TI 2014-038/V
Tinbergen Institute Discussion Paper

Audit Rates and Compliance: A Field Experiment in Long-term Care

Maarten Lindeboom
Bas van der Klaauw
Sandra Vriend

Faculty of Economics and Business Administration, VU University Amsterdam, and Tinbergen Institute, the Netherlands.

Tinbergen Institute is the graduate school and research institute in economics of Erasmus University Rotterdam, the University of Amsterdam and VU University Amsterdam.

More TI discussion papers can be downloaded at http://www.tinbergen.nl

Tinbergen Institute has two locations:

Tinbergen Institute Amsterdam
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 1600

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Duisenberg school of finance is a collaboration of the Dutch financial sector and universities, with the ambition to support innovative research and offer top quality academic education in core areas of finance.

DSF research papers can be downloaded at: http://www.dsf.nl/

Duisenberg school of finance
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 8579

Audit rates and compliance: A field experiment in long-term care

Maarten Lindeboom* Bas van der Klaauw† Sandra Vriend‡

March 18, 2014

Abstract

We provide evidence from a large-scale field experiment on the causal effects of audit rules on compliance in a market for long-term care. In this setting care should be provided quickly and, therefore, the gatekeeper introduced ex-post auditing. Our results do not show significant effects of variations in random audit rates and switching to a conditional audit regime on the quantity and quality of applications for care. We also do not find evidence for heterogeneous effects across care providers differing in size or hospital status. Our preferred explanation for the lack of audit effects is the absence of direct sanctions for noncompliance. The observed divergence of audit rates in the conditional audit regime is the consequence of sorting and thus identifies the quality of application behavior of providers.

Keywords: auditing, field experiment, compliance, feedback, long-term care
JEL-code: C93, H51, I18

*VU University Amsterdam, and Tinbergen Institute.
†VU University Amsterdam, and Tinbergen Institute.
‡VU University Amsterdam, and Tinbergen Institute.
Address: Department of Economics, VU University Amsterdam, De Boelelaan 1105, NL–1081 HV Amsterdam, The Netherlands.

The authors would like to thank the Centrum Indicatiestelling Zorg (CIZ) for facilitating the field experiment and providing the data. We gratefully acknowledge seminar and conference participants in Aarhus, Alicante, Amsterdam, Barcelona, Dortmund and Grindelwald for valuable comments.

1 Introduction

The design of audit policies, using instruments such as the audit rate and sanctions, is concerned with a trade-off between the size of the budget for auditing and incentives for agents. This trade-off is relevant in many settings, including taxation, social insurance, pollution and health care. This paper studies the effects of audit rules on behavior of long-term care providers. Like in most countries, long-term care expenditures in the Netherlands are rapidly increasing and currently account for almost 5% of GDP. Since the operational budget for auditing is limited, the design of audit policies is essential. In our field experiment, we consider both unconditional and conditional changes in audit rates. In the conditional case, audit rates depend on the provider's recent performance.

To the best of our knowledge the effectiveness of audit rules has not been studied yet in health-care markets. These markets are often characterized by a mixture of public funding and private provision. Therefore, there is not only the usual trade-off between strictness of audits and compliance, but also between efficient spending of resources and providing care when needed. The need for quick provision has made the Dutch gatekeeper, who manages long-term care, to introduce ex-post auditing. This allows care provision to start before the result of an audit is known and hence no direct sanction is possible in case of noncompliance. The gatekeeper traditionally applies a higher audit rate to high-risk applications and finds that these have a slightly higher approval rate than low-risk applications. The gatekeeper uses this to justify ex-post auditing without direct sanctions. Only in the long run care providers can lose their contract with the gatekeeper and thereby their opportunity to apply for public funding for long-term care. The audit regime thus mainly relies on trusting care providers, but uses auditing to provide feedback. Receiving feedback is often argued to be important (e.g., Bandiera et al., 2009; Jamtvedt et al., 2006).

Our study relates to theoretical work on random auditing (Allingham and Sandmo, 1972) and conditional audit rules (Landsberger and Meilijson, 1982; Greenberg, 1984; Harrington, 1988; Friesen, 2003). This literature has focused on tax compliance and environmental regulation.[1] Empirical evidence on the performance of audit rules primarily comes from laboratory experiments.[2] The use of field data has proven to be more difficult because of the typical hidden nature of noncompliance and issues of endogeneity of the

[1] Compliance behavior of firms has been studied in particular by Crocker and Slemrod (2005), Bayer and Cowell (2009) and Hoopes et al. (2012). For extensive reviews of the taxation literature see Andreoni et al. (1998), Alm and McKee (1998), Slemrod and Yitzhaki (2002), Slemrod (2007), Kirchler et al. (2007) and Alm (2012). Empirical studies focusing on environmental regulation are e.g., Gray and Deily (1996), Laplante and Rilstone (1996) and Helland (1998).

[2] Compliance has been found to increase (slightly and non-linearly) with audit rates and modestly with sanction rates (Alm and McKee, 1998). Clark et al. (2004) experimentally compare random auditing, a conditional audit rule, and the optimal audit rule proposed by Friesen (2003). They find the latter two to be associated with fewer audits, whereas the compliance rate is maximized under random auditing. Cason and Gangadharan (2006) analyze the conditional audit rule suggested by Harrington (1988) and find participants to behave broadly in line with theoretical predictions.

audit rate (e.g. subsection 6.3 in Andreoni et al. (1998)).

Recently there have been some studies on the effects of auditing on compliance behavior using field experiments. Iyer et al. (2010) consider construction firms and find a significant effect of increasing the awareness of detection risk on the reported tax base. Slemrod et al. (2001) find a significant positive effect of informing individuals that they will be audited closely on tax payments of low and middle-income taxpayers. Finally, Kleven et al. (2011) randomly sent threat-of-audit letters to individuals in Denmark. They find that prior audits and threat-of-audit letters have a significant positive effect on self-reported income, as opposed to third-party reported income. However, all these studies generate variations in beliefs rather than true variations in audit rates.

The aforementioned studies are concerned with 'carrot-and-stick' type of policies including sanctions for noncompliance. External interventions, like sanctions, may crowd out intrinsic motivation (e.g., Bénabou and Tirole, 2003). Some studies looked at the effect of warnings as an alternative (Nyborg and Telle, 2004; Eckert, 2004). Another alternative to deterrence is the incorporation of trust into enforcement policy.[3] Mendoza and Wielhouwer (2013) build a theoretical model in which a trusted agent faces a lower audit rate and remains being trusted until found not complying. They find feasibility of trust-based regulation to hinge on agents' discount rates being sufficiently low and the existence of some costs of screening to the agent.

Our study contributes to the literature in three ways. First, whereas Slemrod et al. (2001) and Kleven et al. (2011) study the effect of announcing certain audits, our field experiment introduces true variation in audit rates, including also conditional variation. Second, audits in the market for long-term care are much more frequent than in tax auditing, due to both higher application rates and higher audit rates. Conditional audit rate updates occur, therefore, more often and are based on more information. Third, whereas direct monetary sanctions are common in many of the applications, these are absent in our setting. Instead, the audit regime relies on providing feedback.

We have detailed administrative data on all long-term care applications filed by the providers participating in our field experiment. This includes the type and amount of care services and some basic patient characteristics. Furthermore, we observe whether an application was selected for audit and, if so, the audit decision and motivation. We can thus study effects on both application quantity and quality (i.e., compliance rate). Our results do not show significant effects of an exogenous change in the unconditional audit rate on the number of applications nor on the audit approval rate. We also do not find significant effects on these two outcome variables of switching to a conditional audit regime. Even though we do observe divergence in audit rates among care providers in the conditional audit regime, we do not find much evidence for behavioral responses of care providers. The resulting audit rate changes can be caused by sorting based on

[3]Rousseau et al. (1998) give a cross-discipline overview of how the concept of trust is defined. A huge (experimental) literature on trust and reciprocal behavior in (repeated) interactions has developed, starting with the work on trust games by Berg et al. (1995) (see Fehr and Gächter (2000) for a review and Fehr and List (2004) for more recent evidence).

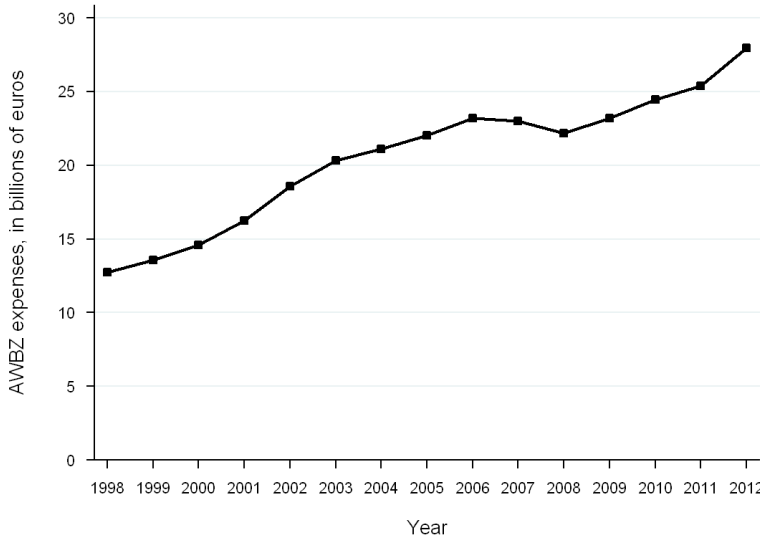
pre-experiment average approval rates of care providers. Finally, we do not find evidence for the presence of treatment effects that are heterogeneous across care providers.

The remainder of this paper is organized as follows. The next section provides details on the Dutch long-term care setting. Section 3 describes the experiment and discusses hypotheses on the effects of the various treatments (audit regime changes). Section 4 describes the data. The empirical results are presented in section 5. Finally, section 6 concludes.

2 Institutional background

Each inhabitant of the Netherlands is publicly insured for long-term care by means of the Exceptional Medical Expenses Act (AWBZ).[4] On July 1, 2013, 4.8% of the Dutch population qualified for receiving long-term care, of which 24% were aged 85 or above. Long-term care expenditures are financed by means of general taxation and co-payments. In most developed countries, long-term care expenditures are rapidly increasing. Figure 1 shows that the Netherlands is no exception. Long-term care expenditures amounted to 28 billion euros (i.e., 4.7% of GDP) in 2012, a 51% increase compared to the 2002 level. Correcting for price level changes a 60% increase between 1998 and 2012 remains, which can only partly be explained by demographic trends.

Figure 1: AWBZ expenses in the Netherlands, 1998 - 2012



Note: Data are from CBS StatLine; figures for 2011 and 2012 are preliminary.

When a patient wants to receive long-term care services, (s)he contacts a care provider. The care provider then files a request to the gatekeeper. The gatekeeper makes an as-

[4]This section draws on Mot (2010), Nederlandse Zorgautoriteit (2012a,b) and information from the Center for Care Assessment (CIZ).

sessment of the needs, stating the type and amount of care services.[5] Next, regional care offices provide the financial resources to the care provider, who can then provide the care to the patient. The patient can also file an application directly to the gatekeeper, but usually (around 85% of cases) the application is filed by the care provider. The patient can choose his/her preferred care provider and is allowed to switch care provider during care provision. This gives care providers an incentive to provide good-quality care. In 2011, there were almost 1,400 extramural care providers and 800 intramural care providers active (a provider may provide both types of care). Intramural care providers are required to be non-profit institutions; extramural care providers can also be for-profit organizations.

Annually, around one million long-term care applications are submitted. These are categorized in several types. We focus on one type, the so-called 'standard assessment guideline' (SIP) application, which accounts for approximately 34% of all applications. SIP care concerns relatively cheap care that lasts for a short period and that needs to be provided quickly. This contains, for example, wound care and/or personal care after hospitalization. Note that such care replaces more expensive hospital care at the end of a hospitalization. A substantial share of the SIP applications are filed by hospitals in the process of discharging a patient from the hospital.

To be able to file SIP applications, a care provider needs to have a contract with the gatekeeper. Because SIP care is quickly needed, care provision can start immediately after filing. Usually a random sample of 6% or 16%, depending on the risk category, is audited ex-post. Applications selected for audit are randomly divided among specialized assessors. An audit may result in an approval or disapproval. A disapproval can be either an adjustment or denial of the application. Several reasons for disapproval exist. For instance, the type of limitations making the patient eligible for long-term care may differ, different services may be needed or the necessary amount of care may be different. Ex-post auditing implies that no adjustments in the type and/or amount can be made as result of an audit. A disapproval decision hence does not have direct implications: the care provider does not have to repay anything. But, care providers do receive feedback on the audit.

3 The experiment

Our field experiment started September 17, 2012 and ran until April 7, 2013. In the experiment we varied the audit rates. We distinguish four groups in the experiment. In the *control group*, audit rates remained at the pre-experiment levels of 6% and 16% for low and high-risk types of applications, respectively. In the *low rate group*, audit rates were set to 2% and 10%. Care providers in the *high rate group* faced higher audit rates of 10%

[5]Types of care services distinguished are: personal care (e.g., help with showering), nursing (e.g., wound care), assistance (help in organizing practical matters in daily life), treatment (e.g., rehabilitation) and short-term stay or stay for an extended period of time in an institution (e.g., nursing home). For each service the amount of care (usually in hours per week) the individual qualifies for is determined.

and 26%. These are all random audit rates. Finally, for care providers in the *conditional audit group*, the audit rate depended on previous performance.

Care providers in the conditional group started at pre-experiment audit rate levels of 6% and 16%. For each approval the audit rate was reduced by 0.2 percentage point. On the other hand, if an application was disapproved, the audit rate was increased by three percentage points.[6] These adjustments imply a constant audit rate when 93.75% of audited applications are approved, which is about the target of the gatekeeper. The audit rates are subject to a minimum of 2% and a maximum of 26%. We updated audit probabilities in this group once every two weeks during the experiment.[7] For updates implemented in week t, data for weeks $t-2$ and $t-1$ were used. With each series of adjustments, care providers were informed about the outcomes of conducted audits and the resulting audit rate.

3.1 Implementation

We selected 226 care providers that filed at least four SIP applications in April 2012. These care providers accounted for 78% of all approximately 22,000 SIP applications filed in this month. We randomly assigned the care providers to one of the four groups.[8] The managing boards of the participating care providers were informed, by letter, early September 2012 about the purpose and set-up of the experiment and the assignment of the audit policy applicable for their provider. Furthermore, case managers of the care providers were informed via e-mail. Finally, a short notice of the experiment was posted in the online information bulletin at the end of October.

To show that the groups are balanced, Table 1 provides descriptive statistics on pre-experiment outcomes and some other characteristics. The table shows no significant differences across groups in the number of applications, the number of audits and the approval rate. On average, care providers filed 18 applications per week in the pre-experiment period of which about 80% are low-risk applications. Of the audited applications, 83% are approved. Around 38% of care providers are hospitals, which is balanced across groups.

3.2 Hypothesized effects

Earlier literature has mainly looked at effects on compliance rates of audits. Additionally, we are interested in the number of applications. In the Allingham-Sandmo model an unconditional increase in the (random) audit rate implies an increase in compliance. This predicts a decline in the approval rate in the low audit rate group and an improvement of

[6] Our conditional audit regime is a variant of the rule proposed by Greenberg (1984) and Harrington (1988). In their setting agents switch between a small number of groups with varying audit intensity. We implement a continuous audit rate adjustment scheme.

[7] In the start-up phase of the experiment the procedures surrounding these updates took some more time. The first (second) adjustments were in place four (three) weeks after the start of the experiment (first update).

[8] The original proposal for the field experiment including power analysis is available at http://personal.vu.nl/b.vander.klaauw/OpzetCIZOnderzoek.pdf [in Dutch].

Table 1: Descriptive statistics on outcomes in the pre-experiment period (Jan 2012 - Sept 2012) and care provider characteristics, by treatment group.

	control	**low**	**high**	**conditional**	**p-value**
Outcomes					
applications (per week)	18.61	18.21	18.73	17.13	0.969
	(2.57)	(2.43)	(2.64)	(2.11)	
audits (per week)	1.39	1.39	1.44	1.33	0.998
	(0.19)	(0.19)	(0.19)	(0.17)	
approval rate	0.84	0.83	0.80	0.88	0.102
	(0.03)	(0.02)	(0.03)	(0.02)	
Care provider characteristics					
hospital	0.32	0.44	0.38	0.39	0.669
	(0.06)	(0.07)	(0.06)	(0.07)	
frac. low risk	0.82	0.84	0.80	0.79	0.528
	(0.03)	(0.03)	(0.03)	(0.03)	
observations	1001	987	1026	1017	
# two-week periods	18	18	18	18	
# care providers	56	55	58	57	

Standard errors of the means are in brackets. Reported p-values are for Kruskal-Wallis rank tests for equality of populations.

application quality in the high audit rate group. If higher audit rates discourage invalid applications, we may also expect the total number of applications to be lower in the high audit rate group.

For the conditional audit regime there may be both a direct and an indirect effect on the approval rate. The direct or selection effect is that, if care providers do not change their application behavior, poorly performing providers will be audited more frequently decreasing the overall application quality. An indirect or behavioral effect arises when providers improve the quality of their application due to the incentives in the conditional audit regime. This has an opposite effect on the total approval rate than the selection effect. Hence, the sign of the overall effect on application quality is ambiguous upfront. However, when looking at individual care providers we identify the indirect effect. When the indirect effect is important, we may expect the number of applications to decline.

Aforementioned predictions stem from evidence on 'carrot-and-stick' type of audit policies. Our setting does not involve direct sanctions, but only provides feedback. Therefore, it relates to trust-based regulation, which according to Mendoza and Wielhouwer (2013) may work when there is a cost associated with audits. Such costs exist when auditors contact care providers for more information, in particular in case of disapprovals. The audit rate signals the degree of trust. If trust-based regulation works, unconditional high audit rates and thus less trust can have adverse effects on compliance. The conditional audit rate regime includes trust dynamics, trust can be built and dissolved based on observed performance. To the extent that providers care about the extent of trust, this regime may provide additional incentives and lead to higher approval rates.

4 Data

We use administrative data on all 269,142 SIP applications filed by the 226 participating care providers between January 1, 2012 and April 7, 2013. For each application we observe the care provider filing the application, the amount and type of care services, the application date and whether or not the application has been audited. If there was an audit, we observe the date of the audit, the assessor who performed the audit, the result of the audit (approved or disapproved) and the reasons for this audit decision. In total, 8.3% of all applications (22,279 applications) have been subject to an audit. The overall approval rate was 0.88.

For most analyses we aggregated the data at the period and care provider level, with a period length of two weeks.[9,10] We normalize the number of applications to weekly averages for ease of interpretation. The panel data set is unbalanced, because some care providers became inactive.[11]

Table 2: Descriptive statistics on outcomes during the experiment.

	control	**low**	**high**	**conditional**	**p-value**
applications (per week)	20.34	17.17	19.04	18.74	0.931
	(3.01)	(2.24)	(2.41)	(2.22)	
low risk	16.81	13.30	15.25	13.89	0.928
	(2.99)	(2.00)	(2.35)	(2.09)	
high risk	3.48	3.85	3.76	4.80	0.661
	(0.84)	(0.88)	(0.79)	(1.06)	
audits (per week)	1.52	0.64	2.43	2.34	0.000
	(0.20)	(0.10)	(0.29)	(0.31)	
approval rate	0.83	0.79	0.84	0.87	0.876
	(0.02)	(0.04)	(0.02)	(0.01)	
approval rate low risk	0.83	0.79	0.83	0.87	0.657
	(0.02)	(0.04)	(0.02)	(0.02)	
approval rate high risk	0.96	0.89	0.87	0.89	0.013
	(0.01)	(0.03)	(0.04)	(0.04)	
observations	754	756	774	752	
# two-week periods	14	14	14	14	
# care providers	54	54	56	54	

Standard error of the mean in brackets. Reported p-values are for Kruskal-Wallis rank tests for equality of populations. Recall that fraction of audits approved is missing in case of either zero applications filed or zero audits performed. Fraction of audits approved is defined in 4,857 out of 7,067 observations. For low risk and high risk fraction approved is defined in 3,949 and 2,321 cases, respectively.

Table 2 shows descriptive statistics on outcome variables during the experiment, by

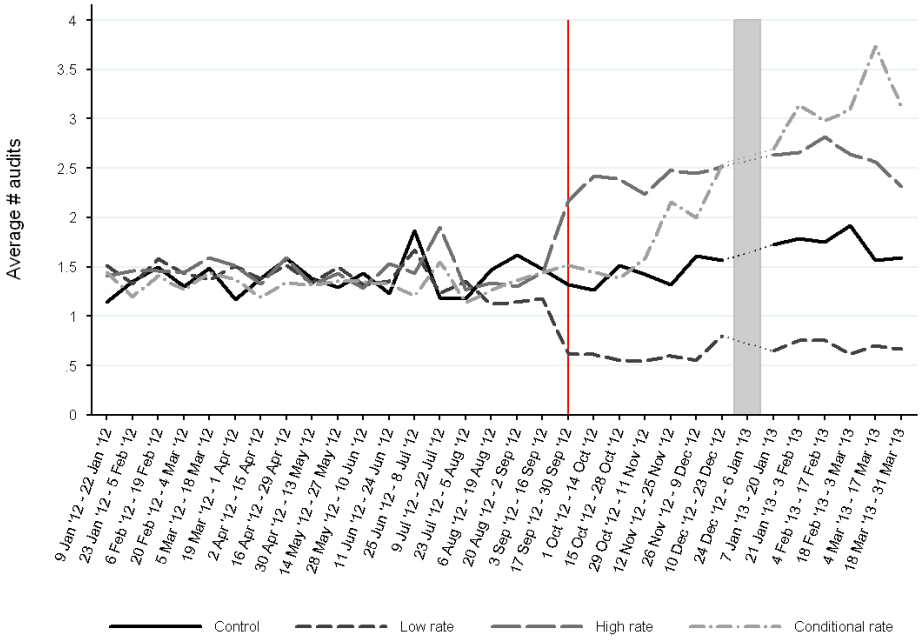
[9]The approval rate is constructed as the aggregated number of approved audits divided by the number of audits, so that a missing value results whenever no audits have been undertaken for a particular period-provider observation.

[10]Table 9 in the appendix shows that the results are robust to aggregating into one week or four week periods.

[11]Inactivity arises, for example, due to closing down or merging. Alternative approaches could be to set the number of applications to zero in case of inactivity, or to exclude all care providers from the data that became inactive during the observation period. Table 7 in the appendix shows that our results are robust against these choices.

treatment group. On average, 17 to 20 applications are filed by each care provider on a weekly basis. Most of these are low-risk applications. A Kruskal-Wallis test shows that there are no significant differences in the number of applications across groups. The significant difference across groups in the number of audits is a direct result of the imposed audit rate variation. This is illustrated in more detail in Figure 2 showing the average number of audits over time. We see a clear jump at the start of the experiment for the low and high audit rate groups, and a gradual increase in the conditional audit rate group. We return to the dynamics in the conditional audit rate group in subsection 5.1. Finally, Table 2 shows the overall approval rate to be similar across groups (around 0.84). The approval rate is larger for high risk applications.

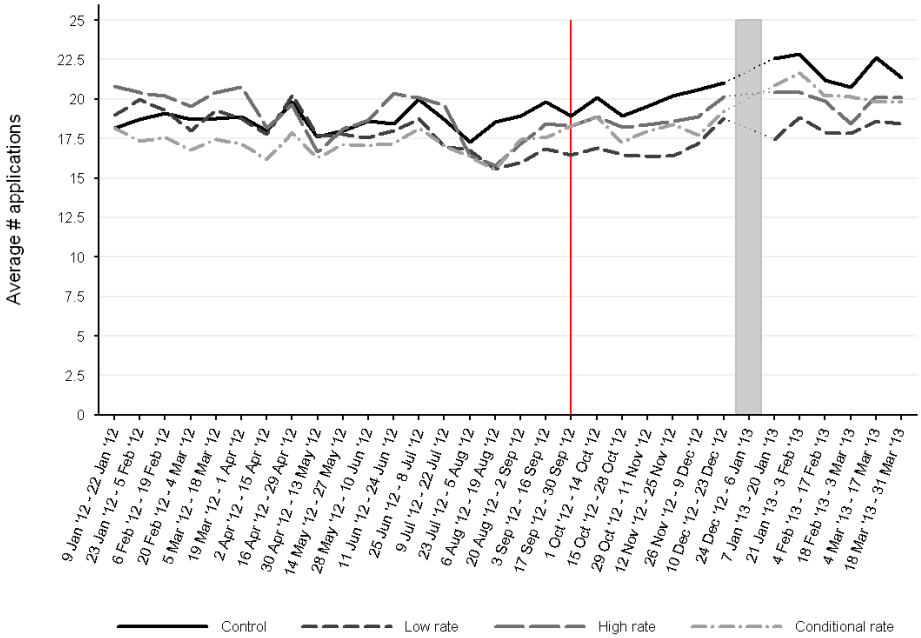
Figure 2: Average number of audits.



Notes: The vertical line indicates the start of the experiment. The grey band shows an interpolated value as around Christmas much fewer applications have been filed.

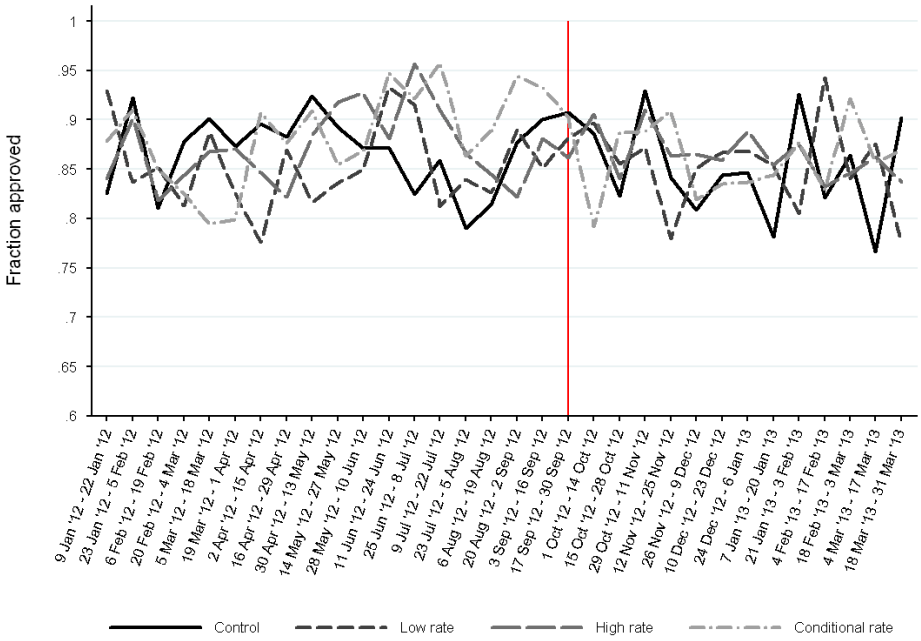
We are interested in the effect of the audit regime on application behavior of care providers. We consider effects on both quantity (number of applications) and quality (audit approval rate). Figure 3 shows the trend in the average number of applications per week, by treatment group. The pattern in the number of applications appears similar across groups, both before and during the experiment, with an increasing trend emerging from around August 2012 onwards. Finally, Figure 4 shows the trend in the average approval rate of audits. The figure illustrates that approval rates fluctuate considerably between 0.76 and 0.96. However, no systematic patterns emerge across groups.

Figure 3: Average number of applications.



Notes: The vertical line indicates the start of the experiment. The grey band shows an interpolated value as around Christmas much fewer applications have been filed.

Figure 4: Average approval rate across care providers.



5 Results

We estimate a linear panel data model to obtain the causal effects of the audit regime.

$$Y_{i,t} = \alpha_i + \gamma_t + \sum_{\substack{g \in \{low, high, \\ conditional\}}} \delta^g T^g_{i,t} + \varepsilon_{i,t} \tag{1}$$

The outcome variables $Y_{i,t}$ are the number of applications and the approval rate of provider i in time period t and $T_{i,t}$ describes the audit regime to which the provider was exposed.[12] When estimating effects on the number of applications, we normalize the outcome by dividing through the pre-treatment provider-specific average. Therefore, we leave out care provider fixed effects α_i from this specification. The treatment effects δ^g's (multiplied by 100) represent percentage changes. For regressions having the approval rate as the outcome, the causal effect estimates represent percentage points changes in the approval rate. Finally, we include time fixed effects γ_t to account for common time trends.[13] We report standard errors clustered by care provider. Bertrand et al. (2004) discuss the case of serially correlated standard errors. We have followed their suggestions to collapse the time series data into one pre- and one post-treatment period. These results are presented in Table 10 in the appendix, and are virtually the same as the results discussed below.

The baseline estimation results in the first column of Table 3 show a 3.6% and 5% decrease in the number of applications for the low audit rate group and the conditional audit rate group, respectively. For the high audit rate group we find a small increase of 2.5%. But, none of these treatment effect estimates are significantly different from zero. Results for the effect on the approval rate are shown in the fifth column. We do not see any effect on the quality of applications. The estimated coefficients are almost zero for each treatment group and insignificant.[14]

The results contradict results from laboratory experiments, where higher audit rates are found to induce slightly more compliance (Alm and McKee, 1998). The sign of the estimates for the unconditional audit rate changes is consistent with what we would expect when less trust (i.e., more frequent auditing) crowds out intrinsic motivation. But again we did not find significant effects. The lack of effects of switching to a conditional audit regime is not in line with previous studies that found positive effects on performance when using a conditional instead of random auditing rule (Clark et al., 2004).

[12] As outcome we also considered other types of long-term care applications, but we did not find any evidence for spillover effects due to changes in the audit regime. These results are available on request.

[13] We investigate whether results are robust to different time trend specifications, such as a polynomial in time or quarter dummies. Results do not change considerably, as shown in Table 8 in the appendix.

[14] If there are learning effects, it may take some time for providers to change behavior. Therefore, we estimated separate effects for the first 14 weeks of the experiments and the final 14 weeks, but we did not find evidence for different effects. These results are available on request.

Table 3: Estimation results: baseline and heterogeneous effects.

	applications			**approval rate**		
	baseline	*size*	*hospital*	*baseline*	*size*	*hospital*
	(1)	(2)	(3)	(4)	(5)	(6)
low	−0.036			−0.001		
	(0.057)			(0.022)		
low × small		−0.039			−0.085	
		(0.095)			(0.060)	
low × large		−0.051			0.026	
		(0.067)			(0.022)	
low × hospital			−0.046			−0.038
			(0.051)			(0.024)
low × non-hospital			−0.010			0.032
			(0.089)			(0.036)
high	0.025			0.000		
	(0.126)			(0.021)		
high × small		0.062			−0.021	
		(0.254)			(0.065)	
high × large		−0.027			0.005	
		(0.043)			(0.019)	
high × hospital			0.301			−0.018
			(0.290)			(0.025)
high × non-hospital			−0.146*			0.011
			(0.078)			(0.031)
conditional	−0.050			−0.000		
	(0.064)			(0.020)		
conditional × small		−0.115			−0.034	
		(0.123)			(0.053)	
conditional × large		−0.006			0.010	
		(0.039)			(0.018)	
conditional × hospital			−0.014			−0.036*
			(0.047)			(0.019)
conditional × non-hospital			−0.066			0.028
			(0.100)			(0.032)
large		0.028				
		(0.059)				
hospital			−0.055			
			(0.054)			
mean control	1.000			0.835		
mean control, small		1.000			0.756	
mean control, large		1.000			0.897	
mean control, hospital			1.000			0.893
mean control, non-hospital			1.000			0.807
F-test p-value, low		0.918	0.730		0.081	0.105
F-test p-value, high		0.731	0.138		0.702	0.460
F-test p-value, conditional		0.398	0.637		0.432	0.085
observations	7,067	7,067	7,067	4,857	4,857	4,857
# providers	226	226	226	224	224	224

Notes: Pre-experiment means of the outcome in the control group are reported as a reference, separately for different types of providers. Cluster-robust standard errors are in brackets, clustered by care provider. * significant at 10%, ** significant at 5% and *** significant at 1%. F-test p-values are for testing the null hypothesis of no heterogeneity in treatment effects.

Next, we study the presence of heterogeneous effects across care providers. The key sources of heterogeneity are whether the care provider is a hospital (38.5% of care providers) or a non-hospital (e.g., nursing homes and home care institutions), and the size of the provider measured in terms of number of SIP applications.[15] We estimate heterogeneous treatment effects by interacting the time trend and treatment indicators in equation (1) with a dummy variable for care provider type.[16]

Treatment effects on the number of applications for small compared to large care providers are shown in column (2) of Table 3. We see a slight decrease for all but small care providers in the high audit rate group, but none of the treatment effects is significantly different from zero. Furthermore, F-tests indicate that effects for small and large care providers are similar for all treatment groups. Effects on the approval rates are positive for large care providers in all treatment groups, whereas effects are negative for small care providers. However, at a 5% significance level, we do not find significant heterogeneity in treatment effects nor is any of the treatment effects estimates significantly different from zero.

Finally, we also do not find evidence for the presence of heterogeneity in effects on the number of applications for hospitals and non-hospital. For the approval rate we see that hospitals in the conditional audit rate group tend to perform worse during the experiment, in comparison to the control group, whereas non-hospitals perform somewhat better on average. However, this difference in effects is only significant at a 10% level.

While we obviously need aggregated data when having the number of applications as the outcome, causal effects on the approval rate can also be estimated using individual application-level data. In such a model, where the outcome measure is an approval dummy variable, we can add additional control variables for type and amount of care and basic patient characteristics. Estimation results are presented in Table 11 in the appendix. In each of the specifications we find small and insignificant effects of the treatments on the probability of approval during an audit. Compared to our baseline results, the magnitude of the effect in the conditional audit rate group is somewhat larger.

5.1 Sorting effects in the conditional audit rate group

This subsection describes the dynamics in the conditional audit rate group caused by audit rate updates during the experiment. In total, twelve updates of the audit probability have been implemented. Table 4 provides descriptives on these adjustments. The first column shows a gradual increase in the average low-risk audit rate from 6% to 13.5%. On average, care providers are scoring below the target approval percentage of 93.75%. In each update there are changes in both directions and for 40 to 50% of care providers the audit rate remains constant. During the experiment, the fraction of care providers for which the

[15] Other measures of size strongly correlate with this and, therefore, give very similar results.

[16] The model specification we estimate is $Y_{i,t} = \alpha_i + \gamma_t + \gamma_t \times D_i + \sum_{\substack{g \in \{low, high,\\ conditional\}}} \delta^g T^g_{i,t} \times (1 - D_i) + \sum_{\substack{g \in \{low, high,\\ conditional\}}} \delta^g T^g_{i,t} \times D_i + \varepsilon_{i,t}$, with D_i the dummy for care provider type. This is equivalent to estimating the baseline model separately for different types of care providers.

high-risk audit rate reached the maximum of 26%, increased. After the final update, this was the case for almost two-fifth of care providers, as illustrated in the penultimate column. For 73% of those, also the low-risk rate reached this maximum. As shown in the final column, only for a few care providers the audit rate hit the floor of 2% at some point.

Table 4: Descriptive statistics on the periodic adjustments in the conditional audit rate group.

Update round	**Mean (st.dev.) audit rate**				**Direction of change**			**Thresholds**	
	low risk		*high risk*		*% ↓*	*% =*	*% ↑*	*% at max*	*% at min*
0 (17 Sep '12)	6.0%	(0.0)	16.0%	(0.0)	-	-	-	-	-
1 (15 Oct '12)	5.8%	(1.1)	15.8%	(1.1)	36.8%	45.6%	17.5%	0.0%	0.0%
2 (5 Nov '12)	7.6%	(3.4)	17.5%	(3.1)	21.1%	24.6%	54.4%	1.8%	3.5%
3 (19 Nov '12)	8.2%	(4.3)	17.9%	(3.5)	28.1%	43.9%	28.1%	5.3%	1.8%
4 (3 Dec '12)	8.5%	(5.4)	18.0%	(4.3)	33.3%	43.9%	22.8%	12.3%	10.5%
5 (17 Dec '12)	9.0%	(5.8)	18.3%	(4.5)	28.1%	47.4%	24.6%	12.3%	7.0%
6 (31 Dec '12)	9.9%	(7.3)	18.6%	(5.2)	29.8%	36.8%	33.3%	21.1%	8.8%
7 (14 Jan '13)	10.5%	(7.5)	18.9%	(5.1)	17.5%	54.4%	28.1%	22.8%	7.0%
8 (28 Jan '13)	11.4%	(7.9)	19.4%	(5.2)	19.3%	50.9%	29.8%	28.1%	5.3%
9 (11 Feb '13)	12.1%	(8.1)	19.8%	(5.2)	15.8%	52.6%	31.6%	31.6%	5.3%
10 (25 Feb '13)	12.7%	(8.3)	20.1%	(5.4)	24.6%	43.9%	31.6%	36.8%	3.5%
11 (11 Mar '13)	12.9%	(8.8)	20.0%	(5.5)	21.1%	50.9%	28.1%	35.1%	5.3%
12 (25 Mar '13)	13.5%	(9.4)	20.1%	(5.5)	17.5%	54.4%	28.1%	38.6%	5.3%

Standard deviation in brackets. Averages and percentages are computed over all 57 care providers in the conditional audit rate group. In the first column the start date of the updated audit rates is shown.

The table clearly shows the presence of dynamics in the conditional audit rate group. Our earlier results indicated the absence of behavioral responses to the change in audit regime. Therefore, these changes should describe the selection effect discussed in subsection 3.2, where pre-experiment performance heterogeneity among care providers determines audit rates. To study sorting of care providers we regress the audit rate at the end of the experiment on the average approval rate of each care provider in the pre-experiment period. The estimation results in Table 5 show a significant negative relationship between pre-experiment application quality and the final audit rate. That is, the lower the pre-experiment approval rate, the higher the final audit rate.

Next, we use the control group to simulate audit rate updates under the conditional audit rate regime. Behavioral responses are, of course, absent in the control group, so simulated adjustments only reflect sorting. We take the realized number of applications and the long-run average approval rate for each care provider in the control group. In the simulations we select applications for audit based on draws from a binomial distribution with success probability equal to the audit rate in that period. Each application selected for audit is approved with a probability equal to the long-run average approval rate of the care provider. Then, we use the resulting number of approved and disapproved audits to compute the updated audit rate according to the rules in the conditional audit rate regime. We updated audit rates in twelve periods, as we did in the experiment.

We simulate the complete adjustment process 10,000 times. For each care provider,

we then compute the average final audit rate over all simulations. Table 6 compares some distributional statistics for the simulated and the actual audit rate distributions. Although there are some differences in the percentages of care providers arriving at the ends of the distribution, the mean and standard deviation are quite close. Histograms of the simulated and the actual audit rate distributions are shown in Figure 5, separately for the low-risk and high-risk rates. Testing for the equality of the simulated and the actual audit rate distributions by means of a Kolmogorov-Smirnov test leads us to conclude that the null hypothesis of equal distributions cannot be rejected (exact p-value is 0.215 for the low risk rate and 0.210 for the high risk rate). Pure sorting based on pre-experiment application quality, without behavioral responses to implemented performance incentives, could thus yield the audit rate distribution as realized for the conditional audit rate group. This confirms again the lack of behavioral responses and that divergence of audit rates in the conditional audit rate regime is solely driven by selection effects.

Table 5: Pre-experiment performance and final audit rates in the conditional audit rate group.

	audit rate
pre-experiment performance	-0.356^{**}
	(0.134)
constant	0.172^{***}
	(0.022)
R^2	0.074
observations	57

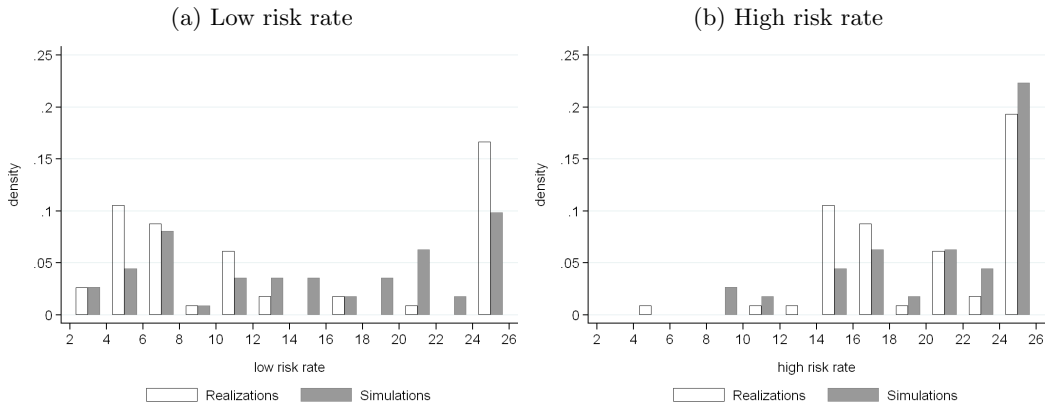
Notes: The outcome is the uncapped low risk audit rate. Only the constant term differs when using the high risk audit rate as the outcome variable. Robust standard errors are in brackets. * significant at 10%, ** significant at 5%, *** significant at 1%.

Table 6: Actual and simulated audit rate distribution.

	actual	**simulated**
mean low risk rate	13.5%	14.7%
st. dev. low risk rate	9.4	7.8
mean high risk rate	20.1%	20.8%
st. dev. high risk rate	5.5	5.2
% at maximum	38.6%	46.0%
% at minimum	5.3%	10.5%

Notes: **Actual** refers to the resulting, end of experiment, distribution in the conditional audit rate group. **Simulated** concerns the audit rates from simulating conditional audit rate updates for the control group.

Figure 5: Histograms of realized and simulated end-of-experiment audit rates.



6 Conclusion

The design of audit policy is concerned with a trade-off between the size of the auditing budget and the incentives for agents. In this paper, we reported on a field experiment studying the effects of the audit selection rule and the audit rate on compliance behavior of care providers. We considered both unconditional changes to random audit rates as well as conditional, performance-related changes.

We studied this in the Dutch market for long-term care, where an additional trade-off between efficient spending of public resources and quick provision of care exists. The need for quick provision made the gatekeeper introduce the ex-post auditing considered in this paper. In such a setting possibilities for correction and repayment in case of, for instance, errors in the amount of care provided relative to the care needs, are missing. In case of a disapproved application in an audit, the care provider will be contacted. Because of the lack of a stick, care providers are trusted to some extent. The gatekeeper justified ex-post auditing from the higher approval rate among high-risk applications which traditionally have a higher audit rate than low-risk applications.

Our results do not show evidence for unconditional changes in the audit rate to influence the number of applications for long-term care and the approval rate of applications selected for audit. We investigated the presence of heterogeneous effects across various types of care providers, distinguishing them for instance based on size and hospital-status. We did not find evidence for the presence of heterogeneous effects. Finally, even though we did observe divergence in audit rates in the conditional audit regime, this is the result of sorting rather than of behavioral responses of care providers.

The lack of effects on the number of applications and the approval rate suggests that when auditing ex-post, the frequency of audit does not provide additional performance incentives for care providers. Our preferred explanation is that lack of direct (financial) sanctions in case of disapproval of an audit does not provide incentives to care providers to comply. Badly performing care providers can lose their contract with the gatekeeper

allowing them to apply for public funding for long-term care, but this happens only very rarely.

References

Allingham, M. G. and Sandmo, A. (1972). Income tax evasion: A theoretical analysis. *Journal of Public Economics*, 1(3 - 4):323 – 338.

Alm, J. (2012). Measuring, explaining, and controlling tax evasion: Lessons from theory, experiments, and field studies. *International Tax and Public Finance*, 19(1):54 – 77.

Alm, J. and McKee, M. (1998). Extending the lessons of laboratory experiments on tax compliance to managerial and decision economics. *Managerial and Decision Economics*, 19(4 - 5):259 – 275.

Andreoni, J., Erard, B., and Feinstein, J. (1998). Tax compliance. *Journal of Economic Literature*, 36(2):818 – 860.

Bandiera, O., Larcinese, V., and Rasul, I. (2009). Blissful ignorance? Evidence from a natural experiment on the effect of individual feedback on performance. Policy Research Working Paper Series No. 4122.

Bayer, R. and Cowell, F. A. (2009). Tax compliance and firms' strategic interdependence. *Journal of Public Economics*, 93(11-12):1131 – 1143.

Bénabou, R. and Tirole, J. (2003). Intrinsic and extrinsic motivation. *Review of Economic Studies*, 70(3):489 – 520.

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, 10(1):122 – 142.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics*, 119(1):249 – 275.

Cason, T. N. and Gangadharan, L. (2006). An experimental study of compliance and leverage in auditing and regulatory enforcement. *Economic Inquiry*, 44(2):352 – 366.

Clark, J., Friesen, L., and Muller, A. (2004). The good, the bad, and the regulator: An experimental test of two conditional audit schemes. *Economic Inquiry*, 42(1):69 – 87.

Crocker, K. J. and Slemrod, J. (2005). Corporate tax evasion with agency costs. *Journal of Public Economics*, 89(9 - 10):1593 – 1610.

Eckert, H. (2004). Inspections, warnings, and compliance: The case of petroleum storage regulation. *Journal of Environmental Economics and Management*, 47(2):232 – 259.

Fehr, E. and Gächter, S. (2000). Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives*, 14(3):159 – 181.

Fehr, E. and List, J. A. (2004). The hidden costs and returns of incentives: Trust and trustworthiness among CEOs. *Journal of the European Economic Association*, 2(5):743 – 771.

Friesen, L. (2003). Targeting enforcement to improve compliance with environmental regulations. *Journal of Environmental Economics and Management*, 46(1):72 – 85.

Gray, W. B. and Deily, M. E. (1996). Compliance and enforcement: Air pollution regulation in the U.S. steel industry. *Journal of Environmental Economics and Management*, 31(1):96 – 111.

Greenberg, J. (1984). Avoiding tax avoidance: A (repeated) game-theoretic approach. *Journal of Economic Theory*, 32(1):1 – 13.

Harrington, W. (1988). Enforcement leverage when penalties are restricted. *Journal of Public Economics*, 37(1):29 – 53.

Helland, E. (1998). The enforcement of pollution control laws: Inspections, violations, and self-reporting. *Review of Economics and Statistics*, 80(1):141 – 153.

Hoopes, J. L., Mescall, D., and Pittman, J. A. (2012). Do IRS audits deter corporate tax avoidance? *Accounting Review*, 87(5):1603 – 1639.

Iyer, G. S., Reckers, P. M. J., and Sanders, D. L. (2010). Increasing tax compliance in Washington state: A field experiment. *National Tax Journal*, 63(1):7 – 32.

Jamtvedt, G., Young, J., Kristoffersen, D. T., O'Brien, M. A., and Oxman, A. D. (2006). Does telling people what they have been doing change what they do? A systematic review of the effects of audit and feedback. *Quality and Safety in Health Care*, 15(6):433 – 436.

Kirchler, E., Muehlbacher, S., Kastlunger, B., and Wahl, I. (2007). Why pay taxes? A review of tax compliance decisions. International Studies Program Working Paper 07-30.

Kleven, H. J., Knudsen, M. B., Kreiner, C. T., Pedersen, S., and Saez, E. (2011). Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark. *Econometrica*, 79(3):651 – 692.

Landsberger, M. and Meilijson, I. (1982). Incentive generating state dependent penalty system. *Journal of Public Economics*, 19(3):333 – 352.

Laplante, B. and Rilstone, P. (1996). Environmental inspections and emissions of the pulp and paper industry in Quebec. *Journal of Environmental Economics and Management*, 31(1):19 – 36.

Mendoza, J. P. and Wielhouwer, J. L. (2013). Only the carrot, not the stick: Incorporating trust into the enforcement of tax regulation. SSRN Working Paper No. 2065252.

Mot, E. (2010). The Dutch system of long-term care. CPB Document No. 204.

Nederlandse Zorgautoriteit (2012a). Marktscan extramurale AWBZ: Weergave van de markt 2008 - 2011. Utrecht: NZa.

Nederlandse Zorgautoriteit (2012b). Marktscan intramurale AWBZ: Weergave van de markt 2010 - 2011. Utrecht: NZa.

Nyborg, K. and Telle, K. (2004). The role of warnings in regulation: Keeping control with less punishment. *Journal of Public Economics*, 88(12):2801 – 2816.

Rousseau, D. M., Sitkin, S. B., Burt, R. S., and Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *Academy of Management Review*, 23(3):393 – 404.

Slemrod, J. (2007). Cheating ourselves: The economics of tax evasion. *Journal of Economic Perspectives*, 21(1):25 – 48.

Slemrod, J., Blumenthal, M., and Christian, C. (2001). Taxpayer response to an increased probability of audit: Evidence from a controlled experiment in Minnesota. *Journal of Public Economics*, 79(3):455 – 483.

Slemrod, J. and Yitzhaki, S. (2002). Tax avoidance, evasion, and administration. In Auerbach, A. J. and Feldstein, M., editors, *Handbook of Public Economics, Volume 3*. Elsevier Science B.V.

Appendix

Table 7: Robustness: various ways of dealing with inactive care providers.

	# applications			**approval rate**	
	(1)	(2)	(3)	(4)	(5)
low	−0.036	−0.013	−0.029	−0.001	0.003
	(0.057)	(0.064)	(0.057)	(0.022)	(0.022)
high	0.025	0.029	0.041	0.000	0.006
	(0.126)	(0.127)	(0.128)	(0.021)	(0.021)
conditional	−0.050	−0.059	−0.033	−0.000	0.003
	(0.064)	(0.075)	(0.063)	(0.020)	(0.019)
mean control	1.000	1.000	1.000	0.835	0.838
	(0.013)	(0.013)	(0.013)	(0.027)	(0.028)
observations	7,067	7,232	6,830	4,857	4,721
# providers	226	226	214	224	212

Notes: For the control group the mean and standard error of the outcome variable over the pre-experiment period are reported as a reference. Cluster-robust standard errors in brackets, clustered by care provider. * significant at 10%, ** significant at 5% and *** significant at 1%. Note that by construction no distinction has to be made between unbalanced and balanced panel data estimates for the approval rate. Columns (1) and (4) repeat the baseline results, column (2) shows balanced panel estimates, and columns (3) and (5) present results when excluding inactive care providers.

Table 8: Robustness: time trend specification.

	# applications			approval rate		
	(1)	(2)	(3)	(4)	(5)	(6)
low	−0.036	−0.020	−0.004	−0.001	0.005	−0.007
	(0.057)	(0.051)	(0.046)	(0.022)	(0.020)	(0.019)
high	0.025	0.040	0.057	0.000	0.007	−0.005
	(0.126)	(0.122)	(0.121)	(0.021)	(0.018)	(0.017)
conditional	−0.050	−0.035	−0.018	−0.000	0.007	−0.005
	(0.064)	(0.060)	(0.056)	(0.020)	(0.017)	(0.015)
mean control	1.000	1.000	1.000	0.835	0.835	0.835
	(0.013)	(0.013)	(0.013)	(0.027)	(0.027)	(0.027)
time trend	2-week dum.	quarter dum.	2nd order pol.	2-week dum.	quarter dum.	2nd order pol.
observations	7,067	7,067	7,067	4,857	4,857	4,857
# providers	226	226	226	224	224	224

Notes: For the control group the mean and standard error of the outcome variable over the pre-experiment period are reported as a reference. Each specification contains a separate dummy for the time period including Christmas. Cluster-robust standard errors in brackets, clustered by care provider. * significant at 10%, ** significant at 5% and *** significant at 1%. Columns (1) and (4) repeat the baseline results.

Table 9: Robustness: level of aggregation.

	# applications			approval rate		
	(1)	(2)	(3)	(4)	(5)	(6)
low	−0.036	−0.037	−0.036	−0.001	−0.000	−0.004
	(0.057)	(0.057)	(0.057)	(0.022)	(0.020)	(0.029)
high	0.025	0.022	0.023	0.000	−0.002	0.007
	(0.126)	(0.126)	(0.126)	(0.021)	(0.017)	(0.027)
conditional	−0.050	−0.054	−0.049	−0.000	−0.002	0.010
	(0.064)	(0.066)	(0.064)	(0.020)	(0.016)	(0.025)
mean control	1.000	1.000	1.000	0.835	0.833	0.835
	(0.013)	(0.012)	(0.014)	(0.027)	(0.026)	(0.027)
period length	2 weeks	1 week	4 weeks	2 weeks	1 week	4 weeks
observations	7,067	14,559	3,538	4,857	8,090	2,792
# providers	226	226	226	224	224	224

Notes: For the control group the mean and standard error of the outcome variable over the pre-experiment period are reported as a reference. Cluster-robust standard errors in brackets, clustered by care provider. * significant at 10%, ** significant at 5% and *** significant at 1%. Columns (1) and (4) repeat the baseline results.

Table 10: Robustness: accounting for serial correlation.

	# applications		**approval rate**	
	(1)	(2)	(3)	(4)
low	−0.036	−0.037	−0.001	−0.033
	(0.057)	(0.057)	(0.022)	(0.054)
high	0.025	0.018	0.000	0.043
	(0.126)	(0.125)	(0.021)	(0.053)
conditional	−0.050	−0.056	−0.000	−0.008
	(0.064)	(0.065)	(0.020)	(0.042)
mean control	1.000	1.000	0.835	0.835
	(0.013)	(0.000)	(0.027)	(0.027)
observations	7,067	444	4,857	428
# providers	226	226	224	224

Notes: For the control group the mean and standard error of the outcome variable over the pre-experiment period are reported as a reference. Baseline results have cluster-robust standard errors in brackets, clustered by care provider; for collapsed estimation heteroskedasticity robust standard errors are reported. * significant at 10%, ** significant at 5% and *** significant at 1%. Columns (1) and (3) repeat the baseline results. Columns (2) and (4) show results when collapsing the data into one pre-treatment and one post-treatment period.

Table 11: Estimation results using individual application-level data.

	(1)	(2)	(3)	(4)	(5)
low	0.011	0.001	0.003	0.003	0.004
	(0.016)	(0.016)	(0.015)	(0.015)	(0.015)
high	−0.001	−0.000	−0.000	−0.000	0.003
	(0.014)	(0.014)	(0.014)	(0.014)	(0.013)
conditional	0.016	0.017	0.018	0.019	0.019
	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)
mean control	0.835	0.835	0.835	0.835	0.835
	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)
care provider f.e.	yes	yes	yes	yes	yes
time dummies	yes	yes	yes	yes	yes
care type characteristics	no	yes	yes	yes	yes
care amount characteristics	no	no	yes	yes	yes
patient characteristics	no	no	no	yes	yes
assessor dummies	no	no	no	no	yes
observations	22,279	22,279	22,279	22,279	22,279

Notes: Cluster-robust standard errors in brackets, clustered by care provider. * significant at 10%, ** significant at 5% and *** significant at 1%. Care type characteristics include indicators for high risk category applications, an indicator for personal care, nursing care, assistance, treatment and intramural care. Care amount characteristics include constructed indicator variables for low, moderate and high intensity personal care, and similarly for nursing care. Patient characteristics include a gender dummy and dummies for age categories younger than 50, 50 to 60, 60 to 70, and so on up to older than 90. For almost 50% of observations, the assessor is unknown. We include a separate dummy for this.