DIW BERLIN

2168

Discussion Papers

Deutsches Institut für Wirtschaftsforschung 2026

AI Adoption by Human Experts
Evidence from Primary Care Physicians

Shan Huang, Renke Schmacker, and Hannes Ullrich

IMPRESSUM

DIW Berlin
German Institute for Economic Research
Anton-Wilhelm-Amo-Str. 58
10117 Berlin

Tel. +49 (30) 897 89-0
Fax +49 (30) 897 89-200
http://www.diw.de

ISSN electronic edition 1619-4535

Papers can be downloaded free of charge from the DIW Berlin website:
http://www.diw.de/discussionpapers

Discussion Papers of DIW Berlin are indexed in RePEc and Econstor:
http://ideas.repec.org/s/diw/diwwpp.html
https://www.econstor.eu/handle/10419/10

AI adoption by human experts: Evidence from primary care physicians*

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*

June 22, 2026

Abstract

AI can raise productivity by extracting information from rich data, yet little is known about how experts weigh AI-generated signals against established decision-support tools. We conduct a nationwide survey experiment with 372 Danish primary care physicians (21.5% of all clinics), who make diagnostic and treatment decisions on urinary tract infection vignettes before and after receiving a diagnostic signal. Holding accuracy constant, we randomize between-subjects whether the signal appears as an AI prediction or a commonly used dipstick test result. Physicians update beliefs 41% less in response to AI than to dipstick signals, consistent with AI skepticism. Roughly one-third of physicians ignore the AI tool; linked administrative data show that these non-adopters resemble adopters on a range of observables, including clinical practice and prescribing measures, except for lower baseline technology use at their clinics. When physicians use the AI tool, they ignore asymmetry in informativeness between positive and negative signals and, when shown both the AI and a redundant signal, exhibit correlation neglect. These frictions in information processing lead to increased antibiotic prescribing with the AI signal. Our findings highlight the importance of training and information design for AI implementation.

Keywords: expert decision-making, artificial intelligence, healthcare, mental models
JEL codes: I11, D81, D83, J24, O33

*We thank Nikhil Agarwal, Nicola Fuchs-Schündeln, Jon Kolstad, Anders Munk-Nielsen, Ziad Obermeyer, Mujaheed Shaikh, Sebastian Schweighofer-Kodritsch, Heiner Schumacher, Janus Laust Thomsen, Hans-Martin v. Gaudecker, Tobias Werner, Joachim Winter, and seminar participants at Max-Planck-Institute for Human Development in Berlin, University of Copenhagen, WZB Berlin, the Danish National Center for AI in Society (CAISA) Research Seminar, the Digital Economy Workshop in Athens, the symposium "DTx, AI, and Data" at FU Berlin, the NBER Conference on Applications of AI in Healthcare, the Digital Health and Precision Medicine Conference in Toulouse, the American-European Health Economics Workshop in 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 their expertise on UTI diagnosis and treatment in primary care in the design of the survey experiment. We are very grateful to Pernille Bang, Ivan Lyngsaa, Johan Sæverud, and the Center for Economic Behavior and Inequality for invaluable help in setting up and running the survey, as well as 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

Recent advances in artificial intelligence (AI) have raised expectations of substantial productivity gains across a wide range of tasks, but existing evidence is concentrated in settings where tasks are relatively general or routine. Studies on customer support, professional writing, and software development show that AI assistance increases both the quantity and the quality of output, especially among less experienced and lower-skilled workers (Noy and Zhang 2023; Peng et al. 2023; Brynjolfsson et al. 2025). Evidence from broader workplace adoption finds smaller productivity effects and no earnings gains (Humlum and Vestergaard 2025). Less is known about high-stakes expert tasks, where decisions carry significant consequences and errors are costly. In these settings, professionals have substantial domain knowledge and potentially strong prior beliefs. If they hold private information unavailable to the AI, deviations from algorithmic recommendations do not necessarily reflect errors (Rambachan 2022). Optimal decisions require combining private information with algorithmic signals. The value of AI therefore depends on how professionals weigh algorithmic recommendations against their own assessments and how expert judgment is influenced by AI, yet evidence on this issue remains limited.

Primary healthcare is one such expert setting, where timely diagnostic information is noisy and scarce. By extracting and evaluating diagnosis-relevant signals from complex patient characteristics, clinical histories, and electronic health records, AI-based diagnostic support promises productivity gains through improvements in the quality of clinical decision-making (Bayati et al. 2014; Lin et al. 2019; Mullainathan and Obermeyer 2022; Shapiro Ben David et al. 2025). Despite rapid technological progress and growing policy interest, little is known on how human decision-makers integrate new, AI-generated signals with existing sources of information. A small body of experimental work suggests that experts fail to realize the positive potential of AI tools because they disagree with its recommendations or fail to use the information provided correctly (Agarwal et al. 2023; Stevenson and Doleac 2024).

This paper contributes new evidence by studying how physicians update diagnostic beliefs and treatment decisions when assisted by AI, relative to a familiar and well-understood diagnostic test. We study primary care diagnosis and treatment of urinary tract infections (UTIs), one of the most common conditions encountered in routine practice and characterized by substantial diagnostic uncertainty.[1] Patients often present with overlapping or nonspecific symptoms, while definitive confirmation is possible but delayed by several days under current diagnostic technologies. This delay creates a need for faster diagnostic information, which AI tools may help provide (Yelin et al. 2019; Huang

[1] UTIs are highly prevalent and impose substantial healthcare costs, with nearly half of women experiencing at least one episode during their lifetime (Foxman 2002; Flores-Mireles et al. 2015).

et al. 2022; Ribers and Ullrich 2024). Such need is particularly relevant for suspected UTIs: antibiotic treatment is effective for bacterial infections but inappropriate for many 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). The UTI setting thus provides a natural environment to study belief updating, treatment choices under uncertainty, and the integration of multiple noisy diagnostic signals.

Crucially, the UTI setting offers a natural diagnostic benchmark, the urine dipstick test, to compare against the AI tool. To make a UTI diagnosis, primary care physicians during an initial consultation rely on patient-reported symptoms, medical history, and point-of-care tests such as urine dipsticks. Dipstick tests provide immediate but imperfect information and are familiar to physicians 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 may appear as a black box. We exploit this contrast in an experimental setting that compares how physicians update their beliefs when receiving information from a traditional diagnostic test versus an AI tool, while holding signal quality constant.

We conduct a nation-wide survey experiment with 372 primary care physicians in Denmark, representing 21.5% of all primary care clinics. Participants make diagnostic and prescribing decisions for standardized UTI cases after sequentially observing patient symptoms and a diagnostic signal. The signal is described either as a standard point-of-care dipstick test or as the output of an AI tool predicting whether the patient has a bacterial UTI as would be later confirmed by laboratory tests. The experimental design allows us to measure differences in belief updating, assess deviations from Bayesian updating, and study how physicians combine information from distinct sources. We also assess the relevance of correlation neglect, a common behavioral error in processing information from multiple correlated sources (Enke and Zimmermann 2019). The experimental design allows us to vary whether AI and dipstick signals provide independent information, or whether the AI signal already incorporates the dipstick signal, inducing signal correlation.

We find that physicians update their diagnostic beliefs less in response to AI signals than to traditional dipstick test results, even when both signals are equally informative. In fact, roughly one-third of physicians do not meaningfully update to the AI signal. Linked administrative data show that AI adoption is more common in clinics with greater overall technology use, but is otherwise unrelated to clinic characteristics and to measures of practice styles, physician preferences, prescribing behavior, or the quality of care. 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. We find that adoption

is predicted by physicians' acknowledgment that the tool uses more data than they have available, as well as by its statistical properties; non-adoption, by contrast, is predicted by skepticism toward the tool and concerns about its lack of transparency.

Examining how physicians incorporate the AI signal relative to the dipstick signal, we find evidence of two systematic deviations from Bayesian updating. First, we use the fact that both signals have asymmetric informativeness: a negative signal is more informative than a positive one. Physicians account for this asymmetry with the dipstick, but not with the AI signal. Although they update approximately as Bayesians after a positive AI signal, they significantly underweight negative AI signals, effectively treating both signal realizations as equally informative. Second, even when the correlation between the AI and dipstick signals is explicitly communicated, physicians exhibit correlation neglect by updating to correlated signals almost exactly as much as to uncorrelated ones, corroborating a core result in Agarwal et al. (2023).

These belief-updating deviations also affect treatment decisions. Consistent with the failure to account for asymmetry in the AI signal, physicians are more likely to prescribe antibiotics after a negative AI signal than a dipstick signal; after positive signals, prescribing does not differ by signal source. Consistent with correlation neglect, physicians are also more likely to prescribe after seeing both a positive AI signal and a positive dipstick result than after seeing only the positive AI signal, even though the dipstick provides no additional information in this setting.

These findings have implications for implementing AI tools with humans in the loop. Training and experience play an important role for humans to correctly interpret information provided by new diagnostic technologies. One approach to designing AI may be to include multiple diagnostic signals in its recommendations without revealing these individual signals, hence avoiding the need for humans to extract information that is difficult to interpret and synthesize. This would imply a departure from current clinical practice and counter explainability efforts. An alternative approach is to design AI tools to provide only diagnostic information that are complementary to any information physicians already have access to. Under such a design, transparency, explanation, and training become crucial.

Our findings speak to the growing economics literature on human–AI collaboration and complementarity (Agrawal et al. 2022; Mullainathan and Obermeyer 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. 2026; Humlum and Vestergaard 2024; Brynjolfsson et al. 2025). Our setting features physicians as high-skilled experts for whom AI promises an expansion of the productivity frontier by making complex, high-dimensional information cheap to process (Kleinberg et al. 2018). We find that, despite potential productivity gains, AI adoption comes with frictions. Physicians respond less to AI-generated signals than to familiar diagnostic tools with equivalent statistical properties and many voice skepticism toward the tool.

Finally, by documenting systematic differences in how individuals respond to novel versus familiar information sources, our findings speak to the design and implementation of AI decision-support tools in professional settings (Bansal et al. 2019; Tschandl et al. 2020; Vaccaro et al. 2024; Zöller et al. 2025; Abaluck et al. 2026). We highlight that adoption does not only depend on predictive accuracy, but also on how users interpret algorithmic information and what mental models are at play.

The remainder of the paper is organized as follows. Section 2 describes our survey design and section 3 describe the data. Section 4 presents the results on AI signal adoption. Section 5 describes deviations from Bayesian updating, potential underlying mechanisms and consequences for treatment decisions. Section 6 discusses policy implications and section 7 concludes.

2 Survey Design

2.1 Context

We study AI diagnostic support in the context of urinary tract infection (UTI) diagnosis in primary care, a well-suited setting to examine expert use of AI diagnostic support.

UTIs are among the most frequent conditions in primary care, and often a patient's initial point of contact with the healthcare system. UTIs occur when bacteria infect the urinary tract, bladder, or kidneys. Symptoms such as dysuria, pain, and urinary urgency can be debilitating, and untreated infections may lead to serious complications, including sepsis and death. Nearly half of women experience at least one episode of UTI during their lifetime (Foxman 2002). In the United States, community-acquired UTI imposes estimated annual healthcare costs of $1.6–3.5 billion (Foxman 2002; Flores-Mireles et al. 2015).

Diagnosing UTI is a prediction policy problem (Kleinberg et al. 2015): conditional on a bacterial infection being present, the optimal action is to prescribe an antibiotic. The difficulty is that antibiotics are effective only against bacterial UTIs, while similar symptoms can arise from a wide range of alternative conditions for which antibiotics are ineffective or inappropriate, including sexually transmitted urethritis or vaginitis, noninfectious urethritis, early pyelonephritis, or fungal infections (Wilson and Gaido 2004;

Gupta and Trautner 2012). Prescribing for suspected UTI therefore trades off timely treatment against unnecessary antibiotic use.

Primary care physicians diagnose UTI under considerable uncertainty. Definitive confirmation by urine culture analysis, either in-practice or at specialized laboratories, takes two to four days to yield results; however, physicians have to make diagnostic and treatment decisions at the point of care. Physicians thus have to rely on symptoms, history, and rapid tests that are immediate but have imperfect accuracy. The most common point-of-care tests are urine dipstick tests for nitrites and leukocyte esterase, usually used jointly. Nitrite tests have high specificity (90%) but low sensitivity (47%), whereas leukocyte esterase tests have lower specificity (58%) but higher sensitivity (81%) (Chernaya et al. 2022). Throughout the experiment, we fix nitrite test results as negative; unless otherwise noted, we therefore use *dipstick test* to refer to the leukocyte esterase result. Because the leukocyte esterase test is more sensitive than it is specific, its false-negative rate is lower than its false-positive rate. A negative result is therefore the more informative realization, making the test better suited to ruling out infection than to ruling in.

In this setting, the relevant AI supports diagnostic classification rather than generating text or treatment recommendations. Generative AI or Large Language Models (LLMs) may affect care by reducing administrative burden or retrieving guidelines and medical knowledge. In contrast, the diagnostic support tool we study statistically processes patient data using Machine Learning to predict a bacterial infection. Such tools may be combined with generative AI, but their value still depends on the statistical properties of the diagnostic signal, including sensitivity, specificity, and complementarity with physicians' clinical information.

Such AI-based support for UTI diagnosis can add value through two channels. First, by using delayed urine culture results as a high-quality ground truth, a tool trained on large-scale administrative data can predict baseline infection risk from orders of magnitude more cases than any individual physician observes (Huang et al. 2022; Ribers and Ullrich 2024). Second, patient records contain information that is predictive of UTI risk, including recurrent infections, recent procedures, catheterization, kidney stones, pregnancy, and urinary tract abnormalities. This high-dimensional information may be difficult or time-consuming for physicians to process at the point of care and is largely *complementary* to their clinical information, such as patients' current symptoms, thereby adding informative value. This contrasts with settings such as radiology, where AI and physicians evaluate the same image information (e.g., Agarwal et al. 2023).

2.2 Survey Flow

The experimental survey has two main goals. First, we study whether and which physicians adopt a hypothetical AI diagnostic tool into their decisions, relative to a familiar diagnostic tool, when quality is held fixed. Second, we examine whether physicians exhibit systematic biases when processing information from the AI tool, in particular when combining it with other diagnostic signals.

To achieve these goals, we design a survey experiment involving 372 Danish primary care physicians, covering more than 20% of all Danish primary care clinics. We show in Section 3 that participating clinics are broadly comparable to the population of Danish primary care clinics on observable characteristics. The survey experiment consists of several stages, described below in the order presented to respondents. Figure 1 illustrates the survey flow, and Appendix D shows the wording of the survey questions.

Baseline questions. We first ask questions about respondents' background characteristics, such as age and gender. Respondents are also asked to report their beliefs about the sensitivity and specificity of dipstick tests before being provided with our fixed experimental values. Afterward, respondents enter the experimental part of the survey.

Experimental Stage 1: AI Tool vs Dipstick treatment. The main experimental stage consists of six patient vignettes. Each vignette describes a hypothetical patient presenting with symptoms typical of a urinary tract infection (UTI). Respondents assess the probability that a UTI is present and decide whether to initiate further diagnostics and antibiotic treatment.

Figure 2 shows an example vignette. Each vignette reports the patient's age and gender, primary symptoms, comorbidities, and active prescriptions. Table 1 lists the possible vignette attributes. Patient age is randomized independently across respondent-vignette observations and takes one of four values. Symptoms, which reflect baseline UTI risk, and comorbidities, which reflect the relative benefit of antibiotic prescribing, are balanced across vignettes.[2]

Each vignette is presented in two steps: pre-signal and post-signal. First, before the diagnostic signal, physicians evaluate the prior probability of UTI based on symptoms and patient background characteristics and decide whether they would prescribe antibiotics. Second, after observing the diagnostic signal, they estimate the posterior UTI probability,

[2]In three out of six vignettes, the patient has type 1 diabetes, which increases the risk of a complicated UTI without changing baseline UTI risk, thereby increasing the benefit of antibiotic prescribing holding diagnosis fixed. Our main regressions include both vignette fixed effects and patient-age fixed effects to account for systematic differences in belief updating associated with these characteristics.

decide whether to prescribe antibiotics, and indicate whether to seek additional diagnostics for the urine sample (culture or laboratory testing).

We randomly assign respondents, in a between-subjects design, to receive one of two diagnostic signals. Before seeing the first vignette, respondents receive a detailed description of the signal assigned to their treatment arm.

- In the **AI Tool treatment**, respondents receive a binary signal from a hypothetical AI support tool. We describe that the tool uses artificial intelligence to predict the outcome of urine culture tests from microbiological laboratories for bacterial urinary tract infections among primary care patients. We ask respondents to assume that the tool is clinically approved. We explain that the tool uses the patient's clinical history as data inputs, including prior general practice consultations, diagnoses, referrals, prescriptions, laboratory test results, hospitalizations, *as well as* the current dipstick test result.

- In the **Dipstick treatment**, respondents receive a binary urinary dipstick result indicating positive or negative leukocytes.[3]

In both treatments, we randomize the direction of the diagnostic signal, that is whether the result is positive or negative, at the respondent-vignette level. Both the AI and dipstick signals are reported to have a sensitivity of 80% and a specificity of 60%, calibrated to the clinical performance of leukocyte esterase dipstick tests in Danish primary care (Chernaya et al. 2022). The experiment thus compares physicians' responses to two signals with the same quality, a novel AI signal and a familiar dipstick test, in a setting where treatment decisions have to be made based on imperfect point-of-care diagnostic signals (see Section 2.1). Holding signal accuracy fixed implies that a Bayesian updater should respond identically in both treatments, providing a benchmark for testing whether physicians use a novel AI-based signal differently from an established diagnostic test, without requiring us to observe patients' true infection states.

For the AI tool, a positive signal corresponds to a predicted UTI probability above a threshold chosen to generate a sensitivity of 80% and a specificity of 60%.[4] We use a binary rather than continuous AI output for two reasons: first, the AI signal is directly comparable to many conventional diagnostic tests in medicine reported or interpreted as binary signals (e.g., positive or negative); second, it avoids requiring physicians to

[3]Because updating to two dipstick results would not be comparable to updating to a single AI signal, we use only the leukocyte result as the diagnostic dipstick signal. We prominently state that the nitrite result, which is typically used jointly with leukocytes in clinical practice (see Section 2.1), is negative before and after the signal in all treatments and vignettes to ensure that the absence of the nitrite result does not affect belief updating.

[4]Varying the threshold selects a point on the receiver operating characteristic (ROC) curve. Because specificity (60%) is lower than sensitivity (80%) in our setting, this would select a comparatively low threshold, reducing the rate of false negatives at the cost of a higher false-positive rate.

interpret intermediate risk scores, as Agarwal et al. (2023) show that such intermediate AI assessments can reduce decision quality.

Mental models. After Stage 1, we elicit respondents' mental models of the diagnostic tool in their treatment arm: the AI tool or the dipstick test. This survey block captures physicians' subjective understanding of how the respective tool works, following recent work that measures mental models using surveys (Andre et al. 2022, 2024; Haaland et al. 2025).

We first ask respondents, in a free-text field, which factors they considered when deciding whether to incorporate the diagnostic signal into their decisions. We elicit these open-ended responses before any closed-form items to avoid priming, and later categorize them (Haaland et al. 2025). Afterwards, participants indicate their agreement with seven closed-form statements. Each statement is evaluated on a Likert scale from 1 (strongly disagree) to 5 (strongly agree). Appendix C.1 provides details on the construction of the mental model measures.

Experimental Stage 2: Correlated and uncorrelated signals. In Stage 2, physicians evaluate three additional vignette cases with low baseline UTI risk (weak symptoms) and low prescribing benefit (non-diabetic patients). In contrast to the first stage, respondents observe *both* diagnostic signals in these three vignettes, the dipstick test and the AI tool, and we randomize between-subjects whether the two signals are independent or correlated.

- In the **Uncorrelated treatment**, respondents are informed that the AI tool does *not* incorporate the dipstick result, so that the two signals provide independent information.
- In the **Correlated treatment**, respondents are informed that the AI tool *does* incorporate the dipstick result. If the tool uses that input optimally, the dipstick provides no information beyond the AI signal, so a Bayesian would not update further on the dipstick once the AI signal has been observed. Any additional updating in response to the dipstick therefore reflects correlation neglect.

We hold signal accuracy fixed across the Correlated and Uncorrelated treatments. As shown in Figure 1, signal accuracy depends on the respondent's first-stage treatment arm to reflect that an AI tool using both dipstick results *and* additional patient data should be more accurate than the dipstick alone.[5] In the Uncorrelated treatment,

[5]For respondents previously assigned to the AI Tool treatment, the AI tool retains its Stage 1 accuracy, while the additional dipstick signal has sensitivity of 70% and specificity of 50%. For respondents previously assigned to the Dipstick treatment, the dipstick retains its Stage 1 accuracy, while the AI tool has sensitivity of 90% and specificity of 70%.

respondents are told that the AI tool is trained on a larger dataset, allowing it to maintain the same accuracy without incorporating the dipstick result.

For simplicity, the two signals always point in the same direction: both are either positive or negative. Since respondents observe only three vignettes in this stage, we consider it unlikely that respondents learn about the correlation structure from the realized signals.

Additional questions. The final block includes questions about respondents' attitudes toward AI and broader behavioral preferences. We ask about respondents' preferred presentation of the AI tool and dipstick test, and elicit their willingness to pay for the AI tool. We also ask about respondents' familiarity with generative AI and LLMs, and their perceived usefulness of generative AI in clinical practice. Finally, we elicit preferences related to antimicrobial resistance, along with general risk preferences, time preferences, and social preferences toward in-group relative to out-group patients.

3 Data

Our main data come from the survey experiment described above, conducted among primary care physicians in Denmark. We link responding clinics to national administrative registers containing information on clinical practice, prescriptions, patient populations, and physician characteristics. These linked data allow us to characterize the analysis sample and compare it to the population of Danish primary care clinics.

Participant recruitment. We implemented the survey among primary care physicians in Denmark at the end of 2025.[6] We invited all physicians from Danish outpatient primary care clinics listed in a national register maintained by MedCom, a publicly funded, non-profit organization owned by the Danish Ministry of the Interior and Health, Danish Regions, and Local Government Denmark (MedCom 2025).

We recruit respondents by sending survey invitations to each clinic's official digital mailbox (Digital Post). Since 2013, Digital Post has been the legally mandated channel used by all medical clinics in Denmark for secure communication with public authorities and businesses, such as laboratories and insurers.[7] It thus constitutes the standard and universal channel for official correspondence with primary care clinics and ensures full population coverage.

[6] The survey was open from 30 October to 10 December 2025. We sent reminders on 18 November and on 2 December.

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

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).

Out of 1607 invitations sent, 383 physicians from 356 clinics completed the survey, corresponding to a response rate of about 22% of all clinics.

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.[8] 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. This suggests that respondents do not systematically overestimate dipstick accuracy and therefore discount the AI tool's accuracy in comparison. If anything, respondents' stated beliefs imply that the experimental signals should appear more accurate than expected.

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.

Representativeness of the analysis sample. To examine the representativeness of respondents, we link clinics in the analysis sample to national administrative registers

[8] We preregistered that we will exclude participants who repeatedly update in the wrong direction.

covering 2024 and 2025, maintained by Statistics Denmark and the Danish Health Data Authority. Appendix C.2 describes the administrative data sources in detail.

Out of the 346 analysis clinics, we match 333 clinics to the administrative data for information on their clinical practice in 2024, including consultations, diagnostic use, antibiotic prescriptions, and patient pools. For a subset of 320 clinics, we additionally observe clinic characteristics, including location, size, and average physician characteristics.

Table 3 compares clinics in the analysis sample to the population of Danish primary care clinics observable in the administrative data. To assess the magnitude of differences between the analysis sample and the population, we report standardized mean differences (SMDs, also referred to as normalized mean differences). As a rule of thumb, SMDs with absolute values above 0.25 are commonly interpreted as indicating potentially problematic imbalance (Imbens and Rubin 2015; Baker et al. 2026). Across all observable characteristics, SMDs are below 0.25 in absolute values, suggesting that the analysis sample is broadly comparable to the population of Danish primary care clinics on observable characteristics.

While we cannot rule out selection into the survey experiment on unobservables, we find little evidence of large observable differences between participating clinics and the population. We assess the relevance of remaining observable differences in robustness checks using population weights based on clinic characteristics and by restricting the sample to single-physician clinics.

4 AI signal adoption

We first examine whether physicians' diagnostic beliefs respond to signals from the AI support tool relative to the dipstick test, using the results from Experimental Stage 1. We begin with a descriptive analysis of absolute belief updating, measured as the average difference between posterior and prior beliefs induced by each signal. We then characterize individuals who consistently do or do not incorporate the AI signal into their posterior beliefs, based on two sets of predictors: (1) survey-elicited mental models of AI adoption; and (2) clinical practice, clinic characteristics, and antibiotic prescribing intensities constructed from the administrative data and additional survey questions. We defer the analysis of differential effects on antibiotic treatment decisions to Section 5, where we examine how physicians use the AI signal.

4.1 Average belief updating

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

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

where $|(P - P_0)|_{iv}$ denotes the absolute change from prior to posterior belief for physician i in vignette v, α_v are vignette fixed effects, and $\lambda_{a(iv)}$ are fixed effects for the patient age assigned in vignette v. The vignette fixed effects absorb differences across symptoms and comorbidities in the vignette case descriptions, while the patient-age fixed effects account for the separately randomized patient age shown in each vignette. β is the average treatment effect of presenting a diagnostic signal as coming from the AI Tool rather than a Dipstick test on belief updating.

Figure 3 plots a histogram of 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.

The figure also displays the estimated treatment effect on absolute belief updates. The AI Tool treatment reduces average belief updates by 7.63 percentage points (significant at the 1% level) compared to the Dipstick treatment (41% from a baseline of 18.48). Physicians thus update substantially less when the signal comes from the AI support tool.

Robustness By design, the AI and dipstick signals have identical diagnostic accuracy, so the weaker updating in the AI Tool treatment should reflect physicians' response to the signal *source*. A natural concern is that it instead reflects a misunderstanding of the accuracy values we provide, or a perception that the AI tool is of low quality. Several pieces of evidence speak against this.

First, physicians appear well informed about the dipstick's diagnostic properties: average reported beliefs about its sensitivity and specificity are close to, if slightly below, the values fixed in the experiment (Table 2), and a substantial share of prior beliefs clusters around these values (Appendix Figure A1). Second, the estimated effect is nearly unchanged when we restrict the sample to physicians whose prior beliefs about the dipstick's sensitivity, its specificity, or both lie below the experimental values (Appendix Table A4, Columns 2–4). For these physicians, the AI represents an improvement over their own view of the dipstick, so a perception that the AI tool adds little over and above the dipstick alone cannot explain the underweighting we observe. Finally, physicians attend to the accuracy values we provide rather than disregarding them: Appendix Table A5 shows that

they update more strongly when the signal's stated accuracy is higher.[9]

The treatment effect is also robust to a range of alternative specifications (Appendix Table A4, Columns 5–9). We reweight the sample to match the population of Danish primary care clinics on clinical-practice, patient-pool, and clinic characteristics (Section 3); we restrict the sample to single-physician clinics, where responses can be linked to one physician; and we vary the set of controls, including survey-elicited risk preferences to assess whether the results are driven by differential risk aversion toward a novel tool. The estimated effects remain close to the baseline throughout; where the absolute point estimate declines, the sample is substantially smaller and the standard errors correspondingly larger.

Taken together, this evidence suggests that the treatment effect is unlikely to stem from a general inability to interpret sensitivity and specificity. Instead, it is consistent with physicians responding more cautiously to information from an AI tool than to a familiar diagnostic test of identical diagnostic quality.

4.2 Adopters and non-adopters

Classifying AI adoption. A noticeable pattern in Figure 3 is the mass of zero in the AI Tool treatment as opposed to the Dipstick treatment. Specifically, in 17.3% of AI signal cases, physicians do not update their beliefs at all, compared to only 6.4% in the Dipstick treatment. Similarly, in 36.3% of the AI signal cases, physicians update by less than three percentage points, compared to 17.9% in the Dipstick treatment.

To examine consistency in updating patterns across vignettes, we define individuals who update by less than three percentage points in the majority of vignettes as *non-adopters*.[10] Following this classification, we find that 34.1% of physicians update by less than three percentage points in the majority of vignettes in the AI Tool treatment, compared to only 6.7% of physicians in the Dipstick treatment. This pattern suggests that the weaker average response in beliefs to the AI signal is not only driven by physicians updating less to the AI tool. Instead, a substantial number of physicians do not update their beliefs in response to the AI signal. Below, we investigate differences in characteristics between AI adopters and AI non-adopters, as well as the potential mechanisms of why a large share of physicians disregard the AI signal altogether.

[9]The table shows absolute belief updating to Experimental Stage 2 by treatment assignment, leveraging that the sensitivity and specificity vary across participants. As explained in Section 2.2, the AI signals in the second stage vignettes either have a sensitivity of 80% and specificity of 60% when the AI signal was presented in the first stage, or a sensitivity of 90% and specificity of 70% when the dipstick signal was shown in the first stage. The 'Positive signal' updating coefficients are significantly higher when the signal has a sensitivity/specificity of 90%/70% than when it has a sensitivity/specificity of 80%/60% (Chow test, $p < 0.01$ for both Uncorrelated and Correlated treatment, relative to similar updating under a negative signal ('Constant').

[10]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.

Correlates of AI adoption. Figure 5 shows descriptive, bivariate correlations of physician and clinic characteristics with AI adoption in the experiment from combining the survey data with Danish administrative registers and additional survey items. The outcome in each row is an indicator for being classified as an AI adopter, and each coefficient comes from a separate regression on the listed characteristic. For continuous characteristics, the regressor is an indicator for whether the physician or clinic is above the sample median. Appendix Table A3 describes the construction of these variables.

Most strikingly, Panel 5a shows that AI adoption is related to clinics' general technological and diagnostic orientation. AI adopters are less likely to work in clinics with a high share of phone consultations, an outdated technology in Denmark, and more likely to work in clinics with high shares of email and video consultations, as well as higher overall consultation volume. They are also more likely to work in clinics with more microscopy use, a common additional diagnostic in the UTI context (see Section 2.1). In contrast, adoption is not meaningfully related to patient-pool characteristics, such as average patient age, the share of patients from Denmark, or the share of female patients.

In contrast, Panel 5b shows that AI adoption is not systematically related to physician demographics (including physicians' age or gender), clinic characteristics (including clinic location and staff size), or the respondent's general familiarity with generative AI.[11]

Finally, Panel 5c shows little evidence that AI adoption reflects differences in antibiotic prescribing quality or preferences. Adoption is not meaningfully correlated with physicians' survey-based prescribing thresholds, including thresholds adjusted for elicited preferences.[12] Appendix C.3 discusses how, under additional assumptions, the survey-based prescribing thresholds can be interpreted as proxies for diagnostic skill. Adoption is also not related to several quality measures based on observed antibiotic prescribing: overall antibiotic prescribing rates; second-line antibiotic use for UTI, a measure of guideline adherence based on the recommendations of the Danish College of General Practitioners (Dansk Selskab for Almen Medicin 2020); and avoidable hospitalizations for UTI, also referred to as ambulatory-care sensitive conditions, capturing hospitalizations that appropriate primary care may prevent (Carey et al. 2017). Finally, AI adopters and non-adopters do not differ in stated concern about antibiotic resistance or in the weight placed on own patients relative to others, our proxy for the cost of prescribing externalities.

Overall, these patterns suggest that AI adoption is more common in clinics that already make greater use of digital modes of care and diagnostic technologies. By contrast, adoption does not appear to reflect lower prescribing quality, weaker concern about

[11] We elicit physicians' attitudes toward generative AI directly in the survey, where we explain the concept to distinguish it from the diagnostic AI tool we study and give examples of LLMs.

[12] Appendix Figure A2 provides external validation for the survey prescribing measures: physicians with higher prescribing rates in the vignette experiment also tend to practice in clinics with higher observed UTI prescribing rates among comparable patients in the administrative data.

antibiotic resistance, or broader familiarity with generative AI. AI adoption may therefore reinforce existing differences in technology use across clinics, without necessarily reflecting differences in underlying antibiotic prescribing preferences or practice quality.

Mental models. To understand potential mechanisms behind AI adoption, we examine physicians' mental models of the AI tool. We view these mental models as (potentially misspecified) representations of the tool's 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).

Figure 4 summarizes the open-text comments from respondents in the AI Tool treatment. These comments describe the factors physicians considered when deciding whether to incorporate the AI signal into their diagnostic assessment. The left panel reports the share of physicians mentioning each factor. The right panel reports, for each factor, the coefficient from a separate regression of the AI adoption indicator on an indicator for mentioning that factor.

The main pattern is that AI adopters emphasize the informational and statistical properties of the AI signal, whereas AI non-adopters express general concerns about the tool and AI technologies. Physicians are more likely to adopt the AI signal when they mention that the tool provides additional information or uses more data than they have at the point of care, or when they refer to its accuracy. By contrast, non-adoption is predicted by general concerns about the tool and its transparency. Physicians who express skepticism towards the tool, its results, or AI technology in general or a lack of transparency in how the signal is generated, are less likely to update their beliefs in response to the AI signal.[13]

Notably, none of the respondents mention accountability or liability concerns, which are commonly raised in broader discussions about AI adoption. This may reflect the institutional setting in Denmark, where medical devices approved for use by the authorities do not generate individual liability issues, and malpractice litigation is practically absent compared to settings such as the United States.[14] In our setting, AI non-adoption may reflect less legal concerns and more uncertainty about whether the AI tool contains useful diagnostic information and processes this information appropriately.

[13]We classified responses as skepticism towards the tool if physicians mention that they have no trust in the AI tool and the data/statistics it is based on. Examples include: "I don't trust its validity", "I don't trust AI yet", or "It it not a real tool but a computer program based on data and not clinical practice".

[14]In particular, medical-injury compensation in Denmark does not assign responsibility or sanction individual health professionals (Patienterstatningen n.d.). Individual sanctions are handled through a disciplinary system and are also rare, with authorization withdrawn by court judgment for four physicians and 75 sanctions involving physicians recorded in total in 2024 (Styrelsen for Patientsikkerhed 2026).

Because the open-text comments were elicited before the closed-form mental model statements, they capture physicians' unprompted reasoning about the tool. Appendix Figure A3 shows that the closed-form statements point to similar mechanisms: AI adopters are more likely to view the AI tool as based on informative data and as processing this information appropriately.

5 Deviations from Bayesian Updating

The previous section showed that many physicians update their beliefs less when a diagnostic signal comes from an AI tool compared to a dipstick signal, even when both are equally informative by our experimental design. We next compare physicians' belief updating to the Bayesian benchmark and examine two systematic deviations from Bayesian updating: first, whether they take into account signal asymmetry; and second, whether they account for signal correlation. Such deviations are policy-relevant because they inform how AI support should be designed and implemented.

5.1 Empirical strategy: Belief updating

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

$$\log\left(\frac{P}{1-P}\right) = \log\left(\frac{P^0}{1-P^0}\right) + \log\left(LR^{\text{signal}}\right),$$

where P is the posterior, P^0 is the prior, and LR^{signal} is the likelihood ratio of the signal, measuring the relative probability of observing that signal when the patient has a UTI versus when she does not. We estimate the following empirical analogue:

$$\log\left(\frac{P_{iv}}{1-P_{iv}}\right) = \delta^{\text{prior}} \cdot \log\left(\frac{P^0_{iv}}{1-P^0_{iv}}\right) + \beta^{\text{pos}} \cdot \log\left(LR^+_{iv}\right) + \beta^{\text{neg}} \cdot \log\left(LR^-_{iv}\right) + \varepsilon_{iv}, \quad (2)$$

where δ^{prior} captures the weight placed on the prior, while β^{pos} measures responsiveness to a positive signal and β^{neg} to a negative signal. The term ε_{iv} captures idiosyncratic updating errors. In our setup, the likelihood ratio depends on the direction of the signal: a positive signal has $LR^+ = 2$ and a negative signal has $LR^- = 1/3$.[15]

[15] For a positive signal, the likelihood ratio is $\text{LR}^+ = \frac{\text{sensitivity}}{1-\text{specificity}}$. For a negative signal it is $\text{LR}^- = \frac{1-\text{sensitivity}}{\text{specificity}}$.

For a Bayesian updater, the coefficients δ^{prior}, β^{pos}, and β^{neg} equal one. Coefficients below one indicate underupdating, that is, placing less weight on the prior or signal than Bayes' rule prescribes. Coefficients above one indicate overupdating. A difference between β^{pos} and β^{neg} reflects asymmetric updating. Because negative signals are relatively more informative ($|\log(LR^-)| > \log(LR^+)$), a Bayesian updater would react more strongly to them. If physicians fail to account for this asymmetry, they place too little weight on negative signals relative to positive ones, implying that $\beta^{\text{neg}} < \beta^{\text{pos}}$.

5.2 Neglect of the signal asymmetry of the AI signal

In this section, we study how physicians' belief updating compares to the Bayesian benchmark, and whether deviations from that benchmark have consequences for antibiotic prescribing.

Belief updating Table 4 reports the estimated updating coefficients from Equation 2 by treatment arm. We show results for both the overall sample, as well as for the subsample of signal adopters, with adoption defined as in Section 4.2.

For the dipstick signal (Column 1), the updating coefficients for positive and negative signals are statistically indistinguishable from one another ($\beta_{pos} = 1.07$, $\beta_{neg} = 1.12$). Both lie slightly above but close to the Bayesian benchmark of one, with only β_{neg} statistically different from one ($p < 0.05$). Crucially, the positive and negative signal realizations are not equally informative. The negative signal is more informative because the leukocyte esterase dipstick has higher sensitivity than specificity and is therefore better suited as a rule-out than as a rule-in test. Thus, identical updating coefficients for both signal realizations imply that physicians' posteriors move more after the more informative signal. In other words, physicians correctly account for the asymmetry of the dipstick signal.

For the AI tool (Column 2), belief updating patterns differ markedly. Both updating coefficients are substantially and significantly below one ($\beta_{pos} = 0.87$, $\beta_{neg} = 0.55$), consistent with our earlier finding that many physicians do not adopt the AI signal. In addition, updating to the negative AI signal is significantly weaker than updating to the positive signal. In fact, the effective weight physicians place on the two signals is essentially identical.[16]

To assess whether the failure to incorporate the AI signal's asymmetry is driven by non-adopters, we restrict the sample to adopters of the respective signal (Columns 3 and

[16]The change in the log-odds of the posterior attributable to a signal equals the updating coefficient times the signal's log-likelihood ratio. For the AI tool this gives $\beta_{pos} \cdot \log(LR^+) = 0.87 \cdot 0.69 \approx 0.60$ for a positive signal and $\beta_{neg} \cdot |\log(LR^-)| = 0.55 \cdot 1.10 \approx 0.61$ for a negative signal. Although the negative signal is objectively more informative ($|\log(LR^-)| > \log(LR^+)$), physicians move their posterior by roughly the same amount in either direction.

4). Among adopters, physicians update just as strongly to the positive dipstick signal as to the positive AI signal. This suggests that the underadoption of the AI tool is driven mainly by AI non-adopters rather than by attenuated updating among physicians who do engage with the signal. However, AI adopters still do not account for the AI signal's asymmetry: even in this subsample, updating to negative AI signals is significantly weaker than to positive ones.

Prescription decisions We next ask whether physicians' underupdating in response to negative AI signals carries over to antibiotic prescribing. Antibiotics are the first-line treatment for UTIs, which account for a substantial share of antibiotic prescriptions in primary care. Even modest changes in how physicians respond to diagnostic signals can therefore have meaningful implications for antibiotic stewardship.

We remain agnostic about whether prescribing decisions in the survey experiment are correct or not, as doing so would require assigning infection status to the vignette cases and thus imposing strong assumptions about the underlying ground truth. Instead, we focus on 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 decrease antibiotic prescribing relative to dipstick testing.

Table 5 reports the results. Columns 1 and 2 compare empiric antibiotic prescription rates after positive and negative signals across the AI TOOL and DIPSTICK treatments. After a positive signal, the difference is not statistically significant: prescriptions are approximately 14% higher after an AI than after a dipstick signal. After a negative signal, however, prescriptions are 31% higher after an AI than after a dipstick signal ($p = 0.09$). Columns 3 and 4 estimate the treatment effect on a binary indicator for whether the prescription decision moves from pre- to post-signal in the direction the signal implies. There is no difference between AI and dipstick in initiating a prescription after a positive signal ($p = 0.922$), but physicians are 38% (15 percentage points) less likely to *stop* prescribing after a negative signal under AI than under dipstick ($p = 0.037$).

Taken together, these prescription results mirror the belief updating results. Physicians underappreciate the negative signals provided by the AI tool relative to the dipstick signal. This leads them to adjust their diagnostic beliefs less and leads them to prescribe more antibiotics when the signal points against a bacterial infection.

Mechanisms We interpret the contrast between the two signals in accounting for signal asymmetry as reflecting that a trained interpretive rule for the familiar test is available but not for the novel AI signal. Clinicians are explicitly taught to translate a test's operating characteristics into differential diagnostic conclusions: mnemonics such as *SnNout* (a

highly *sensitive* test, when *negative*, rules a condition *out*) and *SpPin* (a highly *specific* test, when *positive*, rules it *in*) encode exactly the principle that the two realizations of a test can carry different diagnostic weight. For the dipstick, this rule is well established and reinforced in medical training in the context we are focusing on ("leukocytes rule out"). Our results show that simply providing the sensitivity and specificity values in the AI treatment is insufficient to ensure correct updating. In the absence of an established heuristic, physicians move their posterior by roughly the same amount in either direction, as if positive and negative AI signals were equally informative.

Our evidence speaks against two alternative mechanisms. First, physicians may be cognitively imprecise about signal strength. Augenblick et al. (2025) show that people overinfer from weak signals and underinfer from strong ones. In isolation, the results from the AI treatment are consistent with this pattern: physicians underinfer less from the less informative positive signal than from the more informative negative one. However, this mechanism cannot account for the comparison between dipstick and AI treatment. Both AI and dipstick signals have identical signal strengths ($LR^+ = 2$, $LR^- = 1/3$), so a cognitive force that compressed perceived informativeness would have to affect both signal sources identically. Yet, physicians fully reflect the dipstick's asymmetry ($\beta_{neg} = 1.12$, if anything above one), while only compressing the AI signal's asymmetry.

Second, the AI tool's novelty may generate ambiguity about its reliability, prompting physicians to discount its signal and to update cautiously. The simplest version of this mechanism is uniform discounting: physicians treat the AI as an unfamiliar source and respond less to its signals overall. While this mechanism may explain underupdating to a negative AI signal, it is contradicted by results from the adopter subsample (Table 4, Columns 3 and 4), where physicians update to positive AI signals as strongly as to positive dipstick signals, suggesting that adopters are not cautious to signals from the AI tool overall. A more nuanced possibility is that ambiguity regarding the new signal interacts with the fear of missing an infection. In that case, physicians would understand the AI signal's asymmetry but deliberately override it. However, such a directional safety motive should make physicians respond less to the reassuring negative signal than to the positive signal, rather than producing similar responses to both signal realizations. Instead, the observed posterior movements are nearly identical after positive and negative AI signals, which is more naturally explained by a failure to register the asymmetry, and physicians behaving as if the two AI signal realizations were equally informative.

5.3 Correlation neglect 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 multiple pieces of information, many of which the physician may also observe directly. This raises scope for correlation neglect: physicians may treat multiple signals as independent even when one signal partly incorporates the other. In this section, we test whether physicians exhibit correlation neglect in their belief updating and prescribing decisions.

Belief updating In Experimental Stage 2, respondents observe both an AI *and* a dipstick signal, and we randomize whether the two are correlated. In the CORRELATED treatment, the AI 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 AI and dipstick signals can be treated as separate sources of information. Comparing both treatment arms does not require knowing the exact correlation structure of the AI and dipstick signals: The information structure is mechanically more informative in the UNCORRELATED treatment, where the AI tool and the dipstick test provide independent information, than in the CORRELATED treatment, where the dipstick signal is incorporated in the AI tool already. A Bayesian updater should thus respond more strongly in the UNCORRELATED treatment.

Figure 6 shows that physicians update their beliefs very similarly under correlated and uncorrelated signal structures; 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.[17] 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 (2) separately for correlated and uncorrelated signals within a Bayesian updating framework (see Section 5.1). The likelihood ratio of the signal differs across treatment arms: In the AI TOOL treatment arm, the joint likelihood ratio (LR) with uncorrelated signals is $LR^{++,unc} = 2.8$ and $LR^{--,unc} = 1/5$; while it is only $LR^{++,cor} = 2$ and $LR^{--,cor} = 1/3$ with correlated signals. In the DIPSTICK treatment arm, the joint likelihood ratio with uncorrelated signals is $LR^{++,unc} = 6$ and $LR^{--,unc} = 1/21$, whereas it is only $LR^{++,cor} = 3$ and $LR^{--,cor} = 1/7$ with correlated signals.[18] Hence, holding signal accuracy fixed, the correlated signals are generally less informative. This framework therefore benchmarks updating against the information implied by the correlation structure.

Second, we estimate a simple specification, analogous to Equation (1), in which we

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

[18] With two independent signals, the joint likelihood ratio is given by the product of the likelihood ratios of both signals. When the AI tool incorporates the dipstick signal, we make the simplifying assumption that it optimally uses the information provided by the dipstick test, so that the joint likelihood ratio is determined by the AI tool's likelihood ratio.

regress the absolute belief change on signal direction, separately for the CORRELATED and UNCORRELATED treatment:

$$|(P - P_0)_{iv}| = \alpha_v + \beta \cdot \text{Positive Signal}_{iv} + \varepsilon_{iv}. \tag{3}$$

This second specification allows us to compare the magnitude of belief updating across correlation conditions without imposing a Bayesian structure, while still accounting for whether the signal is positive or negative.

Table 6 reports the results for both estimation approaches, separately for the correlated and uncorrelated treatment arms. Panel A shows the results for the Bayesian updating regression based on Equation (2). The updating coefficients β^{pos} and β^{neg} are substantially smaller in the UNCORRELATED than in the CORRELATED treatment arm (Chow test, $p < 0.01$ for both positive and negative signals). This pattern is consistent with correlation neglect: respondents do not sufficiently adjust their updating to reflect different informativeness implied by the signals' correlation structures. In fact, the belief movement implied by the updating coefficients is almost the same in the UNCORRELATED treatment compared to the CORRELATED treatment.[19]

Panel B of Table 6 shows the same pattern directly using the absolute belief-updating specification in Equation (3). The updating coefficients are *almost identical* across correlation conditions. This pattern corroborates that individuals update similarly in both conditions, even though uncorrelated signals provide substantially more information.

Prescription decisions Next, we study whether correlation neglect also impacts antibiotic prescription decisions. Since signals in the UNCORRELATED treatment are substantially more informative than those in the CORRELATED treatment, responses in line with the Bayesian benchmark imply that prescribing decisions should adjust more strongly in the UNCORRELATED treatment. Table 7 shows regression results of empiric antibiotic prescribing on treatment arm indicators. Contrary to the Bayesian benchmark, but consistent with our belief updating results, we find no significant treatment effect of the CORRELATED treatment on antibiotic prescriptions. Physicians consider the correlated and uncorrelated signals equally informative for their treatment decisions.

Finally, comparing the first and second stages of the experiment yields a tight within-physician test of whether correlation neglect affects prescription decisions. We focus on vignettes with identical symptom descriptions in which physicians saw an AI signal

[19] To show that belief movements are almost identical across correlated and uncorrelated treatments for both positive and negative signals, we can calculate the change in the log-odds of the posterior, which is given by the product of the updating coefficient and the (average) signal's log-likelihood ratio. For positive signals, this gives $\beta_{pos,cor} \cdot \log(LR^{+}) = 1.28 \cdot 0.89 \approx 1.14$ in the CORRELATED treatment and $\beta_{pos,unc} \cdot |\log(\overline{LR^{--}})| = 0.76 \cdot 1.43 \approx 1.09$ in the UNCORRELATED treatment. For negative signals, it gives $\beta_{neg,cor} \cdot \log(\overline{LR^{++}}) = 0.91 \cdot 1.54 \approx 1.40$ in the CORRELATED treatment and $\beta_{neg} \cdot |\log(LR^{-})| = 0.57 \cdot 2.37 \approx 1.35$ in the UNCORRELATED treatment.

in the first stage and an equally informative AI signal in the second stage. The only difference is that in the second stage they additionally received a dipstick signal. Because the dipstick information is already embedded in the AI signal, this dipstick signal is redundant and conveys no new information.[20] A physician who correctly accounts for the correlation between the two signals should therefore leave their prescription decision unchanged, whereas a physician who neglects this correlation will treat the dipstick result as independent confirmation and adjust their decision in the direction of the (redundant) signal.

Table 8 shows in Column 1 that physicians prescribe significantly more antibiotics after seeing the positive but redundant dipstick signal ($p < 0.05$). Prescriptions almost double relative to seeing the AI signal alone, even though the dipstick adds no information. Thus, physicians treat the redundant signal as independent confirmation and update as if it were genuinely new evidence.

Note that the design imposes an asymmetry on what we can detect. Because the second-stage vignettes are all low-probability cases, prescribing rates in the first stage are low to begin with, leaving ample room to detect an *increase* in prescriptions but little room to detect a *decrease* (Column 2). We are therefore well-powered to detect the upward response we find, but cannot rule out downward responses in other parts of the case distribution. With that caveat, the results provide clear proof-of-concept evidence that correlation neglect affects treatment decisions.

Mechanisms We interpret the updating patterns above as reflecting a genuine belief that the dipstick signal carries information beyond what the AI tool already conveys, even though the dipstick result is one of the inputs the AI signal is built on. Under this interpretation, physicians treat the two signals as if they were (at least partly) independent sources of evidence and therefore fail to discount the redundant component. Supporting evidence for this interpretation comes from physicians' own stated preferences (Appendix Figure A4). A large majority (more than 75%) report that they consider it important to see the dipstick signal separately, even though the dipstick result is already incorporated into the AI tool's assessment. Notably, this is not driven by a belief that they can interpret the raw dipstick better than the algorithm (fewer than 30% agree with this statement).

Two alternative explanations could reproduce these updating patterns without reflecting genuine correlation neglect, and we rule out both. First, the result could be an artifact of how the second, additional signal is introduced, that is, confusion about the new signal, rather than a failure to account for correlation. However, Appendix Table A5 shows that

[20]The AI signal has a sensitivity of 80% and a specificity of 60% in both stages. In the second stage, physicians additionally observe a dipstick signal (sensitivity 70%, specificity 50%), but because the dipstick result is one of the inputs the AI tool already uses to generate its signal, it carries no information beyond what the AI signal conveys.

correlation neglect is present regardless of whether the respondent saw the AI or the DIPSTICK treatment in the first stage. This robustness check confirms that the result does not depend on the way the "new" additional signal is introduced, corroborating that it is a robust failure to fully account for the correlation between signals.[21] Second, the patterns 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. Appendix Table A6 addresses this by excluding non-adopters, as classified from the first stage of the experiment; the estimates remain very similar, indicating that our evidence is not driven by participants disregarding the AI signal altogether.

6 Discussion and Policy Implications

The documented frictions in how primary care physicians incorporate AI-generated diagnostic information in the experiment have implications for how AI decision-support may need to be designed and accompanied by training.

Observing that physicians' belief updating is systematically related to their mental models of whether the tool extracts diagnostic information from patient data without bias, adoption may be improved by addressing these beliefs rather than emphasizing predictive accuracy alone. Furthermore, in the experiment, physicians explicitly state that they prefer to see both the AI signal and the dipstick signal separately, while asserting that they cannot interpret the dipstick signal better than the AI. Coupled with our finding of correlation neglect, this need for transparency poses a dilemma. If the human in the loop sees a collection of individual signals but processes these less than optimally, transparency would need to be coupled with detailed and potentially complex explanations. Such solutions might prove challenging as growing experimental evidence finds that efforts to provide transparency, explainability, and feedback can be hampered by cognitive biases (Poursabzi-Sangdeh et al. 2021; Bauer et al. 2023; Serra-Garcia and Gneezy 2025; Von Zahn et al. 2025; Chan 2026).[22]

In the context of antibiotic prescribing, where a false positive includes a public health cost component, the implications of AI adoption are particularly policy-relevant. Because, in the experiment, both the dipstick and the AI test have a higher false positive than false

[21] Section 2.2 describes the specific ways the new additional signal is explained.

[22] An alternative design, though implying a departure from current diagnostic practices and reducing transparency, would let an AI tool summarize multiple diagnostic signals for the physician. In some situations, this might help reduce physicians' time required to arrive at an initial diagnosis and treatment decision, while effectively reducing overuse of antibiotics. This would resemble an implementation which was trialed by the UK National Health Service for cases of suspected UTI in 2019. Here, a smartphone app allowed patients to obtain an antibiotic prescription based on symptom reports and a dipstick result without seeing a physician. Unfortunately, this pilot study lacked an evaluation of prescription quality and health outcomes. See Thornley et al. (2020) and https://www.bbc.com/news/uk-england-derbyshire-49031625, accessed on 22/6/2026.

negative rate, correlation neglect as well as the neglect of signal asymmetry would lead to increased antibiotic use with the introduction of an AI tool. This illustrates how the consequences of AI adoption for decision outcomes depend on the quality and design of existing diagnostic information as well as the information contributed by AI.

In medical settings, diagnostic information is commonly presented in binary, or discrete, scales which often have asymmetric signal strength, where the degree of asymmetry is determined by a trade-off between false positive and false negative test results. Figure 7 shows the location of the dipstick and AI signal on the ROC curve, which contains all combinations of false positive rates and true positive rates permitted by a given classifier (e.g. a prediction technology or a decision-maker). The position of a signal on a given ROC curve is determined by the threshold on an underlying continuous probability measure, which is chosen to weigh the cost of false negative classifications against false positive classifications. This threshold is decreasing in the cost of a false negative over the cost of a false positive. The figure also shows how physicians' location of the dipstick signal and the AI signal, implied by their observed belief updating, deviates from their signal properties given in the experiment, overweighting the dipstick and underweighting the AI while also shifting the classification threshold upwards with the AI signal.

Consequently, a signal might be designed to take into account known cognitive biases while maintaining its predictive accuracy, for example by presenting a signal with symmetric sensitivity and specificity. Thus, the threshold determining these properties of a discrete signal might need to be selected not only based on costs of classification errors but also on how humans make optimal use of such signals. Alternatively, directly presenting the continuous prediction to decision-makers would provide most information. However, Agarwal et al. (2023) show how this can lead to increased diagnostic errors in intermediate risk ranges, impying lower prediction accuracy, whereas predictions close to the bounds $[0, 1]$ lead to the best outcomes. One possible solution could be to delegate decisions to the AI if its prediction accuracy is high, and delegate to the physician without providing an AI signal for cases in which prediction accuracy is low (Agarwal et al. 2023; Ribers and Ullrich 2024).

7 Conclusion

In this paper, we examine how physicians may integrate AI-based diagnostic information into their clinical decision-making, and what our findings imply for the broader prospects of AI adoption in primary care. We describe several patterns across our analyses.

First, physicians update their beliefs less in response to AI signals than in response to a conventional dipstick test, even when the informational content of the two is comparable.

Our finding is consistent with a general skepticism toward algorithmic recommendations. Rather than being treated as just another piece of evidence, AI output appears to be discounted in a way that established diagnostics are not.

Second, the average effect is driven by a substantial number of physicians who do not adopt the AI in a meaningful way. Roughly one-third of physicians are such non-adopters. We find that non-adopters exhibit lower technology use more generally, but do not differ from adopters on a range of other observable factors, including measures related to physicians' practice styles, physician preferences, and quality of care. Instead, non-adopters express skepticism toward the tool and concerns about its lack of transparency. Even though some physicians appear more open to new technologies, we find little evidence in this setting that AI adoption depends on individuals' location in the productivity or skill distribution.

Third, adopters deviate from Bayesian updating in two ways. First, they ignore the asymmetry of the AI signal, even though they account for it correctly when using the dipstick signal. Second, when physicians are presented with the AI signal and the dipstick test, they insufficiently account for the correlation between the two information sources. This points to a specific cognitive friction in multi-signal environments that is likely to become more, not less, relevant as AI tools proliferate alongside existing diagnostics.

We find that clinical decision-making might be affected by these frictions, leading to increased antibiotic use, where overprescribing is a concern due to the threat of antibiotic resistance. For implementation, this implies that how information is presented to physicians, and how they are trained to interpret asymmetric and correlated signals, may meaningfully shape the value realized from AI tools.

Our findings should be interpreted in light of the limitation that we observe physicians in essentially a one-shot encounter with the AI tool. Repeated use, accompanied by feedback on the tool's performance and practical usefulness, might reshape adoption patterns and updating behavior over time, possibly attenuating some of the skepticism we document and possibly reinforcing it where early experiences disappoint. Goldsmith-Pinkham et al. (2026) provide early evidence that, with feedback over time, radiologists become more productive in analyzing lung scans and rejection of AI information decreases, although without disappearing completely. Studying these dynamics is an important next step for further research. Even so, understanding behavior at the point of early adoption is important: during this early phase, norms, trust, and implementation choices around medical AI are being formed, and the frictions we identify are likely to affect this adoption process into clinical practice.

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.

Augenblick, Ned, Eben Lazarus, and Michael Thaler, "Overinference from weak signals and underinference from strong signals," *Quarterly Journal of Economics*, 2025, *140* (1), 335–401.

Baker, Andrew, Brantly Callaway, Scott Cunningham, Andrew Goodman-Bacon, and Pedro H. C. Sant'Anna, "Difference-in-Differences Designs: A Practitioner's Guide," *Journal of Economic Literature*, June 2026, *64* (2), 498–557.

Bansal, Gagan, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz, "Beyond accuracy: The role of mental models in human-AI team performance," in "Proceedings of the AAAI conference on human computation and crowdsourcing," Vol. 7 2019, pp. 2–11.

Bauer, Kevin, Moritz Von Zahn, and Oliver Hinz, "Expl (AI) ned: The impact of explainable artificial intelligence on users' information processing," *Information Systems Research*, 2023, *34* (4), 1582–1602.

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," *Management Science*, 2026.

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

Carey, Iain M, Fay J Hosking, Tess Harris, Stephen DeWilde, Carole Beighton, and Derek G Cook, "An evaluation of the effectiveness of annual health checks and quality of health care for adults with intellectual disability: an observational study using a primary care database," *Health Services and Delivery Research*, September 2017, *5* (25), 1–170.

Chan, Alex, "Preference for Explainable AI," NBER Working Paper 35240, May 2026.

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.

Dansk Selskab for Almen Medicin, "FAQta-ark om urinvejsinfektioner i almen praksis," 2020. https://content.dsam.dk/guides/vejlednings-pdf/uvi_faqta-ark.pdf. Accessed: June 15, 2026.

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," *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.

Goldsmith-Pinkham, Paul, Chenhao Tan, and Alexander K Zentefis, "Human-AI Collaboration in Radiology: The Case of Pulmonary Embolism," *arXiv preprint arXiv:2601.13379*, 2026.

Grether, David M., "Bayes Rule as a Descriptive Model: The Representativeness Heuristic," *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.

Haaland, Ingar, Christopher Roth, Stefanie Stantcheva, and Johannes Wohlfart, "Understanding Economic Behavior Using Open-Ended Survey Data," *Journal of Economic Literature*, December 2025, *63* (4), 1244–1280.

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.

_ and _ , "Still waters, rapid currents: Early labor market transformation under generative AI," NBER Working Paper 33777, May 2025.

Imbens, Guido W. and Donald B. Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, 1 ed., Cambridge University Press, April 2015.

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

_ , Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer, "Prediction Policy Problems," *American Economic Review*, 2015, *105* (5), 491–495.

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.

MedCom, "Lægepraksis i Danmark," 2025. https://medcom.dk/standarder/ydere-lokationsnumre/laegepraksis-i-danmark/. Accessed: June 15, 2026.

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," *The Quarterly Journal of Economics*, April 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.

Patienterstatningen, "Oplysningspligt." https://patienterstatningen.dk/for-professionelle/oplysningspligt. Accessed: June 15, 2026.

Peng, Sida, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer, "The impact of AI on developer productivity: Evidence from github copilot," *arXiv preprint arXiv:2302.06590*, 2023.

Poursabzi-Sangdeh, Forough, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach, "Manipulating and measuring model interpretability," in "Proceedings of the 2021 CHI conference on human factors in computing systems" 2021, pp. 1–52.

Rambachan, Ashesh, "Identifying Prediction Mistakes in Observational Data," Working Paper October 2022.

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.

Serra-Garcia, Marta and Uri Gneezy, "Improving human deception detection using algorithmic feedback," *Management Science*, 2025, *71* (12), 10289–10307.

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.

Styrelsen for Patientsikkerhed, "Tilsyn i tal," May 2026. https://stps.dk/sundhedsfaglig/tilsyn/tilsyn-med-sundhedspersoner/tilsyn-i-tal. Accessed: June 15, 2026.

Thornley, Tracey, Charlotte L Kirkdale, Elizabeth Beech, Philip Howard, and Peter Wilson, "Evaluation of a community pharmacy-led test-and-treat service for women with uncomplicated lower urinary tract infection in England," *JAC-Antimicrobial Resistance*, 2020, *2* (1), dlaa010.

Tschandl, Philipp, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy et al., "Human–computer collaboration for skin cancer recognition," *Nature medicine*, 2020, *26* (8), 1229–1234.

Vaccaro, Michelle, Abdullah Almaatouq, and Thomas Malone, "When combinations of humans and AI are useful: A systematic review and meta-analysis," *Nature Human Behaviour*, 2024, *8* (12), 2293–2303.

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.

Zahn, Moritz Von, Lena Liebich, Ekaterina Jussupow, Oliver Hinz, and Kevin Bauer, "Knowing (not) to know: Explainable artificial intelligence and human metacognition," *Information Systems Research*, 2025.

Zöller, Nikolas, Julian Berger, Irving Lin, Nathan Fu, Jayanth Komarneni, Gioele Barabucci, Kyle Laskowski, Victor Shia, Benjamin Harack, Eugene A Chu et al., "Human–AI collectives most accurately diagnose clinical vignettes," *Proceedings of the National Academy of Sciences*, 2025, *122* (24), e2426153122.

Figures

Figure 1. Survey flow

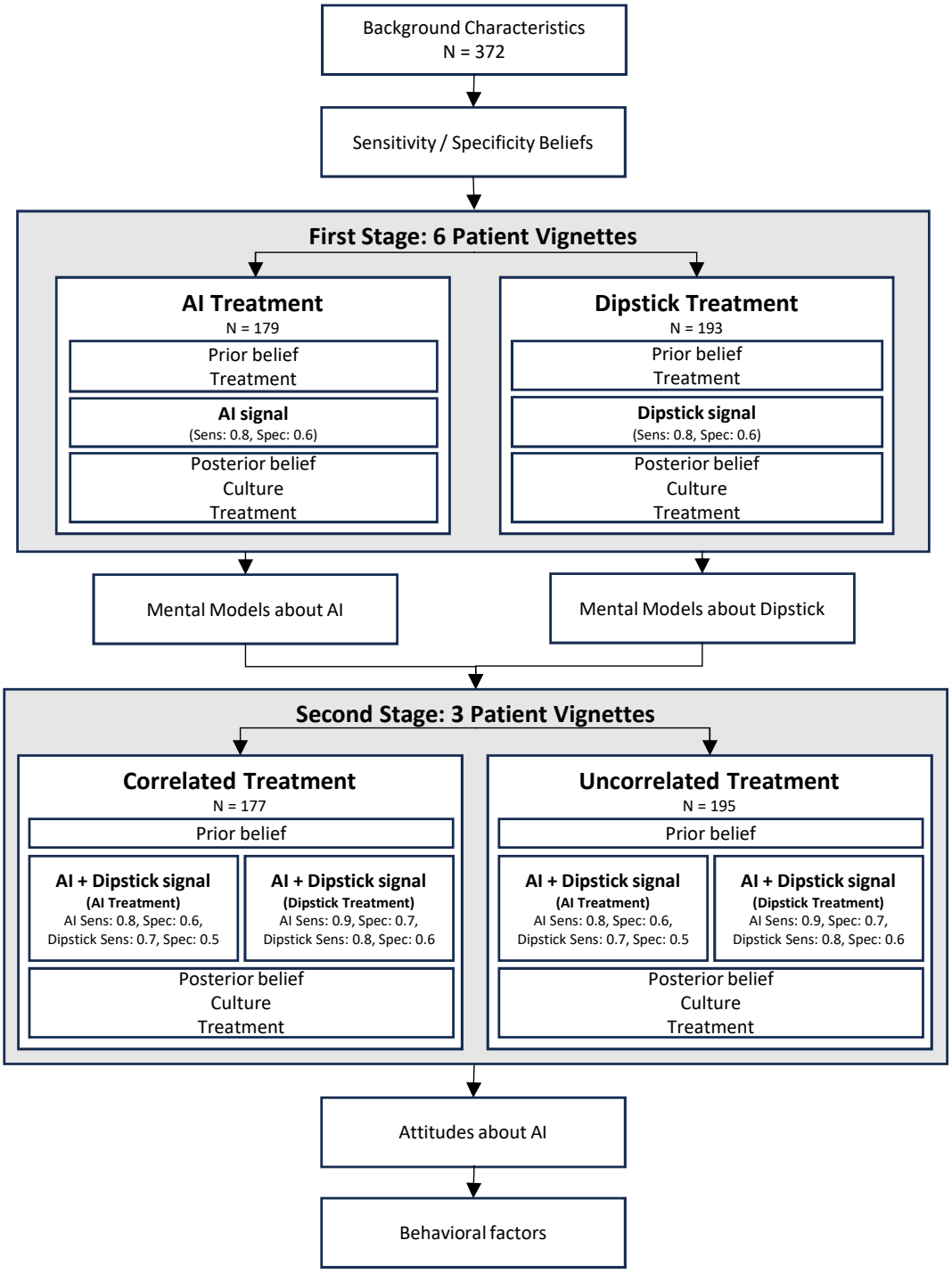


Figure 2. Screenshot of vignette

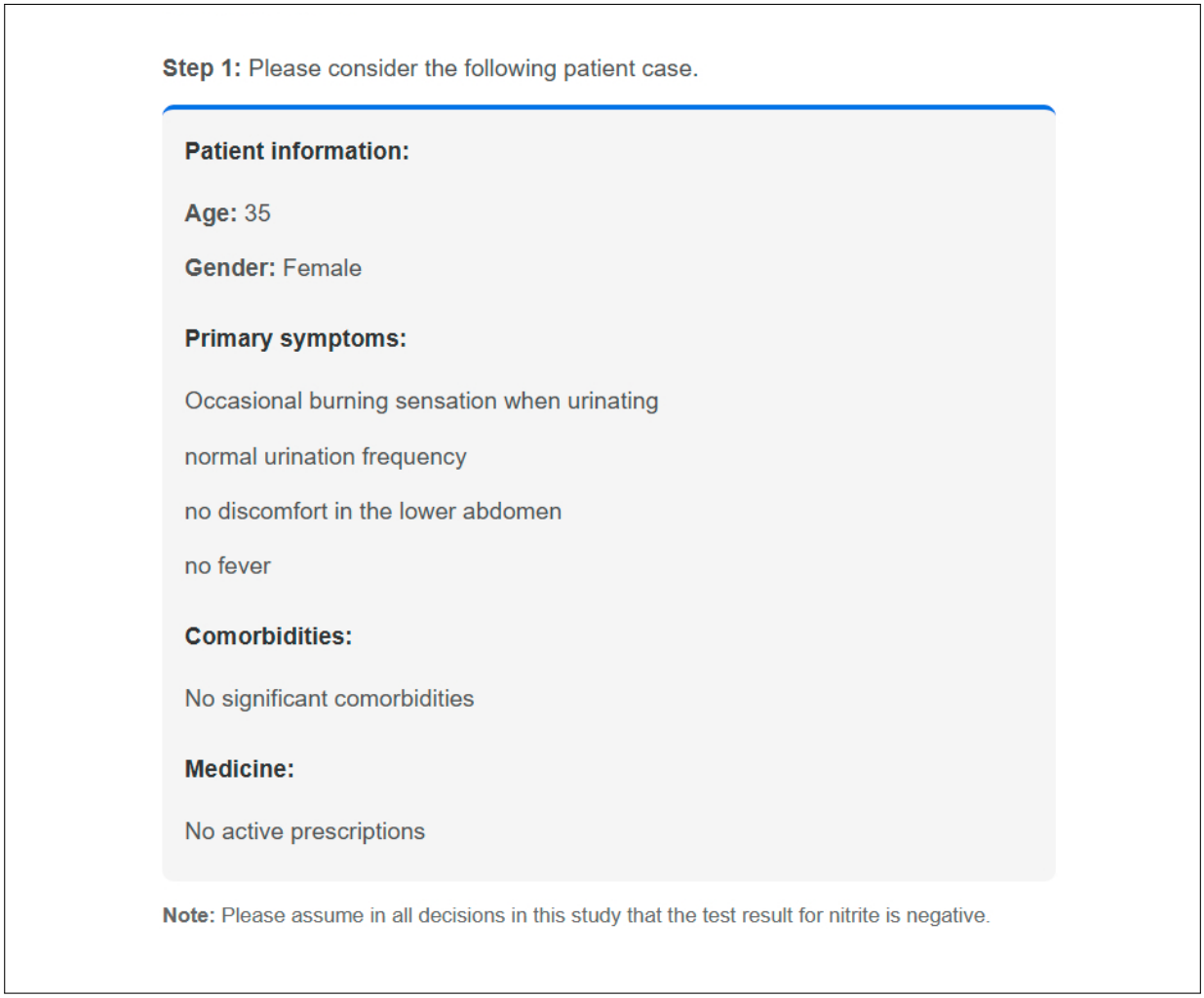
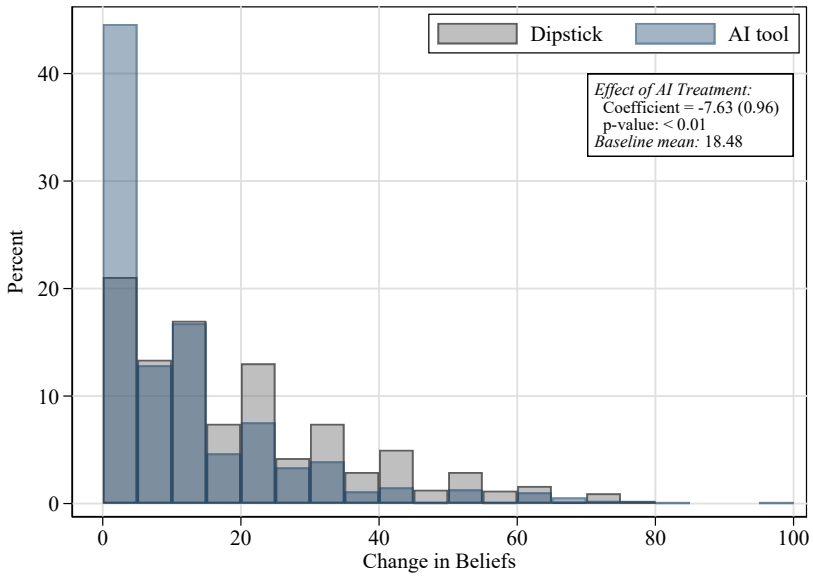
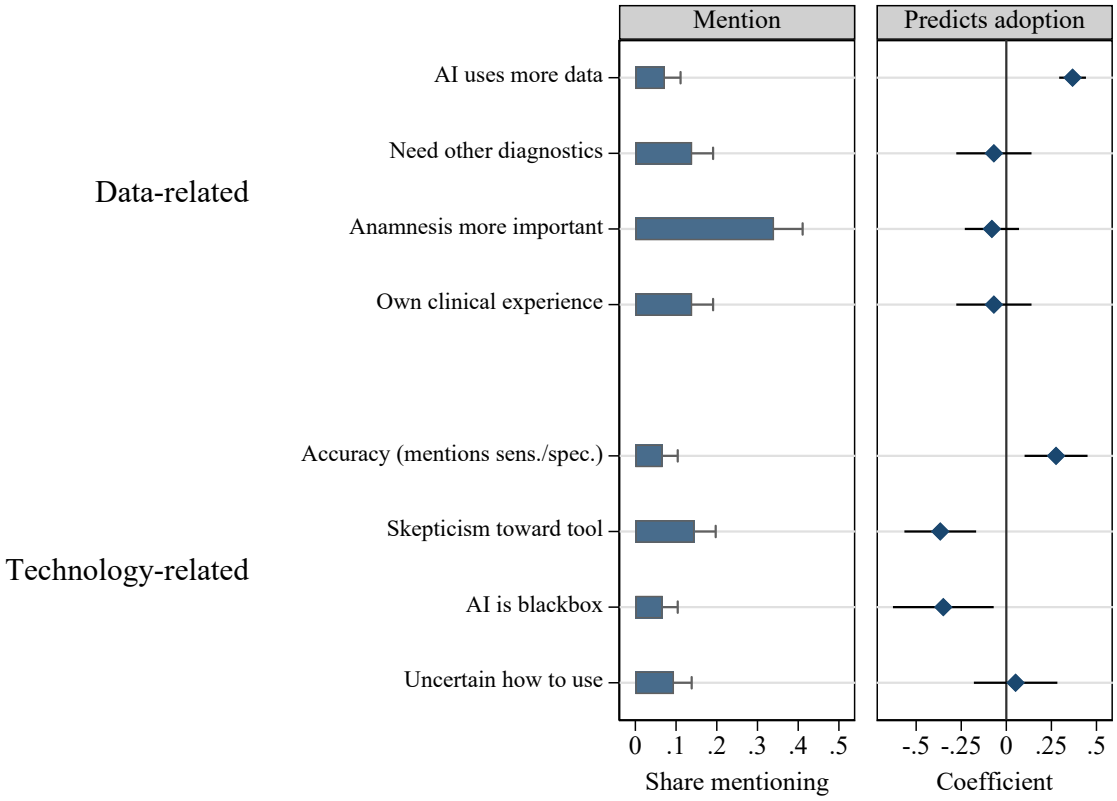


Figure 3. Absolute belief updating: Signal source



Notes: The figure shows the distribution of absolute belief updating, defined as the absolute difference between a physician's prior and posterior UTI beliefs, in Experimental Stage 1 (vignettes 1–6). It compares respondents assigned to the AI Tool treatment with respondents assigned to the Dipstick treatment (baseline). The figure also reports the coefficient on the AI Tool indicator from the specification in Equation (1), estimated with absolute belief updating as the outcome and including vignette and patient-age fixed effects. Standard errors are clustered at the physician level and shown in parentheses; the corresponding p-value is reported. Number of observations (physician-vignettes): 2,232.

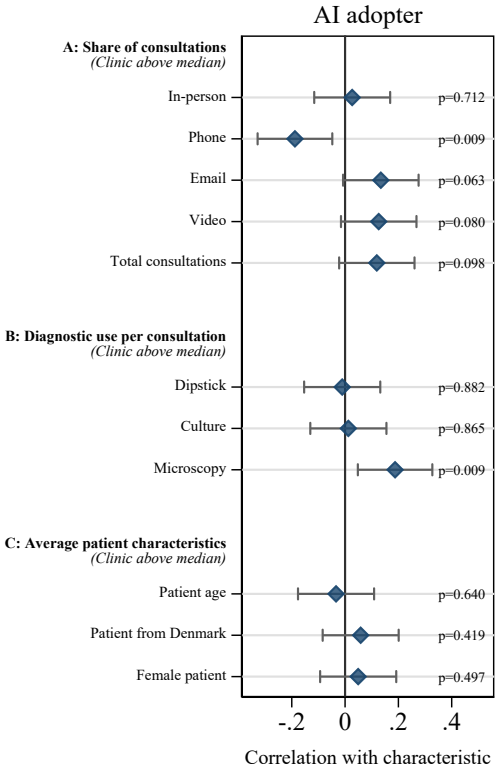
Figure 4. Mental models about the AI tool



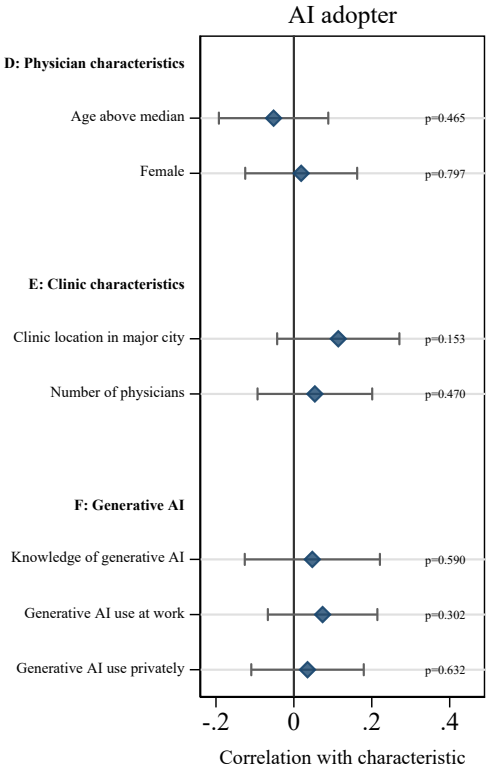
Notes: Sample restricted to respondents assigned to the AI Tool treatment. The left panel shows the share of physicians whose open-text response mentions each factor. The right panel reports coefficient estimates from separate OLS regressions of an AI-adoption indicator on an indicator for mentioning the respective factor. AI adoption is defined as updating beliefs by more than three percentage points in at least half of the vignettes (see Section 4.2). Categories are non-mutually exclusive. Horizontal lines denote 95% confidence intervals. Number of observations (physicians): 179.

Figure 5. Correlates of AI adoption

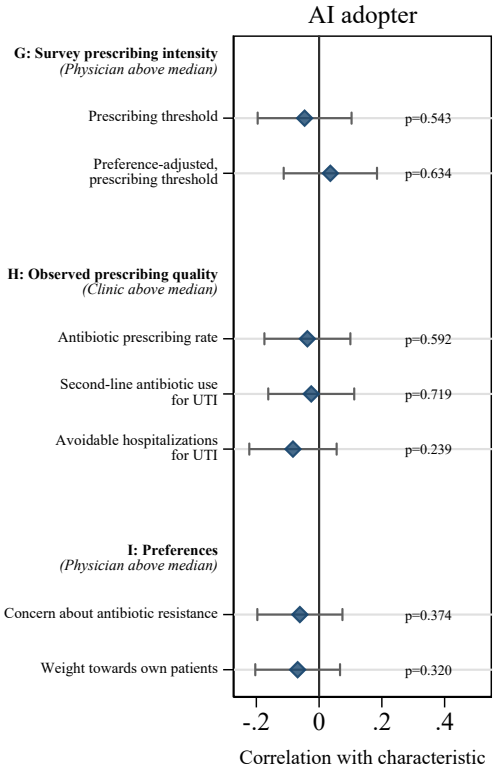
(a) Clinical practice and patient pools



(b) Physician and clinic characteristics

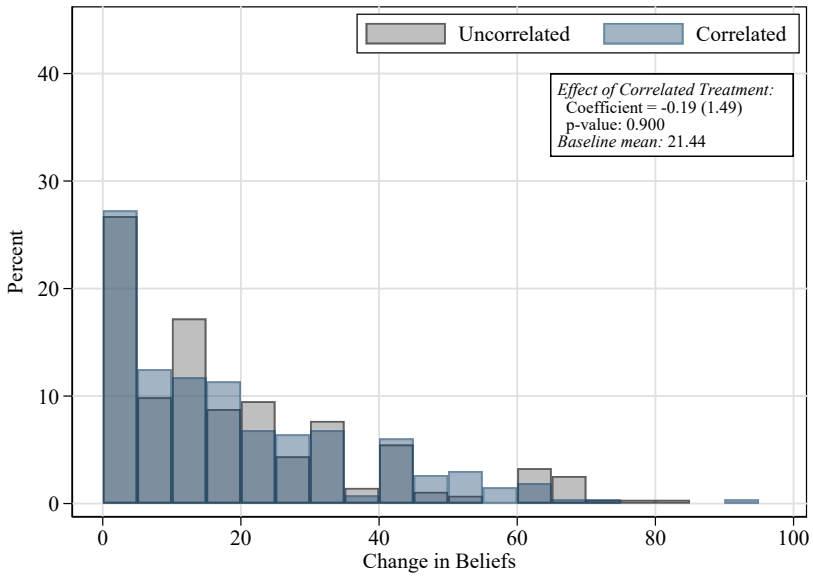


(c) Antibiotic prescribing



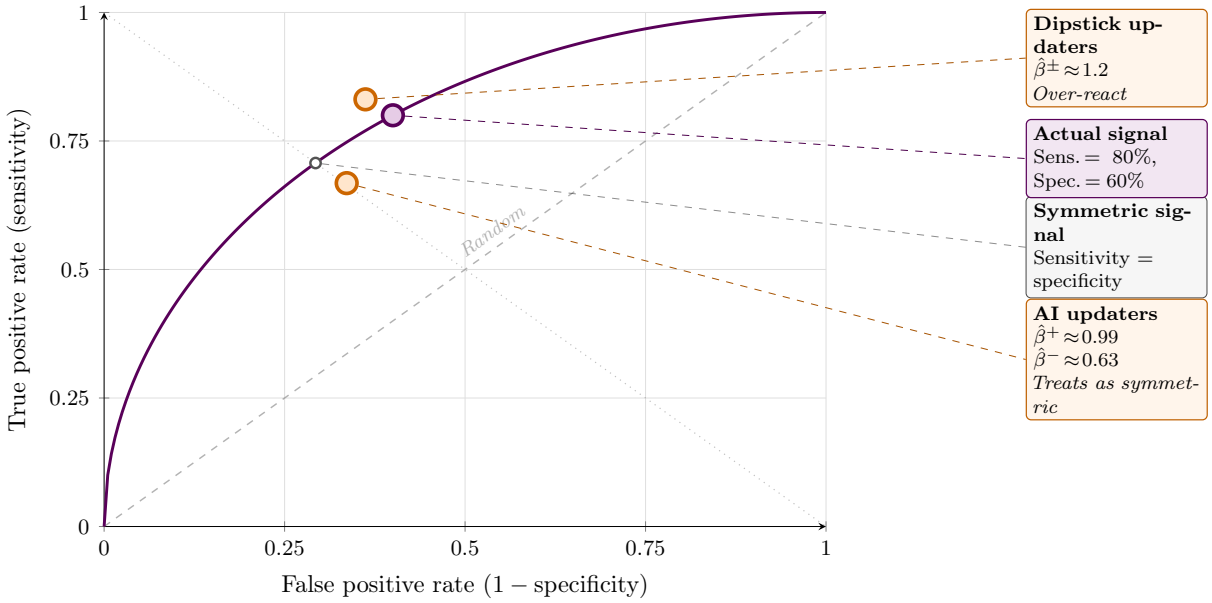
Notes: Sample restricted to respondents assigned to the AI Tool treatment and matched to the relevant register data. Each coefficient is from a separate OLS regression of an AI-adoption indicator on the listed characteristic. Appendix Table A3 describes the construction of all characteristics. AI adoption is defined as updating beliefs by more than three percentage points in at least half of the vignettes (see Section 4.2). All regressions are at the physician level; clinic-level characteristics are assigned to all respondents from the same clinic. Characteristics shown as clinic above-median variables are indicators for whether the clinic is above the sample median; characteristics shown as physician above-median variables are defined analogously at the respondent level. Horizontal lines denote 95% confidence intervals; p-values are reported next to each estimate.

Figure 6. Absolute belief updating: Signal correlation



Notes: The figure shows the distribution of absolute belief updating, defined as the absolute difference between a physician's prior and posterior UTI beliefs, in Experimental Stage 2 (vignettes 7–9). It compares respondents assigned to the CORRELATED treatment with respondents assigned to the UNCORRELATED treatment (baseline). The figure also reports the coefficient on the CORRELATED indicator from the specification in Equation (1), estimated with absolute belief updating as the outcome and including vignette and patient-age fixed effects. Standard errors are clustered at the physician level and shown in parentheses; the corresponding p-value is reported. Number of observations (physician-vignettes): 1,116.

Figure 7. Actual and implied signal asymmetry



Notes: The figure shows a simulated ROC curve on which a signal with sensitivity of 80% and a specificity of 60% would be located. The depicted off-curve positions show the (Bayesian) sensitivities/specificities implied by the physicians' estimated $\hat{\beta}$ for the dipstick signal and the AI signal.

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
Comorbidity conditions	
Comorbidities and active prescriptions	None /
	Type 1 diabetes and Insulin aspart (Novolog) via insulin pump

Notes: The table lists the patient characteristics used to construct the vignette cases. In Experimental Stage 1, primary symptoms and comorbidity conditions are balanced across six vignettes. In Experimental Stage 2, three additional vignettes are constructed with weak primary symptoms and no comorbidities. Age is randomized independently at the respondent-vignette level.

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
# **Clinics**	179	167	346

	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
# **Clinics**	185	161	346

Notes: This table reports summary statistics for the analysis sample by signal source treatment (Panel A) and signal correlation treatment (Panel B). The first and second columns show means for each treatment group. The third 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

	Sample		Population		Difference
	Mean	*SD*	*Mean*	*SD*	*SMD*
Panel A: Clinical practice and patient pools					
In-person consultation share	0.51	(0.082)	0.52	(0.090)	[-0.11]
Phone consultation share	0.20	(0.085)	0.22	(0.093)	[-0.15]
Email consultation share	0.27	(0.087)	0.25	(0.093)	[0.24]
Video consultation share	0.017	(0.020)	0.015	(0.020)	[0.13]
Dipstick claims rate	0.033	(0.015)	0.034	(0.017)	[-0.04]
Culture claims rate	0.013	(0.016)	0.014	(0.018)	[-0.06]
Microscopy claims rate	0.0062	(0.011)	0.0054	(0.010)	[0.08]
Antibiotic prescribing rate	0.049	(0.014)	0.052	(0.017)	[-0.16]
Average patient age	47.4	(5.12)	48.3	(5.37)	[-0.17]
Patients from Denmark	0.87	(0.085)	0.85	(0.11)	[0.13]
Female patients	0.61	(0.049)	0.61	(0.051)	[0.09]
Total consultations	22001.7	(12399.8)	20170.8	(12100.0)	[0.15]
# **Clinics**	333		1,625		
Panel B: Physician characteristics					
Number of physicians	3.44	(2.30)	3.16	(2.31)	[0.13]
Single physician clinics	0.22	(0.42)	0.31	(0.46)	[-0.20]
Average age of physicians	52.6	(6.49)	53.5	(7.10)	[-0.12]
Share of female physicians	0.57	(0.35)	0.59	(0.37)	[-0.06]
Clinic location in major city	0.31	(0.46)	0.27	(0.44)	[0.11]
# **Clinics**	320		1,559		

Notes: This table reports clinic-level summary statistics for clinical practice and patient pools (Panel A) and clinic characteristics (Panel B). The population of clinics is constructed from the national clinic register, *Yderregister*. We restrict the analysis sample to clinics that can be matched to the administrative register data. Consultations are identified using clinic claim codes; antibiotic prescriptions are identified using Anatomical Therapeutic Chemical (ATC) codes and indication codes. The *Population* columns report means and standard deviations (SDs) for all clinics in the population, and the *Sample* columns report the corresponding statistics for clinics in the analysis sample. The final column reports the standardized mean difference (SMD) between the analysis sample and the population, defined as $(\bar{X}_S - \bar{X}_P)/\sqrt{(SD_S^2 + SD_P^2)/2}$, where $\bar{X}$ denotes the group mean and SD denotes the standard deviation for analysis clinics (S) or population clinics (P). SMDs above 0.25 in absolute value are commonly interpreted as indicating potentially problematic imbalance (Imbens and Rubin 2015; Baker et al. 2026).

Table 4. Signal source: Belief updating (Bayesian)

	All physicians		Only adopters	
	(1) Dipstick b/se	(2) AI tool b/se	(3) Dipstick b/se	(4) AI tool b/se
δ^{prior}	0.82***	0.87***	0.79***	0.82***
	(0.027)	(0.029)	(0.025)	(0.040)
β^{Pos}	1.07***	0.87***	1.15***	1.26***
	(0.059)	(0.083)	(0.058)	(0.11)
β^{Neg}	1.12***	0.55***	1.17***	0.77***
	(0.047)	(0.047)	(0.046)	(0.060)
$\delta^{\text{prior}} = 1$ (*p-value*)	0.000	0.000	0.000	0.000
$\beta^{\text{pos}} = 1$ (*p-value*)	0.231	0.127	0.009	0.016
$\beta^{\text{neg}} = 1$ (*p-value*)	0.013	0.000	0.000	0.000
$\beta^{\text{pos}} = \beta^{\text{neg}}$ (*p-value*)	0.518	0.000	0.746	0.000
Adj. R^2	0.761	0.737	0.772	0.740
N	1158	1074	1080	708
Subjects	193	179	180	118

Notes: This table reports estimates from the Bayesian updating specification in Equation (2), which regress posterior beliefs on the prior and the likelihood ratio of a signal, separately for the Dipstick and AI Tool treatments. Columns (1)–(2) use all physicians; Columns (3)–(4) restrict to adopters of the respective signal, defined as updating beliefs by more than three percentage points in at least half of the vignettes (see Section 4.2). Standard errors are clustered at the physician level and shown in parentheses.

Table 5. Effect of AI Treatment on Antibiotic Prescribing Behavior

	Final prescriptions		Prescription change	
	(1)	(2)	(3)	(4)
	After pos. signal	After neg. signal	Start prescribe after pos. signal	Stop prescribe after neg. signal
	b/se	b/se	b/se	b/se
AI	0.048	0.042*	-0.003	-0.152**
	(0.035)	(0.025)	(0.033)	(0.072)
Constant	0.304***	0.117***	0.195***	0.468***
	(0.033)	(0.023)	(0.032)	(0.082)
Vignette FE	✓	✓	✓	✓
Patient age FE	✓	✓	✓	✓
Baseline mean	0.336	0.136	0.189	0.400
Adj. R^2	0.208	0.095	0.093	0.057
N	1080	1152	879	210

Notes: Columns (1) and (2) regress a dummy variable indicating whether a physician ultimately prescribes empirical antibiotic treatment on signal dummies, estimated separately for positive and negative signals. Column (3) restricts the sample to cases where the physician did not prescribe pre-signal, and the outcome indicates whether they switched to prescribing after a positive signal. Column (4) restricts the sample to cases where the physician prescribed pre-signal, and the outcome indicates whether they switched to not prescribing after a negative signal. Standard errors clustered at the respondent level in parentheses.

Table 6. Signal correlation: Belief updating

Panel A: Bayesian updating		
	(1) Correlated b/se	(2) Uncorrelated b/se
δ^{prior}	0.71***	0.70***
	(0.045)	(0.039)
β^{Pos}	1.28***	0.76***
	(0.099)	(0.054)
β^{Neg}	0.91***	0.57***
	(0.057)	(0.033)
Adj. R^2	0.748	0.728
N	531	585
Subjects	177	195
Panel B: Absolute updating		
	(1) Correlated b/se	(2) Uncorrelated b/se
Positive signal	42.1***	41.6***
	(2.13)	(1.98)
Constant	-12.8***	-13.2***
	(1.81)	(1.54)
Vignette FE	✓	✓
Patient age FE	✓	✓
Adj. R^2	0.554	0.532
N	531	585
Subjects	177	195

Notes: This table reports belief updating in Experimental Stage 2, separately for the CORRELATED and UNCORRELATED treatments. Panel A reports estimates from the Bayesian updating specification in Equation (2), which regresses posterior beliefs on the prior and the likelihood ratio implied by the joint AI and dipstick signals. Panel B reports estimates from the absolute belief-updating specification in Equation (3), which regresses absolute belief updating on an indicator for a positive signal (baseline category: negative signal). Standard errors are clustered at the physician level and shown in parentheses.

Table 7. Effect of CORRELATED Treatment on Antibiotic Prescribing Behavior

	(1) After positive signal b/se	(2) After negative signal b/se
Correlated	0.035	0.007
	(0.036)	(0.007)
AI	-0.052	0.008
	(0.036)	(0.008)
Constant	0.120***	-0.007
	(0.038)	(0.007)
Patient age FE	✓	✓
Baseline mean	0.144	0.004
Adj. R^2	0.006	0.000
N	576	540

Notes: The table reports estimates from a regression of a dummy variable indicating whether a physician ultimately prescribes empirical antibiotic treatment on the CORRELATED treatment indicator and the AI treatment indicator. The sample is restricted to the second stage vignettes, in which both AI and dipstick signal were displayed. In the CORRELATED treatment, the AI signal contains the dipstick signal, while they are independent in the UNCORRELATED treatment. Hence, the signals in the CORRELATED treatment are substantially less informative. Standard errors clustered at the respondent level in parentheses.

Table 8. Effect of the redundant signal in CORRELATED on Antibiotic Prescribing Behavior

	(1) After positive signal b/se	(2) After negative signal b/se
Redundant dipstick signal	0.084**	-0.007
	(0.037)	(0.007)
Constant	-0.028	0.004
	(0.050)	(0.004)
Patient age FE	✓	✓
Baseline mean	0.085	0.015
Adj. R^2	0.040	-0.017
N	130	134

Notes: The table reports estimates from a regression of a dummy variable indicating whether a physician ultimately prescribes empirical antibiotic treatment on the "Redundant Dipstick Signal" indicator. The sample is restricted to respondents who received the AI signal in the first stage and were assigned to the CORRELATED treatment in the second stage. The sample is further limited to vignettes with identical symptom descriptions. The estimation compares prescription decisions when only the AI signal is observed to decisions when the AI signal and the dipstick result are observed jointly. The only change in the vignettes relative to the earlier vignettes is that the dipstick signal is now displayed, whereas previously it was not; the signal itself is redundant because it is already incorporated into the AI signal (sensitivity and specificity are the same). Hence, physicians who take the correlation structure of the signals into account should not change their prescription decisions. 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 additional variables

Variable	Description	Data source
Panel A: Clinical practice and patient pools		
Share of in-person consultations	Share of all consultations conducted face-to-face.	SSSY
Share of tele-consultations	Share of consultations conducted by telephone.	SSSY
Share of email consultations	Share of consultations conducted via email.	SSSY
Share of video consultations	Share of consultations conducted via video.	SSSY
Total consultations	Total number of consultations.	SSSY
Dipstick claims per consultation	Number of dipstick tests billed divided by total consultations.	SSSY
Culture claims per consultation	Number of urine culture tests billed divided by total consultations.	SSSY
Microscopy claims per consultation	Number of microscopy tests billed divided by total consultations.	SSSY
Antibiotic prescribing rate	Total antibiotic prescriptions divided by total consultations.	LMDB
Matched antibiotic prescribing rate	Antibiotic prescribing rate matched to the population of vignette patients: Total number of UTI-indicated antibiotic prescriptions divided by total consultations. Both prescriptions and consultations are restricted to female patients aged 20 to 35.	LMDB
Second-line antibiotic use for UTI	Number of UTI-indicated antibiotic prescriptions that are not guideline-recommended first-line treatments (pivmecillinam, nitrofurantoin, or trimethoprim), divided by the total number of UTI-indicated antibiotic prescriptions. Guideline recommendations are provided by the Danish College of General Practitioners (Dansk Selskab for Almen Medicin 2020).	LMDB
Avoidable hospitalizations for UTI	Number of hospitalizations with UTI-related primary care sensitive diagnostic codes from patients whose primary care clinic had billed a consultation claim up to two weeks prior to the hospitalization event, divided by total consultations. We only consider the closest match between consultation and hospitalization for each clinic, patient, and hospitalization event; however, the same hospitalization event may be matched to multiple primary care clinics. To identify UTI-related avoidable hospitalizations, we use the International Classification of Diseases (ICD) codes from Carey et al. (2017)	SSSY, LPR
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
Panel B: Physician characteristics		
Number of physicians	Total number of physicians affiliated with the clinic.	YDER
Single physician clinics	Indicator equal to 1 if the clinic has only one physician.	YDER
Average age of physicians	Mean physician age within the clinic.	YDER, CPR
Share of female physicians	Fraction of physicians who are female.	YDER, CPR
Clinic location in major city	Indicator for clinics located in municipalities with more than 100,000 inhabitants: Copenhagen/Frederiksberg, Aarhus, Odense, or Aalborg.	YDER

Clinical practice variables are constructed from primary care claims (SSSY) and prescription records (LMDB). We identify consultations and diagnostic use as follows using SSSY claim codes. Dipstick tests: 807101, 808155; urine culture: 807105, 807189; microscopy: 807102, 807103, 807122, 807123, 807169, 808156, 808157, 808165, 808166; in-person consultations: 800101, 800106; telephone consultations: 800109, 800201, 808294, 808618, 800501; email consultations: 800105, 800110; and video consultations: 800125, 804436. Antibiotic prescriptions are identified by Anatomical Therapeutic Chemical (ATC) code J01, and UTI-related prescriptions are identified by indication codes 103 and 104. Avoidable hospitalizations for UTI are identified from ICD-10 diagnostic codes: N10–N12, N13.6, N39.0.

Table A3 – continued

Variable	Description	Data source
Panel C: Elicited preferences		
Concern about antibiotic resistance	Self-reported concern about antimicrobial resistance when making prescription decisions, on a scale from 0 (not concerned at all) to 10 (very concerned).	Survey
Weight towards own patients	Amount (in 1,000 DKK) a physician claims for improving care in their own practice out of a hypothetical 100,000 DKK fund, with the remainder donated to a charity supporting vulnerable patients nationwide. Higher values indicate greater weight on own patients relative to others, proxying the perceived cost of the prescribing externality.	Survey
Risk preferences	General willingness to take risks, on a scale from 0 (completely unwilling) to 10 (very willing).	Survey
Time preferences	General willingness to forgo a present benefit for a larger future benefit (patience), on a scale from 0 (completely unwilling) to 10 (very willing).	Survey
Panel D: Additional variables constructed from the survey		
Survey prescribing rate	Rate at which the physician chooses to prescribe an empiric antibiotic treatment across survey vignettes at the pre-signal stage, before observing any diagnostic signal.	Survey
Prescribing threshold	Physician-level posterior-belief threshold above which the physician prescribes, recovered from post-signal beliefs and prescribing decisions across non-diabetic vignettes (rounds 1–3 and 7–9). Measured as the lower bound of the implied threshold interval, i.e., the highest posterior belief at which the physician does not prescribe (see Appendix C.3).	Survey
Preference-adjusted prescribing threshold	Residual from a linear regression of the prescribing threshold on fixed effects for survey-elicited preferences: concern about antibiotic resistance (0–10), weight towards own patients (in 10,000 DKK bins), risk preferences (0–10), and time preferences (0–10). The preference-adjusted prescribing threshold accounts for variation in the threshold attributable to elicited preferences (see Appendix C.3).	Survey

Table A4. Robustness checks

	(1) Main	(2) Low sens.	(3) Low spec.	(4) Both	(5) Reweighted: clinical practice	(6) Reweighted: clinical practice clinic char.	(7) Only single physician clinics	(8) None	(9) Baseline + risk pref.
AI Tool	-7.63***	-7.86***	-7.31***	-5.53***	-7.31***	-7.00***	-5.00**	-7.62***	-6.78***
	(0.96)	(1.41)	(1.22)	(1.85)	(1.45)	(1.56)	(2.09)	(0.96)	(1.57)
Vignette FE	✓	✓	✓	✓	✓	✓	✓		✓
Patient age FE	✓	✓	✓	✓	✓	✓	✓		✓
Risk pref. FE									✓
Baseline mean	14.81	14.32	14.55	13.66	14.77	14.66	14.27	14.81	14.66
Observations	2232	990	1404	564	2154	2070	426	2232	2070
Physicians	372	165	234	94	359	345	71	372	345

Notes: This table reports robustness checks for the estimated effect of the AI Tool treatment on absolute belief updating in Experimental Stage 1 (vignettes 1–6). Each column estimates the specification in Equation (1), with absolute belief updating as the outcome and the Dipstick treatment as the baseline. Column (1) reports the main specification. Columns (2)–(4) restrict the sample to physicians whose prior beliefs about dipstick sensitivity, specificity, or both are below the experimentally provided values. Columns (5)–(6) reweight respondents using stabilized inverse-probability weights, estimated from a clinic-level logit model of inclusion in the analysis sample using the full clinic population. Column (5) uses weights based on clinical-practice and patient-pool characteristics; Column (6) additionally includes clinic characteristics. Column (7) restricts the sample to single-physician clinics. Columns (8)–(9) vary the control set, with Column (8) omitting the baseline fixed effects and Column (9) adding risk-preference fixed effects. Baseline mean refers to the Dipstick treatment in the corresponding estimation sample. Standard errors are clustered at the physician level and shown in parentheses.

Table A5. Signal correlation: Belief updating by CORRELATED vs UNCORRELATED and AI vs DIPSTICK treatments

	Uncorrelated treatment		Correlated treatment	
	(1) b/se	(2) b/se	(3) b/se	(4) b/se
Positive signal	35.1***	46.8***	35.6***	49.0***
	(2.56)	(2.80)	(2.70)	(3.09)
Constant	-13.0***	-13.8***	-13.9***	-11.2***
	(1.78)	(2.44)	(2.27)	(2.77)
Vignette FE	✓	✓	✓	✓
Patient age FE	✓	✓	✓	✓
1st-stage treatment	AI	DIPSTICK	AI	DIPSTICK
2nd-stage AI sens/spec	80/60	90/70	80/60	90/70
Adj. R^2	0.504	0.570	0.502	0.624
N	273	312	264	267
Subjects	91	104	88	89

Notes: This table reports belief updating in Experimental Stage 2, separately for the Correlated and Uncorrelated treatments and separately for whether an individual was in the AI or the DIPSTICK treatment in the first stage of the experiment. For individuals who were in the AI treatment in the first stage, the sensitivity/specificity of the AI signal in the second stage is 80%/60%. For those who were in the DIPSTICK treatment in the first stage, the sensitivity/specificity of the AI signal in the second stage is 90%/70%. The table reports estimates from the absolute belief-updating specification in Equation (3), which regresses absolute belief updating on an indicator for a positive signal (vs. negative). Standard errors are clustered at the physician level and shown in parentheses.

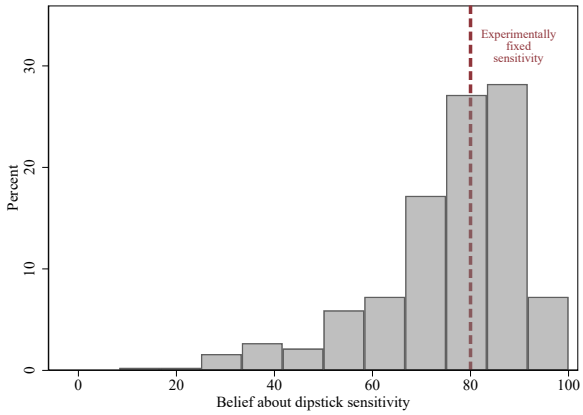
Table A6. Signal correlation: Belief updating (only adopters)

Panel A: Bayesian updating		
	(1) Correlated b/se	(2) Uncorrelated b/se
δ^{prior}	0.67***	0.68***
	(0.051)	(0.042)
β^{Pos}	1.34***	0.77***
	(0.11)	(0.057)
β^{Neg}	0.93***	0.57***
	(0.061)	(0.035)
Adj. R^2	0.752	0.730
N	432	462
Subjects	144	154
Panel B: Absolute updating		
	(1) Correlated b/se	(2) Uncorrelated b/se
Positive signal	47.0***	43.8***
	(2.31)	(2.20)
Constant	-12.9***	-13.8***
	(2.05)	(1.78)
Vignette FE	✓	✓
Patient age FE	✓	✓
Adj. R^2	0.610	0.548
N	432	462
Subjects	144	154

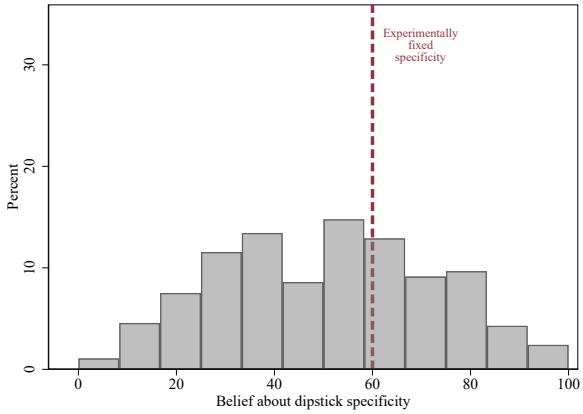
Notes: This table reports belief updating in Experimental Stage 2, separately for the CORRELATED and UNCORRELATED treatments, restricted to respondents who are classified as adopters. Adoption is defined as updating beliefs by more than three percentage points in at least half of the first-stage vignettes (see Section 4.2). Panel A reports estimates from the Bayesian updating specification in Equation (2), which regresses posterior beliefs on the prior and the likelihood ratio implied by the joint AI and dipstick signals. Panel B reports estimates from the absolute belief-updating specification in Equation (3), which regresses absolute belief updating on an indicator for a positive signal (vs. negative). Standard errors are clustered at the physician level and shown in parentheses.

B Appendix Figures

Figure A1. Distribution of prior beliefs about dipstick sensitivity and specificity



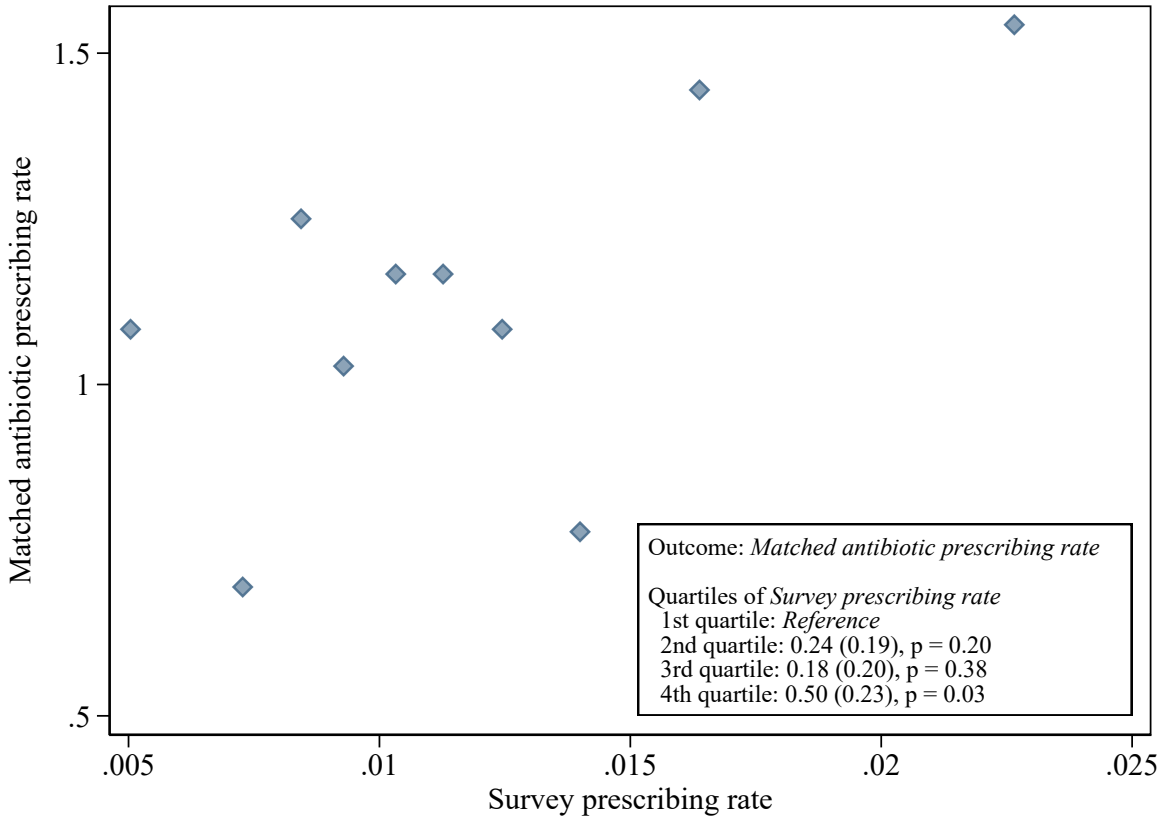
(a) Dipstick sensitivity



(b) Dipstick specificity

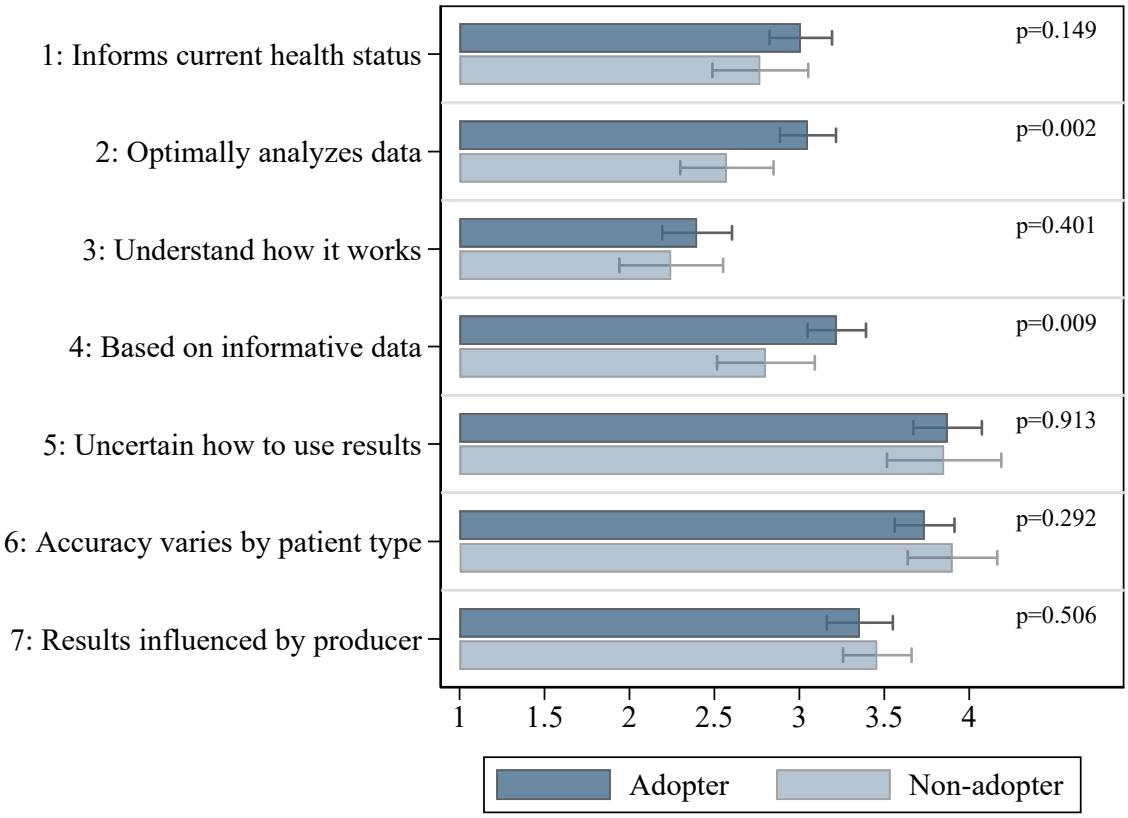
Notes: These figures show histograms of the distribution of participants' prior beliefs about the dipstick test's sensitivity and specificity. Vertical lines represent experimentally fixed values (sensitivity 80%, specificity 60%). Number of observations (physicians): 372.

Figure A2. Relationship between survey prescribing and observed antibiotic prescribing



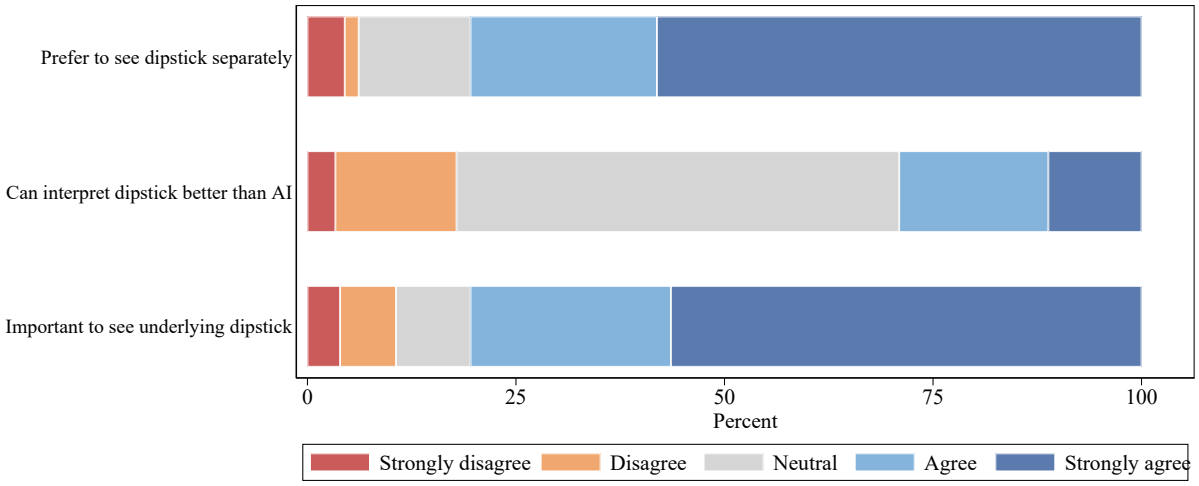
Notes: The figure shows a binscatter plot of the relationship between the physicians' antibiotic prescribing in the survey experiment and observed clinic-level antibiotic prescribing constructed from the administrative data. Survey prescribing is constructed as the rate of prescribing an empirical antibiotic treatment in the survey vignettes, prior to receiving any diagnostic signal. The observed prescribing rate is measured as the number antibiotic prescriptions for urinary tract infections (UTI) per consultation for the vignette patient population of female patients aged 20 to 35. Each dot in the figure shows the average observed prescribing within a decile of survey prescribing. The box reports estimates from one regression of observed prescribing on quartiles of the survey prescribing rate, with the first quartile as the reference group. Standard errors are shown in parentheses. Number of observations: 359 physicians (from 333 clinics).

Figure A3. Mental models of the AI tool



Notes: Sample restricted to respondents assigned to the AI Tool treatment. The figure reports average agreement with each closed-form mental-model statement, separately for AI adopters and AI non-adopters. Agreement is measured on a scale from 1 (strongly disagree) to 5 (strongly agree). AI adoption is defined as updating beliefs by more than three percentage points in at least half of the vignettes (see Section 4.2). Full statement wording is reported in Appendix Table A2. Number of observations (physicians): 179.

Figure A4. Preference for receiving AI and dipstick signals separately



Notes: The figure shows the distribution of physicians' agreement with three statements about receiving the AI and dipstick signals separately. Responses are measured on a scale from 1 (strongly disagree) to 5 (strongly agree). Number of observations (physicians): 372.

C Data Appendix

C.1 Survey mental models

We elicit participants' mental models of the diagnostic signals first using open-ended responses, followed by closed-form survey items.

We categorize the open-ended responses using structured codebooks developed separately for the AI and dipstick treatment arms. We use the same coding categories across treatment arms whenever possible. In both the AI and dipstick arms, we code whether respondents describe the signal as informative; refer to its accuracy; raise concerns about bias, heterogeneous performance across patients, or non-optimal tool performance; use the signal only to confirm their own assessment; or instead emphasize anamnesis and own clinical experience. In addition, we code tool-specific considerations: for the AI tool, uncertainty about how to incorporate the prediction, use of additional patient data, black-box prediction, convenience, a preference for seeing the dipstick result, and the need for other diagnostics; and for the dipstick tool, sampling problems and a preference for other dipstick indicators such as nitrite, blood, glucose, or ketones. The categories are not mutually exclusive. Each response was coded independently by three coders, and we classify that a category is mentioned if at least two coders assigned it.

The closed-form statements probe respondents' mental models in a more structured way, including their perceptions about 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. All closed-form statements are listed in Table A2.

C.2 Administrative data sources

Our administrative data cover the full population of Danish citizens and primary care clinics, provided by Statistics Denmark and the Danish Health Data Authority (*Sundhedsdatastyrelsen*). We construct the population of primary care clinics using the national clinic register in September 2025 (*Yderregister*, YDER).

We use the Central Person Register (*Det Centrale Personregister*, CPR) to obtain basic demographic information on patients in 2024. For clinics in the *Yderregister*, we observe 2024 patient purchases of prescribed antibiotics from *Lægemiddeldatabasen* (LMDB) and 2024 primary care insurance claims from *Sygesikringsregisteret* (SSR). Data can be linked using unique individual and clinic identifiers.

The *Yderregister* is maintained by the Danish Health Data Authority. Note that we sent survey invitations using a different register, which is maintained by Medcom (MedCom

2025), an organization owned by the Danish Ministry of the Interior and Health, Danish Regions, and Local Government Denmark. However, more than 95% of invited clinics can be matched to the *Yderregister*. For survey respondents matched to the administrative data (. respondents from 333 clinics), we construct clinic-level measures of clinical practice. For a subset of 345 respondents from 320 clinics, we can additionally construct clinic-level average physician characteristics. Variable definitions are reported in Table A3.

C.3 Prescribing thresholds as proxies of diagnostic skill

We combine physicians' treatment decisions and their stated beliefs from the survey experiment to construct physician-level prescribing thresholds. Below, we present a framework under which these thresholds measure diagnostic skill.

Utility setup. Physician i faces patient j with true infection status $\omega_j \in \{0, 1\}$, where $\omega_j = 1$ denotes bacterial UTI, and chooses whether to prescribe an antibiotic, $d_{ij} \in \{0, 1\}$. Let $p_{ij} = \Pr(\omega_j = 1 \mid I_{ij})$ be the physician's posterior belief that the patient has a bacterial UTI, given information I_{ij}. Normalizing the loss from a correct decision to zero, a false positive incurs costs $c_i^{\text{FP}} > 0$ from unnecessary use and contribution to antimicrobial resistance, while a false negative incurs costs $c_i^{\text{FN}} > 0$, reflecting harm from a missed infection. This implies the following loss function:

$$L_i(d_{ij}, \omega_j) = \begin{cases} 0 & \text{if } d_{ij} = \omega_j, \\ c_i^{\text{FP}} & \text{if } d_{ij} = 1,\ \omega_j = 0, \\ c_i^{\text{FN}} & \text{if } d_{ij} = 0,\ \omega_j = 1. \end{cases}$$

Expected losses are $c_i^{\text{FP}}(1 - p_{ij})$ from prescribing and $c_i^{\text{FN}} p_{ij}$ from not prescribing. The physician prescribes if and only if $c_i^{\text{FP}}(1 - p_{ij}) \leq c_i^{\text{FN}} p_{ij}$, yielding the following threshold rule:

$$d_{ij} = 1 \iff p_{ij} \geq \tau_i^{\text{pref}}, \qquad \tau_i^{\text{pref}} = \frac{c_i^{\text{FP}}}{c_i^{\text{FN}} + c_i^{\text{FP}}}.$$

This preference threshold τ_i^{pref} rises in c_i^{FP} and falls in c_i^{FN}. Thus, physicians who place greater weight on the costs of antibiotic resistance have higher thresholds, while those who place greater weight on the costs of untreated infection have lower thresholds.

Diagnostic uncertainty and cost asymmetry. Assume that a physician with imperfect skill does not know her most informative posterior p_{ij}^* with certainty. Her stated belief p_{ij} is a point estimate, and she treats p_{ij}^* as lying in a symmetric ambiguity set, $\mathcal{P}_{ij}(\delta_i) = [p_{ij} - \delta_i,\ p_{ij} + \delta_i]$, where the radius $\delta_i \geq 0$ is smaller for more skilled physicians. Assume

the physician minimizes worst-case expected loss over $\mathcal{P}_{ij}(\delta_i)$. The worst case under prescribing is $p_{ij} - \delta_i$ (unnecessary prescription most likely), and under no prescribing $p_{ij} + \delta_i$ (missed infection most likely):

$$\max_{p \in \mathcal{P}_{ij}} L(1, p) = c_i^{\text{FP}}(1 - (p_{ij} - \delta_i)), \qquad \max_{p \in \mathcal{P}_{ij}} L(0, p) = c_i^{\text{FN}}(p_{ij} + \delta_i).$$

Finally, assume cost asymmetry, $c_i^{\text{FN}} > c_i^{\text{FP}}$ for every i. Such cost asymmetry is natural in the UTI context, where an untreated infection can progress to serious consequences for the patient, such as pyelonephritis or sepsis, while an unnecessary antibiotic course carries harm primarily through its contribution to population-level antibiotic resistance.

Decisions under diagnostic uncertainty. The physician prescribes when the worst-case loss from not prescribing exceeds that from prescribing, $c_i^{\text{FN}}(p_{ij} + \delta_i) \geq c_i^{\text{FP}}(1 - p_{ij} + \delta_i)$. Rearranging, the physician prescribes whenever her posterior is above her effective prescribing threshold, $p_{ij} \geq \tau_i^{\text{eff}}$, where:

$$\tau_i^{\text{eff}} = \underbrace{\frac{c_i^{\text{FP}}}{c_i^{\text{FN}} + c_i^{\text{FP}}}}_{\tau_i^{\text{pref}}} - \underbrace{\frac{c_i^{\text{FN}} - c_i^{\text{FP}}}{c_i^{\text{FN}} + c_i^{\text{FP}}}}_{>0 \text{ under } c_i^{\text{FN}} > c_i^{\text{FP}}} \cdot \delta_i.$$

The effective threshold thus equals the preference threshold minus a term proportional to diagnostic uncertainty, where that term is strictly positive under cost asymmetry. For any $\delta_i > 0$, the effective threshold is above the preference threshold, $\tau_i^{\text{eff}} < \tau_i^{\text{pref}}$, and the gap between both is increasing in δ_i: a low-skill physician with higher uncertainty prescribes more aggressively than her stated preferences imply, lowering the prescribing threshold to avoid the costs of an underestimated true infection probability. This result relates to Chan et al. (2022), who show that less-skilled agents optimally operate at lower effective decision thresholds conditional on preferences. In our setting, we observe beliefs but not prior beliefs (i.e., baseline prevalence) or the ground-truth infection state.

Empirical counterpart. We use post-signal beliefs and prescribing decisions from all vignettes j from non-diabetic patients (vignettes 1–3 and 7–9), where the relative benefit of treatment is similar across cases. The observed pairs (p_{ij}, d_{ij}) provide bounds around the prescribing threshold:

$$\tau_i \in \Big(\max_{j:\, d_{ij}=0} p_{ij},\ \min_{j:\, d_{ij}=1} p_{ij} \Big],$$

where the lower bound is the highest posterior belief without a prescription, and the upper bound is the lowest posterior at which the physician prescribes. If a physician always (never) prescribes, we set the lower (upper) bound to zero (one). When the bounds cross

by more than 10 percentage points, we leave the threshold unidentified; if they cross by less, we adjust them by widening each bound by 5 percentage points (truncated at $[0, 1]$), as such crossing likely reflects imprecision in the response (23 respondents). As our empirical measure of the prescribing threshold, $\hat{\tau}_i$, we use the lower observed bound, as a large number of respondents never prescribe and thus have an uninformative upper bound of one (68 respondents). We obtain valid threshold measures $\hat{\tau}_i$ for 279 physicians, of whom 138 are in the AI arm.

As prescribing thresholds also reflect preferences, we additionally report results using a prescribing threshold that is residualized by survey-elicited preferences (concern about antibiotic resistance, weight towards own patients, risk preferences, and time preferences), in order to obtain a preference-adjusted measure.

D Survey Instructions

Consent to the processing of your personal data by a research project

In connection with your participation in a trial as part of a research project at the University of Copenhagen, we hereby request your consent to us processing your personal data. We do so under the rules in the General Data Protection Regulation (GDPR).

See the information sheet (Link) for more details about the project and the processing of your personal data.

Title of the research project:
Choice of treatment decisions in primary care

I confirm that I have read the information sheet and that this forms the basis on which I consent to the processing of my personal data by the project.

I hereby give my consent that the University of Copenhagen may register and process my personal data in the abovementioned research project.

I do not consent

Your consent to the processing of personal data is voluntary, and may be withdrawn at any time. You may at any time change or withdraw your consent by contacting Hannes Ullrich (Associate Professor of Economics), +45 35 33 76 59, hannes.ullrich@econ.ku.dk.

If you withdraw consent, it will take effect from that point in time, and will not affect the legality of our work with your data in connection with the project up to that point. Your data will therefore continue to be included in the work done by the project up until the point at which you withdraw your consent

Background variables

What is your gender?

- Male
- Female
- Non-binary / third gender

What is your birth year?

Please select the option that best describes your role within your practice.

- General Practitioner (GP)
- Trainee physician
- Other medical staff
- Non-medical staff

Do you make prescription decisions in your current role?

- Yes
- No

Beliefs About Diagnostic Accuracy

First, we would like to ask for your perceptions of how accurate dipstick tests are in diagnosing urinary tract infections (UTIs).

Imagine you are performing a dipstick test on a urine sample. Consider a test commonly used in general practice, such as the Siemens Multistix.

> Please assume in all decisions in this study that the dipstick test result for nitrite is negative.

In patients who **actually have a urinary tract infection (UTI)**, what percentage do you believe will test **positive for leukocytes** on the dipstick test?
(This represents the test **sensitivity**, or its ability to correctly identify those **with** a UTI.)

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

In patients who **do not have a urinary tract infection (UTI)**, what percentage do you believe will test **negative for leukocytes** on the dipstick test?
(This represents the test **specificity**, or its ability to correctly identify those **without** a UTI.)

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

First Stage Vignette Introduction

[AI Treatment:]

> Please assume in all decisions in this study that the dipstick test result for nitrite is negative.

On the next pages, you will be presented with 6 patient files. Assume that these are new patients whom you have not treated before. For each patient file, we ask you to assess the probability that the patient has a UTI and to make a treatment decision.

We will ask for your assessment in two steps:

> **Step 1:** You will review the patient's demographics, primary symptoms, and medication.
> **Step 2:** You will then receive additional information: **the result of a data-driven AI support tool.**

How does the AI support tool work?
The AI support tool uses artificial intelligence to predict bacterial urinary tract infections in primary care patients. It is trained on data from patient records and test results provided by microbiological laboratories (KMA). Suppose the tool is clinically approved.

Patient information used:

- **clinical history** incl. the history of general practice consultations (diagnoses, referrals, prescriptions), laboratory test results, and hospitalizations
- **current dipstick test result**

Prediction result:

- **binary prediction of UTI**: UTI positive or negative

How accurate is the AI support tool?
The tool classifies a patient as UTI positive when the estimated UTI probability is above a certain threshold probability.

For your assessment, assume that in clinical trials, the AI support tool achieved a **sensitivity (true positive rate) of 80%** and a **specificity (true negative rate) of 60%**. This means that when the patient truly has a UTI, the tool will correctly give a positive result with 80% probability. When the patient truly does not have a UTI, the tool will correctly give a negative result with 60% probability.

▾ More information

The tool was developed by researchers at the University of Copenhagen and trained to predict the outcome of urine culture tests for UTI from microbiological laboratories in Denmark. Based on this training data, the tool calculates the probability of a UTI for new cases.

Similar AI support tools have been developed in health care systems in Israel, the Netherlands, and the USA.

- Herter W.E., Khuc J., Cina G., et al. (2022). Impact of a Machine Learning-Based Decision Support System for Urinary Tract Infections: Prospective Observational Study in 36 Primary Care Practices. *JMIR Medical Informatics* 10(5), e27795.
- Kanjilal S., Oberst M., Boominathan S., et al. (2020), A decision algorithm to promote outpatient antimicrobial stewardship for uncomplicated urinary tract infection. *Science Translational Medicine* 12(568).
- Yelin I., Snitser O., Novich G., et al. (2019), Personal clinical history predicts antibiotic resistance of urinary tract infections, *Nature Medicine* 25(7), 1143-1152.

[Dipstick Treatment:]

> Please assume in all decisions in this study that the dipstick test result for nitrite is negative.

On the next pages, you will be presented with 6 patient files. Assume that these are new patients whom you have not treated before. For each patient file, we ask you to assess the probability that the patient has a UTI and to make a treatment decision.

We will ask for your assessment in two steps:

> **Step 1:** You will review the patient's demographics, primary symptoms, and medication.
> **Step 2:** You will then receive additional information: **the dipstick test result for leukocytes.**

How does the dipstick test work?
The dipstick test gives a binary result: either leukocytes positive (leukocytes detected) or leukocytes negative (no leukocytes detected).

How accurate is the dipstick test?
The dipstick test yields a positive result when there is a trace or greater reaction to leucocyte esterase (LE).
For your assessment, assume that in clinical trials, the dipstick test for leukocytes achieved a **sensitivity (true positive rate) of 80%** and a **specificity (true negative rate) of 60%**. This means that when the patient truly has a UTI, the test will correctly give a positive result with 80% probability. When the patient truly does not have a UTI, the test will correctly give a negative result with 60% probability.

First Stage Vignettes: Rounds 1–6

Case No 1 of 6

Step 1: Please consider the following patient case.

Patient Information:

Age: 25

Gender: Female

Primary Symptoms:

Occasional burning sensation when urinating

normal urination frequency

no discomfort in the lower abdomen

no fever

Comorbidities:

Type 1 diabetes

Medications:

Insulin aspart (Novolog) via insulinpumpe

Note: Please assume in all decisions in this study that the dipstick test result for nitrite is negative.

How do you assess the probability that the patient has a urinary tract infection?

Definitely no UTI **Definitely a UTI**

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

What treatment would you select?

Start immediate antibiotic therapy (with Pivmecillinam, Trimethroprim, or Nitrofurantoin)

Prescribe antibiotic but ask patient to wait-and-see

No antibiotic treatment for now

[AI Treatment:]

Case No 1 of 6

Step 2: Please consider the patient case again.

Patient Information:

Age: 25

Gender: Female

Primary Symptoms:

Occasional burning sensation when urinating

normal urination frequency

no discomfort in the lower abdomen

no fever

Comorbidities:

Type 1 diabetes

Medications:

Insulin aspart (Novolog) via insulinpumpe

Note: Please assume in all decisions in this study that the dipstick test result for nitrite is negative.

AI Support Tool (Sensitivity: 80%, Specificity: 60%)

AI Support Tool Result:

UTI: negative

Note: The AI support tool uses the full clinical history available in the patient record along with the current dipstick test to predict the UTI probability.

How do you assess the probability that the patient has a urinary tract infection?

Your previous assessment: %

Definitely no UTI **Definitely a UTI**

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

Please make a treatment decision for this patient.

Would you perform a urine culture?

Please select all that apply.

- Yes, microscopical examination
- Yes, flexicult at point-of-care
- Yes, send the sample to a microbiological laboratory
- No

What treatment would you select?

- Start immediate antibiotic therapy (with Pivmecillinam, Trimethroprim, or Nitrofurantoin)
- Prescribe antibiotic but ask patient to wait-and-see
- No antibiotic treatment for now

[Dipstick Treatment:]

Case No 1 of 6

Step 2: Please consider the patient case again.

Patient Information:

Age: 25

Gender: Female

Primary Symptoms:

Occasional burning sensation when urinating

normal urination frequency

no discomfort in the lower abdomen

no fever

Comorbidities:

Type 1 diabetes

Medications:

Insulin aspart (Novolog) via insulinpumpe

Note: Please assume in all decisions in this study that the dipstick test result for nitrite is negative.

Dipstick Test (Sensitivity: 80%, Specificity: 60%)

Dipstick Test Result:

Leukocytes: positive

How do you assess the probability that the patient has a urinary tract infection?

Your previous assessment: 77%

Definitely no UTI **Definitely a UTI**

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

Please make a treatment decision for this patient.

Would you perform a urine culture?

Please select all that apply.

- Yes, microscopical examination
- Yes, flexicult at point-of-care
- Yes, send the sample to a microbiological laboratory
- No

What treatment would you select?

- Start immediate antibiotic therapy (with Pivmecillinam, Trimethroprim, or Nitrofurantoin)
- Prescribe antibiotic but ask patient to wait-and-see
- No antibiotic treatment for now

Mental Models

[AI Treatment:]

What factors made you use or ignore the AI support tool prediction? Please mention your thoughts about the tool.

Please respond in 2-3 sentences.

We now ask you to think about how you interpret the results provided by the AI support tool.

Please indicate your agreement with the following statements.

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
I am uncertain how to incorporate the AI-generated prediction result into my diagnostic assessment.	○	○	○	○	○
I am concerned that the accuracy of the AI tool may vary across different types of patients.	○	○	○	○	○
I have a good understanding of how the AI model analyzes patient data to generate its prediction.	○	○	○	○	○
The AI tool can provide me with information about an individual patient's current health status.	○	○	○	○	○
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 data used by the AI tool (patient records, dipstick result) is informative about a patient's current UTI status.	○	○	○	○	○
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.	○	○	○	○	○

[Dipstick Treatment:]

What factors made you use or ignore the dipstick test result? Please mention your thoughts about the test.

Please respond in 2-3 sentences.

→

We now ask you to think about how you interpret the results provided by the dipstick test.

Please indicate your agreement with the following statements.

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
I am concerned that the accuracy of the dipstick test may vary across different types of patients.	○	○	○	○	○
I believe that the dipstick test optimally analyzes information about a patient's UTI status from the urine sample.	○	○	○	○	○
I have a good understanding of how the dipstick test analyzes a urine sample to produce its result.	○	○	○	○	○
I am uncertain how to incorporate the dipstick test result into my diagnostic assessment.	○	○	○	○	○
I believe that the result of the dipstick test is influenced by the manufacturer to influence my treatment decision according to their objectives.	○	○	○	○	○
The dipstick test can provide me with information about an individual patient's current health status.	○	○	○	○	○
I believe that the urine sample is informative about a patient's current UTI status.	○	○	○	○	○

Second Stage Vignette Introduction

[Correlated Treatment (AI Treatment):]

> Please assume in all decisions in this study that the dipstick test result for nitrite is negative.

On the next pages, you will be presented with 3 additional patient files. As before, we ask you to assess the probability that the patient has a UTI and to make a treatment decision.

> In these new cases, you observe two indicators:
> (i) the result of the **dipstick test for leukocytes**
> (ii) the result of the **AI support tool** that you have already seen before (sensitivity: 80%, specificity 60%)

How does the dipstick test work?
The dipstick test gives a binary result: either leukocytes positive (leukocytes detected) or leukocytes negative (no leukocytes detected).

How accurate is the dipstick test?
The dipstick test yields a positive result when there is a trace or greater reaction to leucocyte esterase (LE).
For your assessment, assume that in clinical trials, the dipstick test for leukocytes achieved a **sensitivity (true positive rate) of 70%** and a **specificity (true negative rate) of 50%**. This means that when the patient truly has a UTI, the test will correctly give a positive result with 70% probability. When the patient truly does not have a UTI, the test will correctly give a negative result with 50% probability.

Note that the dipstick test result is also used by the AI prediction tool to calculate its prediction.

[Correlated Treatment (Dipstick Treatment):]

> Please assume in all decisions in this study that the dipstick test result for nitrite is negative.

On the next pages, we will present you with 3 more patient files. Again, we ask you to assess the probability that the patient has a UTI and to make a treatment decision.

> In these new cases, you observe two indicators:
> (i) the result of the **dipstick test for leukocytes** (sensitivity: 80%, specificity 60%)
> (ii) the result of an **AI support tool**

How does the AI support tool work?
The AI support tool uses artificial intelligence to predict bacterial urinary tract infections in primary care patients. It is trained on data from patient records and test results provided by microbiological laboratories (KMA). Suppose the tool is clinically approved.

Patient information used:

- **clinical history** incl. the history of general practice consultations (diagnoses, referrals, prescriptions), laboratory test results, and hospitalizations
- **current dipstick test result**

Prediction result:

- **binary prediction of UTI**: UTI positive or negative

How accurate is the AI support tool?
The tool classifies a patient as UTI positive when the estimated UTI probability is above a certain threshold probability.

For your assessment, assume that in clinical trials, the AI support tool achieved a **sensitivity (true positive rate) of 90%** and a **specificity (true negative rate) of 70%**. This means that when the patient truly has a UTI, the tool will correctly give a positive result with 90% probability. When the patient truly does not have a UTI, the tool will correctly give a negative result with 70% probability.

▾ More information about the AI support tool

The tool was developed by researchers at the University of Copenhagen and trained to predict the outcome of urine culture tests for UTI from microbiological laboratories in Denmark. Based on this training data, the tool calculates the probability of a UTI for new cases.

Similar AI support tools have been developed in health care systems in Israel, the Netherlands, and the USA.

- Herter W.E., Khuc J., Cina G., et al. (2022). Impact of a Machine Learning-Based Decision Support System for Urinary Tract Infections: Prospective Observational Study in 36 Primary Care Practices. *JMIR Medical Informatics* 10(5), e27795.
- Kanjilal S., Oberst M., Boominathan S., et al. (2020), A decision algorithm to promote outpatient antimicrobial stewardship for uncomplicated urinary tract infection. *Science Translational Medicine* 12(568).
- Yelin I., Snitser O., Novich G., et al. (2019), Personal clinical history predicts antibiotic resistance of urinary tract infections, *Nature Medicine* 25(7), 1143-1152.

[Uncorrelated Treatment (AI Treatment):]

> Please assume in all decisions in this study that the dipstick test result for nitrite is negative.

On the next pages, you will be presented with 3 additional patient files. As before, we ask you to assess the probability that the patient has a UTI and to make a treatment decision.

> In these new cases, you observe two indicators:
> (i) the result of the **dipstick test for leukocytes**
> (ii) the result of the **AI support tool**, which does not include the dipstick test anymore but is trained on a larger dataset, such that it achieves the same accuracy as before (sensitivity: 80%, specificity 60%)

How does the dipstick test work?
The dipstick test gives a binary result: either leukocytes positive (leukocytes detected) or leukocytes negative (no leukocytes detected).

How accurate is the dipstick test?
The dipstick test yields a positive result when there is a trace or greater reaction to leucocyte esterase (LE).
For your assessment, assume that in clinical trials, the dipstick test for leukocytes achieved a **sensitivity (true positive rate) of 70%** and a **specificity (true negative rate) of 50%**. This means that when the patient truly has a UTI, the test will correctly give a positive result with 70% probability. When the patient truly does not have a UTI, the test will correctly give a negative result with 50% probability.

[Uncorrelated Treatment (Dipstick Treatment):]

> Please assume in all decisions in this study that the dipstick test result for nitrite is negative.

On the next pages, we will present you with 3 more patient files. Again, we ask you to assess the probability that the patient has a UTI and to make a treatment decision.

> In these new cases, you observe two indicators:
> (i) the result of the **dipstick test for leukocytes** (sensitivity: 80%, specificity 60%)
> (ii) the result of an **AI support tool**

How does the AI support tool work?
The AI support tool uses artificial intelligence to predict bacterial urinary tract infections in primary care patients. It is trained on data from patient records and test results provided by microbiological laboratories (KMA). Suppose the tool is clinically approved.

Patient information used:

- **clinical history** incl. the history of general practice consultations (diagnoses, referrals, prescriptions), laboratory test results, and hospitalizations

Prediction result:

- **binary prediction of UTI**: UTI positive or negative

How accurate is the AI support tool?
The tool classifies a patient as UTI positive when the estimated UTI probability is above a certain threshold probability.

For your assessment, assume that in clinical trials, the AI support tool achieved a **sensitivity (true positive rate) of 90%** and a **specificity (true negative rate) of 70%**. This means that when the patient truly has a UTI, the tool will correctly give a positive result with 90% probability. When the patient truly does not have a UTI, the tool will correctly give a negative result with 70% probability.

▾ More information about the AI support tool

The tool was developed by researchers at the University of Copenhagen and trained to predict the outcome of urine culture tests for UTI from microbiological laboratories in Denmark. Based on this training data, the tool calculates the probability of a UTI for new cases.

Similar AI support tools have been developed in health care systems in Israel, the Netherlands, and the USA.

- Herter W.E., Khuc J., Cina G., et al. (2022). Impact of a Machine Learning-Based Decision Support System for Urinary Tract Infections: Prospective Observational Study in 36 Primary Care Practices. *JMIR Medical Informatics* 10(5), e27795.
- Kanjilal S., Oberst M., Boominathan S., et al. (2020), A decision algorithm to promote outpatient antimicrobial stewardship for uncomplicated urinary tract infection. *Science Translational Medicine* 12(568).
- Yelin I., Snitser O., Novich G., et al. (2019), Personal clinical history predicts antibiotic resistance of urinary tract infections, *Nature Medicine* 25(7), 1143-1152.

Second Stage Vignettes: Rounds 7–9

Case No 1 of 3

Step 1: Please consider the following patient case.

Patient Information:

Age: 25

Gender: Female

Primary Symptoms:

Occasional burning sensation when urinating

normal urination frequency

no discomfort in the lower abdomen

no fever

Comorbidities:

Type 1 diabetes

Medications:

Insulin aspart (Novolog) via insulinpumpe

Note: Please assume in all decisions in this study that the dipstick test result for nitrite is negative.

How do you assess the probability that the patient has a urinary tract infection?

Definitely no UTI **Definitely a UTI**

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

What treatment would you select?

Start immediate antibiotic therapy (with Pivmecillinam, Trimethroprim, or Nitrofurantoin)

Prescribe antibiotic but ask patient to wait-and-see

No antibiotic treatment for now

[Correlated Treatment:]

Case No 2 of 3

Step 2: Please consider the patient case again.

Patient Information:

Age: 25

Gender: Female

Primary Symptoms:

Occasional burning sensation when urinating

normal urination frequency

no discomfort in the lower abdomen

no fever

Comorbidities:

Type 1 diabetes

Medications:

Insulin aspart (Novolog) via insulinpumpe

Note: Please assume in all decisions in this study that the dipstick test result for nitrite is negative.

Dipstick Test (Sensitivity: 70%, Specificity: 50%)

Dipstick Test Result:

Leukocytes: negative

Note: Please consider that the dipstick test is also included in the AI support tool prediction.

AI Support Tool (Sensitivity: 80%, Specificity: 60%)

AI Support Tool Result:

UTI: negative

Note: The AI support tool uses the full clinical history available in the patient record along with the current dipstick test to predict the UTI probability.

How do you assess the probability that the patient has a urinary tract infection?

Your previous assessment: 65%

Definitely no UTI **Definitely a UTI**

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

Please make a treatment decision for this patient.

Would you perform a urine culture?

Please select all that apply.

- Yes, microscopical examination
- Yes, flexicult at point-of-care
- Yes, send the sample to a microbiological laboratory
- No

What treatment would you select?

- Start immediate antibiotic therapy (with Pivmecillinam, Trimethroprim, or Nitrofurantoin)
- Prescribe antibiotic but ask patient to wait-and-see
- No antibiotic treatment for now

[Uncorrelated Treatment:]

Case No 1 of 3

Step 2: Please consider the patient case again.

Patient Information:

Age: 25

Gender: Female

Primary Symptoms:

Occasional burning sensation when urinating

normal urination frequency

no discomfort in the lower abdomen

no fever

Comorbidities:

Type 1 diabetes

Medications:

Insulin aspart (Novolog) via insulinpumpe

Note: Please assume in all decisions in this study that the dipstick test result for nitrite is negative.

Dipstick Test (Sensitivity: 80%, Specificity: 60%)

Dipstick Test Result:

Leukocytes: positive

AI Support Tool (Sensitivity: 90%, Specificity: 70%)

AI Support Tool Result:

UTI: positive

Note: The AI support tool uses the full clinical history available in the patient record to predict the UTI probability.

How do you assess the probability that the patient has a urinary tract infection?

Your previous assessment: 51%

Definitely no UTI **Definitely a UTI**

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

Please make a treatment decision for this patient.

Would you perform a urine culture?

Please select all that apply.

- Yes, microscopical examination
- Yes, flexicult at point-of-care
- Yes, send the sample to a microbiological laboratory
- No

What treatment would you select?

- Start immediate antibiotic therapy (with Pivmecillinam, Trimethroprim, or Nitrofurantoin)
- Prescribe antibiotic but ask patient to wait-and-see
- No antibiotic treatment for now

We now ask you about your preferences regarding the way the AI support tool and the dipstick result are presented.

Please indicate your agreement with the following statements.

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
I would prefer to see the dipstick result separately, even if it is already incorporated into the AI tool's prediction.	○	○	○	○	○
I believe my interpretation of the dipstick test result is more accurate than how the AI tool incorporates it into its prediction.	○	○	○	○	○
Even if the AI tool uses the dipstick result optimally, it is important for me to see the underlying dipstick test result separately.	○	○	○	○	○

In general, how concerned are you about antimicrobial resistance when making prescription decisions in your clinical practice?

Please use a scale from 0 to 10, where 0 means “not concerned at all” and a 10 means you are “very concerned.”

Not concerned at all **Very concerned**

0 1 2 3 4 5 6 7 8 9 10

→

Out of 100 physicians randomly selected from all over Denmark, how many do you think prescribe fewer antibiotics per patient visit for suspected UTI than you do?

0 10 20 30 40 50 60 70 80 90 100

→

In the following, we would like to ask you some general questions about yourself.

In general, how willing or unwilling are you to take risks?

Please use a scale from 0 to 10, where 0 means “completely unwilling to take risks” and a 10 means you are “very willing to take risks”.

Completely unwilling to take risks **Very willing to take risks**

0 1 2 3 4 5 6 7 8 9 10

→

How willing are you to give up something that is beneficial for you today in order to benefit more from that in the future?

Please again indicate your answer on a scale from 0 to 10, where 0 means you are "completely unwilling to do so" and a 10 means you are "very willing to do so".

Completely unwilling to do so										**Very willing to do so**
0	1	2	3	4	5	6	7	8	9	10

→

Imagine there is a national health initiative where general practices can access up to 100,000 DKK to improve care for their own patients (e.g., enhanced services, better equipment, additional staff support). You can decide how much of this funding to claim for improving patient care in your own practice. Any funds you do not claim will be donated to a charity that supports the most vulnerable patients across the broader Danish population.

How much of the funds (in 1,000 DKK) would you claim to improve patient care in your own practice? The remaining funds would be donated to a charity.

0K	10K	20K	30K	40K	50K	60K	70K	80K	90K	100K

→

Imagine you have the option to purchase a clinical decision support system similar to the AI support tool described to you before. The tool synthesizes the complete patient file and gives a highly accurate diagnostic recommendation (high sensitivity and specificity).

Suppose a fee has to be paid for each consultation in which this support system is used.

How much would you be willing to pay for the tool per consultation in US Dollar?

If you are not interested in using such a tool, please enter 0.

→

Generative AI is a type of artificial intelligence that creates output (e.g., text or images) in response to prompts. Examples of Generative AI include ChatGPT or Claude. Some medical journal systems also offer the possibility to assist in journal taking using generative AI.

Have you heard about generative AI?

- Yes
- No

Do you use generative AI for your clinical practice (administrative or medical purposes)?

- Yes, integrated in the medical journal system
- Yes, outside of the medical journal system (e.g., ChatGPT)
- Both
- No

Do you use generative AI outside your job?

- Yes
- No

In general, how helpful do you think new generative AI tools can be for the following tasks:

	Not at all	A little	A moderate amount	A lot	A great deal	Don't know
Administrative tasks	○	○	○	○	○	○
Extract and summarize information from patient journals	○	○	○	○	○	○
As a decision support system for diagnosis and treatment	○	○	○	○	○	○

Appendix References

Carey, Iain M, Fay J Hosking, Tess Harris, Stephen DeWilde, Carole Beighton, and Derek G Cook, "An evaluation of the effectiveness of annual health checks and quality of health care for adults with intellectual disability: an observational study using a primary care database," *Health Services and Delivery Research*, September 2017, *5* (25), 1–170.

Chan, David C, Matthew Gentzkow, and Chuan Yu, "Selection with Variation in Diagnostic Skill: Evidence from Radiologists," *The Quarterly Journal of Economics*, April 2022, *137* (2), 729–783.

Dansk Selskab for Almen Medicin, "FAQta-ark om urinvejsinfektioner i almen praksis," 2020. https://content.dsam.dk/guides/vejlednings-pdf/uvi_faqta-ark.pdf. Accessed: June 15, 2026.

MedCom, "Lægepraksis i Danmark," 2025. https://medcom.dk/standarder/ydere-lokationsnumre/laegepraksis-i-danmark/. Accessed: June 15, 2026.