The Foerder Institute for Economic Research
at Tel Aviv University

The Eitan Berglas School of Economics

מכון למחקר כלכלי על שם ד"ר ישעיהו פורדר
על יד אוניברסיטת תל אביב

בית-הספר לכלכלה ע"ש איתן ברגלס

Challenging Encounters and Within-Physician Practice Variability

Gabriel Chodick, Yoav Goldstein, Ity Shurtz, Dan Zeltzer

Discussion Paper No. 2-2023

The Foerder Institute for Economic Research
and
The Sackler Institute of Economic Studies

Challenging Encounters and Within-Physician Practice Variability

Gabriel Chodick, Yoav Goldstein, Ity Shurtz, Dan Zeltzer*

February 9, 2023

*Gabriel Chodick: Sackler School of Medicine, Tel-Aviv University, Israel and Maccabitech, Maccabi Healthcare Services, chodick@tauex.tau.ac.il. Yoav Goldstein: Berglas School of Economics, Tel Aviv University, yoavg2@mail.tau.ac.il. Ity Shurtz: Department of Economics, Ben Gurion University, Israel, shurtz@bgu.ac.il. Dan Zeltzer: Berglas School of Economics, Tel Aviv University, and IZA Institute of Labour Economics dzeltzer@tauex.tau.ac.il. An earlier version of this paper titled "Emotional Events and Physician Behavior" constituted Goldstein's master's thesis at the Hebrew University of Jerusalem. We thank Michael Gilead, Ran Spiegler, and participants at the 2021 European Winter Meeting of the Econometric Society, the 2022 Meeting of the American Society of Health Economists, and The 9th Annual European Health Economics Assoication PhD Conference for helpful comments. Shurtz gratefully acknowledges support from the Israel Science Foundation (Grant 456/21). Zeltzer gratefully acknowledges support from the Pinhas Sapir Center for Development and the Foerder Institute for Economics Research.

Abstract

We examine how physician decisions are impacted by difficult cases—encounters with newly diagnosed cancer patients. Using detailed administrative data, we compare primary care physicians' decisions in visits that occurred before and after difficult cases and matched comparison cases by the same physicians on other dates. Immediately following a difficult case, physicians increase referrals for common tests, including diagnostic tests unrelated to cancer. The effect lasts only for about an hour and is not driven by patient selection or schedule disruption. The results highlight difficult encounters as a source of variability in physician practice.

Keywords: primary care, practice variation, intra-rater reliability

JEL Classification: I1, D91

1 Introduction

A physician may reach different conclusions when considering similar, or even identical, cases at different times. Such within-physician inconsistency is a cause for concern (Kraemer et al., 2012; Kraemer, 2014), but, it is both hard to document in the field and not well understood.[1]

Physicians commonly find themselves in difficult clinical encounters, after which they immediately continue to see other patients. This paper explores whether such encounters are a source of variability in subsequent medical practice. Specifically, we study how physicians' practice deviates from their own baseline after they encounter patients newly diagnosed with cancer ("difficult cases"). We focus on cancer because it is a fairly common, yet serious and often terminal condition. Further, unlike other conditions, cancer has a clear diagnosis date, which we accurately observe.

We draw on administrative data from Maccabi Healthcare Services, a large Israeli HMO, that cover about a quarter of the Israeli population. Particularly appealing for our purpose, these data provide a comprehensive description of all clinical encounters for each physician over the entire study period of 2012–2015, including the precise timing of visits, patients' medical histories, and outcomes. Therefore, they allow us to observe decisions in great detail, at the baseline and following difficult cases. We supplement these data with data from the Israel National Cancer Registry, to which reporting of every new cancer case is mandatory.

To evaluate the impact of difficult cases on subsequent physician decisions, we match each difficult case with a non-cancer encounter of the same physician in other weeks, at the same time of the year, weekday, and time of day. We refer to these matched cases as *index* cases. We then compare *treated* visits that occurred before and after difficult

[1] Within-expert variability (which Kahneman et al. (2021) call *occasion noise*) is much harder to document than between-expert variability. Still, within-expert variability has been demonstrated in some medical areas, including, for example, the assessment of coronary angiograms, emergency imaging, and lower-limb spasticity, (Detre et al., 1975; Robinson et al., 1999; Banky et al., 2019).

index cases to *comparison* visits that occurred before and after matched (non-cancer) index cases. Matching the time of index cases aims to eliminate seasonality and weekly periodicity issues.

The assumption underlying our approach is that there are no systematic baseline differences between treated and comparison visits. This assumption is plausible given the centralized and semi-automated scheduling system of the HMO, in which patients choose their time slots from those available. It is also supported by the evidence: even though we only match the time of index visits, we observe no systematic differences between treated and comparison visits in the baseline (pre-treatment) patient mix. Further, placebo analyses using pre-determined characteristics as outcomes yield (desired) null effects.

We find that in visits that occur shortly after a difficult case, physician utilization of common medical tests increases by 5% relative to their baseline, pre-treatment rate. Other visit outcomes, including drug prescriptions, referrals to specialists, and referrals to the emergency room do not significantly change. This increase in testing is transient, persisting only for about an hour. These results are significant and robust to alternative definitions of both the set of tests considered and the choice of comparison cases. The magnitude of the increase in testing does not vary with physicians' clinical experience.

To gain additional insight into which testing decisions are being most affected by difficult cases, we evaluate the impact of difficult cases on the congruence of physicians' testing decisions with the predicted testing decisions of their colleagues, a benchmark for the prevailing practice. We find that difficult cases increase the correlation of physicians' testing decisions with the propensity of other physicians to test. This suggests that difficult cases do not simply induce a uniformly higher rate of testing but rather increase testing that conforms to the prevailing practice.

Considering potential explanations, we argue that it is unlikely that physicians learn from difficult encounters information that is pertinent just to a few (predominantly non-cancer) cases that immediately follow them, regardless of their prior clinical experience. In addition, we find that the duration of subsequent visits does not change, implying

that the increase in testing is not due to physicians substituting testing for time due to a schedule disruption. Finally, cancer screening tests alone do not make up for the increase in testing. Taken together, these results highlight difficult cases as a potential source of a subtle and temporary change in physician practice.

What might explain such temporary change in practice? Difficult cases may make the prospects of other serious conditions loom larger for a while. They may also make physicians nervous about missing a serious diagnosis. Both changes would increase the expected value of testing, particularly among marginal patients. While we cannot separately identify these (and other) potential mechanisms, we discuss how others might do so in future work.

Regarding contribution to the literature, a large body of work documents inconsistencies *between* physicians.[2] Our work contributes to a small but growing literature that focuses on potential sources of *within*-physician variability. Existing works show an association between time of day and physician decision making. For example, physicians have been shown to be more likely to prescribe opioids, skip preventive health measures, skip handwashing, and lower the probability of inpatient admission to the ER at the end of the day (Dai et al., 2015; Neprash and Barnett, 2019; Hsiang et al., 2019; Jin et al., 2020). Recent works also highlight heuristic thinking as an alternative source of within-physician variability (Singh, 2021; Jin et al., 2021; Shurtz et al., 2021; Ly, 2021; Wang et al., 2022). Our results highlight a new source of within-physician practice variability: challenging encounters, which are weaved into physicians' clinical work routine. And while we focus on cancer, which we accurately observe, other types of challenging encounters may have similar impacts.

Finally, our work is also related to existing literature that shows that different arbitrary events, such as sports matches, weather, pollution, and the news influence expert decisions in various domains (Eren and Mocan, 2018; Chen and Loecher, 2020; Heyes and Saberian, 2019; Kahn and Li, 2019; Geerling et al., 2020). For a recent

[2]For examples of between-physician variability in practice, see Van Parys and Skinner (2016), Abaluck et al. (2016), Currie and MacLeod (2017), Molitor (2018), and Currie and MacLeod (2018), Chan et al. (2022).

review, see Kahneman et al. (2021).

The remainder of the paper is organized as follows. Section 2 discusses the data. Section 3 lays out our empirical strategy. Section 4 presents our main results. Section 5 discusses potential explanations and related evidence. Section 6 concludes.

2 Data

Data source. Our data come from Maccabi Healthcare Services (in short, Maccabi), one of Israel's four non-profit HMOs that provide universal tax-funded healthcare coverage to all Israeli residents. Maccabi is the second largest of these four HMOs, covering approximately two million patients nationwide. Coverage largely resembles that of Medicare Parts A, B, and D. Maccabi is an integrated payer-provider that either directly employs or contracts with a national network of physicians and outpatient clinics. It owns three hospitals, and it procures services for its members from external providers. All of its primary care physicians (PCPs) are connected through a unified electronic health records system.

Our data cover three of the country's five districts in which three-quarters of Maccabi's patients reside.[3] The population we draw our sample from includes all 30 million visits to 1,133 of Maccabi's PCPs made by 1.5 million patients between 2012 and 2015. We observe the exact timing of every visit, physician and patient identifiers, and a visit summary, which includes diagnoses, orders of laboratory and imaging tests, and drug prescriptions. We also observe patient and physician characteristics, including patient demographic information and existing chronic conditions, as well as physician age, gender, and experience.

Difficult cases. We define a difficult case as the first encounter between a PCP and her patient within 30 days after the patient was diagnosed with cancer.[4] To validate

[3]Smaller regions were excluded by Maccabi for confidentiality reasons.

[4]We include all cancer sites, except for rare types excluded under a cell-suppression policy to preserve privacy. See Table A1 for details and descriptive statistics.

that the cancer diagnosis in Maccabi records indicates a newly diagnosed condition (as opposed to a record of an old diagnosis during a follow-up visit), we cross-check against the Israeli Cancer Registry, to which all new cancer diagnoses must be reported by law. We exclude Maccabi diagnoses that did not have a corresponding registry fewer than 30 days before or after them. The median physician in our sample was exposed to five difficult cases during our study period (Figure A1).

Nearly all patients are informed about a cancer diagnosis by an oncology specialist, not their regular PCP (even if the latter sometimes updates the records), so these encounters do not involve the PCPs breaking the news to patients. However, the PCP is the patient's main point of contact to discuss the news and subsequent treatment, and the patient and the PCP typically have a preexisting, often long-standing, relationship, making it likely that these encounters are noteworthy and potentially challenging for the physician. Rich medical literature documents that it is difficult for physicians to discuss bad news with patients. The literature documents affective responses by the physicians, such as fear and anxiety, to such encounters (Ptacek et al., 2001; Fallowfield and Jenkins, 2004; Amiel et al., 2006; Martin Jr et al., 2015).

Comparison cases. We match each difficult case with similar cases by the same PCP in other years, in the same period of the year, day of the week, and time of day. We start by matching each difficult case with all visits by the same physician in other years that did not involve newly diagnosed cancer patients. To account for seasonality and weekly periodicity differences in visits, we restrict the sample to visits that occurred in the five-week period that includes the week of the year of the difficult case and two weeks before and after it, on the same weekday as the difficult case. Finally, to match on time of day, we select only visits with the same sequential number during the day as the difficult case.[5] This method matches each difficult case with up

[5]We use the sequential number of the visit during the day as a proxy for the time of day of the difficult case because it is easier to implement. However, as shown in Figure A2, the distribution of the time of day of difficult and comparison cases is very similar. This figure also shows that the timing of difficult cases is

to 15 visits during our four-year study period (up to five weeks in each of the three alternative years). We exclude 82 difficult cases that we were not able to match to any comparison case. Thanks to the regularity of physician schedules, we end up with an average of 12 comparison cases for each difficult case.

To check the robustness of our results to our choice of comparison cases, we also reproduce our findings using two alternative sets of comparison cases. First, instead of using visits from different years, we create a new set of comparison cases using visits from the same year as the difficult case, two weeks before and after it (*Alternative I*). This approach is more robust to potential long-term trends in outcomes. Second, we restrict the original set of comparison cases to only those in the exact week of the year in other years (*Alternative II*). This approach more accurately accounts for seasonality. Figure A3 shows the frequency of cases in each comparison group over time. For all three definitions, the cases are evenly spread over time, exhibiting no time trend and no significant difference in frequency relative to the set of difficult cases and the unrestricted set of all visits.

Sample construction. We construct the sample of PCP visits in two steps. First, we pool all difficult cases and their matches over the period of July 2012 through December 2015. This yields 5,368 difficult cases and 64,042 matched comparison cases, handled by 747 physicians (excluding 23 index cases that are the only ones recorded for the physician on a given date). Together, we refer to these difficult cases and matched non-cancer cases as the *index* cases. In the second step, we keep all visits that occurred shortly before or after each index case. In the baseline analysis, we focus on a window of up to eight visits before and after each index case (N = 971,943 visits, of which 73,821 are associated with difficult cases). For studying the dynamic of the effects, we use a wider window of up to 12 visits before and 18 visits after each index event (N = 1,660,257 visits, of which 124,227 are associated with difficult cases).

very similar to the timing of all primary care visits in our data.

Outcomes. Our main outcomes of interest are the physician's actions during a visit. We record indicators for whether the physician used any of the five most common lab tests or any of the five most common imaging and other medical tests (see Table A2 for the list of most common tests); an indicator for any drug prescription; an indicator for any referrals to specialists or (separately) to the emergency room; and visit duration in minutes. We explore the robustness of the results to the chosen measure of tests. Alternative measures include an indicator for each of the following: the five most common lab tests, the five most common imaging and other tests, the seven most common tests of each type, and the three most common tests of each type. An additional outcome that we use as a robustness test is the total number of tests (among the five most common tests of each type) that were given during the visit.

3 Empirical Strategy

Our main analysis consists of estimating a series of difference-in-differences (DD) specifications to examine the impact of difficult cases on different outcomes. Using our sample of eight-visit windows around the index event (excluding the index case itself), we estimate:

$$Y_{imt} = \eta Treat_{imt} + \tau Post_{imt} + \mu Post_{imt} \cdot Treat_{imt} + \xi_m + \nu_{imt}, \tag{1}$$

where the subscripts i, m, and t denote physician, match, and time; Y is one of several outcomes; $Treat$ is an indicator for treated visits (i.e., visits occurring before or after difficult cases, as opposed to matched index cases); $Post$ is an indicator for visits occurring after the index case; ξ denotes the match fixed effect; and ν_{imt} is the error term. All standard errors calculated throughout the analysis are clustered at the match level. The parameter of interest is μ, which captures the average treatment effect of difficult cases on the outcome. We also plot estimates from a more flexible event-study

DD:

$$Y_{imt} = \beta Treat_{imt} + \sum_{r} \left(\gamma_{r(imt)} + \delta_{r(imt)} Treat_{imt}\right) + \psi_m + \varepsilon_{imt}, \tag{2}$$

where $r(i, m, t)$ is the visit number relative to the index case (we bin pairs of adjacent visits, to reduce noise); γ_r and δ_r are indicator variables for the visit number and its interaction with $Treat$, respectively ($r = -1$ is the omitted level); ψ_m denotes the match fixed-effect; and ε_{imt} is the error term. The parameter of interest is δ_r, which captures the treatment effect for visits occurring at different times relative to the index case. This more flexible specification allows us to evaluate the pre-trends (which should be zero under the parallel trends assumption) and examine the persistence in treatment effects.

Identification and supporting evidence. The key identification assumption is that within matches, outcomes of treated and comparison visits have parallel time trends absent the treatment. This assumption is supported by the fact that Maccabi's scheduling system allows patients to select their visit time from all available slots of a physician (through a web portal, a mobile app, or a 24/7 national call center). Because patients do not know the nature of other visits when selecting their slot, presumably the characteristics of visits before and after difficult and comparison cases are no different. We provide supporting evidence for this assumption. First, we compare the pre-period outcomes between the treatment and comparison groups of visits. That is, using the visits with $r(imt) < 0$, we estimate:

$$Y_{imt} = \alpha Treat_{imt} + \zeta_m + u_{imt}, \tag{3}$$

where ζ denotes the match fixed effect. The results of this analysis are shown in Table 1. Indeed, during the pre-period, patient characteristics and visit outcomes are balanced across the treatment and comparison visits. Table A3 shows that patient characteristics and visit outcomes are balanced in the alternative samples as well. Second, we run a set of negative control ("placebo") regressions, where we estimate equation (2) using

various patient and case (pre-determined) characteristics as the outcome. Under the identification assumption, there should be no difference in these characteristics between treated and comparison visits—both before and after the index event. Table A4 shows that indeed these differences are not significantly different from zero.

4 Results

4.1 The Effect of Difficult Cases on Physician Practice

Table 2 shows the estimates from equation (1) for the impact of difficult cases on the outcomes of subsequent visits. Difficult cases increase physicians' use of testing. In the visits that immediately follow the difficult index case, physicians significantly increase their testing rate by 1.23 percentage points (a 4.5% increase over the pre-treatment baseline testing rate of 27.43%). There is no significant impact on other visit outcomes, specifically, prescriptions and referral rates to specialists and the ER.

The increase in testing is robust to a conservative Bonferroni correction for multiple-hypothesis testing (the standard p-value for the increase in testing is 0.00063; the Bonferroni-corrected p-value is 0.0025, i.e., well below 1%). The results are also robust to using alternative comparison group definitions (Table A5), controlling for time fixed-effects and patient characteristics (Table A6), and using alternative sets of tests as the outcome variable (Table A7).

To further investigate the timing of the increase in testing after difficult cases, Figure 1 presents the event-study DD estimates of equation (2). Panel (a) shows the residualized means of both treatment and comparison visits (according to all three definitions) around the time of the index case. Panel (b) shows the point estimates and 90% confidence intervals for the differences between treated and comparison visits (δ_r from equation (2)). As expected, before the exposure to the difficult case, there is no difference between the outcomes of treatment and comparison visits. The testing rate in treated visits sharply increases immediately after the difficult case, whereas

the testing rate in comparison visits smoothly continues its pre-index trend. This divergence persists for about eight visits—approximately one hour. Subsequently, the gap between the treatment and comparison visits closes.

4.2 Elevated Testing and Conformity to the Prevailing Practice

Which patients are being tested more following difficult cases? Do physicians uniformly test more, or do they test more the "marginal" patients—those they (or others) would have likely considered testing anyway? To gain insight, we compare physician decisions against the decisions of all other physicians for observably similar cases, which serve as a proxy for the prevailing professional practice (this method is similar toCurrie et al. (2016), and Currie and MacLeod (2017)).

Our strategy involves three steps. First, we train a standard machine learning model to predict the probability that the average physician would refer each patient to tests based on the observed case characteristics. We do so using data on testing decisions by all physicians at all times, not specifically around difficult cases. Appendix B discusses this model construction in detail. Second, we predict the probability of testing based on each case characteristic, for all visits in our sample. We refer to this predicted probability of testing as the testing *propensity score* (denote by PS_{imt}), as it reflects the propensity of all practicing physicians to test similar cases. Third, we estimate a triple difference model to evaluate how the correlation of physician testing decisions with the propensity score changes after they see a difficult case. Specifically, we interact the baseline specification from (1) with the (continuous) propensity score. We estimate:

$$\begin{aligned} Y_{it} =& \beta_0 \cdot Treat_{imt} + \beta_1 \cdot PS_{imt} + \beta_2 \cdot PS_{imt} \cdot Treat_{imt} + \\ & \gamma_0 \cdot Post_{imt} + \gamma_1 \cdot Post_{imt} \cdot PS_{imt} + \gamma_2 \cdot Post_{imt} \cdot Treat_{imt} \\ & \delta_0 \cdot Post_{imt} \cdot Treat_{imt} \cdot PS_{imt} + \phi_m. \end{aligned} \tag{4}$$

The parameters of interest are γ_2, and δ_0, which capture two respective aspects of the change in physician testing decisions following a difficult case: the change in physicians' *baseline* rate of testing and the change in the *correlation* between a physicians' testing decisions and the predicted propensity score.

Figure 2 shows the estimated effect and 90% confidence interval for every level of the score. The figure shows that the magnitude of the increase in testing referrals induced by difficult cases is increasing in the test propensity score. Namely, physicians' increase in testing following difficult cases is in agreement with the prevailing professional practice.

5 Potential Explanations

Why do physicians order more tests after a difficult case? Considering our results, we argue that it is unlikely that elevated testing following difficult cases reflects learning, schedule disruption, or cancer-specific practice response. Alternative explanations—which we cannot separately identify—include an increase in the salience of rare adverse patient outcomes, increased aversion to missing a diagnosis, or an emotional response to the difficult case. (These explanations are neither exclusive nor exhaustive). The rest of this section discusses how these different explanations fare relative to the evidence and highlight potential directions for future work.

Learning from Experience. Perhaps physicians learn from difficult encounters and that affects their subsequent testing decisions. We argue that this is unlikely for three reasons. First, given the variety of conditions PCPs handle and the quasi-random assignment of patient appointments, it is unlikely that any of the handful of patients seen immediately after a cancer patient has a condition that is materially related to the index cancer case. Second, learning is, by definition, persistent (at least to some extent), whereas the estimated increase in testing is very short-lived, lasting for only about an hour. Finally, learning should be more pronounced among the least

experienced physicians, whereas the effects we document do not vary by physician experience, as measured by either the physician age or the number of previous difficult cases seen during the study period.[6]

Schedule Disruption. Recent studies show that disruptions to the physician's schedule can be associated with shorter subsequent visits, which in turn affect physician behavior—physicians may make up for lost time by testing more (Freedman et al., 2021; Neprash, 2016). Figure A4 compares the distribution of visit duration for the index difficult cases and other visits in our sample. On average, physician encounters with newly diagnosed cancer patients are 40% (three minutes) longer than the average visit.[7] To directly examine this issue, we estimate whether there are any treatment effects on visit duration. Panel B of Table 2 shows no significant changes to the average visit duration in the eight visits following a difficult case, suggesting that the longer index visits do not shorten subsequent visits. Instead, on days with difficult cases, physicians leave the clinic a few minutes later than usual.[8]

[6]To study the heterogeneity of the effect, we estimate a triple-difference specification, fully interacting equation (1) with different observed characteristics (one at a time). Table 3 shows the results. We detect no statistically significant difference in the magnitude of the main effect along any of the physician (Panel A) dimensions (age, gender, and previous exposure to difficult cases).

[7]The one visit *before* a difficult cases is also estimated to be 20% longer. This is most likely driven by measurement error related to the mechanics of time stamping: our measure of the duration of each visit is the time between when the physician swipes the patient's insurance member card to start recording the notes for the visit and the time the card of the next patient in line is swiped. If physicians start an encounter with a newly diagnosed cancer patient by talking to the patient for longer than usual before taking notes, the additional visit time would be (mis)attributed to the previous visit. Consistent with this being the case, there is no difference in the duration of earlier visits.

[8]Table A8 shows that there is a statistically significant difference of 0.064 hours (3.84 minutes) in the end-of-day time between difficult and comparison cases. As expected, there is no difference in the start-of-day time.

Other Explanations. A difficult case may affect physicians' decisions by temporarily drawing their attention to cancer or other adverse outcomes, thus making them more likely to order tests in subsequent visits to avert such outcomes.[9] This explanation is in line with literature that documents salience effects in decision making and establishes connections between choices, attention, and memory (for a review, see Bordalo, Gennaioli and Shleifer, 2021).

To evaluate the direct contribution of elevated cancer screenings to the increase in testing, we estimate the impact of difficult cases on screening tests for any of the most common cancer types—breast, prostate, and colon—as the outcome variable. We find a marginally significant increase in cancer screening tests (Panel B of Table 2). But because their baseline rate is low (only 2% of primary care visits involve a referral to cancer screening), the estimated increase in cancer screenings accounts for at most 15% (0.18/1.23) of the total increase in testing following difficult cases.

It is still possible that following difficult cases, physicians increase testing because they pay more attention to rare or adverse outcomes more generally or, are temporarily more averse to making diagnostic errors. Such explanations are consistent with recent work that suggests that physicians make heuristic decisions, which may rely on recent cases as cues (Singh, 2021; Jin et al., 2021; Shurtz et al., 2021; Ly, 2021; Wang et al., 2022). Future work can further explore such mechanisms by exploiting natural or experimental variation in the order of visit schedules.

It could also be that difficult cases trigger emotional responses that affect subsequent judgments. Discussing difficult medical news with patients has been shown to trigger various negative emotional responses among physicians (such as anxiety, sadness, or pessimism), even experienced physicians (for a survey, see Fallowfield and Jenkins, 2004). And a large literature exists that suggests that emotions influence judgment (for reviews, see Lerner et al., 2014; Meier, 2021).

[9]For example, Shurtz et al. (2021) show that PCPs persistently increase their referrals to colonoscopy in the months following an encounter with a colon cancer patient. For evidence outside the healthcare context, see Malmendier and Nagel (2011), Malmendier and Nagel (2015), and Cameron and Shah (2015).

While we could not directly test whether difficult cases invoke any change in preferences or an emotional response, we test a related hypothesis, that physicians' response to more terminal cases, which intuitively induce a more intensive affective response, is stronger. Panel B of Table 3 shows the results. In line with this explanation, the estimated increase in testing rates is somewhat (though insignificantly) stronger after encounters with patients who are more likely to die from their condition.

To better evaluate the possibility that emotions influence physician decisions, future work could document or manipulate the emotional state of physicians as they practice medicine. This can be done by combining surveys or experimental designs with administrative claims data. Researchers can survey doctors' emotions at different points in time using the oft-used Positive and Negative Affect Schedule (PANAS; Watson et al., 1988) or using ecological momentary assessments (i.e., repeated sampling of subjects' current behaviors and experiences in real time, in subjects' natural environments, such as through a mobile app; see Shiffman et al., 2008). Alternatively, one can randomly trigger emotional responses, for example by priming physicians to negative emotions through exposure to sad or stress-inducing case descriptions or experiments and document subsequent practice patterns (such as Li et al., 2017, who use experimental designs to elicit physicians' social preferences).

6 Conclusion

We examine whether PCPs alter their clinical decision making following difficult cases—encounters with patients who were recently diagnosed with cancer. We find that such cases are followed by an immediate, sharp, and statistically significant increase in doctors' orders for tests. The effect is temporary: on average, it persists for about eight visits (roughly an hour) after the difficult case. The effect is not limited to novice physicians. It is concentrated in patients whom other physicians would also be inclined to test; namely, it conforms with the prevailing professional testing practice.

The evidence is hard to reconcile with physician learning, schedule disruption, or a

cancer-specific increase in tests. Other explanations, more behavioral in nature, such as increased attention to rare events or (possibly emotion-triggered) heightened aversion toward missing a diagnosis, seem more plausible, though we cannot separately identify them in our setting.

However, regardless of the exact mechanism, our results suggest that within-individual variability in professional judgment arises due to transient responses to prior challenging encounters. The presence of such internal variability in practice further underscores the potential scope for decision support tools (such as algorithmic advice, alerts to established guidelines, or information on decisions made by the majority of other experts in similar cases) in improving the consistency in medical practice or in other high-stakes judgments. Promising directions for future work include further documenting the potential sources and nature of physician practice variability and understanding the scope for mitigating it using various interventions.

References

Abaluck, Jason, Leila Agha, Chris Kabrhel, Ali Raja, and Arjun Venkatesh, "The determinants of productivity in medical testing: Intensity and allocation of care," *American Economic Review*, 2016, *106* (12), 3,730–3,764.

Amiel, Gilad E, Lea Ungar, Mordechai Alperin, Zvi Baharier, Robert Cohen, and Shmuel Reis, "Ability of primary care physician's to break bad news: A performance based assessment of an educational intervention," *Patient Education and Counseling*, 2006, *60* (1), 10–15.

Banky, Megan, Ross A Clark, Yong-Hao Pua, Benjamin F Mentiplay, John H Olver, and Gavin Williams, "Inter-and intra-rater variability of testing velocity when assessing lower limb spasticity," *Journal of Rehabilitation Medicine*, 2019, *51* (1), 54–60.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer, "Salience," 2021. NBER Working Paper No. 29274.

Cameron, Lisa and Manisha Shah, "Risk-taking behavior in the wake of natural disasters," *Journal of Human Resources*, 2015, *50* (2), 484–515.

Chan, David C., Matthew Gentzkow, and Chuan Yu, "Selection with variation in diagnostic skill: Evidence from radiologists," *Quarterly Journal of Economics*, 2022, *137* (2), 729–783.

Chen, Daniel L. and Markus Loecher, "Mood and the malleability of moral reasoning," Technical Report 2020.

Currie, Janet M and W Bentley MacLeod, "Diagnosing expertise: Human capital, decision making, and performance among physicians," *Journal of Labor Economics*, 2017, *35* (1), 1–43.

_ and _ , "Understanding physician decision making: The case of depression," 2018. NBER Working Paper No. 24955.

_ , _ , **and Jessica Van Parys**, "Provider practice style and patient health outcomes: The case of heart attacks," *Journal of Health Economics*, 2016, *47*, 64–80.

Dai, Hengchen, Katherine L Milkman, David A Hofmann, and Bradley R Staats, "The impact of time at work and time off from work on rule compliance: The case of hand hygiene in health care," *Journal of Applied Psychology*, 2015, *100* (3), 846.

Detre, Katherine M, Elizabeth Wright, ML Murphy, and T Takaro, "Observer agreement in evaluating coronary angiograms," *Circulation*, 1975, *52* (6), 979–986.

Eren, Ozkan and Naci Mocan, "Emotional judges and unlucky juveniles," *American Economic Journal: Applied Economics*, 2018, *10* (3), 171–205.

Fallowfield, Lesley and Valerie Jenkins, "Communicating sad, bad, and difficult news in medicine," *The Lancet*, 2004, *363*, 312–319.

Freedman, Seth, Ezra Golberstein, Tsan-Yao Huang, David Satin, and Laura Barrie Smith, "Docs with their eyes on the clock? The effect of time pressures on primary care productivity," *Journal of Health Economics*, 2021, *77* (102442).

Geerling, Wayne, Gary Magee, Pul A. Raschky, and Russell Smyth, "Bad news from the front and from above: Bombing raids, military fatalities and the death penalty in Nazi Germany," *Economic Inquiry*, 2020, *58* (3), 1450–1468.

Heyes, Anthony and Soodeh Saberian, "Temperature and decisions: Evidence from 207,000 court cases," *American Economic Journal: Applied Economics*, 2019, *11* (2), 238–265.

Hsiang, Esther Y, Shivan J Mehta, Dylan S Small, Charles AL Rareshide, Christopher K Snider, Susan C Day, and Mitesh S Patel, "Association of

primary care clinic appointment time with clinician ordering and patient completion of breast and colorectal cancer screening," *JAMA Network Open*, 2019, *2* (5), e193403–e193403.

Jin, Lawrence, Rui Tang Han Ye, Junjian Yi, and Songfa Zhong, "Time dependency in physician decision-making," *AEA Papers and Proceedings*, 2020, *110*, 284–288.

_ , _ , _ , and _ , "Path dependency in physician decisions," Technical Report, Mimeo 2021.

Jr, Elliott B Martin, Natalia M Mazzola, Jessica Brandano, Donna Luff, David Zurakowski, and Elaine C Meyer, "Clinicians' recognition and management of emotions during difficult healthcare conversations," *Patient education and counseling*, 2015, *98* (10), 1248–1254.

Kahn, Matthew E. and Pei Li, "The effect of pollution and heat on high skill public sector worker productivity in China," 2019. NBER Working Paper No. 25594.

Kahneman, Daniel, Olivier Sibony, and Cass R Sunstein, *Noise: A flaw in human judgment*, Little, Brown Spark, 2021.

Kraemer, Helena Chmura, "The reliability of clinical diagnoses: State of the art," *Annual Review of Clinical Psychology*, 2014, *10*, 111–130.

_ , David J Kupfer, Diana E Clarke, William E Narrow, and Darrel A Regier, "DSM-5: How reliable is reliable enough?," *American Journal of Psychiatry*, 2012, *169* (1), 13–15.

Lerner, Jennifer S., Ye Li, Piercarlo Valdesolo, and Karim Kassam, "Emotion and decision making," *Annual Review of Psychology*, 2014, *66*, 799–823.

Li, Jing, William H Dow, and Shachar Kariv, "Social preferences of future physicians," *Proceedings of the National Academy of Sciences*, 2017, *114* (48), E10291–E10300.

Ly, Dan P., "The influence of the availability heuristic on physicians in the emergency department," *Annals of Emergency Medicine*, 2021, *5* (78), 650–657.

Malmendier, Ulrike and Stefan Nagel, "Depression babies: Do macroeconomic experiences affect risk taking?," *Quarterly Journal of Economics*, 2011, *126* (1), 373–416.

_ and _, "Learning from inflation experiences," *Quarterly Journal of Economics*, 2015, *131* (1), 53–87.

Meier, Armando N, "Emotions and risk attitudes," Technical Report, SOEPpapers on Multidisciplinary Panel Data Research 2021.

Molitor, David, "The evolution of physician practice styles: evidence from cardiologist migration," *American Economic Journal: Economic Policy*, 2018, *10* (1), 326–56.

Neprash, Hannah T., "Better late than never? Physician response to schedule disruptions," Technical Report, Mimeo 2016.

Neprash, Hannah T and Michael L Barnett, "Association of primary care clinic appointment time with opioid prescribing," *JAMA Network Open*, 2019, *2* (8), e1910373–e1910373.

Parys, Jessica Van and Jonathan Skinner, "Physician practice style variation-implications for policy," *JAMA Internal Medicine*, 2016, *176* (10), 1549–1550.

Ptacek, J.T., John J. Ptacek, and Neil M. Ellison, ""I'm sorry to tell you ..." Rhysicians' reports of breaking bad news," *Journal of Behavioral Medicine*, 2001, *24* (2), 205–217.

Robinson, Philip J, Daniel Wilson, A Coral, Ad Murphy, and P Verow, "Variation between experienced observers in the interpretation of accident and emergency radiographs," *The British Journal of Radiology*, 1999, *72* (856), 323–330.

Shiffman, Saul, Arthur A Stone, and Michael R Hufford, "Ecological momentary assessment," *Annual Review of Clinical Psychology*, 2008, *4*, 1–32.

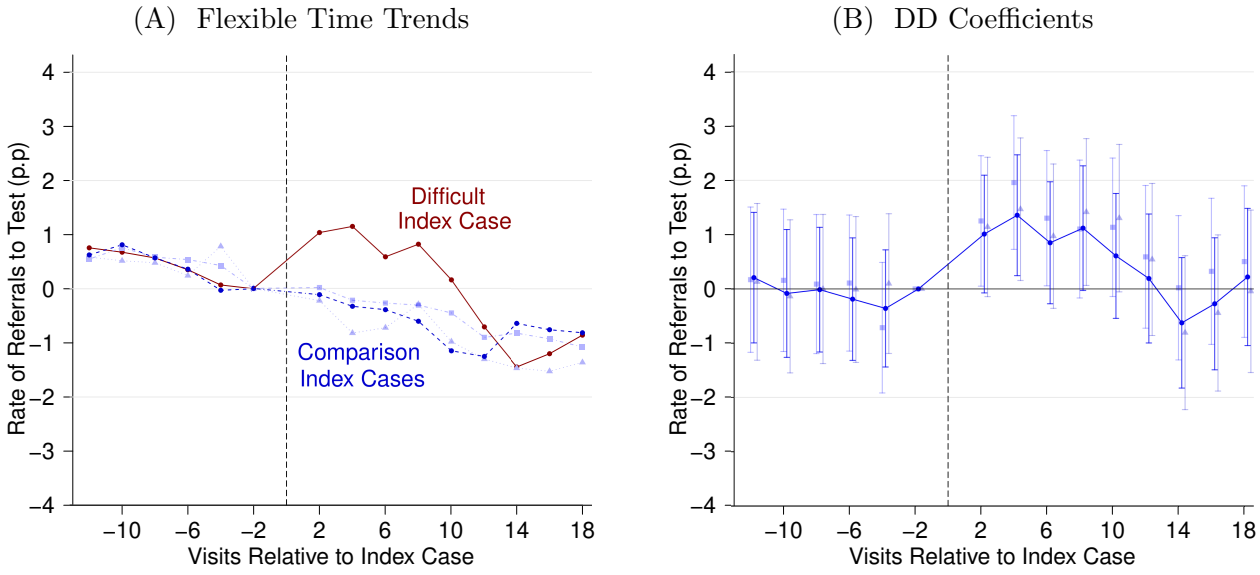
Shurtz, Ity, Yoav Goldstein, and Gabriel Chodick, "Realization of low probability clinical risks and physician behavior: Evidence from primary care physicians," Technical Report, Mimeo 2021.

Singh, Manasvini, "Heuristics in the delivery room," *Science*, 2021, *374* (6565), 324–329.

Wang, Annabel Z., Michael L. Barnett, and Jessica L. Cohen, "Changes in cancer screening rates following a new cancer diagnosis in a primary care patient panel," *JAMA Network Open*, 2022, *5* (7), e2222131–e2222131.

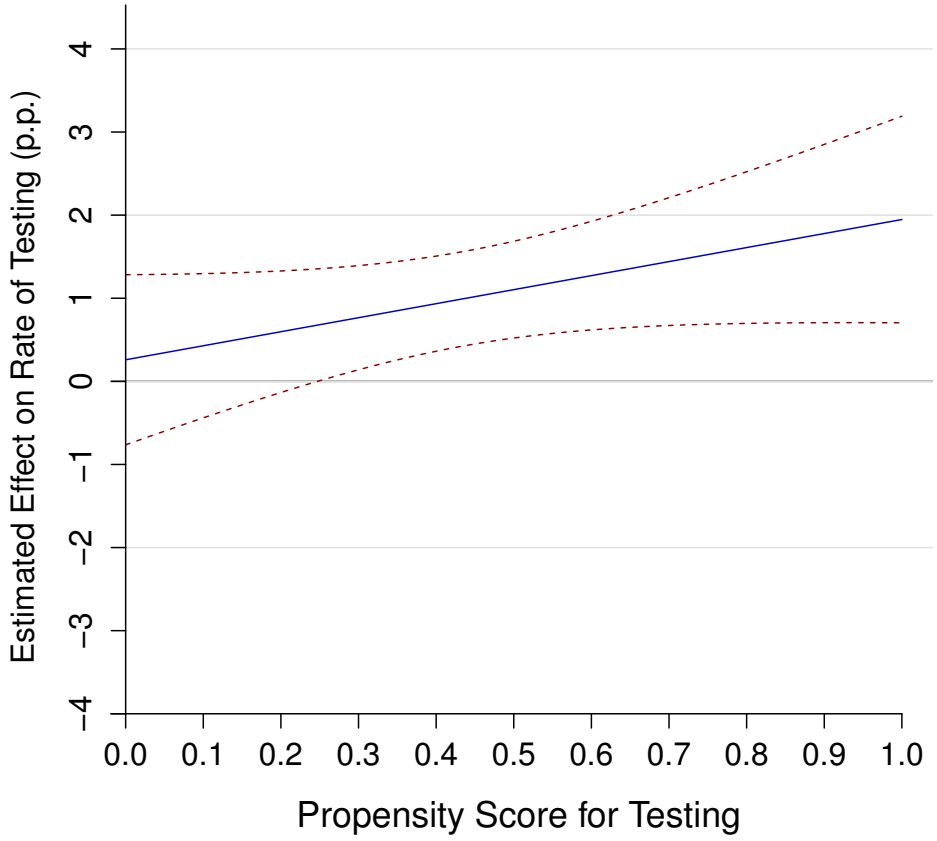
Watson, David, Lee Anna Clark, and Auke Tellegen, "Development and validation of brief measures of positive and negative affect: The PANAS scales," *Journal of Personality and Social Psychology*, 1988, *54* (6), 1063.

Figure 1: The Impact of Difficult Cases on Testing



Notes: The figures show estimates from equation (2) for the change in testing in visits occurring after difficult cases compared to visits occurring after comparison cases by the same physician. The x-axes show the visit number relative to the index case; visits are binned in pairs. The y-axes show estimates for the rate of physician referrals to common tests (lab, imaging, and other) at each visit. Panel (a) shows estimated time trends in the average testing rate around difficult and matched comparison cases. Blue circles, faded squares, and faded triangles represent our main and alternative definitions of comparison cases, respectively. Rates shown are relative to the baseline rate during the two visits immediately preceding the index case (visits number -1 and -2, jointly labeled "-2" for short). The regression includes match fixed effects. The index case (number 0) is excluded from the sample. Panel (b) shows estimates using the same specification for the trend in testing around difficult cases relative to comparison cases of each type. Standard errors are clustered at the match level. Error bars show 90% confidence intervals. The main sample consists of 1,660,257 visits, of which 124,227 are associated with difficult cases and the rest with comparison cases. Alternative samples consist of 504,764 and 363,312 visits.

Figure 2: The Estimated Impact of Difficult Cases on Testing, by Testing Propensity Score



Notes: This figure plots the estimated heterogeneity in the impact of difficult cases on the rate of testing, as a (linear) function of the following cases' testing propensity score. Estimates are calculated using the triple-differences regression equation (4). The patient's propensity score for testing is the probability that a physician will refer the patient to tests, which is predicted based on case characteristics. This score is calculated using a gradient-boosting model in a preliminary step. Section 4.2 discusses the empirical specification in detail, and Figure A5 shows the fit of this model. The blue solid line shows the estimated effect. The red dashed lines show the 90% confidence interval. The y-axis values represent percentage points. Standard errors are clustered at the match level. The sample consists of 971,943 visits, of which 73,821 are associated with difficult cases and the rest with comparison cases.

Table 1: Balance of Pre-Treatment Characteristics Between Treatment and Comparison Visits

	Treatment (1)	*Comparison* (2)	*Difference* (3)	*p-value* (4)
A. Patient				
Age	50.97	51.00	−0.03	0.73
Share male	41.87	41.56	0.31	0.16
Socio-economic	6.64	6.64	<0.01	0.60
Share TIA	1.62	1.56	0.09	0.27
Share Diabetic	15.13	15.28	−0.15	0.37
Share CVD	4.13	3.91	0.19	0.02
Share Obesity	21.42	21.49	−0.07	0.71
Share Cancer	12.24	12.25	−0.01	0.95
B. Visit				
Visit Duration	7.85	7.81	0.04	0.20
Test Referral	27.36	27.51	−0.14	0.51
Cancer Screening	1.93	1.93	<0.01	0.96
ER Referral	1.17	1.20	−0.03	0.53
Specialist Referral	11.28	11.28	<0.01	0.98
C. Physician				
Age	55.67	55.67		
Share Male	52.36	52.36		
Experience (Years)	18.18	18.18		
Number of Physicians	707	707		
Number of Index Cases	5,147	61,261		
Number of (Pre-Index) Visits	51,004	598,289		

Notes: The table compares average characteristics and outcomes between the 12 visits that preceded the index case in the treatment and comparison groups. During this pre-period, we expect no within-match differences in outcomes between these groups. Columns 1 and 2 show means residualized (by including match fixed effects) using equation (3). Columns 3 and 4 show the difference and the p-value for the difference being statistically significant. Standard errors are clustered at the match level. Each row shows data from a separate regression. Panel A shows patient characteristics. CVD stands for cardiovascular disease; TIA stands for transient ischemic attack. Panel B shows visit outcomes. Panel C shows mean physician characteristics, which are identical between the treatment and comparison by construction.

Table 2: The Impact of Difficult Cases on Outcomes of Subsequent Visits

	Baseline Mean	Estimated Effect
	(1)	(2)
A. Main Visit Outcomes		
Test Referral	27.43	1.23***
		(0.36)
Drug Prescription	46.55	-0.23
		(0.37)
Specialist Referral	10.04	0.37
		(0.24)
ER Referral	1.13	-0.09
		(0.08)
B. Additional Outcomes		
Cancer Screening Test Referral	1.94	0.18*
		(0.11)
Visit Duration (minutes)	7.31	0.04
		(0.06)
Number of observations (visits)	971,943	
Number of clusters (matches)	5,368	

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

Notes: The table shows estimates for the impact of difficult cases on outcomes of subsequent visits. Each row shows difference-in-difference estimates of this impact (δ from equation (1)) for a different outcome. The sample includes a balanced panel of eight visits before and eight visits after the index case, covering a window of about two hours. For all outcomes except for duration, numbers in column 1 represent percentages and numbers in column 2 represent percentage points (pp). For duration, numbers represent minutes. When estimating the effect on visit duration, we exclude one visit before the index event to mitigate measurement error related to time stamping (see Appendix A for details); the number of observation in this case is 875,849. Standard errors are clustered at the match level.

Table 3: Triple-Difference Estimates of Heterogeneity in the Impact of Difficult Cases on Testing

	No	Yes	Difference
	(1)	(2)	(3)
A. Physician			
Age $\geq$ 57	1.20**	1.27**	0.07
	(0.51)	(0.50)	(0.72)
Male	1.21**	1.25**	0.05
	(0.52)	(0.49)	(0.72)
Exposure $\geq$ 5	1.05	1.29***	0.24
	(0.91)	(0.39)	(0.99)
B. Case			
Age $\geq$ 63	0.85*	1.67***	0.81
	(0.48)	(0.54)	(0.72)
High-Risk Cancer	0.93*	1.59***	0.66
	(0.49)	(0.53)	(0.72)
Died Within 4 Years	1.01**	2.03**	1.02
	(0.40)	(0.81)	(0.90)
Number of observations (visits)	971,943		
Number of clusters (matches)	5,368		

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

Notes: The table shows the estimated heterogeneity in the impact of difficult cases on the rate of testing, as a function of physician (Panel A) and case (Panel B) characteristics. Estimates were calculated using the triple-differences regression equation (4), using a balanced panel of eight visits before and eight visits after the index case. Columns 1 and 2 show the estimated effect of Difficult Cases on testing for different values of the characteristic. Column 3 shows the difference between these effects (δ_0 in equation (4)) and the corresponding standard error. The comparison group is our main comparison group. Standard errors are clustered at the match level. Numbers represent percentage points (pp).

Challenging Encounters and Within-Physician Practice Variability

On-Line Appendices

Gabriel Chodick, Yoav Goldstein, Ity Shurtz, Dan Zeltzer

Appendix A Sample Construction and Variables Definitions

This appendix provides more details on the construction of the main and alternative samples of difficult and comparison cases, and on the definition of the outcome variables used in the analysis.

Sample Construction

Difficult Cases

We construct the sample of difficult cases in steps. First, we sample all recorded diagnoses for any type of cancer in Maccabi medical records, with no prior diagnosis of the same cancer type for the patient within six months. We identify cancer diagnosis and cancer type using the ICD9 classification and restrict our attention to cancer types with at least 1,000 patients diagnosed with it in our data. The restriction to *new* diagnoses decreases the number of diagnoses that we include from 417,637 to 92,316. This large number of diagnoses reflects the fact that physicians often mention old cancer diagnoses on medical records.

Second, we match Maccabi's new cancer diagnoses with the National Cancer Registry, to which reporting of all new cancer diagnoses is mandatory. We keep only new cancer diagnoses with a registry within a window of 30 days before or after the date of the diagnosis. This restriction allows us to focus attention on new diagnoses rather than diagnosis codes associated with preexisting cancer. This step results in 8,054 diagnoses that are verified

against the registry as new.

Third, we identify difficult cases as the first primary care visit after the date of the new cancer diagnosis. We restrict attention to visits that occurred within 30 days of the date of the new cancer diagnosis, during which it is most likely that such a new cancer will be discussed with a PCP. This resulting sample includes 6,254 diagnoses.

Finally, to observe a baseline period before the index case during which different baseline patient characteristics are measured, we include in our final sample only cases that occurred beginning in July 2012, which is six months after the start of our study period (January 2012). This final sample includes 5,368 index difficult cases.

Alternative Comparison Cases

The main comparison group of cases is defined in Section 2. Here, we discuss two alternative definitions of the comparison group of cases, which we use in robustness analyses to verify that our main results are not sensitive to the specific way we define the comparison group.

Alternative I. We include in this comparison group all cases that occurred in the two weeks before and after the difficult case. This method allows us to assign up to four comparison cases for every difficult case. Note that while the method we use in our main analysis allows us to control for the time variation in physicians' work during the year, the complementary method we use here allows us to control for the unobserved characteristics in physicians' work that are constant in a short period of time. The resulting sample includes 15,994 comparison cases.

Alternative II. We include in this comparison group all cases that occurred in other years, in the same week of the year, weekday, and serial number as the index difficult case. This method is very similar to the main method we use, but it is more restrictive in that it requires that the treatment and comparison case occur in the *same* week of the year. This method allows us to assign up to three comparison cases for every difficult case. The resulting sample

includes 9,953 comparison cases.

Variable Definitions

For each primary care visit, we observe the time stamp (reflecting the exact time of the beginning of the visit), diagnosis codes entered by the physician in the visit summary, as well as drugs, lab tests, and imaging and other tests that were prescribed by the physician to the patient on the date of the visit. For each patient, we observe all visits (to PCPs and specialists) and hospitalizations during the period of 2012–2015. We use this data to construct our outcome and control variables. We also observe demographic and general health details for each patient (based on that patient's file in Maccabi's records).

Outcome Variables. To determine the visit outcomes, we assign indicator variables for each of the following: (1) test referral (one of the five most common lab tests or the five most common imaging and other tests);[1] (2) drug prescription (any); (3) specialist referral (any); (4) ER referral; and (5) cancer screening referrals (PSA, mammography, and colonoscopy).[2] These indicator variables equal 1 if and only if the physician prescribes tests or drugs, or gives a specialist or ER referral to the patient on the date of the visit. The last outcome variable that we define is visit duration. Because we observe only one time stamp per visit, we define visit duration as the difference between two consecutive time stamps, which excludesg 3.6% of visits that are longer than 40 minutes to avoid mismeasuring breaks as visits.

Control Variables. To define control variables, we assign each visit the relevant characteristics of the patient. Three immediate variables are the patient's age, gender, and socio-economic status (on a 1–10 scale). We also define indicator variables for the following

[1]Each test is recognized by its unique identifier in Maccabi's medical records. Blood (urine) tests refer to any test that was mentioned explicitly to be performed by taking a blood (urine) sample.

[2]During the period of our research, Maccabi did not document colonoscopy referrals. However, we use referrals to the Gastro Institute, which is the place in which colonoscopies are performed for Maccabi's patients. The baseline rates of these tests are 0.89% for mammography, 0.48% for PSA, and 0.11% for colonoscopy.

chronic conditions: obesity, diabetes, cardiovascular disease, transient ischemic attack, and cancer. Last, we define three more control variables by calculating the number of hospitalizations, the number of visits to any physician, and the number of visits to a PCP that the patient had during the six months before the index visit.

Appendix B Predicting Testing Propensity

This section discusses the construction of the testing propensity score that we use as a baseline to study the congruence of physician testing decisions to the prevailing practice as discussed in Section 4.2.

For the construction of this score, we sample at random one million visits from the unrestricted study sample of 23 million visits to all Maccabi physicians from January 2012 through December 2015. We refer to this sample as the "prediction sample". Using this sample, we train a standard gradient-boosting algorithm (discussed in detail below) to predict the probability of a referral to common tests based on the observed visit characteristics. The outcome includes the following groups of tests: lab tests that are included in our main outcome, the next five most common lab tests, imaging and other tests that are included in our main outcome, the next five most common imaging and other tests, cancer-related lab tests, and cancer-related screening tests. As potential predictors, we use the following variables: age, sex, and the number of each of any physician visits, PCP visits, and hospitalizations, as well as the total number of days in a hospital during the months six, three, and one before the index visit. We also include 13 indicator variables for chronic conditions and 13 indicator variables for new chronic conditions in the six months before the visit.

Gradient-boosting algorithm. We use a gradient-boosting algorithm (Chen and Guestrin, 2016) with regularization to avoid overfitting. We tune the model hyper parameters using a grid search. That is, we define a grid with six values of η (learning rate): {0.01, 0.02, 0.04, 0.08, 0.16, 0.32}, six values of γ (minimal loss): $\{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$, five values

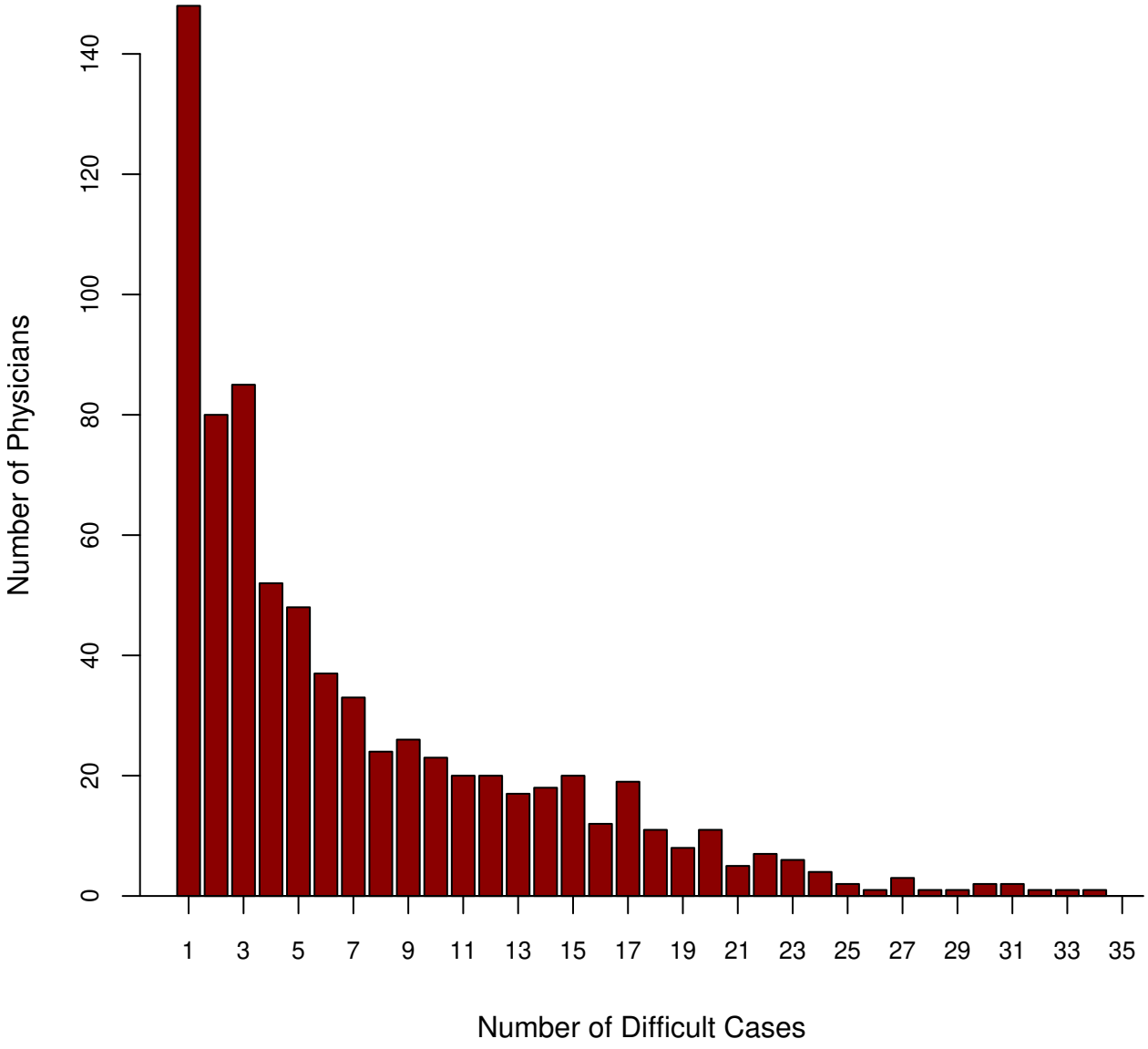
of the number of columns to select in each stage: $\{4, 15, 26, 37, 48\}$, and five values of the maximum depth of the tree: $\{4, 6, 8, 10, 12\}$. The number of trees we chose is 500.[3] Then, we randomly split the prediction sample into a training data set and a test data set, with 800,000 visits and 200,000 visits, respectively. We train a model with each set of parameters on the training set and calculate the cross-entropy using the test set. We choose the set of parameters that leads to the minimal cross-entropy.

Then, we use full prediction sample and the chosen combination of parameters to build the final gradient-boosting model and use the fitted values of this model on our research data set. We further calibrate the model by re-weighting the predicted values using linear regression of the actual testing rate on the predicted testing probability, using the prediction sample. These same weights are later used to re-weight the predictions made using the main study sample.

Finally, we classify predictions into 100 bins, each representing one percentile of the score distribution. We refer to the final predicted values of this model on our research data as the *testing propensity score*. Appendix Figure A5 shows the average testing rate as a function of the predicted rate, using the sample of pre-event visits in our main study sample. Considering that this is an out-of-sample calibration plot as it uses a different sample than the one the model is trained on, the model appears to be reasonably well calibrated.

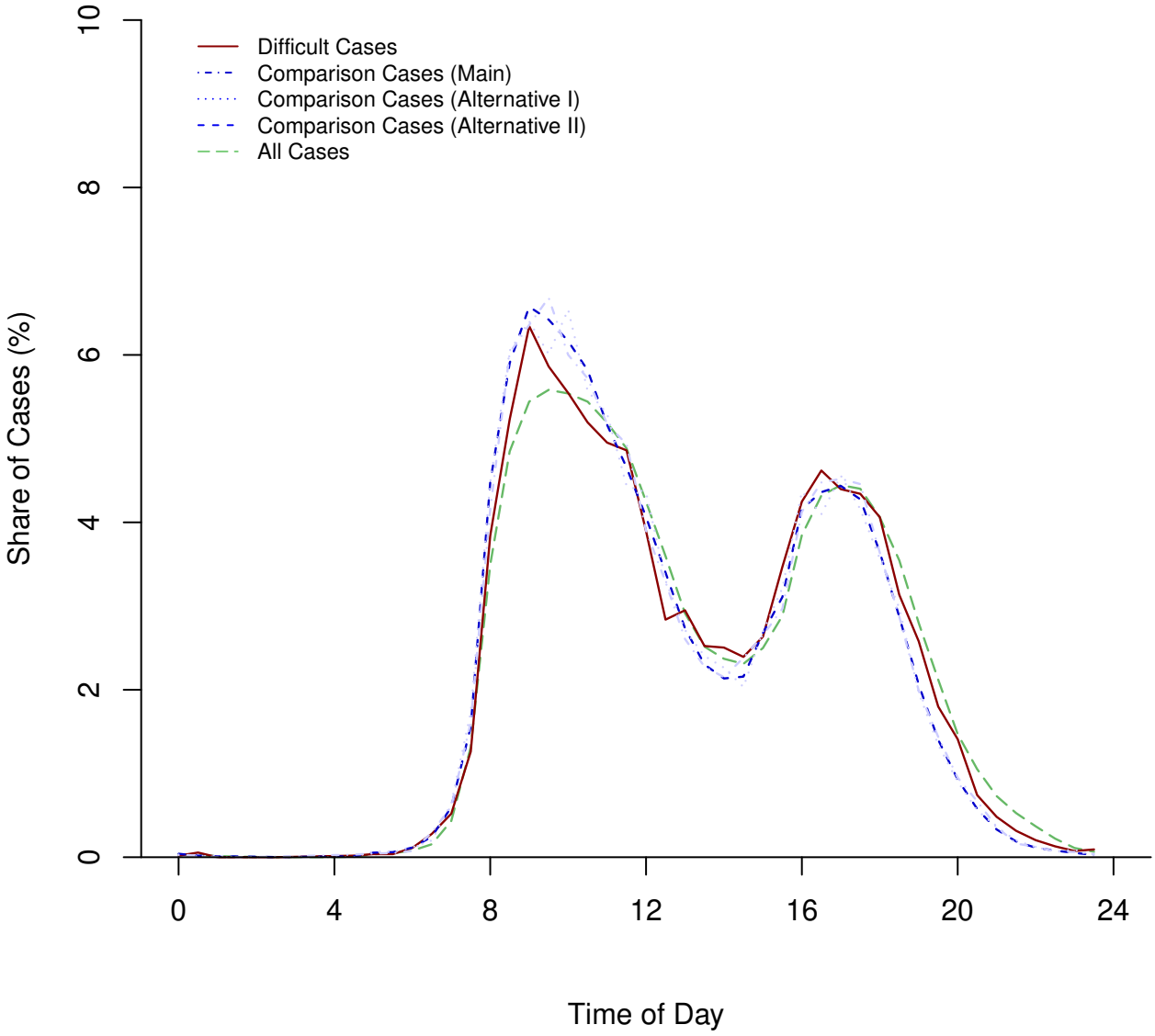
[3]Note that we have checked that the number of trees and the sample size are large enough. We calculated the cross entropy of five different models with 100, 300, 500, 700, 900 tress, and we saw that the loss is quite constant after 500. Similarly, we calculated the cross entropy on data set with different numbers of observations: 25K, 50K, 100K, 200K, 400K, 800K, and we saw that the loss is rather constant after 200K.

Appendix Figure A1: Exposure to Difficult Cases



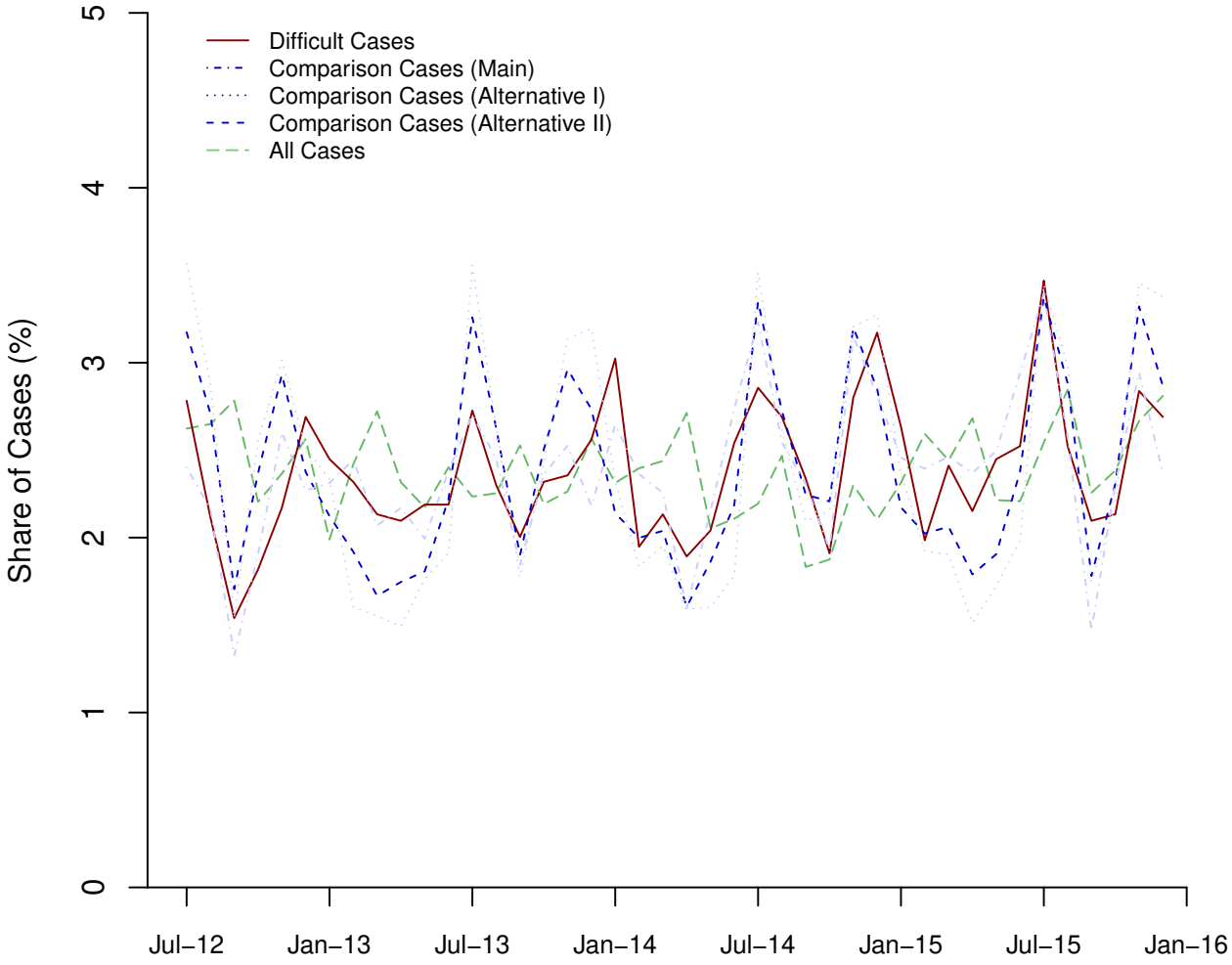
Notes: The figure shows a histogram of the number of distinct difficult cases each physician in our sample was exposed to during the study period of July 2012 through December 2015.

Appendix Figure A2: Distribution of Case Time of Day



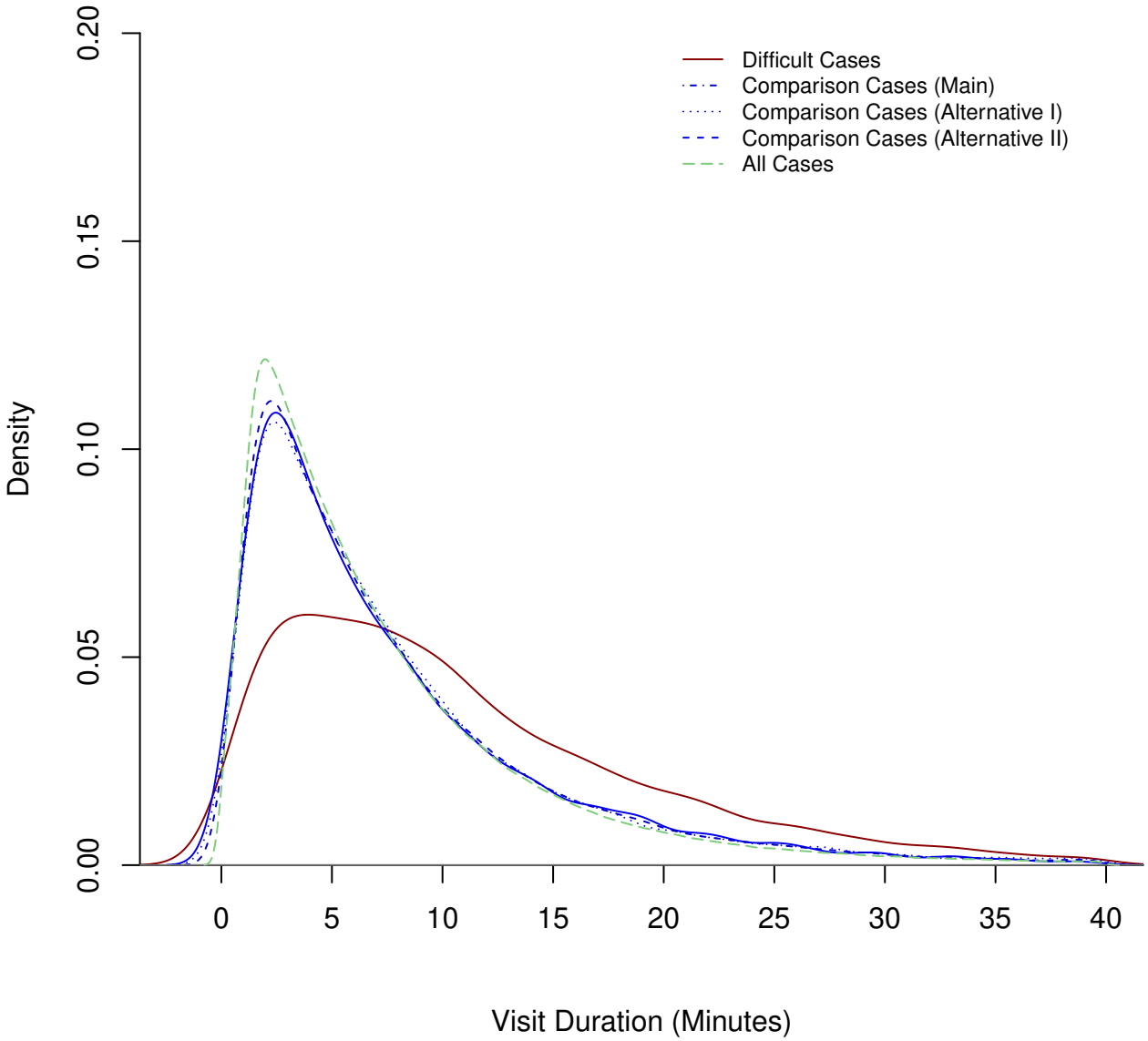
Notes: The figure shows the distribution of the time of day of index cases in the treatment (difficult cases) and alternative comparison groups. For reference, we also include the same distribution for the unrestricted study population (all cases). The x-axis shows the hour of day, using a 24-hour format (e.g., 8 is 8:00 a.m. and 20 is 8:00 p.m.). The sample includes 5,368 difficult cases and 64,042, 15,944, and 9,953 comparison cases (Main, Alternative I, Alternative II). The unrestricted study population includes all 20.2 million Maccabi PCP visits that occurred during the study period.

Appendix Figure A3: Distribution of Case Dates



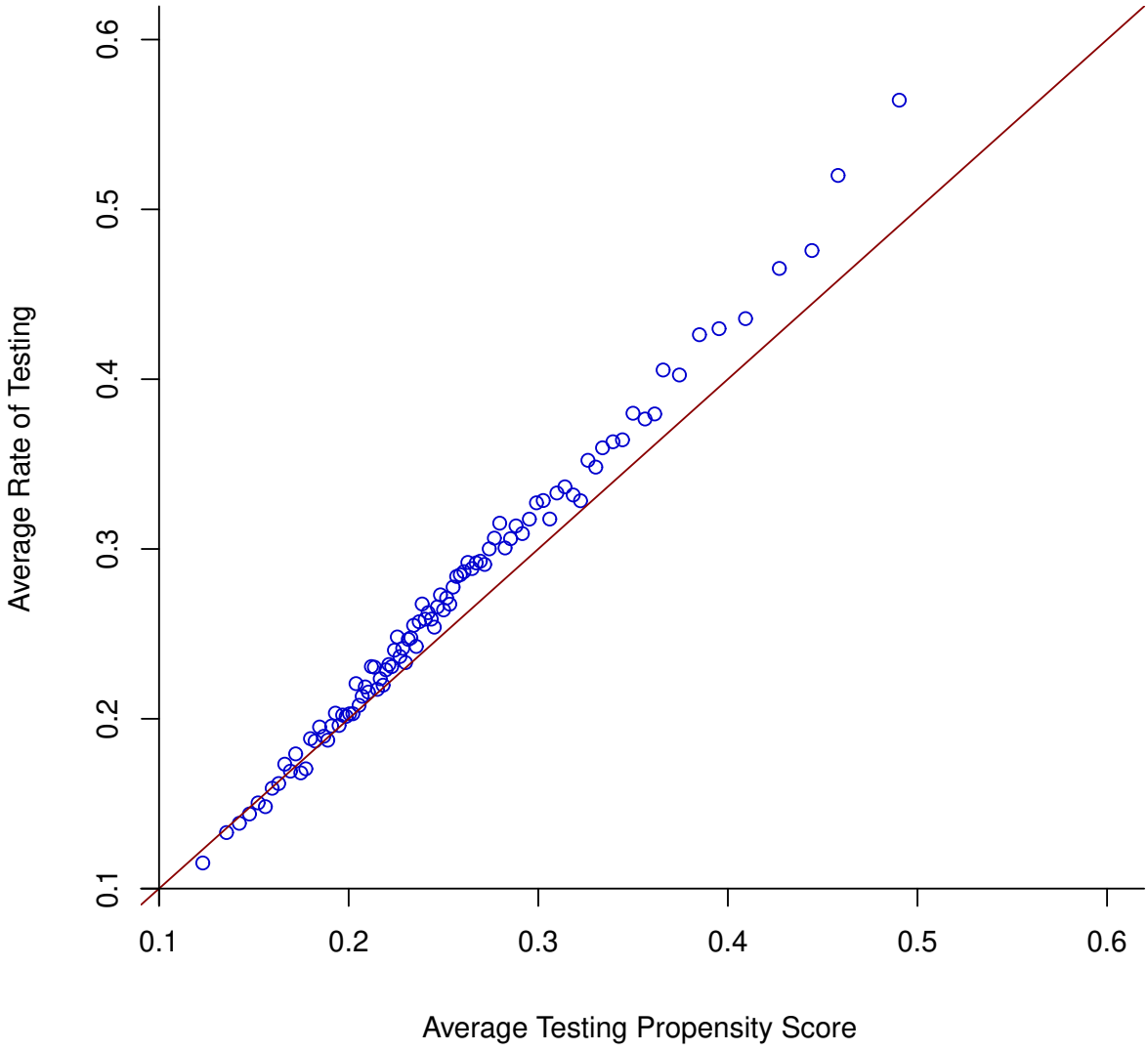
Notes: The figure shows the distribution of dates of index cases in the treatment (difficult cases) and alternative comparison groups. For reference, we also include the same distribution for the unrestricted study population (all cases). The x-axis shows the dates in our research period (July 2012 through December 2015). The sample includes 5,368 difficult cases and 64,042, 15,944, and 9,953 comparison cases (Main, Alternative I, Alternative II). The unrestricted study population includes all 20.2 million Maccabi PCP visits that occurred during the study period.

Appendix Figure A4: The Distribution of Lengths of Difficult and Comparison Cases



Notes: This figure shows kernel density estimates of the distribution of the duration of visits in the treatment (difficult cases) and alternative comparison groups. For reference, we also include the same distribution for the unrestricted study population (all cases). On average, index difficult cases, which involve newly diagnosed cancer patients, last three minutes longer than matched comparison cases. The sample includes 5,368 difficult cases and 64,042, 15,944, and 9,953 comparison cases (Main, Alternative I, Alternative II). The unrestricted study population includes all 20.2 million Maccabi PCP visits that occurred during the study period.

Appendix Figure A5: Calibration of the Testing Appropriateness Score



Notes: This figure shows a calibration plot for the gradient-boosting model used to predict the testing propensity score for the sample that includes only pre-period visits. The x-axis shows the average propensity score, which is a measure of the predicted testing probability, split into 100 bins defined by the percentiles of the predicted value distribution. The y-axis shows the actual test rate for each bin. Appendix B describes the data and procedures used. The sample consists of 971,943 visits, of which 73,821 are associated with difficult cases and the rest with comparison cases.

Appendix Table A1: Difficult Cases, by Cancer Type

Cancer Type	Cases	ICD9 Range	One-Year Mortality (%)
(1)	(2)	(3)	(4)
Pancreas	141	157–157.9	62.4
Liver	43	155–155.9	51.2
Stomach	107	151–151.9	40.2
Lung	404	162–162.9	39.4
Leukemia	24	204–204.9	16.7
Myeloma	98	203–203.9	13.3
Kidney	124	189–189.9	12.1
Bladder	118	188–188.9	10.2
Colon	563	153–153.9	9.1
Lymphoma	389	202–202.9	8.8
Rectum	146	154–154.9	8.2
Ovary & Uterus	289	179–183.9	6.9
Melanome	459	172–172.9	2.6
Prostate	433	185–185.9	2.5
Breast	1,645	174–175.9	1.5
Thyroid	408	193–193.9	0.5

Notes: The table shows the number of difficult cases in our full sample by the index patient's newly diagnosed cancer type. Column 2 shows the number of index cases associated with each cancer type. Column 3 shows the ICD9 code ranges used for defining each type. Column 4 shows the one-year mortality rate based on our data. Cancer types are ordered by mortality rate, in descending order. In the heterogeneity analysis, we consider cancer types with mortality rates greater than 5% as high-risk cancer. We suppressed cancer types with ten or fewer cases in our sample.

Appendix Table A2: Most Common Tests, by Their Setting

	Frequency	Included
Lab Tests		
Blood	19.95%	+
Alanine Transaminase (ALT GPT)	14.18%	+
Alkaline Phospatase	12.36%	+
Cholesterol	11.27%	+
Triglycerides	11.16%	+
Thyroid-Stimulating Hormone (TSH)	10.16%	
Urine	10.01%	
Aspartate Aminotransferase (AST GOT)	9.55%	
Vitamin B12	6.44%	
Ferritin	5.72%	
Imaging and Other Tests		
X-Ray	3.71%	+
Ultrasound	3.25%	+
Other Outpatient Diagnostics	2.58%	+
Other Imaging	1.24%	+
Cardiac	0.94%	+
Mammography	0.83%	
Bone Density	0.55%	
Electrocardiography (EKG)	0.53%	
Echo Doppler	0.21%	
Echo Heart	0.14%	

Notes: The table shows the ten most common laboratory and imaging tests in our data and their frequencies and the share of the visits in which there is a referral for the test. The five most common tests of each kind are included in our main outcome (testing).

Appendix Table A3: Balance of Pre-Treatment Characteristics Between Treatment and Comparison Visits, Alternatives I and II

	Treatment (1)	*Comparison* (2)	*Alternative I* (3)	*Alternative II* (4)
A. Patient				
Age	50.97	51.00	50.95	51.13
Share male	41.87	41.56	41.50	41.61
Socio-economic	6.64	6.64	6.64	6.64
Share TIA	1.62	1.56	1.52	1.52
Share Diabetic	15.13	15.28	15.15	15.41
Share CVD	4.13	3.91	3.88	3.94
Share Obesity	21.42	21.49	21.62	21.61
Share Cancer	12.24	12.25	12.08	12.26
B. Visit				
Visit Duration	7.85	7.81	7.82	7.76
Test Referral	27.36	27.51	27.88	27.45
Cancer Screening	1.93	1.93	1.92	1.95
ER Referral	1.17	1.20	1.20	1.21
Specialist Referral	11.28	11.28	11.38	11.08
Number of Physicians	707	707	699	631
Number of Index Cases	5,147	61,261	9,529	15,236
Number of (Pre-Index) Visits	51,004	598,289	148,744	93,229

Notes: The table compares average characteristics and outcomes between the 12 visits that preceded the index case in the treatment and the alternative comparison groups. Columns 1, 2, 3, and 4 show means, residualized (by including match fixed effects) using equation 3 for the treatment group (1), main comparison group (2) and alternative comparison groups (3 and 4). Panel A shows patient characteristics. CVD stands for cardiovascular disease; TIA stands for transient ischemic attack. Panel B shows visit outcomes. This table extends Table 1. We omitted Panel C because it is identical in all comparison groups by construction.

Appendix Table A4: Placebo Analysis of the "Effect" of Difficult Cases on Pre-determined Patient and Case Characteristics

	Baseline Mean	Estimated Effect
	(1)	(2)
Patient Age (Years)	51.24	0.16
		(0.15)
Previous Cancer Diagnosis (Percent)	12.31	0.05
		(0.25)
Male Patient (Percent)	41.44	0.09
		(0.37)
Any Physician Visit in 6 Months (Percent)	8.09	-0.03
		(0.11)
Any Hospital Admission in 6 Months (Percent)	0.80	-0.06
		(0.13)
Number of observations (visits)	971,943	
Number of clusters (matches)	5,368	

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

Notes: The table shows placebo analyses. In each row, we reproduce our main difference-in-differences estimates using equation 1 for a different pre-determined outcome. Since outcomes are pre-determined, we do not expect them to be affected by the treatment. Indeed, none of them is. Column 1 shows the baseline mean for each outcome. Column 2 shows point estimates of δ and standard errors (in parentheses). Standard errors are clustered at the match level. The sample includes a balanced panel of eight visits before and eight visits after the event.

Appendix Table A5: The Effect of Difficult Cases on Outcomes, DD by Comparison Group

		Specification	
	Baseline Mean	Alternative I	Alternative II
Outcome	(1)	(2)	(3)
A. Main Visit Outcomes			
Test Referral	27.43	1.52***	1.27***
		(0.40)	(0.42)
Drug Prescription	46.54	-0.58	-0.38
		(0.40)	(0.45)
Specialist Referral	10.04	0.42	0.11
		(0.27)	(0.28)
ER	1.13	-0.11	-0.05
		(0.09)	(0.10)
B. Additional Outcomes			
Cancer Screening Test Referral	1.94	0.07	0.16
		(0.12)	(0.13)
Visit Duration (minutes)	7.31	0.05	-0.02
		(0.06)	(0.07)
Number of observations (visits)		297,356	213,555
Number of clusters (matches)		5,368	5,368

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

Notes: The table shows the DD regression results (δ in equation 1) on outcomes for a balanced panel of eight visits before and eight visits after the event (about a two-hours window around the event), controlling for patient characteristics and for time and match fixed effects. Column 1 shows the baseline mean for each outcome. Column 2 shows the results with Alternative I as the comparison group. Column 3 shows the results with Alternative II as the comparison group. Standard errors are clustered at the match level.

Appendix Table A6: The Impact of Difficult Cases on Outcomes of Subsequent Visits, Including Patient Characteristics as Control Variables

	Baseline Mean	Estimated Effect
	(1)	(2)
A. Main Visit Outcomes		
Test Referral	27.43	1.16***
		(0.36)
Drug Prescription	46.55	-0.31
		(0.36)
Specialist Referral	10.04	0.35
		(0.24)
ER Referral	1.13	-0.09
		(0.08)
B. Additional Outcomes		
Cancer Screening Test	1.94	0.16
		(0.11)
Visit Duration (minutes)	7.31	0.03
		(0.06)
Number of observations (visits)	971,943	
Number of clusters (matches)	5,368	

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

Notes: The table shows the estimation results corresponding to Table 2, controlling for time fixed effects and patient characteristics (see Appendix A for details about the control variables that we include).

Appendix Table A7: The Effect of Difficult Cases on Alternative Outcomes, DD

Dependent Variable:	Top 5 Lab Tests	Top 5 Imaging and Other Tests	Top 7 Tests	Top 3 Tests	Number of Tests
	(1)	(2)	(3)	(4)	(5)
Treat × Post	0.68**	0.96***	1.17***	1.23***	0.0372***
	(0.33)	(0.25)	(0.37)	(0.35)	(0.0133)
Baseline Mean	20.67	10.58	29.94	26.29	0.8182
Number of observations (visits)	971,943				
Number of clusters (matches)	5,368				

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

Notes: The table shows estimates for the impact of difficult cases on alternative measures of testing decisions during subsequent visits. $Treat \times Post$ denotes the difference-in-difference estimates of this impact (δ from equation 1). Each column shows results using a different measure. The sample includes a balanced panel of eight visits before and eight visits after the index case, covering a window of about two hours. Lab Tests and Imaging and Other Tests are indicator variables for a referral for one of the five most common lab and imaging and other tests. Top 7 tests and Top 3 tests are indicator variables for a referral to one of the seven or three most common tests of each type (lab or imaging and other). Number of tests is a discrete variable counting the number of referrals to any of the top 5 tests. Standard errors are clustered at the match level.

Appendix Table A8: Comparing Physician Working Hours on Days with Difficult Cases and Comparison Cases

Dependent Variable:	Start	End
	(1)	(2)
Treat	-0.007	0.064***
	(0.020)	(0.025)
Baseline Mean	9.59	17.02
Number of observations (visits)	21,387	
Number of clusters (matches)	5,368	

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

Notes: The table shows the estimated difference in working hours between days with difficult cases and comparison cases. These are obtained by estimating the equation $Y_{imt} = \beta_1 Treat_{imt} + \psi_m^1 + \varepsilon_{imt}^1$, where $Treat_{imt}$ is defined as described in the text, and ψ_m^1 represents match fixed effects. The dependent variables are the time of the first visit the physician had in the specific day (column 1), and the time of the last visit (column 2), and they are measured in hours (1–24).