Agreement Between Patient and Doctor Drug Evaluations

Ethan Deininger
Adviser: Carola Binder

April 27, 2017

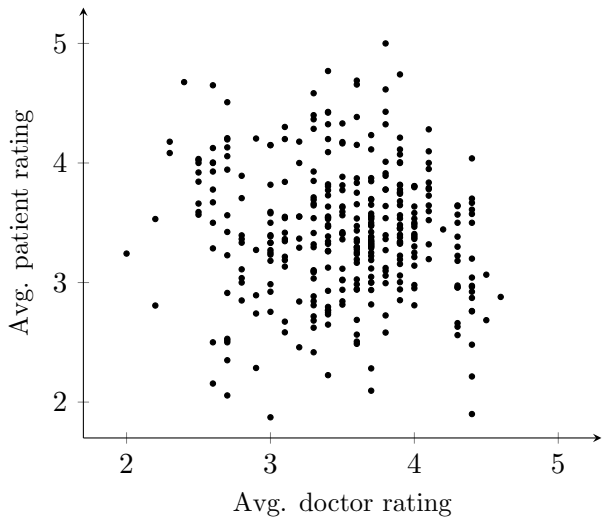
Abstract

Patient ratings of drugs are regressed on doctor ratings and the severity of the patient's condition (measured with QALY) while controlling for demographic variables and fixed effects. Patient rating and demographic data was collected from WebMD, doctor ratings from RateRx, and QALY data from the Tufts CFAR database. Patient and doctor ratings are positively correlated, though not very strongly. Patients with worse conditions rate drugs higher and with lower variance than patients with mild conditions. Doctors rate drugs higher if they have low variance between patient ratings.

Introduction

In medicine, many conditions have a "gold standard" treatment, something (often a prescription drug or class of drugs) that most doctors agree should be the go-to treatment for patients: lithium for bipolar disorder, amphetamines for ADHD, SSRIs for depression, and so on.[1] These medications tend to have the best patient responses in the most patients, and thus are among the most commonly prescribed (and highly rated) drugs within each condition.

Figure 1: Average ratings by drug



Source: WebMD and RateRx.
No. of observations: 358

This observation suggests a model of prescription drugs where patients and doctors should mutually agree on the best and worst drugs. In other words, we should expect that patient and doctors will report similar ratings of prescription drugs for a given condition. However, as shown in Figure 1 - a graph of average patient and doctor rating by drug and condition - a positive relationship relationship is by no means obvious. The lack of a positive trend between patient and doctor ratings is immediately apparent (the correlation is in fact slightly negative). It suggests that the relationship between patient and doctor attitudes towards drugs is not so straightforward. This paper examines the relationship between ratings in more detail, controlling for condition, drug fixed effects, time using the drug, and demographic variables.

I find that doctor ratings are significantly positively correlated with average patient ratings, though not very strongly. The number of patient ratings for a drug (an imperfect proxy for how many people get prescribed the drug) is also positively correlated with average patient ratings. These observations are somewhat reassuring - doctors and patients don't completely disagree overall on what drugs are best - but the small effect sizes should be noted. I additionally find that doctor ratings are negatively correlated with the variance of patient reviews, meaning that doctors prefer drugs that will work moderately well for most people over drugs that will work either very well or very poorly (it's unclear which of these would be better).

In addition to comparing doctor and patient ratings, this paper also looks at the relationship between patient rating and seriousness of condition. If a patient has a condition that doesn't significantly impact their day-to-day life, such as high cholesterol, I expect that their rating behavior will differ significantly from someone whose condition is a constant burden, such as depression. More specifically, I expect that patients with relatively worse conditions will rate drugs higher and with higher variance than patients with less serious conditions. I expect higher average ratings because a patient with depression has an immense amount to "gain" from taking medication, and not very much lower to fall in terms of happiness; compare this to someone with high cholesterol, who only knows about their condition from their doctor and whose daily quality of life can only go down

[1] See e.g. Kanba et al. (2013) on bipolar disorder, Sharma and Couture (2013) on ADHD, and McCarter (2008) on depression.

from medication side effects. I expect higher variance because patients with terrible conditions will either be extremely happy or extremely sad, whereas I expect people with less serious conditions to be more ambivalent about their medication and thus cluster their ratings in the middle. I use quality-adjusted life-year (QALY) as a measure of condition severity, as described in the Data section.

I find that the first hypothesis is correct. Patients with worse conditions rate drugs significantly more highly than those with mild conditions. The effect size is large - the average rating of a drug treating bipolar depression will on average be rated about 0.8 points higher than one for high cholesterol, on a 5 point scale where the vast majority of average scores are within 2 points of each other. The relationship between QALY and variance is significant and equally strong, but in an unexpected direction - patients with worse conditions rate with much *lower* variance than those with more mild conditions.

Literature Review

Theoretical background

One explanation for disagreement between patients and doctors stems from misaligned incentives. Doctors and patients face a principal-agent problem (first described by Jensen and Meckler, 1976), where the agent (physician) makes decisions on behalf of the principal (patient) despite having conflicting self-interested motivations. While doctors benefit from treating patients optimally, their careers may be destroyed if they are brought to trial for malpractice. The current climate of defensive medicine is evidence that the threat weighs heavily on many doctors, and while there is some evidence of recent reforms improving outcomes for doctors (see e.g. Mello et al., 2014), a national survey of neurosurgeons by Nahed et al. (2012) found that 45% reported eliminating risky procedures from their practice due to liability concerns. This means drugs that are relatively riskier but better performing may be rated more highly by patients than doctors.

Asymmetrical information represents another source of friction that may contribute to doctor-patient disagreement. Patients pay doctors in large part for their medical expertise, and are not intimately familiar with the qualities of whatever drug they are prescribed. Thus if a drug is very effective for a small but predictable portion of the population and dangerous for the rest, for instance, doctors may rate it poorly based on its danger, but patients who are able to take it will rate it highly. Additionally, patient ratings may have more noise involved due to the nature of the prescription process. Not all drugs work well for everyone, and so patients will generally try a series of drugs until they find one that works (this behavior may be simulated as a series of updating Bernoulli trials, as in Dickstein, 2014b). Thus doctors will ideally rate drugs based on an estimation of their likelihood to work, but patients will base ratings based on their own experiences; fortunately, however, at large sample sizes the patient ratings will approximate the true rates.[2]

Insurance companies also play large roles in the prescription process. They negotiate with drug manufacturers individually on how much they pay for each type of drug, and classify the drugs they get the best deals on as "preferred" drugs, which they attempt to encourage patients to select over the alternatives. To influence patients, they tend to adopt a multi-tiered pricing approach where

[2]This assumes that the probabilities of each drug working for a given patient are independent. If not, then since doctors tend to prescribe drugs in particular order (i.e. first-line, second-line, etc.), the ratings would be skewed based on the conditional probabilities. I could not find research on this question and it could be an interesting avenue to explore.

patients pay a lower copay for preferred drugs, which has proven effective at influencing patient behavior (see e.g. Gaynor et al., 2007). To influence doctors, they distribute a great deal of up-to-date drug literature to doctors in their network, but select literature which reflects favorably on their preferred drugs. Despite this being a very well-documented phenomenon, it's highly effective; Limbrock (2011) estimates that the impact on doctors of a drug being preferred is equivalent to a $12.11 copay reduction (for context, average copayments are only $13). The impact of these influences on patient and doctor ratings, however, is difficult to estimate; since these effects are primarily driven by the market, their behavior is much more difficult to predict.

Validity of online ratings

One of the major concerns with this paper is the validity of online ratings as sources, particularly the patient ratings (since the doctors must connect their medical license and identity to their ratings and therefore have reputation-based incentives to provide accurate information). However, there are a number of points in WebMD's favor: a wide range of studies find it among the most viewed and most reliable sources on the internet (e.g. Grohol et al., 2014 and Ritchie et al., 2016), as well as that it consistently follows best conduct criteria for posting health information (Morgan and Montagne, 2014). More directly relevant to ratings, Adusumalli et al. (2015) looks at how often differences in WebMD drug ratings are borne out by medical data. Of the drug pairs with at least a half-point rating difference that they find evidence directly comparing efficacy, 62% of the differences were supported by published literature. This indicates that online ratings are often based on actual differences between drugs, but also that there is room for additional concerns such as severity of side effects. There is also a question of selection bias, where patients posting online ratings may not be representative of the patient population as a whole. This concern should be somewhat alleviated by Emmert et al. (2015), which finds no significant differences between online and offline patient ratings of physicians.

Many doctors are forced to care about patient ratings: Zgierska et al. (2014) finds in an anonymous survey that 59% of doctors report that patient evaluations are directly linked to their compensation, and 20% reported their employment being threatened due to patient satisfaction data. It's additionally important that we understand how patient reports relate to expert opinions because patients are increasingly becoming informed from other patients' self-reporting. Jans & Kranzbhler (2015) finds that, while patients trust expert opinion more than individual reports of other patients, this is counterbalanced when patient reports significantly outnumber doctor reports, and even when both are present patients tend to balance expert and patient opinions.

A final concern is that patient self-reporting might not always be perfectly lined up with patient health. For example, Trujols et al. (2014) finds that patient satisfaction self-reports don't always correspond to actual harm-reduction, and Adusmalli et al. (2015) finds that patients tend to rate drugs more highly when they have addictive properties. If the data allowed, I would have liked to control for actual patient health outcomes, and perhaps this is an area for future research.

Data

This project involved three key pieces of data, each collected from a different source: the patient ratings of drugs, the doctor ratings of drugs, and the quality-adjusted life-year (QALY) ratings of conditions. The patient ratings were taken from WebMD, the largest US health website. I wrote a

python program using the Scrapy web crawling framework to visit every review page of every drug on the website and save information about each review. A total of 242,730 reviews were collected, spanning 9,742 drugs and 1,698 conditions.

However, not all reviews were used for analysis. When rating a drug, reviewers may elect to fill out optional demographic information, including their sex, age, time on the medication, and whether they are a patient or caretaker. Since I believed these attributes would be important factors to control, I limited analysis to reviews with every piece of demographic information filled out (this dropped about 14% of reviews). To limit the sample exclusively to self-reporting, I also dropped all reviews made by caretakers or individuals under the age of 13.

I further limited analysis to 17 different conditions. These conditions were selected based on the following process: the pool of possible conditions was limited to those which had at least 500 ratings in my WebMD sample. I dropped conditions that were vague or not specific enough (e.g. Other, Pain, Birth Control), as well as any that did not have doctor or QALY ratings. The full list of conditions can be found in table 1. Finally, I limited analysis to drugs which had at least ten reviews, since when looking at drug-level observations I wanted a reasonable number of ratings to average. This process left a sample of 56,339 reviews spanning 423 drugs and the 17 selected conditions.

Doctor ratings were collected from RateRx, a service by the company HealthTap that solicits doctors to provide feedback on their website. The doctors must go through a verification process to prove they have a medical license, and then use the website to provide feedback on medical issues and answer patient questions. One form of feedback is rating various prescription drugs for different conditions, and the website displays the average doctor rating and number of reviews. I manually collected ratings on the drugs that overlapped with the WebMD data, a total of 326 of the 423 drugs.

Finally, QALY ratings were collected from the Cost-Effectiveness Analysis Registry available through the Tufts Medical Center. The registry provides a searchable database for articles that have performed QALY analyses for given conditions. I tried to find all studies that did so for the typical case of each condition and averaged the results. The QALY ratings range from 0.40 (bipolar depression) to 0.99 (high cholesterol), and the full list may be seen in table 1.

A QALY rating is essentially a utility coefficient on a year of a person's life, where 1 is normal and 0 is dead. Thus someone with high cholesterol is living very close to a normal life (the only reduction being the small increased chance of a heart attack), while someone with bipolar depression is experiencing incredibly reduced utility from everything in their life because they are so unhappy. Calculating QALY is rather complicated - there are unsurprisingly a great number of difficulties and intricacies surrounding the project of converting the effect of a medical condition on a patient's life into a number. Obviously there is room for disagreement about the exact quantities, and objections that a single number cannot possibly capture the complex details of individual conditions are well taken, but I hope that the table of conditions by QALY will roughly conform to the reader's intuition, and thus using it as a predictive measure will have some value even if it's not perfect.

This data took the shape of individual-level reviews, which I look at the beginning of the results section. I additionally collapse the data set down to drug-level observations, for the purposes of predicting average rating and variance of ratings. When doing so, I bundled demographic data together into the best predictive groups possible.

Table 1
QALY values of conditions.

Condition	QALY
High Cholesterol	0.99
Underactive Thyroid	0.99
High Blood Pressure	0.98
Bacterial Urinary Tract Infection	0.93
Overweight	0.81
Type 2 Diabetes Mellitus	0.81
Crohn's Disease	0.77
Repeated Episodes of Anxiety	0.76
Schizophrenia	0.75
Chronic Trouble Sleeping	0.66
Panic Disorder	0.60
Attention Deficit Disorder with Hyperactivity	0.58
Migraine Headache	0.57
Rheumatoid Arthritis	0.55
Depression	0.46
Major Depressive Disorder	0.42
Bipolar Depression	0.40

Source: Cost-Effectiveness Analysis Registry via Tufts Medical Center.

Results

Patients using WebMD are asked to rate drugs based on three criteria: Effectiveness, Satisfaction, and Ease of Use. From reading through a number of comments, it appears that different people interpreted the "Ease of Use" criteria in different ways - some rated it based on the drug's side effects, and some rated it based on how hard the medication was to take (most are pills, and thus very "easy" to take). As some evidence for this observation, the correlation coefficient between Effectiveness and Satisfaction is about 0.8, while the correlation between Ease of Use and either of the other measurements hovers around 0.55. For this reason, I will limit my analysis to the Effectiveness and Satisfaction criteria. Throughout this paper I will be predicting the average of the Effectiveness and Satisfaction ratings combined, though the results are very similar when predicting each of the two ratings individually.

I run models predicting both review-level observations and drug-level observations. I will discuss the review-level analysis first. Because these are individual reviews being predicted, we should expect demographic variables to have a strong effect. The main model of interest is as follows:

$$\text{Rating}_{Patient} = \text{Rating}_{Doctor} + \text{QALY} + \text{Demographics} + Condition + Drug$$

The regression results can be found in table 2, where the above model is regression 2. As expected, the demographic variables are all highly significant. In every regression we see that the longer the patient has taken the medication, the higher they will rate it - this makes sense (and is in fact reassuring), since the more that patients like a medication the longer they will be willing to take it. The age coefficients are less straightforward - teenagers rate drugs the most poorly, but ratings

go up and peak for patients 25-44 before dropping back down. for elderly patients. Finally, men rate *very* slightly higher, but the time on medication is by far the strongest demographic effect.

The variables of interest for this paper - QALY and doctor ratings - are not significant at the 5% level in either of the two regressions they are in. That being said, doctor ratings are positive with a moderate effect size and significant at the 10% level, which is encouraging. Interestingly, the effect size of QALY jumps massively (and becomes significant at the 10% level) after controlling for number of patient and doctor ratings. This is presumably because there are more ratings for high-QALY drugs which were acting as a counterweight on the negative effect of the QALY. Despite these variables being positive, adding them to the regression has virtually no effect on its explanatory power as measured by R-squared and the Bayesian Information Criterion (BIC).

Next I look at drug-level observations, regressing first on average patient rating (table 3) and then on patient rating variance (table 4). Other than the dependent variable, the only change in the model is (obviously) the removal of drug fixed effects. Though sex and age demographics were significant for patient-level ratings, their effect size was very small, and their effect at this level when approximated as percentages disappears is no longer significant. When measuring average review, the time on medication measure remains somewhat significant, though not when measuring variance.

In both regressions, QALY is very significant and has a large effect size. Patients with worse conditions rate drugs more highly and with less variance. The first conforms to my hypothesis: drugs have positive effects (treatment) and negative effects (side effects), and patients with worse conditions have more potential to gain from treatment, while patients with mild conditions will still experience frustrations over side effects. The fact drugs with worse conditions display less variance is very surprising. One hypothesis is that, since the low-QALY conditions tend to be severe psychological illnesses, only patients whose medication is working will be in a functional enough state to post online ratings. However, it's possible that once on a working medication, those patients would rate not only their current medication but also previous medications, which would eliminate this effect.

Doctor rating is also significant and has a moderate effect size. It has a positive relationship with average patient rating, but a negative relationship with patient variance, suggesting that doctors prefer drugs that work consistently (as opposed to trying a patient on a series of drugs that will either be great or terrible in an effort to find a great one). Number of patient and doctor ratings are significant but with very low effect sizes, even considering the number of ratings (doctor ratings top out at around 200, patient ratings at around 2000; taking the log of either one does not improve the model). Once again, despite significant values, adding the doctor rating and QALY ratings do has virtually no effect on the explanatory power of either regression.

Conclusion

While the data shows a positive correlation between doctor and patient ratings, emphasis should be placed on just how low the correlation is. At the drug level, which should be a more accurate representation of agreement, the coefficient on doctor rating is just 0.139 and the adjusted R-squared barely budges when doctor ratings are added. Thus the major finding of this paper is that, while doctor and patient ratings are correlated, their relationship is much weaker than we should expect.

Low explanatory power notwithstanding, the interpretation of the results is still interesting. It

suggests that concerns about misaligned incentives have a great deal of merit - in a perfect world, patients and doctors would agree completely on optimal treatment (with the limitation being lack of patient knowledge). Unfortunately, this paper cannot suggest which of the informational problems have the largest effects. But it does suggest that patients should keep their own desires in mind and try to elicit information about why the doctor is making his or her suggestions

Despite its limitations, the data strongly suggests that the model described in the introduction - that most treatments have a "gold-standard" which is widely agreed upon - is flawed. More effort must be placed into narrowing down our understanding of why these disagreements occur. Given how many different explanations and hypotheses are for why such a gap would occur, further research will be necessary to understand and address the problem. Access to additional data, of course, would be ideal. Measuring actual patient health outcomes, as well as knowing the patient's doctor and insurance plan, would be an immense help towards controlling for the differing incentives discussed in the literature review. The best possible analysis would also control for patient fixed effects as they tried multiple drugs over time. The current gulf between patient and doctors is indicative of substantial fundamental problems with our healthcare system that would provide immense gains if addressed.

References

Adusumalli, S., Lee, H., Hoi, Q., Koo, S., Tan, I. B., & Ng, P. C. (2015). Assessment of Web-Based Consumer Reviews as a Resource for Drug Performance. *Journal of Medical Internet Research*, *17*(8).

Dickstein, M. J. (2014a). Physician vs. Patient Incentives in Prescription Drug Choice. Working Paper.

Dickstein, M. J. (2014b). Efficient Provision of Experience Goods: Evidence from Antidepressant Choice. Working Paper.

Emmert, M., Adelhardt, T., Sander, U., Wambach, V., and Lindenthal, J. (2015). A cross-sectional study assessing the association between online ratings and structural and quality of care measures: results from two German physician rating websites. *BMC Health Services Research. 15*, 114.

Gaynor, M., Li, J., & Vogt, W. B. (2007). Substitution, Spending Offsets, and Prescription Drug Benefit Design. *Forum for Health Economics & Policy*, *10*(2).

Grohol, J. M., Slimowicz, J., & Granda, R. (2014). The Quality of Mental Health Information Commonly Searched for on the Internet. *Cyberpsychology, Behavior, and Social Networking*, *17*(4), 216-221.

Jans, L. C., & Kranzbhler, A. (2015). The influence of rating volume in the effects of expert versus patient online ratings. *Acta Orthopdica Belgica*, *81*(4), 662-667.

Jensen, M. and Meckling, W. (1976). Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure. *Journal of Financial Economics*, *3*(4).

Kanba, S., Kato, T., Terao, T., and Yamada, K. (2013). Guideline for treatment of bipolar disorder by the Japanese Society of Mood Disorders. *Psychiatry and Clinical Neurosciences*, *67*(5), 285-300.

Limbrock, F. (2011). Pecuniary and Non-Pecuniary Incentives in Prescription Pharmaceuticals: The Case of Statins. *The B.E. Journal of Economic Analysis & Policy*, *11*(2).

McCarter, T. (2008). Depression Overview. *American Health & Drug Benefits*, *1*(3), 44-51.

Morgan, M., & Montagne, M. (2011). Drugs on the Internet, Part II: Antidepressant Medication Web Sites. *Substance Use & Misuse*, *46*(13), 1628-1641.

Ritchie, L., Tornari, C., Patel, P. M., & Lakhani, R. (2016). Glue ear: how good is the information on the World Wide Web? *The Journal of Laryngology & Otology*, *130*(02), 157-161.

Sharma, A. and Couture, J (2014). A Review of the Pathophysiology, Etiology, and Treatment of Attention-Deficit Hyperactivity Disorder (ADHD). *Annals of Pharmacotherapy*, *48*(2), 209-225.

Trujols, J., Iraurgi, I., Oviedo-Joekes, E., & Gurdia, J. (2014). A critical analysis of user satisfaction surveys in addiction services: opioid maintenance treatment as a representative case study. *Patient Preference and Adherence*, 107.

Zgierska, A., Rabago, D., & Miller, M. (2014). Impact of patient satisfaction ratings on physicians and clinical care. *Patient Preference and Adherence*, 437.

Table 2
Review-level regressions on patient rating.

	(1)	(2)	(3)
Intercept	3.972***	3.109***	6.660***
	(0.273)	(0.905)	(1.663)
Doctor rating		0.326*	0.353*
		(0.183)	(0.188)
QALY		−0.702	−4.342*
		(0.810)	(2.333)
Number of patient ratings			−0.001*
			(0.001)
Number of doctor ratings			0.001
			(0.001)
Sex			
Male	0.050***	0.050***	0.050***
	(0.014)	(0.014)	(0.014)
Time on medication			
< 1 month	−1.537***	−1.535***	−1.536***
	(0.030)	(0.030)	(0.030)
1-6 months	−0.947***	−0.946***	−0.946***
	(0.029)	(0.029)	(0.029)
6-12 months	−0.726***	−0.726***	−0.725***
	(0.032)	(0.032)	(0.032)
1-2 years	−0.562***	−0.561***	−0.561***
	(0.032)	(0.032)	(0.032)
2-5 years	−0.302***	−0.302***	−0.302***
	(0.030)	(0.030)	(0.030)
5-10 years	−0.167***	−0.166***	−0.166***
	(0.033)	(0.033)	(0.033)
Adjusted R-squared	0.210	0.210	0.210
Bayesian info. criterion	159218	159225.6	159236
F-statistic	37.12	37.03	36.92
No. of observations	47258	47258	47258

Parentheses denote standard errors.
Coefficients are significant at 1% (***), 5% (**), 10% (*).

Table 3
Review-level regressions on average drug rating.

	(1)	(2)	(3)
Intercept	2.843***	3.034***	2.930***
	(0.302)	(0.459)	(0.488)
Doctor rating		0.139**	0.134**
		(0.057)	(0.056)
QALY		−1.377***	−1.514***
		(0.405)	(0.412)
Number of patient ratings			0.0002*
			(0.0001)
Number of doctor ratings			−0.001*
			(0.001)
Sex			
Percent male	0.120	0.171	0.044
	(0.686)	(0.681)	(0.680)
Age			
Percent 13-34	−0.131	−0.045	0.256
	(0.275)	(0.275)	(0.328)
Percent 35-54	−0.006	−0.002	0.356
	(0.506)	(0.503)	(0.535)
Time on medication			
Percent < 6 months	0.762	0.866*	0.895*
	(0.495)	(0.493)	(0.490)
Percent 6 months to 2 years	1.613*	1.666**	1.730**
	(0.838)	(0.831)	(0.827)
Adjusted R-squared	0.312	0.322	0.331
Bayesian info. criterion	658.5	662.9	677.5
F-statistic	8.72	8.72	8.37
No. of observations	358	358	358

Parentheses denote standard errors.
Coefficients are significant at 1% (***), 5% (**), 10% (*).

Table 4
Review-level regressions on average drug rating.

	(1)	(2)	(3)
Intercept	2.063***	1.930***	2.232***
	(0.306)	(0.465)	(0.497)
Doctor rating		−0.163***	−0.155***
		(0.057)	(0.057)
QALY		1.457***	1.442***
		(0.411)	(0.419)
Number of patient ratings			−0.0003**
			(−0.0002)
Number of doctor ratings			−0.0001
			(−0.0008)
Sex			
Percent male	−0.294	−0.353	−0.229
	(0.698)	(0.691)	(−0.696)
Age			
Percent 13-34	0.243	0.142	−0.232
	(0.279)	(0.279)	(0.334)
Percent 35-54	−0.310	−0.315	−0.717
	(0.515)	(0.510)	(0.545)
Time on medication			
Percent < 6 months	−0.170	−0.291	−0.315
	(0.503)	(0.500)	(−0.499)
Percent 6 months to 2 years	−0.380	−0.442	−0.480
	(0.852)	(0.844)	(0.941)
Adjusted R-squared	0.152	0.169	0.175
Bayesian info. criterion	558.4	555.7	563.1
F-statistic	4.05	4.31	4.15
No. of observations	358	358	358

Parentheses denote standard errors.
Coefficients are significant at 1% (***), 5% (**), 10% (*).