AI adoption and decision-making: Evidence from primary care*

Shan Huang[1], Renke Schmacker[2,3], and Hannes Ullrich[2,4]

[1] *Stockholm School of Economics,* [2] *DIW Berlin,* [3] *WZB Berlin Social Science Center,* [4] *University of Copenhagen*

February 21, 2026

PRELIMINARY DRAFT

Abstract

Artificial intelligence (AI) support tools hold great potential to leverage patient records more effectively in diagnosis and treatment. Yet little is known about how physicians use AI-generated signals relative to the diagnostic tools they currently rely on. We study primary care diagnosis and treatment of urinary tract infections (UTIs), a leading cause for antibiotic use, where physicians often prescribe prior to attaining diagnostic certainty. We run a nationwide survey experiment with 372 Danish primary care physicians using patient vignettes. Physicians make diagnostic and prescribing decisions before and after receiving diagnostic signals. We hold signal accuracy constant while varying between-subjects whether the same information is presented as an AI prediction output or as urine dipstick result. In a second stage, physicians see both signals and we randomize whether the signals are independent or correlated, allowing for a test of correlation neglect. Physicians update less in response to AI than to dipstick signals, consistent with AI skepticism and a heuristic use of the dipstick signal. When presented with multiple signals, physicians insufficiently adjust for correlation and overweight correlated signals. Physicians can be separated into one group that adopts the AI tool to varying degrees and one that does not adopt the AI tool at all. Non-adopters' mental models indicate concerns about non-informative data and suboptimal information processing by AI. Linked administrative data further show that non-adopters are more likely to practice in rural clinics with lower use of other diagnostic technologies.

*We thank colleagues and seminar participants at Max-Planck-Institute for Human Development in Berlin, WZB Berlin, and the CESifo Summer Institute in Venice for their insightful comments and feedback. We are indebted to Rune Munck Aabenhus and Lars Bjerrum for providing expertise on UTI diagnosis and treatment in primary care. We are grateful to Nilab Ahmady for excellent research assistance. We thankfully acknowledge financial support from the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 802450) and the Joachim-Herz-Foundation. The research project has been approved by the Research Ethics Committee at the Department of Economics, University of Copenhagen, and was preregistered at the AEA RCT Registry (AEARCTR-0017116).

1 Introduction

Artificial intelligence (AI) is increasingly used for decision support in high-stakes environments characterized by uncertainty and incomplete information. In health care, and particularly in primary care, AI-based diagnostic support promises productivity gains by improving the quality of clinical decision-making through the extraction and evaluation of diagnosis-relevant signals from patient characteristics, clinical histories, and other patient records (Bayati et al. 2014; Lin et al. 2019; Mullainathan and Obermeyer 2022; Shapiro Ben David et al. 2025). Such signals are otherwise costly for decision-makers to process due to their high dimensionality. Despite rapid technological progress and growing policy interest, evidence is lacking on how human decision-makers integrate AI-generated signals with existing sources of information. Existing experimental work suggests that experts fail to implement the positive potential of AI tools because they disagree with its recommendations or fail to use the provided information correctly (Agarwal et al. 2023; Stevenson and Doleac 2024; Abaluck et al. 2026). This paper contributes new evidence by studying how primary care physicians update diagnostic beliefs and treatment decisions when assisted by AI, relative to a familiar and well-understood diagnostic test.

We study this question in the context of primary care diagnosis and treatment of urinary tract infections (UTIs), one of the most common conditions encountered in routine practice. Primary care physicians frequently face diagnostic uncertainty, as patients often present with overlapping or nonspecific symptoms. Even though definitive confirmation is possible, it is delayed by several days under current diagnostic technologies, underscoring the need for faster diagnostic information (Yelin et al. 2019; Huang et al. 2022; Ribers and Ullrich 2024). In the case of suspected UTI, antibiotic treatment is effective for bacterial infections but inappropriate for a range of alternative conditions with similar symptoms, creating a trade-off between timely treatment and the risk of unnecessary antibiotic use (Wilson and Gaido 2004; Gupta and Trautner 2012). UTIs are highly prevalent—nearly half of women experience at least one episode during their lifetime—and impose substantial healthcare costs (Foxman 2002; Flores-Mireles et al. 2015). The UTI setting therefore provides a relevant natural environment to study belief updating and treatment choices under uncertainty, as well as the integration of multiple noisy diagnostic signals.

To decide on a UTI diagnosis during an initial consultation, primary care physicians rely on patient-reported symptoms, medical history, and point-of-care tests such as urine dipsticks. These point-of-care tests provide immediate but imperfect information and are well understood by clinicians through training and experience (Chernaya et al. 2022). By contrast, AI-based diagnostic support tools generate probabilistic assessments by combining multiple inputs, but their internal logic appears as a black-box to users. We

exploit this contrast in an experimental setting that allows us to cleanly compare how physicians update beliefs when receiving information from a traditional diagnostic test versus from an AI system, holding constant the statistical properties of the signals.

We conduct a nation-wide survey experiment with primary care physicians in Denmark, in which participants make diagnostic and prescribing decisions for standardized UTI cases. Physicians receive sequential information consisting of patient symptoms and diagnostic signals, described either as a standard point-of-care dipstick test or as AI-generated. The experimental design allows us to measure differences in belief updating, assess departures from Bayesian updating, and study how physicians combine information from distinct sources. Importantly, we assess the relevance of correlation neglect, a common behavioral error when humans process information from multiple, correlated information sources (Enke and Zimmermann 2019). Our setting allows us to cleanly vary whether AI and dipstick signals are separable by originating from independent sources of information, or whether the AI signal incorporates the dipstick signal, inducing signal correlation.

We find that physicians update their beliefs less strongly in response to AI-generated signals than to traditional dipstick test results, even when both signals are equally informative. In fact, physicians over-update to the dipstick signal relative to a Bayesian benchmark, while they under-update to the AI signal. Physicians report a lack of understanding of the AI tool, how to use it, how it converts data into diagnostic information, and if it does so without bias. However, their observed belief updating is only systematically related to their mental models of the AI's ability to extract diagnostic information from patient data. Linked administrative data show that AI adoption is more common in clinics that use advanced point-of-care diagnostics more frequently and are located in urban areas.

We further find that, even in a setting where the combination of two correlated signals is clearly communicated, physicians show correlation neglect by updating to correlated signals almost exactly as much as to uncorrelated signals, corroborating a core result in Agarwal et al. (2023). This finding has an important implication: with humans in the loop, AI may be used best if it includes multiple diagnostic signals in its recommendation without revealing these individual signals, hence avoiding the need for humans to extract information from multiple signals with potentially non-trivial correlation. This might imply a departure from current clinical practice and counter explainability efforts.

Finally, turning to treatment decisions, we find that physicians are less likely to initiate antibiotic prescribing following a positive AI test signal than after a positive dipstick result. In contrast, they are equally likely to discontinue prescribing when either the AI tool or the dipstick test yields a negative result. Due to correlation neglect, they are also more likely to prescribe an antibiotic after positive signals from the AI tool and the dipstick, which is non-informative because already included in the AI signal, compared to after

receiving a positive signal only from the AI tool. These findings suggest that integrating multiple sources of information into an AI-based decision support tool may help improve decisions and reduce the overuse of antibiotics.

Our findings speak to the growing economics literature on human–AI collaboration and complementarity (Agrawal et al. 2022; Mindell et al. 2023; Agarwal et al. 2023; Ribers and Ullrich 2023; Angelova et al. 2025). We provide novel, micro-level evidence on how professionals integrate algorithmic signals into decision-making. In particular, our experimental design allows us to cleanly identify belief updating, shedding light on the mechanisms underlying human–AI interaction.

We also relate to the broader literature on AI adoption and its disruptive labor market effects (Acemoglu et al. 2022; Noy and Zhang 2023; Bick et al. 2024; Humlum and Vestergaard 2024; Brynjolfsson et al. 2025). Our setting features high-skilled experts – physicians – for whom AI promises an expansion of the productivity frontier by making complex, high-dimensional information cheap to process (Kleinberg et al. 2018). We show that despite potential productivity gains, AI adoption is not automatic: physicians respond less to AI-generated signals than to familiar diagnostic tools with equivalent statistical properties and many voice skepticism toward the tool.

Finally, we contribute to work on belief updating and information processing by documenting systematic differences in how individuals respond to novel versus familiar information sources, even when signals are statistically equivalent. More broadly, our findings speak to the design and implementation of AI decision-support tools in professional settings, highlighting that adoption does not only depend on predictive accuracy, but also on how users interpret algorithmic information and what mental models are at play.

2 Experiment

2.1 Experimental Context

Urinary tract infection (UTI) is one of the most frequent conditions that primary care physicians encounter. UTIs occur when bacteria enter and infect the urinary tract, bladder, or kidneys. Symptoms such as dysuria, pain, and urinary urgency can be highly debilitating, and untreated infections may lead to serious complications, including sepsis and death. UTIs are particularly prevalent among women, with nearly half experiencing at least one episode during their lifetime (Foxman 2002). In the United States alone, the annual healthcare costs attributable to community-acquired UTI are estimated at $1.6–3.5 billion (Foxman 2002; Flores-Mireles et al. 2015).

Antibiotic treatment is effective for bacterial UTIs, but similar symptoms can arise from a wide range of alternative conditions for which antibiotics are ineffective or inappro-

priate, including sexually transmitted urethritis or vaginitis, noninfectious urethritis, early pyelonephritis, or fungal infections (Wilson and Gaido 2004; Gupta and Trautner 2012). As a result, antibiotic prescribing for suspected UTI involves a trade-off between timely treatment of bacterial infections and the risk of unnecessary antibiotic use.

In routine practice, primary care physicians diagnose UTI based on patient-reported symptoms, medical history, and point-of-care diagnostic tests. Diagnostic confirmation can be obtained through urine culture analysis, either in-practice or at specialized laboratories, but this process typically requires two to four days to yield results. Point-of-care tests, by contrast, provide immediate information but have imperfect accuracy. As point-of-care-tests for UTI, urine dipstick tests for nitrites and leukocyte esterase are the most common, and usually tested jointly. Nitrite tests have high specificity (true positive rate) of 85% and low sensitivity (true negative rate) of about 45%, whereas leukocyte esterase tests have low specificity of 60% and high sensitivity of 80% (Chernaya et al. 2022). Throughout the experiment, nitrite test results are fixed as negative, and we refer to leukocyte esterase results as *dipstick tests* unless otherwise specified.

As we do not observe the true state of patients, we cannot direcly assess the quality of diagnosis and treatment decisions. Instead, we compare physicians' use of the AI diagnosis to their use of the know dipstick test, both with known sensitivity and specificity. This setting provides a lens for studying physicians' use of a new AI tool and potential mechanisms that lead to differences in AI adoption between physicians.

2.2 Participant recruitment

We elicit experimental data using a survey of primary care physicians in Denmark conducted in November and December 2025. We recruit participants by sending survey invitations to the official digital mailbox (Digital Post) of all Danish primary care practices. Since 2013, Digital Post has been the legally mandated channel used by all medical practices in Denmark for secure communication with public authorities and businesses, such as laboratories and insurers.[1] It thus constitutes the standard and universal channel for official correspondence with primary care practices and ensures full population coverage.

The invitation describes the study only as a questionnaire on treatment choices for infections in primary care. To avoid priming respondents, it does not mention AI adoption or antibiotic resistance.

As compensation for their participation, physicians receive 500 DKK (approximately $78 in 2025), which corresponds to roughly the fee charged for three routine patient consultations under the standard reimbursement schedule for general practice in Denmark (Overenskomst om almen praksis 2025).

[1] See https://en.digst.dk/systems/digital-post/current-legislation-about-digital-post/.

Out of 1607 invitations sent, 383 physicians from 356 practices completed the survey (response rate: 22%).

2.3 Survey Design

Figure 1 illustrates the survey flow. In the following, we describe the survey design step by step.

The survey begins with questions on background characteristics, such as age and gender. Respondents are then asked to report their beliefs about the sensitivity and specificity of dipstick tests. Following these initial questions, participants enter the experimental part of the survey.

2.3.1 Patient vignettes

The main part of the survey consists of patient vignettes. Each vignette describes a hypothetical patient presenting with symptoms typical of a urinary tract infection (UTI). Physicians' task is to assess the probability that a UTI is present and to decide whether to initiate further diagnostics and treatment.

Figure 2 shows a screenshot of an exemplary patient vignette. The vignette includes the age and gender of the patient, her primary symptoms, comorbidities, as well as active prescriptions. In Stage 1 of the experimental part, respondents see six of these vignettes, in which we vary the attributes of the vignettes. Table 1 shows the possible attributes. While age is varied randomly on the vignette level, symptoms and comorbidities are balanced between cases such that cases can be sorted in terms of UTI probability and benefit of antibiotic prescription.[2] Stage 2 of the experimental part presents respondents with three additional vignettes in which we keep both the UTI probability and benefit of antibiotic prescribing low.

Each vignette consists of two steps:

1. Pre-signal: The physician is asked to estimate the prior UTI probability based on the symptoms and patient background characteristics, and whether they would prescribe antibiotics.

2. Post-signal: After seeing a diagnostic signal (explained below), the physician is asked to estimate the posterior UTI probability. Moreover, the physician is asked whether they would seek additional diagnostics of the urine sample (culture or laboratory) and whether they would prescribe antibiotics.

[2] In three out of six vignettes in Stage 1, the patient has type 1 diabetes, which increases the risk for a complicated UTI without changing the baseline risk of a UTI, thus increasing the benefit of an antibiotic prescription holding diagnosis constant.

We experimentally vary in a between-subjects design which diagnostic signal participants receive. The specific signal types are described next.

2.3.2 AI Tool vs Dipstick treatment

At the outset of the experimental section, we explain in detail the diagnostic signal respondents will receive. The type of signal depends on the treatment they are randomly allocated to:

AI Tool Treatment. In the AI Tool Treatment, they receive a binary signal from an AI support tool that "uses artificial intelligence to predict bacterial urinary tract infections in primary care patients" and that "is trained on data from patient records and test results provided by microbiological laboratories (KMA)." We ask respondents to assume that the tool is clinically approved. We explain that the tool uses "clinical history incl. the history of general practice consultations (diagnoses, referrals, prescriptions), laboratory test results, and hospitalizations," as well as the current dipstick test result.

Dipstick treatment. In the Dipstick treatment, they receive a binary signal from a urinary dipstick test, indicating positive or negative leukocytes.

In both treatments, we fix the informativeness of the signal to have a sensitivity of 80% and a specificity of 60%. Keeping the informativeness of the signals constant has the advantage that a rational Bayesian updater would react in the same way in both treatments. Several further, important design choices underlie the vignettes and treatment comparisons:

- **Signal comparison**: Our goal is to study the adoption of a hypothetical AI decision support tool in diagnostic and treatment decisions. This requires an appropriate benchmark. We therefore compare belief updating under the AI signal to updating under a familiar diagnostic signal: the urinary dipstick test. Dipstick tests for leukocytes and nitrite are widely used in general practice for UTI diagnosis and provide immediate results, but have imperfect sensitivity and specificity, as would a data-driven AI support tool. By contrast, the gold-standard laboratory test for bacterial infection typically provides results only after a few days, meaning that treatment decisions have to be made based on imperfect diagnostic signals. The comparison between two such signals allows us to assess whether physicians are more skeptical toward a novel AI-based signal than toward an established diagnostic test.
- **Nitrites always negative**: In clinical practice, physicians typically rely on dipstick results for both leukocytes and nitrites when assessing UTI probability. Because

updating to two signals would not be comparable to updating to a single AI signal, we present only the leukocyte test result as the diagnostic signal. In all treatments and vignettes, the nitrite test result is explicitly stated to be negative. This ensures that negative nitrites affect prior and posterior beliefs symmetrically.

- **Sensitivity and specificity**: We chose a sensitivity of 80% and specificity of 60% since it resembles the actual validity of dipstick tests used in Danish general practice (Chernaya et al. 2022). The binary signal by the AI support tool is set to have the same sensitivity and specificity to allow direct comparability. The binary signal would be given if the predicted UTI probability exceeds a certain threshold, chosen such that it achieves a certain sensitivity and specificity.[3]

- **Binary AI signal**: In our case, the AI support tool gives a binary signal instead of a continuous risk assessment. This has two advantages: First, it is comparable to the way dipstick tests are typically presented (positive/negative). Second, it sidesteps the issue observed in Agarwal et al. (2023) that physicians struggle with interpreting intermediate probability assessments. In their case, decision quality improves for very low or very high AI probability assessments but worsens for intermediate probability assessments.

2.3.3 Mental models about AI and Dipstick signals

After Stage 1 of the experimental part, respondents evaluate either the AI tool or the dipstick test (depending on the treatment arm) by indicating their (dis-)agreement with seven statements. These statements are aimed to elicit respondents' mental models of the diagnostic tools: their subjective understanding of how AI tools and dipstick tests work.

We view mental models of each tool as (potentially misspecified) representations of its key characteristics that shape respondents' beliefs and decisions. Prior work shows that such misspecification can distort beliefs about AI performance and hinder adoption (Raux and Dreyfuss 2025). Moreover, when behavior is driven by misspecified mental models rather than cognitive frictions, it is unlikely to self-correct through learning or feedback (Esponda et al. 2024).

The statements probe perceptions of the source data (urine sample for the dipstick; dipstick results and patient records for the AI tool), the process by which results are generated (chemical reaction versus algorithmic prediction), the role of third parties (the dipstick producer or the algorithm designer), and the quality of the output. Our structured

[3]Choosing a threshold corresponds to selecting a point on the receiver operating characteristic (ROC) curve, which determines the sensitivity and specificity of the signal for a given prediction technology. Because specificity is lower than sensitivity in our setting, we would select a comparatively low threshold.

elicitation follows recent work that measures mental models using targeted survey questions and links them to beliefs and expectations (Andre et al. 2022, 2024).

Table A2 lists all statements. Respondents evaluate agreement with each statement on a Likert scale from 1 (strongly disagree) to 5 (strongly agree).

2.3.4 Correlated and uncorrelated signals

In Stage 2, physicians provide diagnoses and treatment decisions for three additional vignette cases. In these cases, they observe both signals—the dipstick test and the AI tool—but we exogenously vary the informativeness of the AI signal across two newly randomized treatment arms:

Uncorrelated Treatment. Physicians are informed that the AI tool does *not* incorporate the dipstick test result, so the two signals provide independent information.

Correlated Treatment. Physicians are informed that the AI tool *does* incorporate the dipstick test result. As a result, the AI signal contains less incremental information beyond the dipstick, allowing us to test for correlation neglect.

We keep signal accuracy constant between the Correlated and Uncorrelated treatment. As shown in Figure 1, the accuracy of the respective signals depends on whether an individual was in the AI Tool treatment or in the Dipstick treatment in the first stage:

- For respondents in the AI Tool treatment arm, the AI tool has the same sensitivity and sprecificity as before. The additional dipstick signal that they now receive has a sensitivity/specificity of 70% and 50%.[4] Respondents in the Uncorrelated Treatment are informed that the AI tool is now trained on a larger dataset such that it achieves the same accuracy as before although it does not include the dipstick signal anymore.
- For respondents in the Dipstick treatment arm, the dipstick signal has the same sensitivity and sprecificity as before. The AI tool has a higher sensitivity/specificty of 90% and 70%.[5]

For simplicity, the two signals always go in the same direction, hence, they are both positive or both negative. Since there are only three vignettes in this stage, we consider

[4]This is due to the fact that dipstick signal was included in the AI tool signal in the first stage, hence, it has to have a lower accuracy than the AI tool since it utilizes additional data.

[5]This is due to the fact that in the Correlated treatment the dipstick signal is included in the AI tool, hence, the AI tool has to have a higher accuracy than the dipstick test since it utilizes additional data.

it unlikely that individuals learn about the correlation structure of the signals from the observed signals.

2.4 Analysis Sample

For the analysis, we exclude respondents who do not sufficiently complete the survey or who fail an attention check. The attention check flags participants who update their beliefs by more than 5 percentage points in the direction opposite to the diagnostic signal in at least two vignettes. Appendix Table A1 reports sample sizes by treatment arm under these restrictions. The final analysis sample includes 372 physicians from 346 clinics. Respondents spend a median of 19.7 minutes completing the survey.

Treatment balance. Table 2 reports summary statistics and balance across treatment arms. Overall, 59% of respondents are female, with an average age of 50 years.

Physicians in all treatment arms report beliefs about the leukocyte esterase dipstick test's sensitivity as 76% and its specificity as 50%. These values are close to or, if anything, lower than our experimentally defined values of 80% and 60% for the AI tool and the dipstick test in the first part of the experiment.

Most respondents, above 80%, report knowing generative AI and the majority perceive it as helpful, in particular for summarizing patient records, followed by administrative tasks and, to a lesser extent, clinical decision-making. Survey characteristics are well balanced across treatment arms, with minor exceptions: physicians in the AI Tool treatment are slightly older and report lower perceived usefulness of generative AI for clinical decisions, while the Correlated treatment includes a slightly smaller share of physicians in training.

Comparison of survey clinics and population. We next leverage our administrative data to compare clinics in our analysis sample to the population of primary care clinics in Denmark. Appendix B describes the administrative data in detail.

The administrative data contains information for 1,630 clinics on clinical practice (consultations and prescriptions) and patient pools, and 1,560 clinics for physician characteristics. We successfully match 333 of the 346 analysis clinics on clinical practice and patient pools, and 320 clinics on physician characteristics.

Table 3 compares analysis clinics to the population of Danish primary care clinics observable in the administrative data. Differences are generally small, except that analysis clinics prescribe fewer antibiotics per consultation.

3 How do diagnostic beliefs react to AI signals?

We first examine how physicians' diagnostic beliefs respond to signals from the AI support tool relative to the dipstick test, using the first six vignettes of the experiment. We begin with a descriptive analysis of absolute belief updating, that is, the average difference between posterior and prior beliefs induced by each signal. We then move to a more formal Bayesian belief-updating framework to assess whether physicians weight signals in line with their diagnostic accuracy. Finally, we characterize correlates of belief updating and physicians' mental models and clinic characteristics.

3.1 Belief updating

Figure 3 plots the average change in beliefs (posterior minus prior) by treatment. Beliefs adjust substantially less in response to signals from the AI tool than to dipstick test signals. The distribution of updates under the AI signal is more tightly concentrated around zero, with fewer large revisions.

To test for average treatment effects, we estimate the following linear estimation model:

$$(P - P_0)_{iv} = \alpha_v + \beta \text{AI TOOL}_i + \varepsilon_{iv}, \tag{1}$$

where $(P - P_0)_{iv}$ denotes the change from prior to posterior belief for physician i in vignette v, α_v are vignette fixed effects, and AI TOOL_i indicates the AI Tool treatment arm.

Consistent with the pattern in Figure 3, the estimated absolute belief update is 7.3 percentage points smaller in the AI Tool treatment than in the Dipstick treatment (significant at the 1% level). Physicians update significantly less when the signal comes from the AI support tool, suggesting skepticism towards AI-generated signals.

What does the treatment effect capture? By design, the AI and dipstick signals have identical accuracy. Moreover, physicians are familiar with the dipstick's diagnostic properties: before we provide the experimentally fixed values, their reported beliefs about dipstick sensitivity and specificity are close to—if anything slightly below—the true values used in the experiment (see Table 2). If anything, this would imply that the AI signal should appear at least as accurate as the dipstick. The weaker updating in the AI Tool treatment therefore does not reflect ignorance about signal quality, but speaks toward greater skepticism toward AI-generated information.

AI non-adoption. Apart from an average treatment effect, Figure 3 also shows a noticeable mass of zero updates in the AI Tool treatment as opposed to the Dipstick treatment. Specifically, in 20% of AI signal cases, physicians do not update their beliefs at

all (compared to only 6% in the DIPSTICK treatment), and 39% update by less than three percentage points (compared to 16% in the DIPSTICK treatment).

Examining consistency in updating patterns, we find that 30% of physicians update by less than three percentage points in the majority of vignettes in the AI TOOL treatment, compared to only 5% of physicians in the DIPSTICK treatment. In the following, we refer to these individuals as "non-adopters."[6]

3.2 Bayesian belief updating

Next, we compare updating behavior to the rational Bayesian benchmark. To do that, we estimate belief updating coefficients following the regression framework in Grether (1980). For binary signals, Bayes rule can be written in the following form:

$$\log\left(\frac{P}{1-P}\right) = \log\left(\frac{P_0}{1-P_0}\right) + \log\left(LR^{\text{signal}}\right) \tag{2}$$

where P is the posterior, P_0 is the prior, and LR^{signal} is the likelihood ratio of the signal. To estimate the model, we introduce an error term and attach coefficients to the prior and the signal:

$$\log\left(\frac{P}{1-P}\right) = \delta^{\text{prior}} \cdot \log\left(\frac{P_0}{1-P_0}\right) + \beta^{\text{signal}} \cdot \log\left(LR^{\text{signal}}\right) + \varepsilon \tag{3}$$

where δ^{prior} captures the weight placed on the prior, while β^{signal} measures responsiveness to the signals. The term ε captures idiosyncratic updating errors. The likelihood ratio depends on the valence of the signal: a positive signal has $LR^{+} = 2$ and a negative signal has $LR^{-} = 1/3$.[7] For a Bayesian updater, the coefficients δ^{prior} and β^{signal} would equal 1.

Table 4 reports the estimated coefficients by treatment. Consistent with the evidence shown before, physicians update more in response to the signal in the DIPSTICK treatment than in the AI TOOL treatment. Interestingly, the Bayesian updating framework indicates that this gap is driven by *overreaction* to dipstick signals relative to Bayes' rule (i.e., $\beta^{\text{signal}} > 1$). By contrast, in the AI TOOL treatment, physicians' responsiveness is smaller than the Bayesian benchmark of 1, that is, they underreact to the signals.

3.3 Which mental models predict differential updating?

Next, we explore potential mechanisms underlying why physicians update beliefs less in response to a signal from the AI relative to the dipstick signal. Figure 4 summarizes

[6] We also classify belief changes of one or two percentage points as non-updates, since respondents may have misclicked on the slider. We consider an individual to be a consistent non-adopter if their belief changes fall within this range in at least four out of six vignettes.

[7] For a positive signal the LR is $\text{LR}^{+} = \frac{\text{sensitivity}}{1-\text{specificity}}$ and for a negative signal the LR is $\text{LR}^{-} = \frac{1-\text{sensitivity}}{\text{specificity}}$.

physicians' mental models of the AI tool and the dipstick test, where we simplify the scales for agreement with each statement into binary variables that indicate agreement ('somewhat agree' or 'strongly agree'). The left panel reports the share of physicians who agree with a series of statements about each signal, while the right panel shows coefficient estimates from regressions of average belief updating on agreement with each statement, separately by signal source.

The largest differences in perceptions across signal sources concern understanding and usability. Around 60% of physicians report understanding how the dipstick works, compared to fewer than 20% for the AI tool. Meanwhile, more than half of physicians report being uncertain about how to use the AI tool, compared to fewer than 10% for the dipstick tool.

However, across both signal sources, the strongest predictor of belief updating is whether physicians perceive the signal as being based on informative data. For the dipstick, understanding how the test works is also a strong predictor of updating. For the AI tool, believing that it optimally analyzes data and is informative of patients' current health is weakly predictive of updating, even though agreement with this statement varies less across signal sources.

Taken together, these results suggest that physicians perceive the AI tool as less transparent, harder to use, and more susceptible to producer influence – yet, none of these aspects are strongly correlated with actual changes in physicians' beliefs. Instead, observed updating to the AI tool correlates with perceptions of the tool's technical and statistical aspects: beliefs that data are current, analyzed optimally, and informative.

3.4 Who adopts the AI signal?

In a next step, we examine differences between physicians who rely on the AI signal and those who do not. As before, we define *non-adopters* as physicians who fail to update their beliefs in response to the signal in at least four of six vignettes.

Mental models. Figure 5 plots agreement with the mental model statements among physicians in the AI TREATMENT, separately for adopters and non-adopters. Consistent with earlier results, non-adopters are more skeptical that the AI relies on informative data or uses it optimally. Notably, adopters do not report better understanding of how the model works, nor do they express less uncertainty about interpreting its predictions.

Clinic characteristics. Figure 6 shows average standardized clinic characteristics from the administrative data for physicians in the AI TOOL treatment, by adoption status. Relative to adopters, non-adopters submit fewer microscopy claims, conduct a higher share

of phone consultations, and are less likely to be located in a major city. AI adoption thus appears more common in urban clinics that use other diagnostics and technologies.

4 Do physicians exhibit correlation neglect when presented with multiple signals?

Up to this point, physicians received diagnostic signals either from the AI support tool (which incorporates the dipstick result) or from the dipstick test alone. In practice, AI systems often synthesize information from multiple pieces of information, many of which the physician also sees. In this section, we test whether physicians account for the correlation between signals by adjusting their updating, or whether they instead treat signals as independent and exhibit correlation neglect.

In the second stage of the experiment, respondents observe both signals, and we randomize whether the two are correlated. In the CORRELATED treatment, the AI tool signal again incorporates the dipstick result, so the two signals are mechanically linked. In the UNCORRELATED treatment, the dipstick result is not used by the AI tool; consequently, the two signals can be treated as independent sources of information. We now investigate whether the correlation structure is reflected in physicians' belief updating.

Figure 7 shows that belief updating is similar under correlated and uncorrelated signals; the distributions largely overlap and the estimated effect is small and statistically insignificant, ruling out differences larger than -3.0 or $+2.8$ percentage points at the 5% level.[8] Note that this figure provides only a first descriptive look at the data, since signal informativeness differs across treatment arms.

To account for these differences, we use two complementary approaches. First, we estimate Equation (3) separately for correlated and uncorrelated signals within a Bayesian updating framework (Section 3.2). The likelihood ratio of the signal differs across cases: In the AI TOOL treatment arm, the joint LR with uncorrelated signals is $LR^{++} = 2.8$ and $LR^{--} = 1/5$; in the DIPSTICK treatment arm, the joint LR with uncorrelated signals is $LR^{++} = 6$ and $LR^{--} = 1/21$.[9] This framework therefore benchmarks updating against the information implied by the correlation structure. Second, we estimate a simple specification regressing posterior beliefs on prior beliefs and an indicator for signal valence (1 = positive, 0 = negative). This specification allows a direct comparison of how strongly physicians respond across correlation conditions, without imposing a Bayesian structure,

[8] To estimate the difference in absolute belief updating, we estimate regression equation (1) and replace the treatment dummy with the CORRELATED treatment dummy.

[9] With two independent signals, the joint LR is given by the product of the likelihood ratios of both signals. When the AI tool incorporates the dipstick signal, the LR is given by the LR of the AI tool since we assume that the AI tool optimally uses the information provided by the dipstick test.

but accounting for the direction of the signal.

Table 5 reports the results, separately for the correlated and uncorrelated cases. Panel A shows the results for the Bayesian updating regression. The updating coefficients β^{signal} are different from each other, suggesting that respondents update less to the available information in the UNCORRELATED treatment than in the CORRELATED treatment compared to the Bayesian benchmark. This pattern is consistent with correlation neglect: respondents do not sufficiently adjust their updating to reflect the different informativeness implied by the correlation structure of the signals.

Panel B shows results for the signal valence specification. Physicians raise their posterior by 43.3 percentage points when the signals are correlated and by 42.3 percentage points when they are uncorrelated. Thus, they update *almost identically* across the two conditions, even though uncorrelated signals provide substantially more information than correlated signals.

Finally, Appendix Table A4 addresses the potential concern that our findings could arise even in the absence of correlation neglect if participants simply ignored the AI tool and based their beliefs solely on the Dipstick signal in the correlation-neglect vignettes. In that case, we would observe similar updating patterns, but these would not reflect a failure to account for correlation. To rule out this alternative explanation, we restrict the sample to participants assigned to the AI treatment who actively adopted the tool. The estimates remain very similar, indicating that our evidence of correlation neglect is not driven by participants disregarding the AI signal. This finding has an important implication: with humans in the loop, AI may be used best if it includes other available diagnostic signals in its recommendation, hence avoiding the need for humans to extract information from multiple signals with potentially non-trivial correlation.

5 Treatment decisions

A central policy question is whether adopting an AI tool could improve prescribing quality by reducing unnecessary antibiotic use, a common health policy objective to limit the occurrence of antibiotic resistance. Antibiotics are the standard treatment for UTI, and so these infections are responsible for a large part of antibiotic prescriptions in primary care. We remain agnostic about treatment errors, as assigning infection status to the vignettes would require strong assumptions about the underlying ground truth. Instead of computing error rates, we examine how physicians adjust antibiotic prescription decisions in response to diagnostic signals, assuming that the specified sensitivity and specificity reflect true values against a ground truth. This allows us to assess whether AI adoption would increase or reduce antibiotic prescribing relative to dipstick testing.

We construct binary indicators for whether the prescription decision changes from pre- to post-signal in the direction indicated by the signal, separately for positive and negative signals. Using these outcomes, we estimate linear probability models.

However, interpreting changes in treatment decisions remains challenging because individual prescribing thresholds are unobserved; in many cases, signals may not move beliefs sufficiently to alter decisions. To address this, we differentiate between cases with priors above and below the sample median, where higher priors may lie closer to a prescribing threshold, and decision changes may be more likely.

Table 6 reports the results. Column (1) shows no statistically significant difference in switching into prescribing after a positive signal for low-prior cases. In contrast, Column (2) shows that for high-prior cases, physicians are significantly less likely (5% level) to prescribe antibiotics after a positive AI signal than after a positive dipstick signal. Column (3) shows no difference for low-prior cases. The constants indicate that the baseline probability of switching into prescribing after a positive dipstick signal is 27.1% in the high-prior sample and 16.7% in the low-prior sample, consistent with relatively high prescribing thresholds.

Columns (4)–(6) present results for negative signals. We find no significant differences between AI and dipstick signals. The baseline probability of switching to no prescription after a negative dipstick signal is 16.6% in the high-prior sample and 1.75% in the low-prior sample, again consistent with high decision thresholds.

We also inspect how physicians change their prescription decisions when they see the AI signal, which includes the dipstick result, and the dipstick result separately compared to when they only see the AI signal. Specifically, we focus on individuals in the AI treatment and on individuals in the Correlated treatment who were previously in the AI treatment. Both of these groups see the AI signal with the same sensitivity (80%) and specificity (60%). The only difference is that, in the Correlated treatment, they also see the dipstick treatment, which is not informative because it is already included in the AI signal. We focus on vignettes that are identical in terms of symptoms. Table 7 shows that the treatment decision more often changes to giving a prescription for an antibiotic when the two correlated signals, AI and dipstick, are both observed.

Overall, these results suggest that, relative to the status quo of commonly used dipstick tests, primary care physicians may reduce antibiotic prescribing after adopting an AI tool for diagnosis. Referring to the implications of correlation neglect discussed in section 4, this finding suggests that overuse of antibiotics might be effectively reduced by letting an AI tool summarize multiple diagnostic signals instead of physicians. Even though this would likely imply a departure from current diagnostic practices, it might also help reduce physicians' time required to arrive at an initial diagnosis and treatment decision.

References

Abaluck, Jason, Robert Pless, Nirmal Ravi, Anja Sautmann, and Aaron Schwartz, "Does LLM Assistance Improve Healthcare Delivery? An Evaluation Using On-site Physicians and Laboratory Tests," NBER Working Paper 34660, January 2026.

Acemoglu, Daron, David Autor, Jonathon Hazell, and Pascual Restrepo, "Artificial Intelligence and Jobs: Evidence from Online Vacancies," *Journal of Labor Economics*, April 2022, *40* (S1), S293–S340.

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz, "Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology," NBER Working Paper 31422, July 2023.

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb, *Prediction machines, updated and expanded: The simple economics of artificial intelligence*, Harvard Business Press, 2022.

Andre, Peter, Carlo Pizzinelli, Christopher Roth, and Johannes Wohlfart, "Subjective Models of the Macroeconomy: Evidence From Experts and Representative Samples," *The Review of Economic Studies*, November 2022, *89* (6), 2958–2991.

_ **, Philipp Schirmer, and Johannes Wohlfart**, "Mental models of the stock market," 2024.

Angelova, Victoria, Will Dobbie, and Crystal S Yang, "Algorithmic recommendations and human discretion," *Review of Economic Studies*, 2025, p. rdaf084.

Bayati, Mohsen, Mark Braverman, Michael Gillam, Karen M. Mack, George Ruiz, Mark S. Smith, and Eric Horvitz, "Data-Driven Decisions for Reducing Readmissions for Heart Failure: General Methodology and Case Study," *PLoS ONE*, October 2014, *9* (10), e109264.

Bick, Alexander, Adam Blandin, and David J Deming, "The Rapid Adoption of Generative AI," NBER Working Paper 32966, September 2024.

Brynjolfsson, Erik, Danielle Li, and Lindsey Raymond, "Generative AI at Work," *The Quarterly Journal of Economics*, April 2025, *140* (2), 889–942.

Chernaya, Anastasiya, Christian Søborg, and Mette Midttun, "Validity of the urinary dipstick test in the diagnosis of urinary tract infections in adults," *Danish Medical Journal*, 2022, *69* (1), A07210607.

David, Shirley Shapiro Ben, Roni Romano, Daniella Rahamim-Cohen, Joseph Azuri, Shira Greenfeld, Ben Gedassi, and Uri Lerner, "AI driven decision support reduces antibiotic mismatches and inappropriate use in outpatient urinary tract infections," *npj Digital Medicine*, January 2025, *8* (1), 61.

Enke, Benjamin and Florian Zimmermann, "Correlation neglect in belief formation," *The Review of Economic Studies*, 2019, *86* (1), 313–332.

Esponda, Ignacio, Emanuel Vespa, and Sevgi Yuksel, "Mental Models and Learning: The Case of Base-Rate Neglect," *American Economic Review*, March 2024, *114* (3), 752–782.

Flores-Mireles, Ana L., Jennifer N. Walker, Michael Caparon, and Scott J. Hultgren, "Urinary tract infections: epidemiology, mechanisms of infection and treatment options," *Nature Reviews Microbiology*, May 2015, *13* (5), 269–284.

Foxman, Betsy, "Epidemiology of urinary tract infections: incidence, morbidity, and economic costs," *The American Journal of Medicine*, July 2002, *113* (1, Supplement 1), 5–13.

Grether, David M., "Bayes Rule as a Descriptive Model: The Representativeness Heuristic," *The Quarterly Journal of Economics*, 11 1980, *95* (3), 537–557.

Gupta, Kalpana and Barbara Trautner, "Urinary Tract Infection," *Annals of Internal Medicine*, March 2012, *156* (5), ITC3–1.

Huang, Shan, Michael Allan Ribers, and Hannes Ullrich, "Assessing the value of data for prediction policies: The case of antibiotic prescribing," *Economics Letters*, April 2022, *213*, 110360.

Humlum, Anders and Emilie Vestergaard, "The Adoption of ChatGPT," IZA Discussion Paper 16992, May 2024.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, "Human Decisions and Machine Predictions," *Quarterly Journal of Economics*, 2018, *133* (1), 237–293.

Lin, Steven Y., Megan R. Mahoney, and Christine A. Sinsky, "Ten Ways Artificial Intelligence Will Transform Primary Care," *Journal of General Internal Medicine*, August 2019, *34* (8), 1626–1630.

Mindell, David A, Elisabeth Reynolds et al., *The work of the future: Building better jobs in an age of intelligent machines*, Mit Press, 2023.

Mullainathan, Sendhil and Ziad Obermeyer, "Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care," *Quarterly Journal of Economics*, May 2022, *137* (2), 679–727.

Noy, Shakked and Whitney Zhang, "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence," *Science*, 2023, *381* (6654), 187–192.

Raux, Raphaël and Bnaya Dreyfuss, "Human Learning about AI," in "Proceedings of the 26th ACM Conference on Economics and Computation" Association for Computing Machinery New York, NY, USA 2025, p. 1106.

Ribers, Michael Allan and Hannes Ullrich, "Machine Predictions and Human Decisions with Variation in Payoffs and Skill: The Case of Antibiotic Prescribing," Berlin School of Economics Discussion Paper Nr. 27 November 2023.

_ **and** _ , "Complementarities between Algorithmic and Human Decision-Making: The Case of Antibiotic Prescribing," *Quantitative Marketing and Economics*, 2024, *22* (4), 445–483.

Stevenson, Megan T and Jennifer L Doleac, "Algorithmic risk assessment in the hands of humans," *American Economic Journal: Economic Policy*, 2024, *16* (4), 382–414.

Wilson, Michael L. and Loretta Gaido, "Laboratory Diagnosis of Urinary Tract Infections in Adult Patients," *Clinical Infectious Diseases*, April 2004, *38* (8), 1150–1158.

Yelin, Idan, Olga Snitser, Gal Novich, Rachel Katz, Ofir Tal, Miriam Parizade, Gabriel Chodick, Gideon Koren, Varda Shalev, and Roy Kishony, "Personal Clinical History Predicts Antibiotic Resistance of Urinary Tract Infections," *Nature Medicine*, 2019, *25* (7), 1143–1152.

Figures

Figure 1. Survey flow

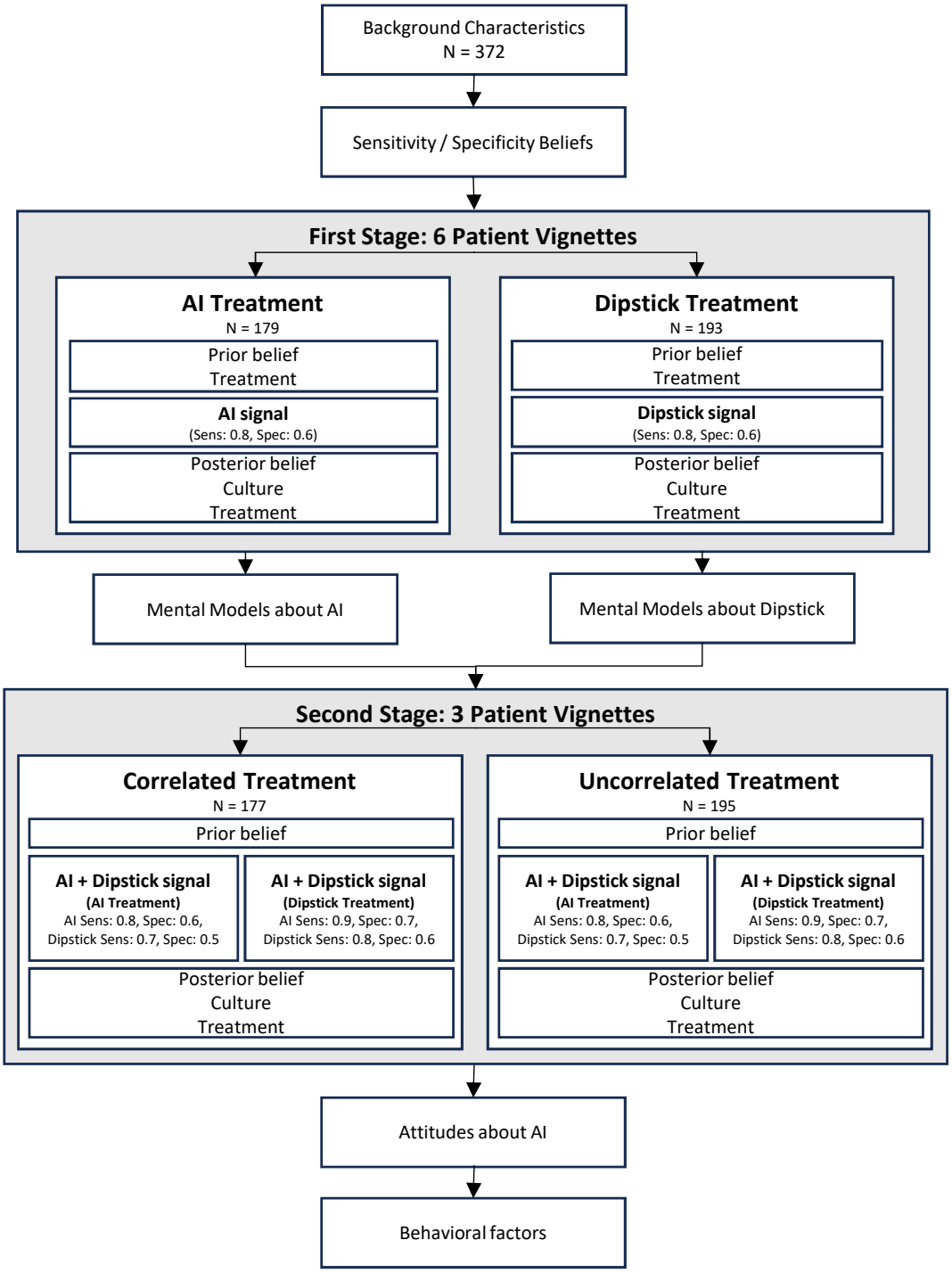


Figure 2. Screenshot of vignette

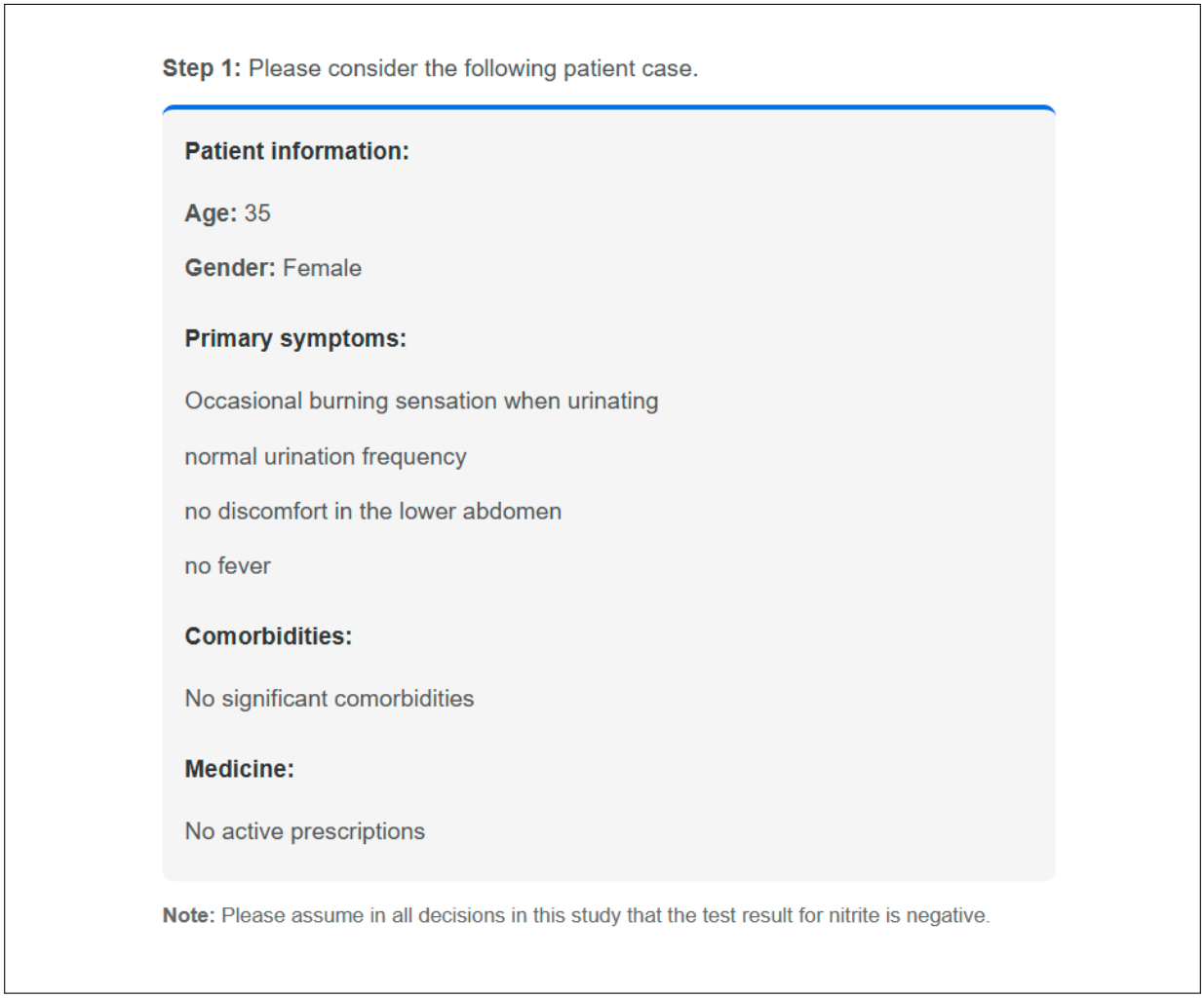
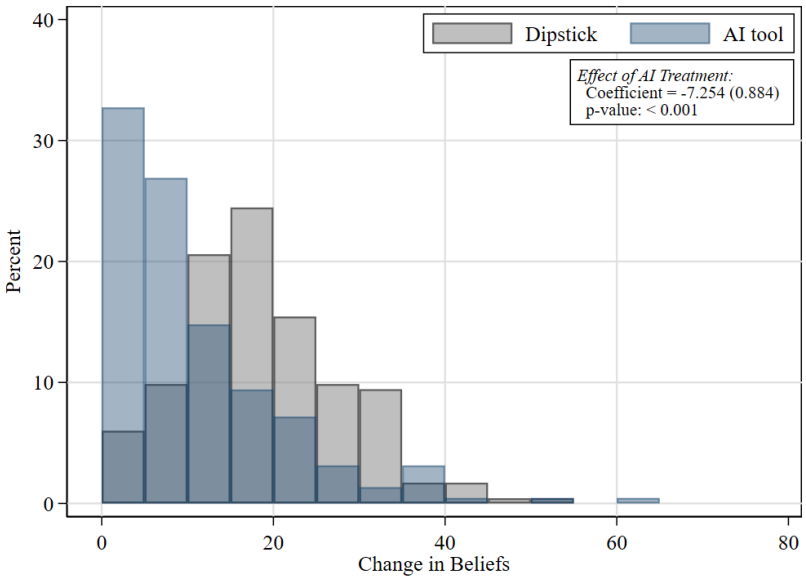
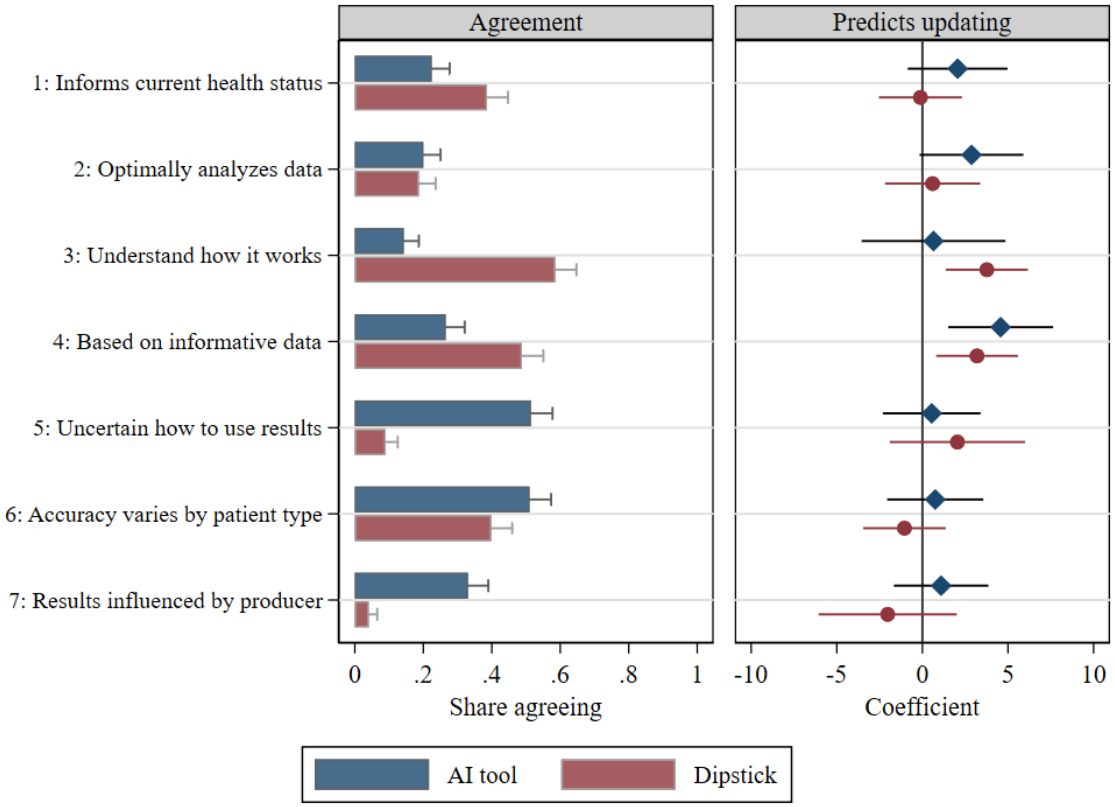


Figure 3. Absolute belief updating: Signal source



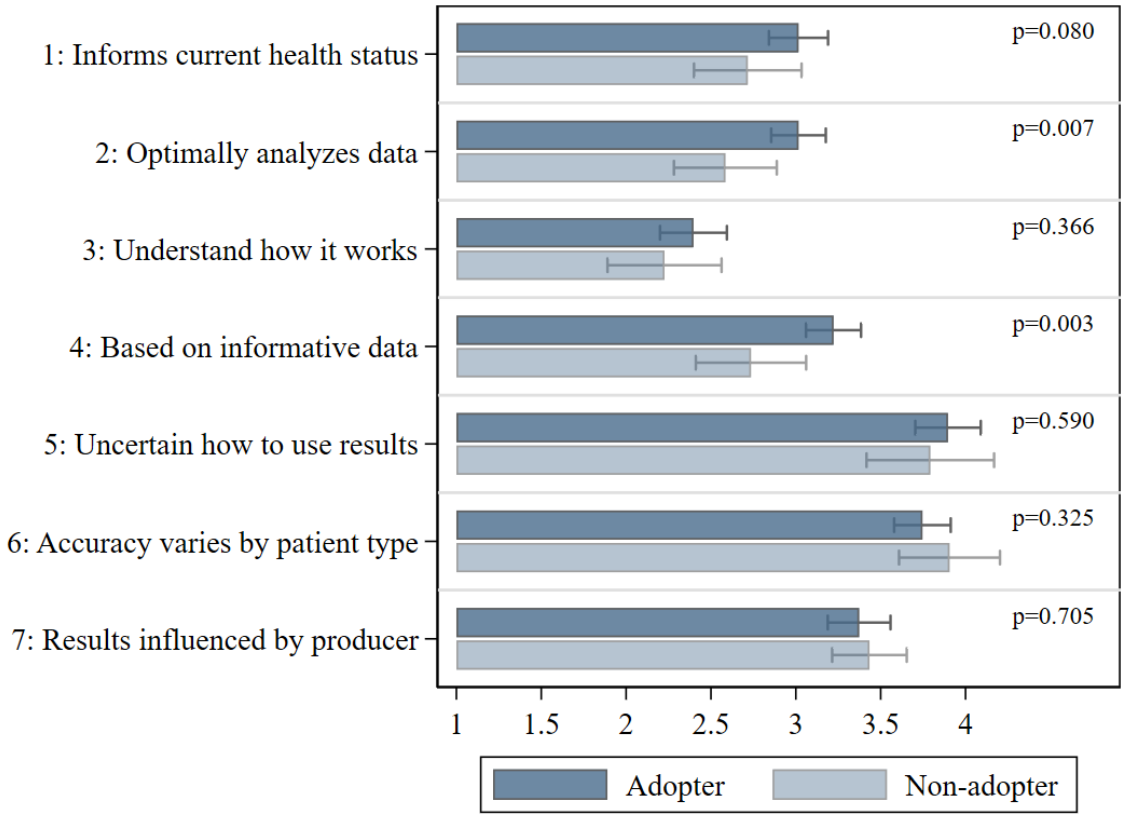
Notes: This figure presents the distribution of absolute belief updating, measured as the absolute difference between prior and posterior beliefs. The figure compares belief updating when physicians receive an AI signal versus a dipstick test (baseline) in vignettes 1–6. The panel also reports the estimated treatment effect from an OLS regression of absolute belief updating on the treatment indicator, including vignette fixed effects. Standard errors (in parentheses) are clustered at the physician level, and corresponding p-values are reported.

Figure 4. Mental models



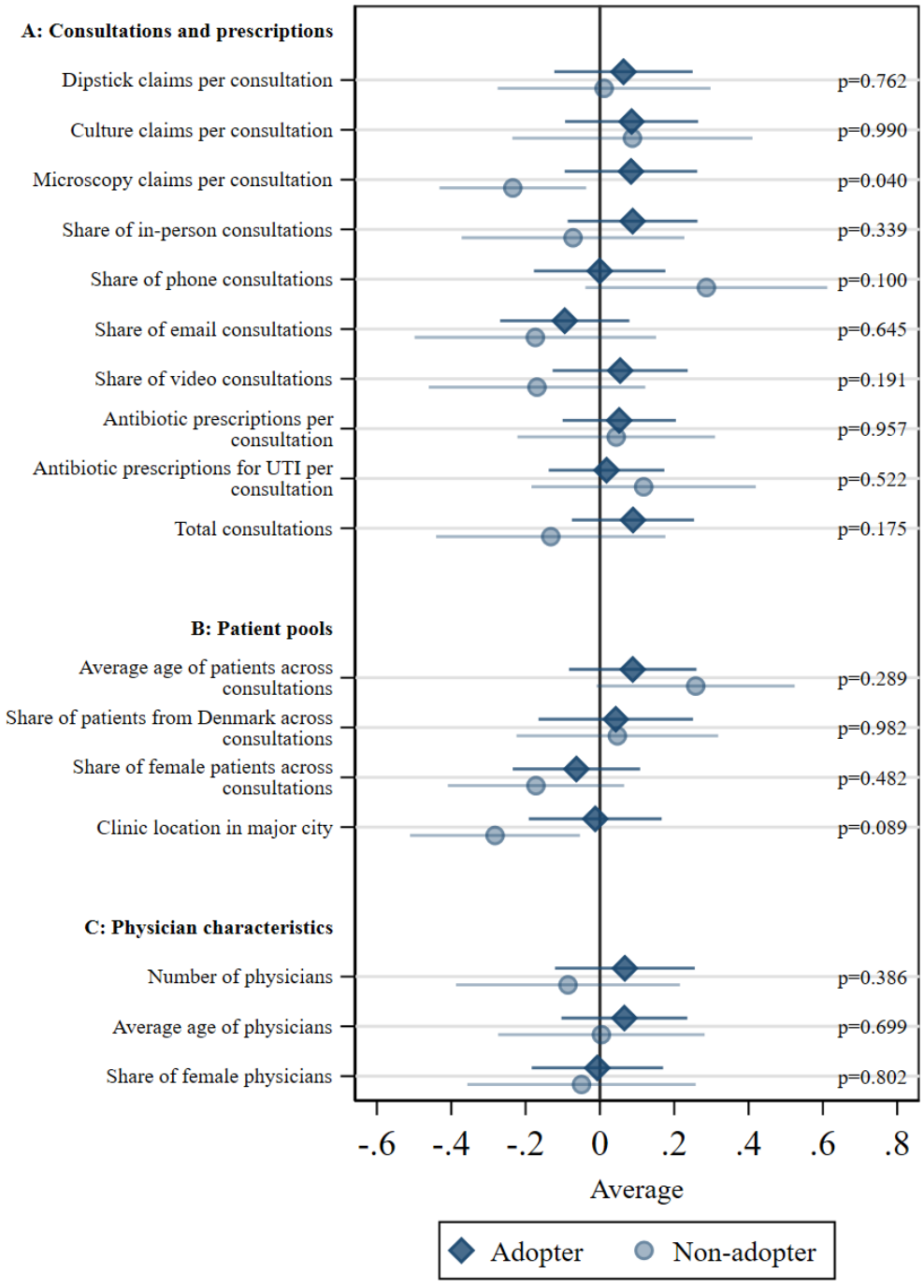
Notes: The left panel shows the share of individuals who agree to the respective item. The right panel shows coefficient estimates from regressing respondents' average updating over vignettes 1-6 on agreement with signal-related statements, separately for the dipstick and AI tool signal.

Figure 5. Mental models (only AI tool, by adopter/non-adopter)



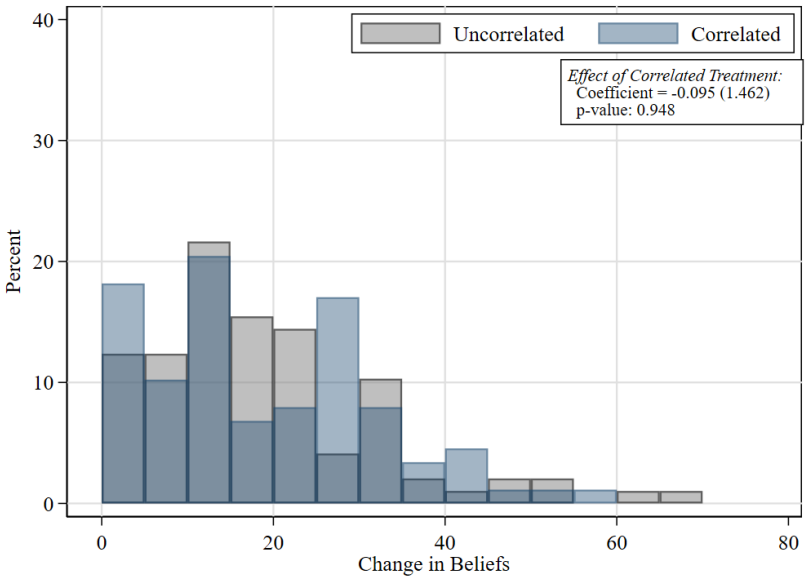
Notes: The figure shows individuals' agreement with the respective items, separately for individuals who consistently update to the AI signal and those who do not.

Figure 6. Clinic characteristics (only AI tool, by adopter/non-adopter)



Notes: The figure shows the average clinic characteristics, standard-normalized within clinics in the analysis sample, separately for individuals who consistently update to the AI signal and those who do not.

Figure 7. Absolute belief updating: Signal correlation



Notes: This figure presents the distribution of absolute belief updating, measured as the absolute difference between prior and posterior beliefs. The figure compares belief updating when AI and dipstick signals are correlated versus uncorrelated (baseline) in vignettes 7–9. The panel also reports the estimated treatment effect from an OLS regression of absolute belief updating on the treatment indicator, including vignette fixed effects. Standard errors (in parentheses) are clustered at the physician level, and corresponding p-values are reported.

Tables

Table 1. Factorial design of vignette attributes

Attribute	Levels
Patient Information	
Gender	Female
Age	22 / 25 / 30 / 35
Primary Symptoms	
Urinary discomfort	Occasional burning sensation during urination /
	Severe burning sensation during urination /
	Severe burning sensation during urination limiting quality of life
Urinary frequency	Normal urinary frequency /
	Increased urinary frequency
Abdominal discomfort	No lower abdominal discomfort /
	Lower abdominal discomfort
Comorbid conditions	
Comorbidities and active prescriptions	None /
	Type 1 diabetes and Insulin aspart (Novolog) via insulin pump

Table 2. Respondent characteristics by treatment

	A: Signal source		
	Dipstick (Baseline)	**AI Tool**	**Difference**
	Mean	*Mean*	*p-value*
Physician characteristics			
Female gender	0.57	0.60	[0.51]
Age in 2025	49.5	51.5	[0.02]
In training	0.016	0.022	[0.63]
Beliefs about dipstick accuracy			
Sensitivity belief	75.3	76.1	[0.64]
Specificity belief	49.3	51.6	[0.31]
Knowledge about generative AI			
Knows generative AI	0.84	0.79	[0.20]
Uses generative AI for work	0.50	0.54	[0.45]
Uses generative AI in private life	0.69	0.61	[0.13]
Opinions about generative AI			
Helpful for administrative tasks (1, not at all - 5)	3.94	3.87	[0.54]
Helpful for summarizing patient journals (1, not at all - 5)	4.23	4.25	[0.84]
Helpful for clinical decisions (1, not at all - 5)	3.59	3.30	[0.01]
# **Respondents**	193	179	372
	B: Signal correlation		
	Uncorrelated (Baseline)	**Correlated**	**Difference**
	Mean	*Mean*	*p-value*
Physician characteristics			
Female gender	0.60	0.57	[0.57]
Age in 2025	50.3	50.7	[0.67]
In training	0.031	0.0056	[0.07]
Beliefs about dipstick accuracy			
Sensitivity belief	76.1	75.2	[0.60]
Specificity belief	48.9	52.0	[0.19]
Knowledge about generative AI			
Knows generative AI	0.85	0.79	[0.10]
Uses generative AI for work	0.51	0.53	[0.73]
Uses generative AI in private life	0.68	0.63	[0.32]
Opinions about generative AI			
Helpful for administrative tasks (1, not at all - 5)	3.93	3.89	[0.73]
Helpful for summarizing patient journals (1, not at all - 5)	4.22	4.26	[0.70]
Helpful for clinical decisions (1, not at all - 5)	3.41	3.50	[0.46]
# **Respondents**	195	177	372

Notes: This table reports summary statistics for the analysis sample by signal source treatment (Panel A) and signal correlation treatment (Panel B). The first column presents full-sample means, while the second and third columns show means for each treatment group. The final column reports p-values from t-tests testing equality of means across the two treatment groups.

Table 3. Comparison of clinics in the analysis sample with the population of primary care clinics

	Population		Sample		Difference
	Mean	*SD*	*Mean*	*SD*	*SMD*
Panel A: Consultations and prescriptions					
Dipstick claims per consultation	0.034	(0.017)	0.033	(0.015)	[0.04]
Culture claims per consultation	0.014	(0.018)	0.013	(0.016)	[0.06]
Microscopy claims per consultation	0.0054	(0.010)	0.0062	(0.011)	[-0.08]
Share of in-person consultations	0.52	(0.091)	0.51	(0.082)	[0.12]
Share of phone consultations	0.22	(0.094)	0.20	(0.085)	[0.17]
Share of email consultations	0.25	(0.094)	0.27	(0.087)	[-0.25]
Share of video consultations	0.015	(0.020)	0.017	(0.020)	[-0.13]
Antibiotic prescriptions per consultation	0.058	(0.24)	0.049	(0.014)	[0.60]
Antibiotic prescriptions for UTI per consultation	0.017	(0.095)	0.014	(0.0044)	[0.66]
Total consultations	20111.4	(12129.0)	22001.7	(12399.8)	[-0.15]
# **Clinics**	1,630		333		
Panel B: Patient pools					
Average age of patients across consultations	48.4	(5.40)	47.4	(5.12)	[0.18]
Share of patients from Denmark across consultations	0.85	(0.11)	0.87	(0.085)	[-0.18]
Share of female patients across consultations	0.61	(0.051)	0.61	(0.049)	[-0.10]
Clinic location in major city	0.24	(0.43)	0.28	(0.45)	[-0.10]
# **Clinics**	1,630		333		
Panel C: Physician characteristics					
Number of physicians	3.16	(2.31)	3.44	(2.30)	[-0.13]
Single physician clinics	0.31	(0.46)	0.22	(0.42)	[0.22]
Average age of physicians	53.5	(7.12)	52.6	(6.49)	[0.13]
Share of female physicians	0.59	(0.37)	0.57	(0.35)	[0.06]
# **Clinics**	1,560		320		

Notes: This table reports clinic-level summary statistics for consultation claims and prescriptions (Panel A), patient composition across consultations (Panel B), and average physician characteristics within clinics (Panel C). The population is constructed from the national clinic register. We restrict to population and sample clinics that can be matched to register data. We identify consultations by clinic claim codes, and antibiotic prescriptions by Anatomical Therapeutic Chemical (ATC) codes and indication codes. The "population" columns report population means and standard deviations (SD) and the "sample" columns report means and SDs for clinics in the analysis sample. The final column reports Cohen's d, the standardized mean difference (SMD), between the population and analysis sample.

Table 4. Signal source: Belief updating (Bayesian)

	(1) Dipstick b/se	(2) AI tool b/se
δ^{prior}	0.86***	0.88***
	(0.026)	(0.030)
β^{signal}	1.21***	0.73***
	(0.038)	(0.050)
$\delta^{\text{prior}} = 1$ (*p-value*)	0.000	0.000
$\beta^{\text{signal}} = 1$ (*p-value*)	0.000	0.000
Adj. R^2	0.800	0.740
N	1076	965

Notes: Bayesian regression regressing posterior beliefs on the prior and the likelihood ratio of the signals (Dipstick vs. AI Tool treatment).

Table 5. Signal correlation: Belief updating

Panel A: Bayesian updating		
	(1) Correlated b/se	(2) Uncorrelated b/se
δ^{prior}	0.64***	0.63***
	(0.038)	(0.034)
β^{signal}	1.07***	0.66***
	(0.049)	(0.029)
Adj. R^2	0.755	0.732
N	506	559
Panel B: Signal valence		
	(1) Correlated b/se	(2) Uncorrelated b/se
Prior belief	0.62***	0.60***
	(0.045)	(0.039)
Positive signal	43.4***	42.5***
	(2.01)	(1.80)
Constant	-4.74***	-3.51***
	(1.17)	(1.06)
Adj. R^2	0.669	0.639
N	506	559

Notes: Table reports belief updating separately for the CORRELATED and UNCORRELATED treatments. Panel A reports Bayesian updating regressions of posterior beliefs on the prior and the likelihood ratio implied by the signals. Panel B reports regressions of posterior beliefs on the prior and an indicator for a positive signal (vs. negative).

Table 6. Effect of AI Treatment on Antibiotic Prescribing Behavior

	Change to prescribe after positive signal			Change to not prescribe after negative signal		
	(1)	(2)	(3)	(4)	(5)	(6)
	All	High Prior	Low Prior	All	High Prior	Low Prior
	b/se	b/se	b/se	b/se	b/se	b/se
AI	-0.0371	-0.105**	0.0365	-0.0115	-0.0342	0.00484
	(0.0290)	(0.0433)	(0.0373)	(0.0183)	(0.0334)	(0.0118)
Constant	0.221***	0.271***	0.167***	0.0891***	0.166***	0.0175**
	(0.0199)	(0.0323)	(0.0241)	(0.0130)	(0.0238)	(0.00772)
Adj. R^2	0.001	0.014	0.000	-0.000	0.000	-0.002
N	1072	557	515	1091	538	553

Notes: Standard errors clustered at the respondent level in parentheses.

Table 7. Effect of the superfluous signal in CORRELATED on Antibiotic Prescribing Behavior

	(1) Change to prescribe after positive signal b/se	(2) Change to not prescribe after negative signal b/se
Non-signal	0.0960*	-0.0160
	(0.0567)	(0.0227)
Constant	0.0870**	0.0417**
	(0.0342)	(0.0205)
Adj. R^2	0.013	-0.004
N	151	174

Notes: Standard errors clustered at the respondent level in parentheses.

Appendix

A Appendix Tables

Table A1. Restrictions on the main analysis sample

	Dipstick (Baseline)	**AI tool**	**Total**
A: Signal source			
No restrictions	50.1%	49.9%	491
+ Exclude if survey is <90% completed	52.0%	48.0%	383
+ Exclude if repeated wrong-direction updating (>5 ppt.)	51.9%	48.1%	372
	Uncorrelated (Baseline)	**Correlated**	**Total**
B: Signal correlation			
No restrictions	50.3%	49.7%	491
+ Exclude if survey is <90% completed	52.7%	47.3%	383
+ Exclude if repeated wrong-direction updating (>5 ppt.)	52.4%	47.6%	372

Note: This table presents the number of respondents after applying our sample restrictions. We exclude respondents who 1) complete less than 90% of the survey; 2) update their beliefs in the opposite direction of the signal by more than 5 percentage points in two or more vignettes. The final analysis sample consists of 372 physician respondents from 346 clinics.

Table A2. Mental model statements by treatment arm

#	Label	AI treatment	Dipstick treatment
1	Informs current health status	The AI tool can provide me with information about an individual patient's current health status.	The dipstick test can provide me with information about an individual patient's current health status.
2	Optimally analyzes data	I believe that the AI tool optimally analyzes information about a patient's UTI status from the dipstick test and the patient records.	I believe that the dipstick test optimally analyzes information about a patient's UTI status from the urine sample.
3	Understand how it works	I have a good understanding of how the AI model analyzes patient data to generate its prediction.	I have a good understanding of how the dipstick test analyzes a urine sample to produce its result.
4	Based on informative data	I believe that the data used by the AI tool (patient records, dipstick result) is informative about a patient's current UTI status.	I believe that the urine sample is informative about a patient's current UTI status.
5	Uncertain how to use results	I am uncertain how to incorporate the AI-generated prediction result into my diagnostic assessment.	I am uncertain how to incorporate the dipstick test result into my diagnostic assessment.
6	Accuracy varies by patient type	I am concerned that the accuracy of the AI tool may vary across different types of patients.	I am concerned that the accuracy of the dipstick test may vary across different types of patients.
7	Results influenced by producer	I believe that the result of the AI tool is influenced by the designer of the algorithm to influence my treatment decision according to their objectives.	I believe that the result of the dipstick test is influenced by the manufacturer to influence my treatment decision according to their objectives.

Notes: This table shows an overview of the statements used in the mental model module, along with the full item wording. Participants are asked to evaluate agreement with each of the seven statements (either regarding the AI or the dipstick signal, depending on treatment arm) on a Likert scale from 1 (strongly disagree) to 5 (strongly agree).

Table A3. Description of outcome and clinic characteristics variables

Variable	Description	Data source	Codes used
Panel A: Consultations and prescriptions			
Dipstick claims per consultation	Number of dipstick tests billed divided by total consultations.	SSSY	807101, 808155
Culture claims per consultation	Number of urine culture tests billed divided by total consultations.	SSSY	807105, 807189
Microscopy claims per consultation	Number of microscopy tests billed divided by total consultations.	SSSY	807102, 807103, 807122, 807123, 807169, 808156, 80157, 808165, 808166
Share of in-person consultations	Share of all consultations conducted face-to-face.	SSSY	800101, 800106
Share of tele-consultations	Share of consultations conducted by telephone.	SSSY	800109, 800201, 808294, 808618, 800501
Share of email consultations	Share of consultations conducted via email.	SSSY	800105, 800110
Share of video consultations	Share of consultations conducted via video.	SSSY	800125, 804436
Antibiotic prescriptions per consultation	Total antibiotic prescriptions divided by total consultations.	LMDB	ATC J01 (all antibiotics)
Antibiotic prescriptions for UTI per consultation	Antibiotic prescriptions associated with UTI diagnoses divided by consultations.	LMDB	Diagnosis (indication) codes 103, 104
Total consultations	Total number of consultations.	SSSY	—
Panel B: Patient pools			
Average age of patients	Mean patient age across consultations.	SSSY, CPR	—
Share of patients from Denmark	Fraction of consultations with patients registered as residents of Denmark.	SSSY, CPR	—
Share of female patients	Fraction of consultations with female patients.	SSSY, CPR	—
Clinic location in major city	Clinic located in a city with a population above 100,000 inhabitants (Copenhagen, Aarhus, Odense, Aalborg).	SSSY, CPR	—
Panel C: Physician characteristics			
Number of physicians	Total physicians affiliated with clinic.	YDER	—
Single physician clinics	Indicator equal to 1 if clinic has only one physician.	YDER	—
Average age of physicians	Mean physician age within clinic.	YDER, CPR	—
Share of female physicians	Fraction of physicians who are female.	YDER, CPR	—

Table A4. Signal correlation: Belief updating for AI adopters

Panel A: Bayesian updating		
	(1) Correlated b/se	(2) Uncorrelated b/se
δ^{prior}	0.75***	0.66***
	(0.048)	(0.055)
β^{signal}	1.59***	1.06***
	(0.092)	(0.068)
Adj. R^2	0.784	0.765
N	204	215
Panel B: Signal valence		
	(1) Correlated b/se	(2) Uncorrelated b/se
Prior belief	0.71***	0.60***
	(0.063)	(0.057)
Positive signal	43.4***	42.4***
	(2.28)	(2.31)
Constant	-9.05***	-4.48**
	(1.79)	(1.73)
Adj. R^2	0.694	0.678
N	204	215

Notes: This table reports belief updating separately for the Correlated and Uncorrelated treatments, restricting to participants in the AI Tool treatment who update their beliefs in at least four of six vignettes during the first part (*AI adopters*). Panel A reports Bayesian updating regressions of posterior beliefs on the prior and the likelihood ratio implied by the signals. Panel B reports regressions of posterior beliefs on the prior and an indicator for a positive signal (vs. negative).

B Data Appendix

B.1 Administrative data

Our administrative data cover the full population of Danish citizens and primary care clinics and are provided by Statistics Denmark and the Danish Health Data Authority (*Sundhedsdatastyrelsen*). The registers rely on unique individual identifiers (*Det Centrale Personregister, CPR*) and clinic identifiers (*Yderregister, YDER*). For each clinic, we observe patient purchases of prescribed antibiotics (*Lægemiddeldatabasen, LMDB*) and primary care insurance claims (*Sygesikringsregisteret, SSR*), as well as information on affiliated physicians and their basic demographics.

Using the national register of primary care clinics, we construct the 2024 clinic population. This register is based on a different system than the list used for survey invitations. However, it matches more than 95% of invited clinics.

For survey respondents matched to the administrative data, we construct clinic-level measures of clinical practice and average physician characteristics in 2024. Variable definitions are reported in Table A3.