pubs.aeaweb.org/doi/10.1257/aer.90.4.1055.

Kautter, John, Gregory C. Pope, Melvin Ingber, Sara Freeman, Lindsey Patterson, Michael Cohen, and Patricia Keenan. 2014. "The HHS-HCC Risk Adjustment Model for Individual and Small Group Markets under the Affordable Care Act." *Medicare & Medicaid research review* 4 (3).

Lamers, Leida M., and Rene C. J. A. van Vliet. 2003. "Health-Based Risk Adjustment: Improving the Pharmacy-Based Cost Group Model to Reduce Gaming Possibilities." *The European Journal of Health Economics* 4 (2): 107–114.

Lamers, Leida M., Rene C. J. A. van Vliet, and Winand PMM van de Ven. 1999. "Farmacie Kosten Groepen: Een Verdeelkenmerk Voor Normuitkeringen Gebaseerd Op Medicijngebruik in Het Verleden (Pharmacy-based Cost Groups: A Risk Adjuster for Capitation Payments Based on the Utilization of Prescribed Drugs)."

Layton, Timothy J., Randall P. Ellis, Thomas G. McGuire, and Richard C. van Kleef. 2018. "Evaluating the Performance of Health Plan Payment Systems." In *Risk Adjustment, Risk Sharing and Premium Regulation in Health Insurance Markets: Theory and Practice,* edited by Thomas G. McGuire and Richard C. van Kleef, 133–180. London: Academic Press (Elsevier).

Layton, Timothy J., E. Montz, and Mark Shepard. 2018. "Health Plan Payment in US Marketplaces: Regulated Competition With A Weak Mandate." In *Risk Adjustment, Risk Sharing and Premium Regulation in Health Insurance Markets: Theory and Practice,* edited by Thomas G. McGuire and Richard C. van Kleef. London: Academic Press (Elsevier).

Pope, Gregory C., John Kautter, Randall P. Ellis, Arlene S. Ash, John Z. Ayanian, Lisa I. Iezzoni, Melvin J. Ingber, Jesse M. Levy, and John Robst. 2004. "Risk Adjustment of Medicare Capitation Payments Using the CMS-HCC Model." *Health Care Financing Review* 25 (4): 119–141. ISSN: 0195-8631, accessed January 26, 2022. pmid: 15493448. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4194896/.