The effect of expert feedback on antibiotics provision in pediatrics: Evidence from a behavioral experiment[*]

Kerstin Eilermann, Katrin Halstenberg, Ludwig Kuntz,
Kyriakos Martakis, Bernhard Roth, Daniel Wiesen

– Manuscript submitted to *Medical Decision Making* –

---

[*] Eilermann: CGS—Cologne Graduate School and Department of Healthcare Management, University of Cologne, Germany. Email: `eilermann@wiso.uni-koeln.de`. Halstenberg: Children's and Adolescents' Hospital, University Hospital of Cologne, Germany. Email: `katrin.halstenberg-scherer@uk-koeln.de`. Kuntz: Department of Healthcare Management, University of Cologne, Germany, Operations Management Group, Judge Business School, University of Cambridge, UK. Email: `kuntz@wiso.uni-koeln.de`. Martakis: Children's and Adolescents' Hospital, University Hospital of Cologne, Germany. Email: `kyriakos.martakis@uk-koeln.de`. Roth: Children's and Adolescents' Hospital, University Hospital of Cologne, Germany. Email: `Bernd.Roth@uk-koeln.de`. Wiesen: Behavioral Management Science and Department of Healthcare Management, University of Cologne, Germany. Email: `wiesen@wiso.uni-koeln.de`.

Abstract

Background. Inappropriate provision of antibiotics is a key driver of antibiotic resistance, particularly in pediatric care. To mitigate antibiotic resistance, antibiotic stewardship programs often suggest the inclusion of feedback targeted at individual providers' behavior. Empirically, it is not well understood how feedback affects individual physicians' antibiotic prescribing behavior. Also, how physicians' antibiotics provision relates to their individual characteristics is largely unexplored.

Objective. The present study was designed to analyze the causal effect of professional norm feedback (expert benchmark) and individual characteristics on physicians' antibiotic prescribing behavior in pediatrics.

Design. We employed a randomized, controlled framed field experiment with pediatricians in Germany (N=73). Pediatricians decided on the length of first-line antibiotic treatment for pediatric routine cases. In the intervention group ($N = 39$), pediatricians received feedback in form of an expert benchmark, which allowed them to compare own therapy patterns with recommendations of experts. The recommendations were elicited in a survey among directors of pediatric departments (N=20), who stated the appropriate length of antibiotic therapies for the routine cases. Pediatricians' individual characteristics were collected in a comprehensive questionnaire.

Results. Providing pediatricians with expert feedback significantly reduced length of antibiotic therapies, by about ten percent (one day) on average. Also, the deviation of the pediatricians' decisions from the expert recommendations significantly decreased. Antibiotic therapy decisions were significantly affected by pediatricians' personality traits, risk attitudes, gender, and clinical experience. The feedback effect was robust towards including these characteristics in our regression analyses.

Conclusions. Our behavioral results indicate that professional norm feedback can be an effective means in guiding pediatricians towards more appropriate antibiotics provision. Therefore, it seems to be a suitable candidate for the inclusion in antibiotic stewardship programs.

Keywords: Framed field experiment, professional norm, expert benchmark, length of antibiotic therapy, gender, personality traits, risk attitudes.

1 Introduction

Antibiotic drug resistance is a serious global public health problem.[1,2] Inappropriate use of antibiotics is widespread and contributes to rapidly increasing bacterial resistances.[3,4] Besides the choice of the antibiotic agent, the dosage, and the correct initiation, the *length of therapy* is relevant for an appropriate antibiotic treatment.[5,6] Excessive use of antibiotics as well as unnecessarily long treatment courses have a significant impact on the development of antibiotic resistances.[3,7] In pediatrics, the inappropriate provision of antibiotics is a particular concern due to antibiotic-related adverse outcomes, such as organ toxicity.[8–10] The need for effective health policies to support physicians practicing in pediatric and neonatal settings is therefore an urgent issue.[11] To support physicians practicing in pediatric and neonatal settings, so-called antibiotic stewardship programs often suggest the inclusion of feedback mechanisms targeted at antibiotic prescribing rates of individual providers.[3,12,13]

Empirically, however, it is not well understood how feedback causally affects *individual* physicians' antibiotics provision. Some studies using cross-sectional data report that feedback can be effective in achieving more appropriate antibiotics provision.[14–16] Nevertheless, cross-sectional data may suffer from multiple confounding effects (e.g., lack of control, self-selection, co-evolvement of institutions) making causal inferences difficult.[17,18] Also, evidence from randomized controlled experiments on the effectiveness of feedback in general medical practice, typically analyzing aggregated outcomes, is rather mixed.[19] Systematic evidence relating to antibiotics provision is scarce.[13] A recent randomized controlled trial in the UK reported evidence that providing feedback which emphasizes a social norm affects general practitioners' antibiotic prescriptions.[20] Similarly, a randomized controlled trial in the US found that peer comparison among primary care practitioners decreases overall antibiotic prescribing rates, aggregated at the practice level.[21] The causal effect of feedback on *individual* physicians' behavior, however, remains not well understood. We add to the existing literature by introducing an incentivized behavioral experiment to investigate the effect of professional norm feedback[22] on individual physicians' behavior.

The main objective of our study was to analyze the causal effect of expert feedback on antibiotics provision in pediatric care. We conducted a randomized, controlled, framed field experiment with 73 pediatricians, in which we implemented a 'simple' feedback intervention. According to Harrison and List's widely-used taxonomy of behavioral experiments—ranging from laboratory experiments to natural field experiments—a framed field experiment is a structured experiment conducted in the subjects' natural environment with the subjects' familiar context of the task, stakes, or information set.[23–25]

In our experiment, participants decided on length of antibiotic therapies for routine cases of pediatric infectious diseases. At a within-subject level, we introduced expert feedback: Pediatricians in the intervention group were given an aggregate expert recommendation on appropriate length of therapies, to which they could compare their own decisions. A control group did not receive any feedback. The expert recommendation was elicited in a survey among directors of pediatric departments in Germany (N=20).

We focused on the length of antibiotic therapy as the number of treatment days is critical for outcomes in children, for the occurrence of adverse effects, and for the development of antibiotic resistance.[5,6] So far, the length of antibiotic therapy has been neglected in studies on feedback interventions to improve antibiotics provision behavior. Existing studies focus on whether antibiotic therapies are initiated or not (e.g., provision of antibiotics for antibiotic-inappropriate diagnoses) or on the choice of antibiotics (narrow-spectrum or broad-spectrum antibiotics). We fill this gap in the literature by focusing on pediatricians' decisions on length of antibiotic therapies.

Further, we investigated how individual characteristics, such as personality traits, risk attitudes, gender, and clinical experience, relate to pediatricians' antibiotic therapy decisions and thereby contribute to the recent stream of literature linking physicians' characteristics to medical treatment decisions. Some evidence suggests that personality traits, such as conscientiousness and neuroticism, play a key role in health care providers' behavior[26,27] and that physicians' risk attitudes affect medical service provision.[28–31] Evidence also shows that gender and experience in medical practice relate to physicians' treatment decisions.[32] Although such individual characteristics seem to be relevant in explaining the behavior of physicians, their association to patterns in pediatricians' antibiotics provision remains rather inconclusive or even unexplored.[33–35] We analyzed how pediatricians' decisions on the length of antibiotic therapies relate to (i) personality traits, (ii) risk attitudes, (iii) gender, and (iv) clinical experience. In a comprehensive post-experimental questionnaire, we elicited the pediatricians' demographics and practice characteristics, personality traits,[36,37] and social and risk preferences,[38–40] and linked them to pediatricians' behavior in the experiment.

Our study addressed the following research questions: (i) How does expert feedback causally affect individual pediatricians' decisions on length of antibiotic therapies? (ii) Does expert feedback affect the appropriateness of antibiotics provision? (iii) How do pediatricians' individual characteristics relate to antibiotics provision?

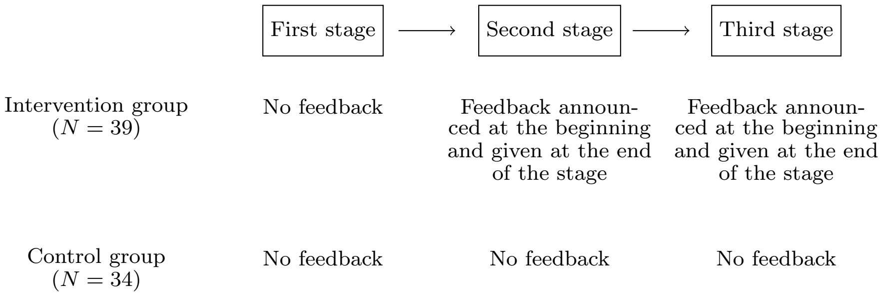
2 Methods

2.1 The experiment: Design

Our framed field experiment comprised three stages. In each stage, pediatricians decided on the length of first-line antibiotic therapies for 40 pediatric routine cases; Appendix A provides the complete list and descriptions of the cases. For all cases, a pediatrician decided on the length of antibiotic therapy (between zero and 28 days). In each of the three stages, the 40 routine cases were shown on the subjects' computer screens in a randomized order. In total, each pediatrician made 120 decisions in the three stages of the experiment. For completing the task, pediatricians received a lump-sum payment of €50.

Pediatricians were randomly allocated to either an intervention or a control group. In the intervention group, we introduced feedback in form of an expert benchmark at the within-subject level (see Figure 1). In the first stage, no feedback was provided. In the second stage, we announced that feedback would be provided at the end of the stage. After the second stage, the feedback was shown (graphically as bar charts and numerically) such that subjects could compare their choices with the expert benchmark. This design allowed us to disentangle the effect of merely announcing feedback (comparing subject behavior in the first and second stages) and the effect of actually receiving feedback which allowed pediatricians to compare their average length of antibiotic therapies for the 40 cases with the expert benchmark (comparing behavior in the second and third stages). In

Figure 1: Stages of the experiment



Notes. In each stage of the experiment, subjects decided on the length of antibiotic therapy for 40 routine cases, which were shown in a randomized order. The first stage ($t = 1$) was the same in the intervention and control groups. At the beginning of the second stage ($t = 2$), the intervention group was told that feedback would be given. After the second stage, feedback was shown such that subjects could compare the average of their chosen length of antibiotic therapies with the expert benchmark. The third stage ($t = 3$) is analogous to the second stage. In the control group, the decision situations in the second and third stages were identical to those in the first stage, and no feedback was announced or given.

a control group, no feedback was announced or provided at any stage. For the instructions of the experiment, please see Appendix B.

2.2 Medical cases and expert benchmark

The 40 pediatric routine cases covered a broad range of typical infections in pediatrics. They comprise six categories: (i) neonatal infections, (ii) infections of the central nervous system, (iii) bone and joint infections, (iv) upper respiratory tract infections, (v) lower respiratory tract infections, and (vi) urinary tract infections. The cases had been validated by pediatricians of the Children's and Adolescents' Hospital at the University Hospital of Cologne (who did not participate in the experiment).

For all cases, the pediatricians decided on the length of *first-line* antibiotic therapy, which could be between zero and 28 days ($d \in \{0, 1, \ldots, 28\}$). Note that besides the length of antibiotic therapy, the class of antibiotics as well as the dosage of an antibiotic agent also play a role for the treatment outcome.[41] We asked the participants in our study to consider a standard antibiotic agent and a standard dosage. We did not specify the agent as we intended to leave the decision on whether any antibiotics should be prescribed to the discretion of the pediatrician. With the option to choose zero days of antibiotic therapy, our cases actually include the decision on whether to prescribe any antibiotics or not. The cases were designed such that a standard antibiotic agent as well as a standard dosage should be available for each case.

For our feedback intervention, we considered a professional norm feedback. More specifically, we used an expert benchmark which was intended to reflect personal expertise, national medical guidelines, and local standards in pediatric departments. To this end, we surveyed directors of German pediatric departments (referred to as 'experts' in the following) on their recommended length of antibiotic therapies for the 40 routine cases that we used in the experiment. In total, 50 randomly chosen directors were contacted by a formal letter, in which they were asked about their willingness to participate in a survey. 20 directors participated in our online survey in September and October 2014.

As the expert benchmark, we chose the length of therapy averaged over all cases and experts. This information was then used in the experiment. The expert benchmark was 6.42 days (SD 4.94, 95% CI 4.26 to 8.59) of antibiotic therapy.

To qualify the experts' decisions and to verify them as a suitable benchmark, we compared the experts' decisions with published recommendations on length of antibiotic therapy for each respective case. It is noteworthy to say that the experts' decisions imply a high compliance recommendations; for descriptive results, see Appendix C. We thus argue that the aggregated expert opinion can be considered a suitable benchmark for the *appropriate* average length of antibiotic therapies. We further chose the 'expert benchmark' as a means of feedback: (i) to maintain the pediatricians' discretion in choosing the cases for which, if at all, they would change their initially chosen length of therapy, (ii) to mimic a simple feedback mechanism which could potentially be implemented in a real clinical setting, as providing feedback on a case-by-case basis seems prohibitively challenging, and, (iii) to provide pediatricians in the experiment with a simple *directional* reference which could guide their own decisions.

Providing pediatricians with an expert benchmark that allows comparison of own decisions with an expert recommendation is distinct from feedback applying peer comparisons (e.g., relative performance compared to peers). The latter may have unintended effects, such as change in behavior of peers performing better than the average towards the average-peers' performance (so-called 'boomerang effect'); see Linder on the importance of the design of feedback mechanisms.[42] Our feedback rather addresses each individual pediatrician's self-image.[43–45]

The aim of interventions in medical practice is usually to increase the appropriateness of care and to enhance adherence to evidence-based recommendations. Our feedback mechanism is distinct from earlier ones as it aims to direct pediatricians' antibiotic therapy decisions towards what would have been an appropriate length of therapy. Only Meeker et al.[21] used a similar approach by using peer comparison with top performers instead of comparison with average-performing peers in their feedback intervention.

2.3 Sample and procedure

The computerized experiment was conducted with mobile tablet computers of the Cologne Laboratory for Economic Research (CLER). The experiment was programmed in z-Tree.[46] Experimental sessions took place at the Children's and Adolescents' Hospital at the University Hospital of Cologne (October and December 2014), the Children's Hospital of the City of Cologne (June 2015), and during the annual conference for pediatricians (*Päd-Ass 2015*) in Cologne (March 2015). Experiments were conducted in hospital seminar rooms, which we equipped with tablet computers and cubicles to ensure anonymous decision-making; for an illustration, see Figure D.1 in Appendix D.

Overall, 73 physicians participated in our experiment. Physicians had to practice pediatrics to be included in the study. Directors of pediatric departments were excluded. Pediatricians were recruited via e-mail and posters, which provided general information about the experiment and the scheduled sessions. Pediatricians were allowed to register only for one of the sessions publicized through an online poll. In total, eight sessions were conducted. In the sessions at the University Children's and Adolescents' Hospital of Cologne and the *Päd-Ass* conference 2015, 22 and 6 subjects participated in the intervention and 14 and 20 in the control group, respectively. At the Children's Hospital of the City of Cologne, 11 subjects participated in the intervention group.

Using a simple coin toss, it was randomly determined whether feedback intervention or control treatment was employed in a particular session. Pediatricians, uninformed about the content of the experiment prior to participation, were therefore allocated to one of the two treatments at random. The baseline characteristics of the participants were well balanced between the two experimental groups; see Table 1. The socio-demographic characteristics, the share of consultants, and the participants' experience (measured in years worked in hospital) were fairly similar in the two treatment groups.

Table 1: Baseline characteristics of the study population

	Intervention ($N = 39$)	Control ($N = 34$)
Sex		
Male	11 (28%)	6 (18%)
Female	28 (72%)	28 (82%)
Share of consultants	15 (39%)	12 (35%)
Years in hospital	5.37 (4.66)	5.05 (5.98)

Notes. Data are N (%) or mean (sd). Areas of specialization within pediatrics were neonatology, nephrology, and hematology/oncology.

We employed a double-blind procedure. The experimenter who conducted the experiment and managed the data was not involved in the recruiting of participants. For each session, an external research assistant, employed by the Department of Personnel Economics of the University of Cologne, facilitated subject recruitment, registration, and remuneration. Upon their arrival, pediatricians drew a number (from one to 20) that indicated their cubicle. Each cubicle was equipped with a tablet computer. Before the experiment started, subjects received written instructions describing the general structure, the decision situation, and the task of the experiment. Decisions on the computer screens were made anonymously; the experimenter was only able to link the randomly assigned computer number to the respective participant's behavioral data. Payment provided in sealed envelopes.

The experimental sessions lasted for about one hour. Before the experiment started, written informed consent was obtained from all subjects. Prior to each stage of the experiment, subjects received written instructions. Subjects were given sufficient time to read the instructions and any upcoming questions were answered in private at the participant's cubicle. After completing the experiment and before receiving their payment, subjects were asked to answer a questionnaire on demographics, personality traits,[36,37] as well as social, risk, and time preferences ('Economic preferences');[38–40] for the questionnaire, see Appendix E. One month after the study had been concluded, debriefings with participating pediatricians and heads of pediatric clinics took place.

2.4 Statistical analyses

First, to analyze the impact of pediatricians' individual characteristics on length of antibiotic therapies chosen in the experiment, we employed ordinary least squares (OLS) and Tobit regressions (as our data was left-censored at the value of 0 days). Second, to address our main research questions, in particular to determine the effect of expert feedback on length of antibiotic therapies and on appropriateness of length of therapies, we employed non-parametric statistical analyses. At the within-subject level, we compared differences in length of therapies and in absolute deviation from the expert recommendations between the three stages. In doing so, we assessed the impact of merely announcing that feedback would be given (differences between the first and second stages) and of actually receiving expert feedback (differences between the second and third stages). We conducted these analyses by using two-sided Fisher-Pitman permutation tests for paired samples. For a between-subject comparison of decisions in the intervention and the control groups, we used two-sided Fisher-Pitman permutation tests for independent samples. To account for heterogeneity in the behavioral data, we also ran a series of OLS and Tobit regressions. To check the robustness of our results on a medical case basis, we conducted non-parametric statistical analyses. STATA 14.1 was used in all the analyses.

2.5 Role of the funding source

The funding source had no role in the study design or implementation.

3 Results

3.1 Pediatricians' characteristics and antibiotic therapy

We first analyzed how the pediatricians' characteristics relate to their antibiotic therapy decisions. We focused on decisions in the first stage of the experiment as pediatricians had neither been exposed to feedback nor had feedback been announced. Pediatricians chose an average length of antibiotic therapy for the 40 routine cases of 7.53 days (95% CI 7.32 to 7.73, $N = 2,920$).

Table 2 provides estimation results from OLS regression analyses. Female pediatricians chose significantly longer treatment courses: Women's length of therapies were on average about one day longer than men's. Also, the pediatricians' experience highly significantly affected decisions on length of antibiotic therapies. The more years pediatricians had practiced in a hospital, the lower the length of therapies. Moreover, the pediatricians' willingness to take risks[38–40] had a significant negative effect on the length of therapies. Considering the Big Five personality traits,[36,37] we found that the more conscientious pediatricians chose a shorter length of therapies. Other personality traits did not significantly relate to pediatricians' decisions. Further, we controlled for individuals' economic preferences, which comprised validated measures for trust, reciprocity, and altruism, as well as risk and time preferences.[38–40] The effects were robust when controlling for medical cases. In all models, we included dummies for each experimental session to control for any session effects. Applying Tobit regression models yielded qualitatively and quantitatively very similar results.

Table 2: Regressions on the impact of characteristics on the length of antibiotic therapies

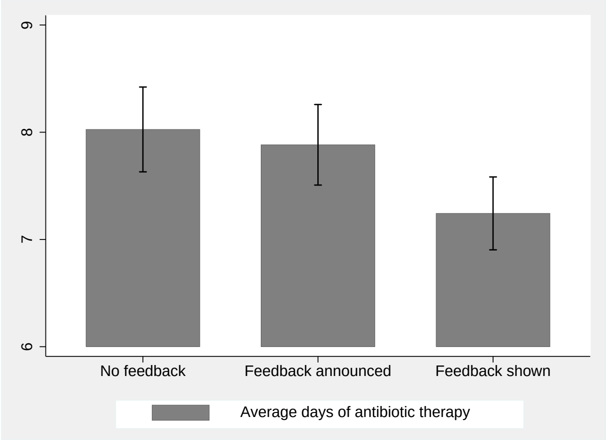
Dependent variable:	Length of antibiotic therapies (in days, d_1)	
Model:	(1) OLS	(2) Tobit
Female	0.919**	1.226***
	(0.397)	(0.449)
Years in hospital	-0.118***	-0.131***
	(0.031)	(0.034)
Willingness to take risks	-0.307***	-0.378***
	(0.087)	(0.096)
Extraversion	0.081	0.109
	(0.165)	(0.180)
Agreeableness	0.137	0.162
	(0.186)	(0.208)
Conscientiousness	-0.633***	-0.685***
	(0.200)	(0.227)
Neuroticism	0.180	0.192
	(0.129)	(0.140)
Openness	0.186	0.240
	(0.135)	(0.152)
Economic preferences	Yes	Yes
Case dummies	Yes	Yes
Session dummies	Yes	Yes
Constant	3.504***	0.183
	(1.256)	(1.497)
Observations	2,920	2,920
Subjects	73	73
R^2/ Pseudo R^2	0.609	0.609

Notes. This table shows parameter estimates from OLS and Tobit regressions, considering the decisions from the first stage of the experiment. The dependent variable is 'length of antibiotic therapies', measured in days. Robust standard errors, clustered at the individual-subject level, are shown in parentheses. 'Willingness to take risks' was measured on a Likert scale ranging from 0 (fully risk-averse) to 10 (fully risk-seeking).[38–40] Besides the Big Five personality traits, we controlled for 'Economic preferences', which comprise validated measures for trust, reciprocity, and altruism, as well as risk and time preferences.[38–40] To control for session effects, we included a dummy for each of the eight sessions. We also added dummies for the 40 medical cases. *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.10$.

3.2 The effect of feedback on the length of antibiotic therapies

We analyzed the effect of feedback on antibiotics provision at a within-subject level. Figure 2 indicates the average days of antibiotic therapy for the three stages of the experiment in the intervention (feedback) group. In the first stage (without feedback), the average length of antibiotic therapy was 7.97 days (95% CI 7.42 to 8.53, $N = 1,560$). After feedback had been announced (in the second stage), the average number of days fell slightly to 7.83 (95% CI 7.31 to 8.35, $N = 1,560$). The announcement effect was not statistically significant (p=0.153). In the third stage, when pediatricians had compared their average length of antibiotic therapy (from the second stage) with the expert benchmark, the mean number of days of antibiotic therapy fell to 7.23 (95% CI 6.93 to 7.53, $N = 1,560$). Providing the subjects with the expert benchmark significantly reduced the length of antibiotic therapies (p=0.000).

Figure 2: The effect of feedback on the length of antibiotic therapies



Notes. This figure shows the length of antibiotic therapies (mean and 95% CI) in the three stages of the experiment in the intervention group. In each stage, 39 subjects decided on the length of antibiotic therapies for 40 routine medical cases presented in a random order on the subjects' computer screens. No feedback was given in the first stage; feedback was announced at the beginning of the second and third stages and shown after the second and third stages.

We also compared subjects' behavior in the Feedback and No-Feedback groups. Table 3 shows the differences in length of antibiotic therapies between the second and first stages and between the third and second stages for both the intervention and control groups. In the intervention group, the mean change in the number of days of antibiotic treatment due to announcing that feedback would be given was -0.15 days (SD 0.63, 95% CI -0.34 to 0.06). The mean change in the number of days was -0.06 (SD 0.25, 95% CI -0.28 to 0.16) for the control group; this change did not differ significantly from the change in the intervention group ($p = 0.577$). After feedback was provided to pediatricians in the intervention group (stage 3), the number of days of antibiotic treatment changed, on average, by -0.60 (SD 0.97, 95% CI -0.91 to -0.29). In the control group, the change in the number of days was -0.06 (SD 0.25, 95% CI -0.15 to 0.03). The difference between the change in stage 3 in the intervention group and the change in the control group was highly significant (p=0.000).

Table 3: Differences in days of antibiotic therapy and absolute deviations from the expert recommendations

	Experimental treatment		
	Feedback (Intervention)	No Feedback (Control)	p-value
A. Change in days of therapy			
$d_2 - d_1$	-0.15 (0.63)	-0.06 (0.63)	0.577
$d_3 - d_2$	-0.60 (0.97)	-0.06 (0.25)	0.000
B. Change in abs. deviation from the expert recommendations			
$\Delta_2 - \Delta_1$	-0.15 (0.56)	-0.09 (0.45)	0.587
$\Delta_3 - \Delta_2$	-0.33 (0.73)	0.00 (0.27)	0.004

Notes. This table indicates changes in average days of antibiotic therapy and in average devaition from the expert recommendations. Standard deviations are in parentheses. Note that Δ_t denotes the average absolute deviation per subject from the expert recommendation B for cases $j = 1, 2, \ldots, 40$ and subjects $i = 1, 2, \ldots, 73$ in stage $t \in \{1, 2, 3\}$ of the experiment. More formally, $\Delta_t = \frac{1}{n}\frac{1}{40}\sum_{i=1}^{n}\sum_{j=1}^{40}|d_{ijt} - B_j|$, with $B_j = \frac{1}{20}\sum_{i=1}^{20} d_{ij}^{\text{experts}}$ and p-values for differences between the groups are shown for two-sided Fisher-Pitman permutation tests for independent samples.

Further, in order to assess the effect of feedback on appropriateness of therapy decisions, we analyzed the pediatricians' absolute deviation from the experts' recommendations on length of therapy for each case. For each of the cases, we use the experts' decisions to define the *appropriate* length of therapy. In stage 1, the pediatricians' absolute deviation from the experts' recommendations was not significantly different between the intervention and the control groups (p=0.301). In the intervention group, the difference between the pediatricians and the expert recommendations was only slightly affected by the announcement of feedback (p=0.085). After providing feedback, however, the difference to the experts significantly decreased ($p = 0.001$) in the intervention group. In the control group, there were no significant differences between the stages (both p-values > 0.05). The change in deviation in the second stage was not significantly different between the intervention and the control groups ($p = 0.587$). In stage 3, however, the reduction in deviation from the expert recommendations was significantly greater in the intervention group than in the control group ($p = 0.004$). This suggests that subjects responded to expert feedback and adapted their decisions in the direction of the given benchmark. Decisions on length of therapies were more appropriate in the intervention group after expert feedback had been given.

Aggregated over the 40 cases, we found a highly significant decrease both in length of antibiotic therapies and in absolute deviation from the expert recommendations after provision of the expert benchmark in the intervention group. To validate this overall finding, we further analyzed the subjects' decisions prior to feedback and afterwards on a base-by-case basis. This analysis supports our major result. We found that, for the vast majority of the cases, the number of days as well as the pediatricians' deviation from the expert recommendations for the respective case decreased. In particular, we observed a decrease or no change in length of therapy for 37 out of the 40 cases and a decrease or no change in absolute deviation from the experts for 35 out of the 40 cases. Changes in the opposite direction for the remaining cases were not significant ($p > 0.10$, Wilcoxon matched-pairs signed-rank tests). For a detailed analysis of the feedback effect on a case-by-case basis, see Tables F.3 and F.4 in Appendix F.

Finally, we employed OLS regressions to investigate how robust our finding is that feedback influences pediatricians' decisions; see Table 4. The feedback effect on pediatricians' behavior (after feedback has been shown) is indicated by the interaction 'Third stage × Feedback'. The effect of announcing feedback is indicated by the interaction 'Second stage × Feedback'. As dependent variables we used 'days of antibiotic therapies' and the 'deviation from the expert recommendations', measured as absolute difference between the pediatricians' choices and the experts' recommended length of therapies (in days). In all models, we included dummies for each experimental session to control for any session effects. The length of antibiotic therapies was not significantly different between both experimental groups in the first stage of the experiment.

The estimates of the regression model support the results of the non-parametric analyses. Providing pediatricians with feedback led to a highly significant reduction in length of antibiotic therapies, and merely announcing feedback did not significantly affect the length of therapies; see Model (1) in Table 4. Also, the absolute deviation from the expert recommendations was significantly reduced by providing feedback; see Model (5) in Table 4. These findings are robust towards including case dummies, individual-specific controls, including gender, years worked in hospital, and personality traits, as well as economic preferences; for length of therapies, see Models (2) to (4), for deviation from the expert recommendations, see Models (6) to (8) in Table 4. Results from Tobit regression models were qualitatively and quantitatively very similar, see Table F.5 in Appendix F. Dummies for each experimental session were included in all models to control for any session effects.

Table 4: OLS regressions on the effect of feedback on antibiotic therapy decisions

Dependent variable	Length of antibiotic therapy (in days)				Abs. deviation from the expert recommendations (in days)			
Model	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Feedback (= 1 if intervention)	-0.048	-0.048	-0.280	-0.526	-0.264	-0.264	-0.422	-0.816**
	(0.758)	(0.760)	(0.682)	(0.660)	(0.479)	(0.480)	(0.377)	(0.370)
Second stage (= 1 if second stage)	0.019	0.019	0.019	0.019	-0.018	-0.018	-0.018	-0.018
	(0.236)	(0.236)	(0.236)	(0.236)	(0.178)	(0.178)	(0.178)	(0.178)
Third stage (= 1 if third stage)	0.521*	0.521*	0.521*	0.521*	0.304	0.304	0.304	0.304
	(0.271)	(0.272)	(0.272)	(0.272)	(0.191)	(0.191)	(0.191)	(0.192)
Announcement effect (Second stage × Feedback)	-0.082	-0.082	-0.082	-0.082	-0.068	-0.068	-0.068	-0.068
	(0.146)	(0.146)	(0.146)	(0.147)	(0.118)	(0.118)	(0.118)	(0.118)
Feedback effect (Third stage × Feedback)	-0.633***	-0.633***	-0.633***	-0.633***	-0.393**	-0.393**	-0.393**	-0.393**
	(0.197)	(0.198)	(0.198)	(0.198)	(0.150)	(0.150)	(0.150)	(0.151)
Pediatricians' characteristics	No	No	Yes	Yes	No	No	Yes	Yes
Economic preferences	No	No	No	Yes	No	No	No	Yes
Case dummies	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Session dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Constant	7.860***	1.600**	3.875***	3.068**	3.027***	1.714***	3.725***	3.587***
	(0.707)	(0.708)	(1.386)	(1.218)	(0.440)	(0.455)	(0.762)	(0.719)
Observations	8,760	8,760	8,760	8,760	8,760	8,760	8,760	8,760
Subjects	73	73	73	73	73	73	73	73
R^2	0.014	0.577	0.588	0.612	0.014	0.180	0.205	0.239

Notes. This table shows parameter estimates from OLS regressions. The interaction 'Third stage × Treatment' indicates the effect of showing feedback to subjects. In Models (1) to (4), the dependent variable is 'length of antibiotic therapy (in days)'. In Models (5) to (8), the dependent variable is 'absolute deviation from the expert recommendations', measured in absolute values of the difference between the pediatricians' choices and the experts' recommended length of therapies (in days). For each case, the subjects' choices were compared to the experts' aggregate opinion for the respective case. Robust standard errors, clustered at the individual-subject level, are shown in parentheses. The variable 'Pediatricians' characteristics' includes subjects' gender, years worked in hospital, and the Big Five personality traits. 'Economic preferences' comprise validated measures for trust, reciprocity, and altruism, as well as time and risk preferences.[38–40] The variable 'case dummies', which are included in models (2) to (4) and (6) to (8), indicates 40 dummies, one for each of the 40 medical cases. Furthermore, dummies for each experimental session were included in all models to control for any session effects. *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.10$.

4 Discussion

In this study, we introduced a novel framed field experiment with pediatricians (N=73) to analyze the causal effect of expert feedback on antibiotics provision in a tertiary pediatric care setting. Pediatricians decided on the initial treatment plan for the length of antibiotic therapies for a series of pediatric routine cases. We found that giving pediatricians simple directional feedback in form of an *expert benchmark*—in particular, the aggregated length of therapy recommended by directors of pediatric departments (N=20), reflecting their personal expertise, national medical guidelines, and local standards in pediatric departments—significantly reduced the length of antibiotics therapy by about ten percent (about one day). Through feedback, the absolute deviation of the pediatricians' decisions from the experts' recommended length of therapies decreased significantly. To further qualify this finding, case-by-case analyses showed that the directors' recommendations were mostly in line with clinical guidelines and treatment recommendations. Our behavioral data thus evidenced that showing the expert benchmark nudged pediatricians towards a more *appropriate* use of antibiotics.

Our study design, which combines an experiment and a comprehensive survey, also allowed us to relate individual pediatricians' decisions on the length of antibiotics therapy to their individual characteristics. We found that pediatricians' individual characteristics play a significant role in antibiotic therapy decisions. Most importantly, pediatricians who were more experienced, more willing to take risks, or more conscientious tended to choose shorter courses of antibiotics therapy. Female pediatricians chose significantly longer courses of therapy than male pediatricians. Further, while pediatricians responded to feedback in a heterogeneous way, the main feedback effect was robust towards pediatricians' main characteristics such as gender, age, practical experience and personality traits. These findings suggest that feedback using an expert benchmarks is a useful means to be included in antibiotic stewardship programs.

The expert benchmark shown to pediatricians transmitted a professional norm[22, 42] which was directed at affecting pediatricians' self-image concerns. That means, capitalizing on the human capacity to reflexive thinking,[47] expert feedback might have triggered pediatricians' self-directed concerns and may refer to the awareness of congruence between a professional norm (expert benchmark) and themselves (own antibiotic therapy decisions). In contrast, social-image concerns, which imply that individuals either like to think positively of themselves or have a preference for being liked and well regarded by others,[43–45] were no candidate to explain the change in pediatricians' behavior, as pediatricians took their decisions in full anonymity and no information on identity or treatment patterns was shared among participants in the experiment. Related experimental studies rather focus on physicians' reputational concerns. Hallsworth et al.,[20] who argued that a letter to British primary care practices from England's Chief Medical Officer with information on the recipient's practice's prescribing rate compared to other practices' prescribing rates in the local area implies a social norm, reported a significant reduction in antibiotics prescription. Two related controlled experiments in primary care found that peer comparison decreased overall antibiotic prescribing and nudges fostered guideline-concordant antibiotic prescribing.[21,48] We add to this recent stream of the literature by providing behavioral evidence on the effect of professional norm feedback which addresses self-image concerns in the absence of peer comparisons at an *individual* decision-maker level.

Limitations. This study has some limitations. First of all, one might argue that the experimental design is too simplistic to be reflective of a real clinical setting. Indeed, our experimental frame was parsimonious, used a set of stylized pediatric routine cases, and did not allow for additional information acquisition (e.g., from lab tests) to better assess the cases. More specifically, one might object, that (i) we were not able to asses pediatricians' response to influencing factors, which are relevant in real-world clinical settings, but might go beyond the stylized case descriptions—for example, actual parental expectations[49] or risk or efficacy perceptions.[50] (ii) Second, different interpretation of the same case information may have affected individual physicians' judgements of the cases and their treatment decisions.[51,52] While we kept the medical case information constant for all subjects, our study design and analyses did not focus on the process by which the therapy decisions were made. (iii) Third, one might argue that in real clinical practice physicians may ask colleagues or search for information in guidelines, or order additional lab test, before making a decision on antibiotic treatment. In our experiment, however, we rather isolated the effect of feedback for a comparable set of information for each case. We made a real *ceteris paribus* variation. The parsimonious experiment enables us to control the decision environment and to avoid deliberately confounding factors potentially affecting pediatricians' behavior. For example, not making additional information acquisition options available, ensured that all participants got exactly the same information when making their decisions, allowing us to draw causal inferences on the feedback effect.

Another concern might relate to the aggregated nature of our feedback mechanism. We used an aggregated benchmark instead of case-by-case recommendations. Related studies on the effect of interventions targeted at physicians' use of antibiotics also employed prescribing rates aggregated over different cases and diagnoses as benchmark for comparison of physicians' prescribing behaviors.[20,21] Our feedback design allowed us to examine whether a 'simple' feedback intervention raised their awareness for appropriate use of antibiotics while maintaining their discretion on antibiotics provision for each medical case. After feedback had been provided, we observed an overall change in length of therapies towards what is appropriate. One might argue that our aggregated results could conceal negative changes in length of therapy for individual cases. Analyses on a case level, however, did not confirm this objection. Our results hence suggest that comparison of own provision behavior with an expert benchmark raises pediatricians' awareness for judicious use of antibiotics, but maintains their individual and case-specific discretion when deciding whether and for which cases to adjust their treatment decisions.

Finally, one might argue our behavioral results might be biased by self-selection into experimental sessions. This is an important concern as we did not randomize between 'feedback' and 'no feedback' within sessions. However, it was randomly determined whether the feedback intervention or the control treatment was employed in a particular session. As we conducted only one single session at one of the participating clinics (the Children's Hospital of the City of Cologne), all participants from this clinic were assigned to the intervention group. To account for this potential concern, we included dummies for each session in our regression analyses. Further, we ran all regression models without the pediatricians from the Children's Hospital of the City of Cologne (N=11). This did not affect our estimation results; see Table F.6 in Appendix F.

In sum, our results provide valuable insights into how feedback—in the form of a benchmark which enables a comparison with experts—affects individual pediatricians' behavior. We show that expert feedback, reflecting a professional norm for appropriate antibiotic therapy durations, can be an effective means in guiding pediatricians towards a more appropriate provision of antibiotics. Using a novel methodology, our behavioral results provide evidence that professional norm feedback works in the context of antibiotics provision. Our findings thus make a strong case for the inclusion of individual feedback addressing the physicians' self-image in antibiotic stewardship programs.

Finally, an appealing feature of our parsimonious design is that it lends itself to further research. For example, future research could consider other feedback mechanisms, such as relative performance benchmarks and different medical settings, to test whether our findings can be generalized to medical areas beyond pediatrics. Future studies could also investigate how physicians' individual characteristics moderate the effect of feedback mechanisms or other antibiotic stewardship measures.

Acknowledgements

We are grateful to Professor Dr Jörg Dötsch, MD, and Professor Dr Michael Weiss, MD, for providing support and the facilities to conduct our experiment at the Children's and Adolescents' Hospital at the University Hospital of Cologne and the Children's Hospital of the City of Cologne, respectively. For helpful comments and suggestions, we thank Pablo Brañas, Matteo M. Galizzi, Robert Hamm, Glenn W. Harrison, Heike Hennig-Schmidt, Bernd Irlenbusch, Hendrik Juerges, Juliane Koeberlein-Neu, Olga Kostopoulou, Joe Newhouse, Andrea Rachow, Martin Roland, Gari Walkowitz, and seminar participants at the Universities of Cologne and Wuppertal as well as participants at the German Association for Health Economics (dggö) conference in Berlin, Germany, in 2016, and the 17th Biennial European Conference of the Society for Medical Decision Making in Leiden, The Netherlands, in 2018. We thank Dr Angela Kribs, MD, and consultants from the University Hospital of Cologne for validating the medical cases. The support of Professor Dr Ronald G. Schmid, MD, and Monika Kraushaar from the *Berufsverband der Kinder- und Jugendärzte, BVKJ e. V.* is gratefully acknowledged. We also thank Emanuel Castillo for programming the online survey and experiment, Uta Richter from the Department of Personnel Economics of the University of Cologne for facilitating the registration and payment of subjects, and Lena Kuhne for helping us to conduct the experiments.

References

1. Levy SB, Marshall B. Antibacterial resistance worldwide: causes, challenges and responses. Nat Med. 2004;10:S122–9.
2. Nathan C, Cars O. Antibiotic resistance – problems, progress, and prospects. N Engl J Med. 2014;371:1761–3.
3. Laxminarayan R, Duse A, Wattal C, Zaidi AK, Wertheim HF, Sumpradit N, et al. Antibiotic resistance—the need for global solutions. Lancet Infect Dis. 2013;13:1057–98.
4. Laxminarayan R, Amábile-Cuevas CF, Cars O, Evans T, Heymann DL, Hoffman S, et al. UN high-level meeting on antimicrobials-what do we need? Lancet. 2016;388(10041):218–20.
5. Haider BA, Lassi ZS, Bhutta ZA. Short-course versus long-course antibiotic therapy for non-severe community-acquired pneumonia in children aged 2 months to 59 months. Cochrane Database Syst Rev. 2008;2, Art. No.: CD005976.
6. Lassi ZS, Imdad A, Bhutta ZA. Short-course versus long-course intravenous therapy with the same antibiotic for severe community-acquired pneumonia in children aged two months to 59 months. Cochrane Database Syst Rev. 2017;10, Art. No.: CD008032.
7. Spellberg B. The new antibiotic mantra–"Shorter is better". JAMA Intern Med. 2016;176(9):1254–5.
8. Hersh AL, Shapiro DJ, Pavia AT, Shah SS. Antibiotic prescribing in ambulatory pediatrics in the united states. Pediatrics. 2011;128:1053–61.
9. Downes KJ, Hahn A, Wiles J, Courter JD, Vinks AA. Dose optimisation of antibiotics in children: application of pharmacokinetics/pharmacodynamics in paediatrics. Int J Antimicrob Agents. 2014;43:223–30.
10. Schulman J, Dimand RJ, Lee HC, Duenas GV, Bennett MV, Gould JB. Neonatal intensive care unit antibiotic use. Pediatrics. 2015;135:826–33.
11. Hyun DY, Hersh AL, Namtu K, Palazzi DL, Maples HD, Newland JG, et al. Antimicrobial stewardship in pediatrics: how every pediatrician can be a steward. JAMA Pediatr. 2013;167(9):859–66.

12. Vodicka TA, Thompson M, Lucas P, Heneghan C, Blair PS, Buckley DI, et al. Reducing antibiotic prescribing for children with respiratory tract infections in primary care: a systematic review. Br J Gen Pract. 2013;63(612):e445–e54.

13. Davey P, Marwick CA, Scott CL, Charani E, McNeil K, Brown E, et al. Interventions to improve antibiotic prescribing practices for hospital inpatients. Cochrane Database Syst Rev. 2017;2, Art. No.: CD003543(2).

14. Welschen I, Kuyvenhoven MM, Hoes AW, Verheij TJ. Effectiveness of a multiple intervention to reduce antibiotic prescribing for respiratory tract symptoms in primary care: randomised controlled trial. Br Med J. 2004;329:431–6.

15. Gjelstad S, Høye S, Straand J, Brekke M, Dalen I, Lindbæk M. Improving antibiotic prescribing in acute respiratory tract infections: cluster randomised trial from norwegian general practice (prescription peer academic detailing (Rx-PAD) study). Br Med J. 2013;347:f4403.

16. Gerber JS, Prasad PA, Fiks AG, Localio AR, Grundmeier RW, Bell LM, et al. Effect of an outpatient antimicrobial stewardship intervention on broad-spectrum antibiotic prescribing by primary care pediatricians: a randomized trial. J Am Med Assoc. 2013;309:2345–52.

17. Falk A, Heckman JJ. Lab experiments are a major source of knowledge in the social sciences. Science. 2009;326:535–8.

18. Herbst D, Mas A. Peer effects on worker output in the laboratory generalize to the field. Science. 2015;350(6260):545–9.

19. Ivers N, Jamtvedt G, Flottorp S, Young JM, Odgaard-Jensen J, French SD, et al. Audit and feedback: effects on professional practice and healthcare outcomes. Cochrane Database Syst Rev. 2012;6, Art. No.: CD000259.

20. Hallsworth M, Chadborn T, Sallis A, Sanders M, Berry D, Greaves F, et al. Provision of social norm feedback to high prescribers of antibiotics in general practice: a pragmatic national randomised controlled trial. Lancet. 2016;387:1743–52.

21. Meeker D, Linder JA, Fox CR, Friedberg MW, Persell SD, Goldstein NJ, et al. Effect of behavioral interventions on inappropriate antibiotic prescribing among primary care practices: a randomized clinical trial. J Am Med Assoc. 2016;315(6):562–70.

22. Kesternich I, Schumacher H, Winter J. Professional norms and physician behavior: homo oeconomicus or homo hippocraticus? J Public Econ. 2015;131:1–11.

23. Harrison GW, List JA. Field experiments. J Econ Lit. 2004;42:1009–55.

24. Galizzi MM, Wiesen D. Behavioural experiments in health: an introduction. Health Econ. 2017;26:3–5.

25. Galizzi MM, Wiesen D. Behavioral experiments in health economics. In: Hamilton J, editor. Oxford research encyclopedia of economics and finance. Oxford(UK): Oxford University Press; 2018. Available from: `http://economics.oxfordre.com/view/10.1093/acrefore/9780190625979.001.0001/acrefore-9780190625979-e-244`.

26. Callen M, Gulzur S, Hasanain A, Khan Y, Rezaee A. Personalities and public sector performance: evidence from a health experiment in pakistan. National Bureau of Economic Research. 2015;NBER Working Paper Series No. 21180.

27. Donato K, Miller G, Mohanan M, Truskinovsky Y, Vera-Hernández M. Personality traits and performance contracts: evidence from a field experiment among maternity care providers in india. Am Econ Rev. 2017;107(5):506–10.

28. Holtgrave DR, Lawler F, Spann SJ. Physicians' risk attitudes, laboratory usage, and referral decisions: the case of an academic family practice center. Med Decis Making. 1991;11(2):125–30.

29. Pearson SD, Goldman L, Orav EJ, Guadagnoli E, Garcia TB, Johnson PA, et al. Triage decisions for emergency department patients with chest pain. J Gen Intern Med. 1995;10(10):557–64.

30. Allison JJ, Kiefe CI, Cook EF, Gerrity MS, Orav EJ, Centor R. The association of physician attitudes about uncertainty and risk taking with resource use in a medicare HMO. Med Decis Making. 1998;18(3):320–9.

31. Michel-Lepage A, Ventelou B, Nebout A, Verger P, Pulcini C. Cross-sectional survey: risk-averse french GPs use more rapid-antigen diagnostic tests in tonsillitis in children. BMJ Open. 2013;3(10):e003540.

32. Franks P, Williams GC, Zwanziger J, Mooney C, Sorbero M. Why do physicians vary so widely in their referral rates? J Gen Intern Med. 2000;15(3):163–8.

33. Hulscher ME, Grol RP, van der Meer JW. Antibiotic prescribing in hospitals: a social and behavioural scientific approach. Lancet Infect Dis. 2010;10(3):167–75.

34. Akkerman AE, Kuyvenhoven MM, Van der Wouden JC, Verheij TJ. Prescribing antibiotics for respiratory tract infections by GPs: management and prescriber characteristics. Br J Gen Pract. 2005;55(511):114–8.

35. Rodrigues AT, Roque F, Falcão A, Figueiras A, Herdeiro MT. Understanding physician antibiotic prescribing behaviour: a systematic review of qualitative studies. Int J Antimicrob Agents. 2013;41(3):203–12.

36. Gosling SD, Rentfrow PJ, Swann Jr WB. A very brief measure of the big-five personality domains. J Res Pers. 2003;37:504–28.

37. Rammstedt B, John OP. Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. J Res Pers. 2007;41:203–12.

38. Falk A, Becker A, Dohmen T, Enke B, Huffman DB, Sunde U. Global evidence on economic preferences. Q J Econ. 2018;1:1–48.

39. Falk A, Becker A, Dohmen T, Huffman DB, Sunde U. The preference survey module: a validated instrument for measuring risk, time, and social preferences. Institute for the Study of Labor (IZA). 2016;IZA Discussion Papers No.9674.

40. Dohmen T, Falk A, Huffman D, Sunde U, Schupp J, Wagner GG. Individual risk attitudes: measurement, determinants, and behavioral consequences. J Eur Econ Assoc. 2011;9(3):522–50.

41. Olofsson SK, Cars O. Optimizing drug exposure to minimize selection of antibiotic resistance. Clin Infect Dis. 2007;45(Supplement_2):S129–36.

42. Linder JA. Moving the mean with feedback: insights from behavioural science. NPJ Prim Care Respir Med. 2016;26:16018.

43. Bénabou R, Tirole J. Incentives and prosocial behavior. Am Econ Rev. 2006;96(5):1652–1678.

44. Bénabou R, Tirole J. Identity, morals, and taboos: beliefs as assets. Q J Econ. 2011;126(2):805–855.

45. Falk A. Facing yourself – a note on self-image. Center for Economic Studies and Ifo Institute (CESifo). 2017;CESifo Working Paper Series No.6428.

46. Fischbacher U. Z-tree: zurich toolboox for readymade economic experiments – experimenter's manual. Exp Econ. 2007;10:171–8.

47. Leary MR, Tangney JP. The self as an organizing construct in the behavioral and social sciences. In: Leary MR, Tangney JP, editors. Handbook of self and identity. 2nd ed. New York(USA): The Guilford Press; 2012. p. 1–18.

48. Meeker D, Knight TK, Friedberg MW, Linder JA, Goldstein NJ, Fox CR, et al. Nudging guideline-concordant antibiotic prescribing: a randomized clinical trial. JAMA Intern Med. 2014;174(3):425–31.

49. Sirota M, Round T, Samaranayaka S, Kostopoulou O. Expectations for antibiotics increase their prescribing: causal evidence about localized impact. Health Psychol. 2017;36(4):402.

50. Broniatowski DA, Klein EY, May L, Martinez EM, Ware C, Reyna VF. Patients and clinicians perceptions of antibiotic prescribing for upper respiratory infections in the acute care setting. Med Decis Making. 2018;38(5):547–61.

51. Kostopoulou O, Russo JE, Keenan G, Delaney BC, Douiri A. Information distortion in physicians diagnostic judgments. Med Decis Making. 2012;32(6):831–9.

52. Nurek M, Kostopoulou O, Hagmayer Y. Predecisional information distortion in physicians diagnostic judgments: strengthening a leading hypothesis or weakening its competitor? Judgm Decis Mak. 2014;9(6):572–85.

53. Deutsche Gesellschaft für Pädiatrische Infektiologie (DGPI). DGPI-Handbuch Infektionen bei Kindern und Jugendlichen. 6th ed. Stuttgart(Germany): Thieme; 2013.

54. Labenne M, Michaut F, Gouyon B, Ferdynus C, Gouyon JB. A population-based observational study of restrictive guidelines for antibiotic therapy in early-onset neonatal infections. Pediatr Infect Dis J. 2007;26(7):593–9.

55. Grol R, Dalhuijsen J, Thomas S, in't Veld C, Rutten G, Mokkink H. Attributes of clinical guidelines that influence use of guidelines in general practice: observational study. Br Med J. 1998;317(7162):858–61.

56. Grilli R, Lomas J. Evaluating the message: the relationship between compliance rate and the subject of a practice guideline. Med Care. 1994;32(3):202–13.

57. American Academy of Pediatrics (AAP). Chlamydia trachomatis. In: Kimberlin DW, Brady MT, Jackson MA, Long SS, editors. Red Book – 2015 Report of the Committee on Infectious Diseases. Elk. Grove Village(IL): Am Acad Pediatrics; 2015. p. 288–94.

58. Shah SS. Mycoplasma pneumoniae. In: Long SS, Pickering LK, Prober CG, editors. Principles and practice of pediatric infectious diseases. 4th ed. Edinburgh(UK): Elsevier Saunders; 2012. p. 993–7.

A Medical cases

Table A.1: Medical cases (with categories of pediatric infectious diseases)

	Case description	Randomized order
Neonatal infections		
1	Newborn at 38 weeks of gestation at the age of four hours after a normal standardized pediatric examination. Spontaneous vaginal delivery, rupture of membranes at birth, maternal fever 38.5°C at birth, C-reactive protein (CRP) $<$ 5 mg/dl (mother), Group B-Streptococal (GBS) status is negative. The child's vital signs and clinical examination are normal.	4
2	Term newborn at the age of six hours after normal postnatal examination. Spontaneous vaginal delivery, rupture of membranes $<$ 18 hours before the onset of labor, positive maternal GBS status two weeks before birth. No antenatal antibiotic treatment. The child's vital signs and clinical examination are normal.	39
3	Term newborn at the age of 12 hours after a normal physical examination. Spontaneous vaginal delivery, rupture of membranes $>$ 18 hours before the onset of labor, positive GBS status two weeks before birth. Maternal antibiotic treatment three hours before birth. The child's vital signs and clinical examination are normal. In the blood test maximal CRP (C-reactive protein) 18 mg/l and Il-6 (Interleukin 6) 10 ng/l.	20
4	Term newborn on the second day of life. Spontaneous vaginal delivery, rupture of membranes at birth, normal postnatal physical examination. In the clinical examination, the child was hypotonic with gray skin colour, impaired microcirculation, tachypnea, and dyspnea. In the blood tests initiated by you, a CRP shows a maximum of 35 mg/l, Il-6 $>$ 8 ng/l. The blood cultures and newborn smears, received after two days, were without pathogen detection.	23
5	Newborn of the 38th gestational week, at the age of two days. Admission to the NICU and start of an antibiotic therapy after an abnormal physical examination. In blood test, maximal CRP 15 mg/l, Il-6 $<$ 8 ng/l. The CSF findings were normal. In the blood culture, detection of Staphylococcus epidermidis. The child's vital signs and physical examination are later normal.	19
6	Newborn with a gestational age of 39 weeks at the age of 20 hours. In the physical examination the child is hypotonic with impaired microcirculation and hypothermia. In blood test, CRP $>$ 75 mg/l, Il-6 150 ng/l. The CSF findings are negative. In the blood culture detection of Staphylococcus epidermidis.	24
7	Newborn at 41 weeks of gestation, at the age of five days. In the clinical examination, the infant shows hyperexcitability and a gray skin color, tachypnea, dyspnea and fever (max. 39°C). In the laboratory analyses initiated by you, the CRP is 9 mg/dl, and the interleukin 6 (Il-6) is 1,450 ng/l. In cerebrospinal fluid (CSF), there were 80 leucocytes/μl. The culture of the CSF remained negative. In the blood culture, E. coli was detected.	5
8	Preterm infant with spontaneous vaginal delivery after 32 weeks of pregnancy. Prenatal maternal antibiotic prophylaxis and a history of rupture in the 29th week of gestation. Mother GBS status negative. Initially slight respiratory distress syndrome. The patient is stabilized by nasal continuous positive airway pressure (nCPAP) quickly, CRP $<$ 5 mg/dl, Il-6 $<$ 8 ng/l. The respiratory support could be terminated at the second day of life.	13
9	Preterm infant after spontaneous vaginal delivery in the 33rd week of pregnancy. Rupture of membranes at birth, positive maternal GBS status, antenatal IV antibiotic treatment three hours before birth. Initial slight respiratory distress syndrome. The patient rapidly stabilizes under nCPAP. The ventilatory support can be terminated at the second day of life. Initiation of the antibiotic therapy in the delivery room. In the blood test initiated by you on the second day of life, CRP 15 mg/l and Il-6 $<$ 8 ng/l.	31
10	Twin preterm infant at the 32nd week of gestation. Spontaneous vaginal birth. The GBS-positive mother received an intravenous antibiotic treatment six hours before birth. Initial respiratory distress syndrome (III°). Surfactant application and further respiratory support with nCPAP in the first hours of life. Initiation of an antibiotic treatment in the labour ward. CRP 30mg/l, Il-6 120 ng/l. Blood cultures and neonatal smears were negative.	15

11	Premature infant with a gestational age of 28 weeks. Caesarean section due to maternal HELLP syndrome. Initial slight respiratory distress syndrome. Rapid stabilization of the respiratory state under nCPAP. Implantation of a silastic catheter. On the fifth day of life deterioration of general condition, gray patchy skin color, capillary refill prolonged and increasing oxygen demand. Removal of the catheter. Improvement of the clinical condition after application of an antibiotic therapy. In blood maximal CRP 35 mg/l and Il-6 148 ng/l. The blood cultures and newborn smears were negative.	32
12	Premature infant at the 25th week of gestation after three cycles of antibiotic therapy because of systemic inflammatory response syndrome (SIRS) and catheter sepsis. O_2-supply via nasal prongs, oral nutrition. In the age of eight weeks poor feeding, vomiting and abdominal distension. With suspected septicaemia or necrotising enterocolitis initiation of an antibiotic treatment. The CRP value was 35 mg/l, Il-6 was 6,480 ng/l. The blood cultures were negative. In the neonatal smears, detection of Staphylococcus epidermidis, Enterobacter species and Candida albicans. Immediate improvement of the clinical condition after the initiation of the therapy.	17
Infections of the CNS		
13	Six-year-old boy with sudden fever between 39°C and 40°C. His temperature cannot be reduced with physical and pharmacological measures. Severe headaches, neck pain, and vomiting. Admission with suspected meningitis and implementation of an antibiotic treatment. In CSF: turbit appearance, leukocyte count$> 1,000/\mu$l. CSF culture: negative.	27
14	Eight-year-old girl with severe headache and neck pain. High fever up to 40°C since the previous day. By suspected meningitis admission in your clinic and initiation of an antibiotic therapy. CSF results: turbit, cell count $> 1,000/\mu$l. In the rapid test and in the CSF culture, detection of meningococcus.	7
15	Ten-year-old boy with infection of the respiratory tract for one week. Fever up to 39°C, headache, and photophobia since the previous night. Admission to the hospital with suspected meningitis. CSF findings: cell count $> 1,000/\mu$l, protein 500 mg/l, lactate 4.5 mmol/l. Pneumococcus species was detected in the blood culture.	6
16	Two-year-old former premature infant with ventriculoperitoneal shunt. High fever up to 40°C, drowsiness, and vomiting since the previous day. CSF after puncture of the shunt valve: cell count $> 1,000/\mu$l. In CSF, detection of Staphylococcus. The ventriculo-peritoneal shunt was explanted shortly after admission.	38
Bone and joint infections		
17	12-year-old boy with pain in his left foot since the previous day. Pain when standing, redness and swelling and effusion in the area of the ankle. Trauma history neagtive and no visible external injury. Hospital admission for puncture and antibiotic therapy. In the puncture, detection of Staphylococcus aureus. Significant improvement of the clinical symptoms and normalization of the inflammation parameters within the first week of antibiotic treatment.	10
Upper respiratory tract infections		
18	Three-year-old child with acute ear pain, infection of the upper respiratory tract, serous rhinitis, and a maximal body temperature of 38.5°C. Otoscopy: redness and withdrawal of the tympanic membrane.	1
19	Eight-month-old infant in poor general condition. Apparent ear pain until the day before. Infection of the upper respiratory tract with purulent rhinitis and temperature up to max. 40°C. Otoscopic findings: purulent otorrhea with perforated eardrum.	12
20	Seven-year-old child with ear pain, infection of the upper respiratory tract, serous rhinitis, and fever up to max. 40°C for three days. Otoscopic findings: redness of the eardrum.	26
21	Ten-year-old girl in good general condition with serous rhinitis and coughing for one week. Frontal headache when tilting the head since the previous day.	3
22	12-year-old girl in good general condition with serous rhinitis and cough for two weeks. Severe facial pain for five days. Fever > 39°C during the clinical examination.	16
23	Eight-year-old boy with purulent rhinitis and cough for one week. Fever> 39°C and strong frontal headache for two days.	35
24	Eight-year-old boy with fever up to 39.8°C, fine maculate, slightly elevated, pale red rash, glossitis, and erythematous tonsils. Positive streptococcal rapid test.	36
25	Five-year-old girl with difficulty in swallowing, red tonsils, and swelling of the cervical lymph nodes without fever. Positive streptococcus A rapid test.	2

26	Ten-year-old girl with rapidly rising fever, pain and malaise. The tonsils are swollen and red, and there is a cervical lymph node swelling. Streptococcal A rapid test positive.	18
Urinary tract infections		
27	Detection of bacterial species $> 10^5$/ml in the investigation of the midstream urine of a 13-year-old female adolescent. The routine clinical examination was unremarkable.	30
28	15-year-old girl with dysuria, pollakisuria, and temperature up to 38.5°C. In the urinary analysis, leucocyturia and bacteriuria.	9
29	16-year-old girl with frequent, imperative urinary urgency and hematuria for two days. On the day of examination strong malaise, fever up to 40.5°C and flank pain. In the urine analysis, 3,000 leukocytes/μl, massive bacteriuria, and 300 eumorphe erythrocytes/μl. In the blood, 19,000 leukocytes/μl and CRP 120 mg/l. In the ultrasound examination, the kidneys were normal and there was no urinary obstruction.	21
30	15-year-old girl with dysuria for the first time, pollakisuria, flank pain, and fever up to 40°C. On the urine strip test (midstream urine) leukocytes ++, nitrite ++. In blood test leukocytosis and a CRP value of 100 mg/l. In the ultrasound examination, the left kidney was enlarged and partly echogenic, no urinary obstruction.	11
31	Four-month-old male infant with fatigue and fever up to 40.5°C. The CSF findings were normal. In the urine probe after catheterisation: 500 leukocytes $/\mu l$. In blood test leukocytes 24,000/μl and CRP 80 mg /l. The renal ultrasound examination revealed a suspected reflux.	22
32	Five-month-old male infant with fever up to 40°C. Poor general condition without a clear infectious focus. In the blood test: 16,400 leukocytes $/\mu l$, CRP 95 mg/l. Urine test strip after bladder puncture: leukocytes + + +, erythrocytes ++, nitrite +, proteins +. Urine culture: Detection of E. coli 10^6/ml.	8
Lower respiratory tract infections		
33	A six-week-old infant has been suffering from rhinitis for three days, fever up to 38°C and increasingly dry cough. The child is pale, with nasal flaring, tachypnea, dyspnea, subcostal chest retractions. Bilateral attenuated respiratory sound, fine crackles, and expiratory wheezing. The rapid test for respiratory syncytial virus (RSV) is positive. In the blood test, 5,300 leukocytes/μl and CRP 20 mg/l.	34
34	A nine-month-old female infant has been suffering from fever up to 38° C for one week, rhinitis and dry cough. Symptomatic therapy with suspected viral infection. Since the previous day deterioration of the general condition and fever up to 40°C. Bilateral attenuated breath sounds and occasional fine crackles and expiratory wheezing in the auscultation. The RSV rapid test is positive. In blood, leukocytes 15,000/μl, CRP 70 mg/l.	40
35	Three-month-old infant with tachypnea and cough resembling whooping cough. Postnatal purulent conjunctivitis. Chlamydia trachomatis pneumonia is suspected.	14
36	Five-year-old boy with fever up to 39.5°C and abdominal pain. Auscultation: inspiratory fine crackles and attenuated breath sounds. Laboratory findings: leukocyte $27,800/\mu l$, CRP 38 mg/dl. The x-ray reveals a lobar pneumonia.	28
37	Seven-year-old girl with severe abdominal pain and fever up to 39°C. In the physical examination: basal attenuated breath sounds in the auscultation and basal damping in the percussion; the abdomen is normal. In blood, 13,500 leukocytes/μl and CRP 77 mg/l. The chest x-ray revealed pneumonia.	37
38	Six-year-old girl with severe cough, purulent rhinitis, and fever up to 40°C. In the auscultation fine inspiratory crackles and expiratory wheezing. Laboratory findings: 17,500 leukocytes/μl, CRP 100 mg /l. Bronchopneumonia in the chest x-ray.	25
39	Six-year-old boy with fever up to 40°C and abdominal pain for ten days. The chest x-ray shows pneumonia with basal pleural effusion. After a seven-day-long antibiotic treatment duration, relapse of fever, and occurrence of increasing dyspnea. In the chest x-ray, an abscedating pneumonia is suspected. Surgical application of an abscess drainage.	29
40	Ten-year-old girl with intermittent fever up to 40°C, cough and rhinitis. A therapy with cefuroxime has not lead to an improvement. The chest x-ray reveals central infiltrates with the involvement and compression of the hili. An atypical pneumonia is suspected.	33

Notes. This table shows the 40 medical cases used in the expert survey and in the experiment. It also shows the six categories of infectious diseases to which the cases can be assigned. The last column provides the randomized order of the cases used in the survey and in the experiment.

B Instructions for the experiment

[*Note that in squared brackets we present instructions from the second and third parts of the experiment.*]

You are taking part in a decision experiment. Please read through the instructions carefully. It is important that you do not talk to other participants for the entire duration of the experiment. If you have any questions, please raise your hand. We will come to your cubicle and answer your questions in person.

In this experiment, all monetary amounts are denoted in 'Taler', at a rate of 1 Taler = €1. Your earnings will be paid in cash at the end of the experiment.

You will make your decisions anonymously in your cubicle. All data will be evaluated anonymously. You have drawn your own cubicle number in order to ensure anonymity.

The experiment will last for approximately 60 minutes and consists of three parts. You will receive detailed instructions prior to each stage of the experiment. Please note: Your decisions in each part of the experiment will not have an impact on any other part of the experiment.

At the end of the experiment, you will receive compensation for the experiment.

We also ask you to answer a few questions at the end of the experiment.

First [Second, Third] part of the experiment

Decision situation

The first part of the experiment relates to a decision situation in the pediatric department of a hospital. You make your decision in the role of the on-duty pediatrician.

In the course of the first part of the experiment, you will be presented with a series of patients, each with different pathologies, symptoms, complaints, or results. If symptoms, complaints, or results are not provided, then they are not considered to be relevant for your decision-making.

In creating an initial treatment plan, you have the task of determining the duration of a course of antibiotics (in days). Here, you can set the length of the course at $0, 1, 2, \ldots, 27$, or 28 day(s). Note that the respective medicines will be administered according to the relevant guidelines. The initial treatment plan can be adjusted through a reevaluation.

Enter the length of the antibiotics course for each patient in the field 'How many days do you prescribe antibiotic therapy for?' on your computer screen. You can enter whole numbers between zero and 28. You confirm your decision by clicking ‘OK’, which will take you to the next screen.

[**After you have made your decision about the length of antibiotic therapy for all patients, you will be informed about an expert opinion on the average length of antibiotic therapy for patients identical to those for whom you have made treatment decisions. The expert opinion is based on responses from 20 leading pediatricians drawn from a representative sample of children’s hospitals in Germany.***]

Earnings
For carrying out the task in the first part of the experiment – determining the length of antibiotic therapy for a series of patients – you will receive a fixed payment of 50 Taler.

Important information:

- Make your decisions anonymously on your computer screen.
- In order that no decision or payout can be matched with a particular participant, an employee of the Department of Business Administration and Personnel Economics at the University of Cologne, who is not involved in conducting the experiment, will place an envelope in your cubicle that is marked only with the cubicle number and contains the total payout for your cubicle.
- Afterwards, please leave the room in which the experiment was conducted.

---

[*] This survey was conducted in August and September 2014 among head physicians in German children’s hospitals. Out of a total of 50 randomly chosen German children’s hospitals, 20 hospitals answered questions about the length of antibiotic therapy in full. The study is archived in the German Clinical Trials Register under the study number DRKS00006782.

C Survey with directors of German pediatric departments

C.1 Descriptive statistics

In total, 20 directors of 50 randomly selected pediatric departments participated in our online survey. The expert sample comprised 19 male pediatricians and one female pediatrician, who were aged between 40 and 62 years. The aggregated expert opinion, i.e., the average length of antibiotic therapy the experts chose for the 40 cases, was 6.42 days (SD 4.94, 95% CI 4.26 to 8.59). This aggregated value served as the 'expert benchmark' in our experiment. See Table C.2 for detailed results of the expert survey.

C.2 Comparison with guidelines

We compared the experts' decisions with published recommendations on the length of antibiotic therapy for each respective case. For this comparison, we only considered recommendations on the length of *first-line therapy with the standard antibiotic agent* to assess the experts' compliance with recommendations, because the participants in our study were asked to decide on the length of first-line antibiotic treatment with the standard antibiotic agent. We primarily used the recommendations published by the German Society for Pediatric Infectious Diseases.[53] The handbook published by this society provides (based on the use of explicitly stated standard antibiotic agents) a recommendation on length of therapy for each case we used in our study. Moreover, it reflects consensus of several leading German pediatricians, which lets us assume that it also reflects local standards of care in pediatric medicine.[†]

Using Fisher-Pitman permutation tests for paired replicates, we analyzed whether the decisions made in the expert survey were significantly different from the recommendations. For each case, we compared the 20 decisions of the experts with the range of recommended numbers of treatment days. We considered decisions as compliant with the recommendations if they were within the range of recommended numbers of treatment days or deviated one day at most (i.e., the recommended intervals were extended by +/- one day). In doing so, we adopted the measure of compliance with recommendations on length of antibiotic therapy that has been applied by other scholars.[54] The interval was not extended by one day, however, if no antibiotic therapy (zero days) or an explicit max-

---

[†] Note that the recommendations are very similar to recommendations from national and international guidelines. The handbook is an aggregate of available evidence, which should also be included in those guidelines.

imum or minimum number of days is recommended (e.g., for the recommendation ‘from one day up to a maximum of two days’, we accepted one or two days as compliant with the recommendation. For ‘at least 10 days up to 14 days’, the range between 10 and 15 days was considered appropriate. Note that we did not extend the interval to zero days if the lower boundary is one day). For cases where the handbook provides no upper or lower boundary (e.g., ‘at least 10 days’), we used recommendations from further national and international guidelines as references (see Table C.2 for details).

The experts’ decisions were in 80 percent of the cases in line with the recommendations, i.e., only in eight out of the 40 cases (20%) did the decisions significantly differ from what guidelines recommend (with a p-value < 0.05). Comparable studies reporting compliance rates with antibiotic prescribing guidelines are rare. Labenne et al.,[54] which is to the best of our knowledge the only study that examines guideline compliance regarding the length of antibiotic therapy for children, reported a compliance rate of 70 percent. Other studies, which lack comparability since they do not consider length of antibiotic therapy, found low average medical guideline compliance rates among physicians of 61 percent[55] or 54.5 percent.[56] Given the experts’ large guideline compliance rate in our survey, we argue the aggregated expert opinion can be considered a suitable benchmark for appropriate length of antibiotic therapy.

Table C.2: Results of the expert survey (n=20) and recommendations on length of therapies

Cases (odered by category)	The experts' decisions (n=20)					Recommended length of therapy (in days)
	min	max	mean	median	s.d.	
Neonatal infections						
1	0	5	0.65	0	1.63	0
2	0	3	0.15	0	0.67	0
3	0	10	2.40	0	3.12	1-2 (max.)
4	0	10	5.50	5	2.42	5 (-7)
5	0	7	2.80	3	2.46	1-2 (max.)
6	3	10	6.85	7	2.16	7 (-10)
7	7	21	11.65	10	4.69	21
8	0	3	1.20	0	1.40	0
9	0	10	3.80	3	3.00	1-2 (max.)
10	2	10	5.25	5	2.12	5 (-7)
11	2	10	6.30	6	2.03	5 (-7)
12	5	14	7.35	7	2.23	5 (-7)
Infections of the CNS						
13	3	21	9.70	10	3.66	7-10
14	5	21	8.35	7	3.92	4-7
15	7	21	10.50	10	3.47	7-10
16	5	21	11.85	14	3.73	at least 10-14
Bone and joint infections						
17	7	28	17.45	17.5	6.58	21
Upper respiratory tract infections						
18	0	7	1.60	0	2.62	0
19	5	14	7.20	7	2.28	10
20	0	7	1.35	0	2.43	5-7
21	0	10	1.25	0	2.75	0
22	0	14	5.65	6	3.63	10 (-14)
23	0	14	5.80	7	4.09	10 (-14)
24	3	10	7.85	7	2.23	10
25	0	10	5.00	6	3.83	5
26	0	10	6.95	7	2.67	10
Urinary tract infections						
27	0	5	0.35	0	1.14	0
28	0	7	3.50	3	2.01	3 (-5)
29	5	14	7.65	7	2.16	(7-) 10
30	7	14	8.40	7	1.96	(7-) 10
31	5	14	8.85	10	2.50	10-14
32	7	14	8.60	7	2.30	10 (-14)
Lower respiratory tract infections						
33	0	7	0.35	0	1.57	0
34	0	10	5.85	7	2.85	7
35	0	21	12.00	14	4.81	at least 10 (10-14)
36	0	14	8.00	7	2.88	7
37	5	14	8.10	7	2.49	7
38	5	14	7.55	7	2.24	7
39	5	21	13.45	14	4.77	at least 21 (21-28)
40	3	14	9.90	10	2.86	10

Notes. This table shows the experts' decisions on the length of antibiotic treatment for each case, as well as the recommendations published by the German Society for Paediatric Infectious Diseases.[53] The experts' decisions on the length of therapy, aggregated over all cases, were used as the 'expert benchmark' in our experiment. We assessed the experts' compliance with the recommendations by comparing the experts' decisions with the recommended length of therapy for each case. We allowed a deviation of one day from the recommended number of days (+/- 1 day). We did not allow a deviation if the recommendation is exactly zero days or if an explicit upper or lower boundary is recommended ('at least' or 'max.'). Note that for cases 35 and 39, the recommendations of the German Society for Paediatric Infectious Diseases provide no upper boundary. They recommend *at least 10 days* for case 35 and *at least 21 days* for case 39. To get an upper boundary, we used recommendations from further guidelines. For case 35, we set the upper boundary to 14 days, since the American Academy of Pediatrics recommends 14 days of antibiotic therapy.[57] For case 39, the upper boundary was set to 28 days because the Centers for Disease Control, the American Academy of Pediatrics, and the Pediatric Infectious Diseases Society recommend a length of therapy between 18 and 28 days.[58]

D Some photographs from the experiments

Figure D.1: Some impressions from the experimental sessions

Notes. This figure shows the cubicles of the mobile computer laboratory. The left picture shows cubicles of the mobile laboratory at the Children's and Adolescent's Hospital at the University Hospital of Cologne. The middle picture shows parts of the laboratory at the Children's Hospital of the City of Cologne. The right picture indicates the laboratory during the annual conference for padiatricians in Cologne (*Päd-Ass 2015*).

E Post-experimental questionnaire

I. Socio-demographics

Your age: _____ years

Your gender: ☐ Male ☐ Female

What is your medical specialty? ____________________

Since when are you a consultant (specialist physician)? ____________________

When did you start practicing in the hospital? ____________________

II. Social and risk preferences

(‘Economic preferences’, according to Falk et al.[38,39] and Dohmen et al.[40])

1. How do you see yourself – Are you a person who is generally willing to take risks, or do you try to avoid taking risks? *Please indicate your answer on a scale from 0 to 10, where a 0 means “not at all willing to take risks”, and a 10 means “very willing to take risks”. You can also use the values in-between to indicate where you fall on the scale.*

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
0	1	2	3	4	5	6	7	8	9	10

2. Please imagine that you have won a prize in a contest. Now you can choose between two dif-ferent payment methods, either a lottery or a sure payment. If you choose the lottery there is a 50 percent chance that you would receive €1,000, and an equally high chance that you would receive nothing.
   What is the smallest sure payment that would make you prefer the sure payment over playing the lottery? Amount €__________

3. How do you see yourself – Are you a person who is generally willing to give up something to-day in order to benefit from that in the future, or are you not willing to do so? *Please use a scale from 0 to 10, where 0 means you are “completely unwilling to give up something today” and a 10 means you are “very willing to give up something today”. You can also use the values in-between to indicate where you fall on the scale.*

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
0	1	2	3	4	5	6	7	8	9	10

4. How well does the following statement describe you as a person? I tend to postpone things even though it would be better to get them done right away. *Please use a scale from 0 to 10, where 0 means "does not describe me at all" and a 10 means "describes me perfectly". You can also use the values in-between to indicate where you fall on the scale.*

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
0	1	2	3	4	5	6	7	8	9	10

5. How would you assess your willingness to trust strangers? *Please indicate your answer on a scale from 0 to 10, where a 0 means "not at all willing to trust strangers", and a 10 means "very willing to trust strangers". You can also use the values in-between to indicate where you fall on the scale.*

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
0	1	2	3	4	5	6	7	8	9	10

6. How well does the following statement describe you as a person? As long as I am not convinced otherwise, I assume that people have only the best intentions. *Please use a scale from 0 to 10, where 0 means "does not describe me at all" and a 10 means "describes me perfectly". You can also use the values in-between to indicate where you fall on the scale.*

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
0	1	2	3	4	5	6	7	8	9	10

7. How do you assess your willingness to share with others without expecting anything in return when it comes to charity? *Please use a scale from 0 to 10, where 0 means you are "completely unwilling to share" and a 10 means you are "very willing to share". You can also use the values in-between to indicate where you fall on the scale.*

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
0	1	2	3	4	5	6	7	8	9	10

8. Imagine the following situation: You won €1,000 in a lottery. Considering your current situation, how much would you donate to charity? *(Values between 0 and 1000 are allowed):*__________

9. How well does the following statement describe you as a person? When someone does me a favor I am willing to return it. *Please use a scale from 0 to 10, where 0 means “does not describe me at all” and a 10 means “describes me perfectly”. You can also use the values in-between to indicate where you fall on the scale.*

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
0	1	2	3	4	5	6	7	8	9	10

10. How do you assess your willingness to return a favor to a stranger? *Please use a scale from 0 to 10, where 0 means you are “not willing to return a favor to a stranger” and a 10 means you are “very willin to return a favor to a stranger”. You can also use the values in-between to indicate where you fall on the scale.*

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
0	1	2	3	4	5	6	7	8	9	10

11. How do you see yourself – Are you a person who is generally willing to punish unfair behavior even if this is costly? *Please use a scale from 0 to 10, where 0 means you are “not willing at all to incur costs to punish unfair behavior” and a 10 means you are “very willing to incur costs to punish unfair behavior. You can also use the values in-between to indicate where you fall on the scale.*

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
0	1	2	3	4	5	6	7	8	9	10

12. How well does the following statement describe you as a person? If someone treats me unjustly, I will try to take revenge at the first occasion. *Please use a scale from 0 to 10, where 0 means “does not describe me at all” and a 10 means “describes me perfectly”. You can also use the values in-between to indicate where you fall on the scale.*

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
0	1	2	3	4	5	6	7	8	9	10

III. Personality traits

(according to Gosling et al.[36] and Rammstedt and John[37])

In the following you can find a number of personality traits that more or less apply to you. Please mark for each statement how well it describes your personality.

	Disagree strongly 1	Disagree 2	Somewhat disagree 3	Neither agree nor disagree 4	Somewhat agree 5	Agree 6	Agree strongly 7
1. I see myself as someone who is reserved.	☐	☐	☐	☐	☐	☐	☐
2. I see myself as someone who is generally trusting.	☐	☐	☐	☐	☐	☐	☐
3. I see myself as someone who tends to be lazy.	☐	☐	☐	☐	☐	☐	☐
4. I see myself as someone who is relaxed, handles stress well.	☐	☐	☐	☐	☐	☐	☐
5. I see myself as someone who has few artistic interests.	☐	☐	☐	☐	☐	☐	☐
6. I see myself as someone who is outgoing, sociable.	☐	☐	☐	☐	☐	☐	☐
7. I see myself as someone who tends to find fault with others.	☐	☐	☐	☐	☐	☐	☐
8. I see myself as someone who does a thorough job.	☐	☐	☐	☐	☐	☐	☐
9. I see myself as someone who gets nervous easily.	☐	☐	☐	☐	☐	☐	☐
10. I see myself as someone who has an active imagination.	☐	☐	☐	☐	☐	☐	☐
11. I see myself as someone who is considerate and kind to almost everyone.	☐	☐	☐	☐	☐	☐	☐

F Robustness checks

Table F.3: The effect of feedback on the length of antibiotic therapies for each case

	The subjects' decisions on days of antibiotic therapy							
Cases (ordered by category)	Stage 2			Stage 3			Change in mean number of days	p-values
	mean	median	s.d.	mean	median	s.d.		
Neonatal infections								
1	1.46	0	2.35	1.00	0	1.75	-0.46	0.43
2	1.46	0	2.51	1.26	0	2.16	-0.21	0.32
3	4.10	5	3.57	3.62	5	2.88	-0.49	0.11
4	6.95	7	2.79	6.64	7	2.36	-0.31	0.31
5	5.90	5	3.37	6.13	5	3.61	0.23	0.22
6	9.64	10	3.78	8.72	7	3.00	-0.92	0.02
7	14.41	14	5.14	12.72	10	4.98	-1.69	0.00
8	3.05	3	3.68	2.82	3	2.83	-0.23	0.80
9	4.59	5	2.56	4.31	5	2.02	-0.28	0.68
10	7.62	7	3.70	6.10	7	2.23	-1.51	0.01
11	8.74	7	4.17	7.69	7	2.59	-1.05	0.13
12	10.62	10	4.83	9.67	7	4.24	-0.95	0.03
Infections of the CNS								
13	11.33	10	4.35	10.87	10	3.74	-0.46	0.77
14	15.18	14	4.07	14.23	14	3.77	-0.95	0.04
15	14.56	14	3.67	14.41	14	3.19	-0.15	0.95
16	13.85	14	4.69	13.64	14	4.31	-0.21	0.96
Bone and joint infections								
17	15.23	14	6.27	14.87	14	6.33	-0.36	0.21
Upper respiratory tract infections								
18	2.97	0	3.79	1.79	0	2.78	-1.18	0.00
19	7.44	7	3.22	6.69	7	1.91	-0.74	0.05
20	2.21	0	3.47	2.05	0	3.15	-0.15	0.65
21	2.05	0	3.53	2.26	0	3.53	0.21	0.19
22	5.21	7	3.25	4.69	5	3.33	-0.51	0.04
23	4.97	5	3.10	4.62	5	3.22	-0.36	0.16
24	8.49	7	3.14	7.82	7	2.21	-0.67	0.03
25	6.38	7	3.41	5.77	7	3.39	-0.62	0.09
26	7.97	7	3.75	7.49	7	2.63	-0.49	0.17
Urinary tract infections								
27	0.46	0	1.55	0.46	0	1.55	0.00	1.00
28	4.90	5	2.39	4.10	5	2.17	-0.79	0.02
29	8.46	7	3.48	7.77	7	2.76	-0.69	0.06
30	9.79	10	2.74	9.36	10	2.91	-0.44	0.11
31	12.03	10	6.37	10.90	10	6.00	-1.13	0.00
32	10.44	10	3.42	9.23	10	3.17	-1.21	0.02
Lower respiratory tract infections								
33	1.74	0	3.38	1.49	0	2.61	-0.26	0.98
34	5.38	7	3.03	5.54	7	2.97	0.15	0.97
35	11.54	10	3.95	10.82	10	3.58	-0.72	0.10
36	9.87	10	3.14	8.77	7	2.49	-1.10	0.01
37	8.46	7	2.01	8.00	7	1.95	-0.46	0.01
38	8.26	7	2.70	7.51	7	1.54	-0.74	0.02
39	14.90	14	5.54	14.03	14	5.18	-0.87	0.03
40	10.49	10	4.41	9.28	10	3.78	-1.21	0.01

Notes. This table shows the effect of feedback on length of antibiotic therapies on case level. It shows the average days subjects in the intervention group (n=39) chose prior to feedback (in stage 2) and after feedback was given (in stage 3). p-values are shown for 2-sided Wilcoxon matched-pairs signed-rank tests.

Table F.4: The effect of feedback on abs. deviation from the experts for each case

Cases (ordered by category)	The subjects' abs. deviation from the experts (in days) Stage 2 mean	 median	 s.d.	 Stage 3 mean	 median	 s.d.	Change in mean absolute deviation	p-values
Neonatal infections								
1	1.68	0.65	1.82	1.28	0.65	1.22	-0.39	0.42
2	1.51	0.15	2.39	1.31	0.15	2.04	-0.20	0.32
3	3.20	2.60	2.28	2.71	2.60	1.50	-0.49	0.11
4	2.17	1.50	2.26	1.83	1.50	1.85	-0.33	0.10
5	3.61	2.20	2.80	3.84	2.80	3.04	0.23	0.51
6	3.08	3.15	3.55	2.44	1.85	2.55	-0.64	0.30
7	4.56	2.35	3.59	4.09	2.35	2.97	-0.47	0.13
8	2.65	1.80	3.13	2.42	1.80	2.16	-0.23	0.27
9	2.05	1.20	1.70	1.75	1.20	1.09	-0.30	0.28
10	2.78	1.75	3.39	1.61	1.75	1.75	-1.17	0.06
11	3.11	1.30	3.69	1.93	0.70	2.21	-1.18	0.02
12	4.10	2.65	4.13	3.31	2.35	3.50	-0.79	0.13
Infections of the CNS								
13	3.34	4.30	3.19	3.02	2.70	2.46	-0.32	0.77
14	6.83	5.65	4.07	6.02	5.65	3.53	-0.81	0.04
15	4.27	3.50	3.42	4.06	3.50	2.99	-0.21	0.95
16	3.99	2.15	3.12	3.60	2.15	2.93	-0.39	0.90
Bone and joint infections								
17	5.84	3.55	3.06	6.02	3.55	3.12	0.18	0.91
Upper respiratory tract infections								
18	3.18	1.60	2.43	2.41	1.60	1.35	-0.77	0.00
19	1.83	0.20	2.64	1.23	0.20	1.54	-0.60	0.17
20	2.66	1.35	2.36	2.50	1.35	1.99	-0.15	0.65
21	2.60	1.25	2.49	2.67	1.25	2.48	0.08	0.65
22	2.47	1.35	2.12	2.56	1.35	2.29	0.10	0.20
23	2.35	1.20	2.15	2.53	1.20	2.29	0.17	0.94
24	2.20	2.15	2.31	1.76	0.85	1.30	-0.43	0.15
25	2.92	2.00	2.19	2.82	2.00	1.99	-0.10	0.98
26	1.82	1.95	3.43	1.79	1.95	1.98	-0.03	0.51
Urinary tract infections								
27	0.74	0.35	1.37	0.74	0.35	1.37	0.00	1.00
28	2.14	1.50	1.74	1.91	1.50	1.16	-0.23	0.25
29	2.50	2.35	2.52	2.03	0.65	1.84	-0.47	0.07
30	2.50	1.60	1.75	2.49	1.60	1.74	-0.02	0.33
31	4.52	1.85	5.47	3.96	1.85	4.92	-0.56	0.78
32	2.74	1.60	2.74	2.39	1.60	2.13	-0.34	0.47
Lower respiratory tract infections								
33	1.91	0.35	3.10	1.66	0.35	2.31	-0.26	0.98
34	2.28	1.15	2.02	2.22	1.15	1.96	-0.06	0.41
35	3.28	2.00	2.18	3.13	2.00	2.04	-0.15	0.96
36	2.69	2.00	2.45	2.00	1.00	1.64	-0.69	0.01
37	1.71	1.10	1.09	1.58	1.10	1.11	-0.12	0.01
38	1.57	0.55	2.30	1.14	0.55	1.01	-0.43	0.33
39	4.04	3.45	4.01	3.75	3.45	3.58	-0.29	0.89
40	2.99	2.90	3.26	2.74	2.90	2.65	-0.26	0.66

Notes. This table shows the effect of feedback on absolute deviation from the expert recommendations on case level. For each case, the pediatricians' choices were compared to the experts' aggregate opinion for the respective case. It shows absolute differences between the pediatricians' choices and the expert recommendations prior to feedback (in stage 2) and after feedback was given (in stage 3). Only the intervention group (n=39) is considered. p-values are shown for 2-sided Wilcoxon matched-pairs signed-rank tests.

Table F.5: Tobit regressions on the effect of feedback on antibiotic therapy decisions

Dependent variable	Length of antibiotic therapy (in days)				Abs. deviation from the expert recommendations (in days)			
Model	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Feedback (= 1 if intervention)	-0.179	-0.168	-0.435	-0.667	-0.263	-0.264	-0.423	-0.815**
	(0.851)	(0.883)	(0.816)	(0.773)	(0.478)	(0.479)	(0.375)	(0.369)
Second stage (= 1 if second stage)	0.014	0.025	0.025	0.022	-0.023	-0.022	-0.022	-0.022
	(0.293)	(0.297)	(0.297)	(0.297)	(0.177)	(0.177)	(0.177)	(0.177)
Third stage (= 1 if third stage)	0.605*	0.616*	0.615*	0.611*	0.304	0.304	0.304	0.305
	(0.325)	(0.333)	(0.333)	(0.331)	(0.191)	(0.191)	(0.191)	(0.191)
Announcement effect (Second stage × Feedback)	-0.081	-0.092	-0.092	-0.089	-0.065	-0.065	-0.065	-0.066
	(0.175)	(0.179)	(0.178)	(0.178)	(0.118)	(0.118)	(0.118)	(0.118)
Feedback effect (Third stage × Feedback)	-0.703***	-0.721***	-0.719***	-0.715***	-0.393***	-0.393***	-0.393***	-0.394***
	(0.229)	(0.234)	(0.234)	(0.232)	(0.150)	(0.150)	(0.150)	(0.150)
Pediatricians' characteristics	No	No	Yes	Yes	No	No	Yes	Yes
Economic preferences	No	No	No	Yes	No	No	No	Yes
Case dummies	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Session dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Constant	7.412***	-1.722*	0.557	-0.455	3.016***	1.711***	3.712***	3.571***
	(0.782)	(1.001)	(1.386)	(1.452)	(0.441)	(0.454)	(0.761)	(0.718)
Observations	8,760	8,760	8,760	8,760	8,760	8,760	8,760	8,760
Subjects	73	73	73	73	73	73	73	73
Pseudo R^2	0.002	0.157	0.161	0.171	0.003	0.040	0.046	0.055

Notes. This table shows parameter estimates from Tobit regressions. The interaction 'Third stage × Treatment' indicates the effect of showing feedback to subjects. In Models (1) to (4), the dependent variable is 'length of antibiotic therapy (in days)'. In Models (5) to (8), the dependent variable is 'absolute deviation from the expert recommendations', measured in absolute values of the difference between the pediatricians' choices and the experts' recommended length of therapies (in days). For each case, the subjects' choices were compared to the experts' aggregate opinion for the respective case. Robust standard errors, clustered at the individual-subject level, are shown in parentheses. The variable 'Pediatricians' characteristics' includes subjects' gender, years worked in hospital, and the Big Five personality traits. 'Economic preferences' comprise validated measures for trust, reciprocity, and altruism, as well as time and risk preferences.[38–40] The variable 'case dummies', which are included in models (2) to (4) and (6) to (8), indicates 40 dummies, one for each of the 40 medical cases. Furthermore, dummies for each experimental session were included in all models to control for any session effects. *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.10$.

Table F.6: OLS regressions on the effect of feedback on antibiotic therapy decisions (without subjects from the Childrens's Hospital of the City of Cologne)

Dependent variable	Length of antibiotic therapy (in days)				Abs. deviation from the expert recommendations (in days)			
Model	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Feedback (= 1 if intervention)	0.283	0.283	-0.161	-0.296	0.601	0.602	0.252	0.188
	(0.906)	(0.908)	(0.852)	(0.619)	(0.576)	(0.577)	(0.510)	(0.390)
Second stage (= 1 if second stage)	0.081	0.081	0.081	0.081	0.057	0.057	0.057	0.057
	(0.236)	(0.236)	(0.237)	(0.237)	(0.188)	(0.189)	(0.189)	(0.189)
Third stage (= 1 if third stage)	0.535*	0.535*	0.535*	0.535*	0.397*	0.397*	0.397*	0.397*
	(0.311)	(0.312)	(0.312)	(0.313)	(0.227)	(0.227)	(0.228)	(0.228)
Second stage × Feedback	-0.144	-0.144	-0.144	-0.144	-0.143	-0.143	-0.143	-0.143
	(0.147)	(0.147)	(0.147)	(0.147)	(0.132)	(0.133)	(0.133)	(0.133)
Third stage × Feedback	-0.647**	-0.647**	-0.647**	-0.647**	-0.486**	-0.486**	-0.486**	-0.486**
	(0.249)	(0.250)	(0.250)	(0.250)	(0.194)	(0.194)	(0.194)	(0.194)
Individual controls	No	No	Yes	Yes	No	No	Yes	Yes
Economic preferences	No	No	No	Yes	No	No	No	Yes
Case dummies	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Session dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Constant	7.860***	1.565**	4.800***	4.421***	3.027***	1.623***	4.006***	4.195***
	(0.708)	(0.709)	(1.577)	(1.218)	(0.440)	(0.457)	(0.840)	(0.774)
Observations	7,440	7,440	7,440	7,440	7,440	7,440	7,440	7,440
Subjects	62	62	62	62	62	62	62	62
R^2	0.016	0.572	0.587	0.614	0.015	0.190	0.219	0.258

Notes. This table shows parameter estimates from OLS regressions. The interaction 'Third stage × Treatment' indicates the effect of showing feedback to subjects. In Models (1) to (4), the dependent variable is 'length of antibiotic therapy (in days)'. In Models (5) to (8), the dependent variable is 'absolute deviation from the expert recommendations', measured in absolute values of the difference between the pediatricians' choices and the experts' recommended length of therapies (in days). For each case, the subjects' choices were compared to the experts' aggregate opinion for the respective case. Robust standard errors, clustered at the individual-subject level, are shown in parentheses. The variable 'Pediatricians' characteristics' includes subjects' gender, years worked in hospital, and the Big Five personality traits. 'Economic preferences' comprise validated measures for trust, reciprocity, and altruism, as well as time and risk preferences.[38–40] The variable 'case dummies', which are included in models (2) to (4) and (6) to (8), indicates 40 dummies, one for each of the 40 medical cases. Furthermore, dummies for each experimental session were included in all models to control for any session effects. *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.10$.