HEDG

HEALTH, ECONOMETRICS AND DATA GROUP

THE UNIVERSITY of York

WP 21/09

Do National Health Guidelines increase coordination level among physicians? An experimental investigation

Massimo Finocchiaro Castro and Domenica Romeo

May 2021

http://www.york.ac.uk/economics/postgrad/herc/hedg/wps/

Do National Health Guidelines increase coordination level among physicians? An experimental investigation

Massimo Finocchiaro Castro[a,b,c*] and Domenica Romeo[a]

[a]*Department of Law, Economics and Social Science, Mediterranean University of Reggio Calabria (Italy)*

[b]*Health Econometrics and Data Group, University of York (UK)*

[c]*Institute for Corruption Studies, Illinois State University (USA)*

Abstract

Coordination is the key to the success of any organization such as the healthcare sector, where higher level of coordination result in greater promptness and quality of care and lower mortality rates. In a framed field experiment, we assess the level of coordination among healthcare providers and monitor whether common practices are adopted in a metropolitan hospital in Italy, by using the Krupka-Weber norm elicitation task. Upon being provided with three clinical vignettes, physicians have been asked to evaluate the appropriateness of each of the possible actions to match the modal judgement. Afterwards, physicians may ask for information on the actions corresponding to national guidelines and eventually change decisions. Data show that the average frequency of coordination across the experiment is 52% and that coordination increases when physicians often exchange opinions and share positive feedbacks with colleagues. Only 23% of participants changed their appropriateness judgment, after realizing they were in contrast with guidelines. In addition, the presence of a leader in the ward facilitates coordination. Finally, the more physicians consult scientific sources, the more willing to accept guidelines suggestions have been. Since guidelines knowledge significantly increases the coordination between physicians, hospitals should implement effective programs to spread guidelines contents.

JEL Classification: C72; C93; D83; I12.

Keywords: coordination game; framed field experiment; vignette; guidelines; physicians' behavior.

[*] Corresponding author: Massimo Finocchiaro Castro, Department of Law, Economics and Social Science, Mediterranean University of Reggio Calabria, Italy, email: massimo.finocchiaro@unirc.it

1. Introduction

Coordination is the key to the success of the organization (Webb, 1991): It ensures unity of action and integrates different activities, resources and structures in an organization context (Beuselinck et al., 2007). Modern day organizations are becoming increasingly dynamic and complex: People working in an organization can have different backgrounds, work attitudes, knowledge and skills. For this reason, managers have to give a common direction to the efforts of all the individuals working in different departments to accomplish the objectives of the business (Lewis, 2006). In the healthcare sector, coordination becomes even more important than in any other organization. Specifically, relational coordination, defined as the "sharing of codes of conduct and procedure[1]" among health professionals (Gittell, 2000) has been shown to improve some dimensions of performance, including emergency and intensity care (Fargason and Haddock, 1992). Moreover, the average length of hospital stay is significantly shortened by the frequency of communication among care providers and, coordination in hospital emergency units improves promptness and quality of care and reduces mortality rates (Argote, 1982; Baggs et al., 1992; Shortell et al., 1994). On the contrary, departments with low levels of coordination result in the worst performance (Country Health Profile, 2017)[2]. In this regard, practice guidelines, which in fact proxy group 'norms', are used in the healthcare sector to achieve the hospitals' desired outcomes, improving patient's health and saving resources while providing care (Shekelle et al., 1999). Physicians agree that not only guidelines improve quality of care, but they could foster coordination (Carrier et al., 2012). Even though they are not actual norms, they can be classified as soft law instruments (Meoli et al., 2018) which assist and uniform practitioners' actions, by identifying recommended courses of actions under certain circumstances (Institute of Medicine, 1990). Practice guidelines, often representing international gold standard reference, result from rigorous clinical research and are largely supported by experts and professionals in the field (Field and Lohr, 1990). This is the reason why they epitomize group norms.

Hence, the purpose of this study is twofold. First, to measure the level of coordination among healthcare providers and, second, to monitor whether common practices are adopted in such a setting, developing a simple, portable and incentive-compatible tool, based on the Krupka-Weber (2008) norm elicitation task[3] (Burks and Krupka, 2012; Gaechter et al., 2013; Barr et al., 2018). For we have

[1] Coordination in a broader sense encompasses frequency and accuracy in communication and even problem-solving capacity (Gittel et al., 2000).

[2] According to the Country Health Profile (2017) cardiovascular diseases which are the leading cause of death among women in Luxembourg and second to cancer for men are attributable to the lack of an integrated and coordinated system of care for chronic patients.

[3] In order to detect whether a group of people share a common understanding of the practices and rules that apply to

designed a framed field experiment composed by two treatments: The Coordination treatment (CT) and the Information treatment (IT), namely. The CT is divided into two stages. In the first stage, physicians have been asked to evaluate the appropriateness of each of the proposed actions to heal a specific disease on a scale of one to four in order to match the modal rating reached in the session. In the second stage, physicians may ask to know which action, among those proposed, correspond to the national guideline with regard to that specific situation. The IT differs from the CT only in an additional feature located in the second stage of the treatment. In particular, if a physician asks to know guideline contents, then, she can change her appropriateness judgments of the actions given in the first stage. Doing so, we can assess the factors leading to changes and if coordination level increases because of such afterthoughts.

Thus, the novelty of our manuscript is twofold. To the best of our knowledge, this is the first work assessing coordination among physicians and the role of national guidelines as coordination tool in an experimental setting. Second, whereas other artefactual field and laboratory experiments have already involved physicians (see e.g. Brosig et al., 2016; Wang et al., 2020), to the best of our knowledge, this is the first framed field experiment conducted in a hospital (their real working environment) employing 52 physicians.

Our results show that the average frequency of coordination across the experiment is 52% and that coordination increases when physicians exchange opinions and share positive feedbacks with colleagues. In addition, the presence of a leader in the medical ward facilitates coordination. Moreover, results suggest that the longer a physician has been working for the same hospital, more likely she is to implement the outcomes prescribed by the guidelines. Looking at the Information treatment, only 8% of the physicians have declined the chance to know national guidelines content. However, just 23% of the participants, whose judgments differed from national guidelines in one or two scenarios, decided to change their appropriateness judgment. Finally, the more physicians consult scientific sources, the more likely to access guidelines information and, to align their decisions to national guidelines they are. Since guidelines dissemination is shown to increase the level of coordination between physicians, hospitals should consider effective programmes to spread their knowledge.

The rest of the paper is organized as follows. Section 2 reviews the related literature. In Section 3, we briefly describe a simple theoretical framework of physicians' behavior and draw some behavioral

specific decision situations, Krupka and Weber (2008), for the first time, adopted a special type of coordination game in which people are incentivized to "coordinate" with others in evaluating what constitutes "appropriate behaviour" in a given situation. A similar approach can be used to investigate whether, in the hospital, physicians coordinate when facing the same scenario.

hypotheses. Section 4 describes the basic setup and the experimental protocol. In Section 5, we discuss the results. Section 6 concludes the study.

2. Literature review

This study contributes to different strands of literature. First, since the main purpose is to test whether physicians agree on what is the most appropriate action to be taken in each situation, this study endorses the trend of the coordination games used in experimental economics to this end. In their seminal study, Krupka and Weber (2008) use a coordination game in which people are incentivized to "coordinate" with others in evaluating the appropriateness of different actions in a given situation, revealing their true beliefs. Results show that people care about social norms and want to behave consistently with them. Similar results are obtained in the experiment by Krupka et al. (2016) whose data show that subjects are significantly affected by social norms in their decisions. The same procedure is applied in a three-person gift exchange game by Gaechter et al. (2013) leading to results consistent with social norms compliance. Similarly, Barr et al. (2018) measure the social appropriateness of discrimination. Their experimental evidence shows that discrimination is perceived as socially inappropriate. Finally, Burks and Krupka (2012) attempt to identify social norms using the well-known elicitation technique to link them to behaviour in the business setting of a large financial service firm in the U.S.A. Their findings show a general tendency for corporate leaders to predict the ethical appropriateness valuations of their peers. To the best of our knowledge, we are the first to apply this framework to health context.

Second, this study integrates the widespread literature on the use of vignettes to learn how people think about a wide range of topics (see e.g Bursztyn. et al., 2018). According to Glassman et al. (2000) and Das and Hammer (2005), vignettes representing dummy patients with common diseases based on realistic clinical situations are a practical and feasible tool to measure the accuracy of physicians' diagnoses and ultimately assess the quality of care. For example, Das et al. (2008) combined medical vignettes with direct observations of the physician-patient relationship to measure providers' knowledge and find that practitioners' competence and effort are very low in low-income countries. This analysis was taken one step further by Das et al. (2015) who resorted to standardised patients (i.e., real or simulated sent anonymously to hospitals) and medical vignettes to assess providers' competence on tuberculosis in India. In the same context, Mohanan et al. (2015) test specific knowledge concerning childhood diarrhoea and pneumonia, providing evidence for Indian practitioners' poor knowledge of key diagnostic questions. Variations in physicians' diagnoses were instead tested by both Yager et al. (1986) and Gorter et al. (2001). Also, clinical vignettes have been

used to assess specific feature such physicians' confidence level, which is often found to be responsible for diagnostic and treatment errors (Berner and Graber, 2008), or their clinical experience (Sandvik, 1995)[4].

Finally, this paper contributes to experimental studies on the role of information on individual's behaviour. King (1974) was one of the pioneers of such literature, showing that providing plant managers with artificial reports about job rotation programs improves their productivity. Beneficial effects of information are also reported in Duflo and Saez (2003), who find that when proposals for retirement plans are complemented with education programmes the enrolment rate significantly increases. This contention is reinforced in the healthcare sector by Bauchner et al. (2001) who report that medical education in the form of discussions, case studies and local opinion leaders improves physicians' behaviour. More specifically, Braddock et al. (1999) show that providing physicians with patients' information about laboratory tests reduces medical errors and encourages preventive care. Supporting this view, Allery et al. (1997) report that any form of information such as that gleaned from medical journals, scientific conferences, discharge letters, contributes to change and improve physician's behaviour[5]. Hence, we contribute to the above-mentioned stream of research by showing the role of information on national guidelines, as a tool to increase coordination among physicians on the action to take when providing health care.

3. Theoretical model and behavioural hypotheses

In this section, we present a simple theoretical framework of the relationship between evidence-based guidelines and personal norms into the physician's utility function. For, we largely adapt the theoretical framework first introduced by Krupka and Weber (2008) and then readjusted by Burks and Krupka (2012) to fit the model to our physicians' behavior.

In particular, we assume that each physician forms his own diagnostic hypothesis, based on his own experience, in the face of a specific clinical circumstance. According to Samuels and Ropper (2005), thanks to their clinical experience, physicians gain subtle skills which might not be captured by evidence-based measures. In fact, physicians seldom resort directly to guidelines in their clinical decision making, preferring to rely on their 'mind-lines' (guidelines-in-the-head) acquired over a lifetime and informed by their conversation with colleagues, their interactions with patients, their

[4] A further review supporting the practice of vignettes can be found in Veloski et al. (2005).
[5] See Beilby and Silagry (1997) and Grimhaw et al. (2001) on the role of information in changing physicians' behaviour..

early training and their readings (Gabbay and le May, 2004). Supporting this view, Elstad et al. (2010) show that clinical experience brings physicians social, behavioural and intuitive skills and knowledge, which allow them to meet patients' needs and to compare present day patients with similar past patients. Moreover, physicians are said to consult outcomes research, depending on their experience level (Tannenbaum, 1994). More experienced physicians are less willing to adhere to practice guidelines for several reasons. First of all, copious, ever-changing and often contradictory guidelines are hardly able to take into account all the factors which come into play when physicians have to make clinical decisions (Gabbay and le May, 2004). Second, the longer physicians have been in practice, the lower their ability to incorporate new guidelines. On the one hand, their experience leads them to understand types of patients as well as the progress of their disease and this contributes to increase physicians' cognitive rigidity (Choudhry et al., 2005). On the other hand, less experienced physicians who cannot depend on such acquired skills are more likely to welcome new information provided by practice guidelines (Elstad et al., 2010). However, since adherence to guidelines reduces the risk of being sued for malpractice, physicians strive to conform to them regardless of their experience level (Elstad et al., 2010). In fact, since practice guidelines define the standard benchmarks for medical treatments, they support the law governing medical malpractice in assessing physicians' conduct and hence eventually identifying cases of negligence (Gabbay and le May, 2004; Havighurst, 1991). In the end, even though physicians can get insurance to reduce their financial risks, they carry many non-insurable costs incurred for malpractice litigation such as psychic, time costs of legal proceeding and the risk of undermining their reputation (Currie and Macleod, 2006), which they try to minimize by following guidelines. Although our focus is not on physician's mind-lines measurement, the likely mismatch between them and the national guidelines, as suggested by the literature, could explain physician's lack of coordination.

Based on the above, we can model the utility function of physicians as depending on monetary payoffs due to the chosen action; on the closeness of chosen action to internal (e.g., hospital or workgroup) guidelines; on the closeness of chosen action to national guidelines. Hence, we assume that a physician's utility function is given by:

$$U(a_k) = V(\pi(a_k)) + \gamma N_g(a_k) + \boldsymbol{\theta} N_n(a_k) \quad (1)$$

- where: a_k refers to the action selected by physician
- $V()$ represents the value given to the monetary payoff which increases in $\pi(a_k)$

- $N_g(a_k)$ reflects the degree of appropriateness of an action according to the hospital internal guidelines[6] which constitute the own workgroup's relevant norms (e.g., norms generally shared by colleagues)
- $\gamma \geq 0$ accounts for physician's sensitivity to the adherence to internal guidelines
- $N_n(a_k)$ refers to the degree of appropriateness of an action following national guidelines, which constitute the standard procedures used by the tort system to investigate on physician's diligence
- $\boldsymbol{\theta}$ refers to physician's sensitivity to the obedience to national guidelines

Even though, internal guidelines and national guidelines ought to coincide, in recent years, heterogeneity in the use of procedures for specific medical conditions has been documented at a national level (Switzer et al., 2003). In particular, the production of clinical guidelines had such an exponential growth that physicians, patients and other stakeholders have to juggle several guidelines which vary in quality and are sometimes discordant (Camera civile di Firenze, 2017). As confirmed by Mapelli and Lucioni (2003), many Italian ASL (Local sanitary units), based on their common practices, have given rise to their own local guidelines to be followed by their employees[7], which in some cases deal with clinical circumstances not covered by national guidelines. For this reason, physicians have to bear in mind both national and internal guidelines, when both exist, while treating the patient, without neglecting their own clinical experience. Whether all these norms do not overlap, physicians decide what is the best strategy to adopt depending on the peculiarity of the clinical case as well as their experience level (Elstad et al., 2010).

Given our assumptions on physician's behaviour, we can derive some hypotheses to be tested. First, physicians take into account their mind-lines, in making important clinical decisions. However, when physicians realize that they are in disagreement with their respected colleagues, their mind-lines morph. By using the ethnographic methods of anthropologists, Gabbay and le May (2004) investigate practitioners' behaviour and beliefs and show that physicians tend to change their mind-lines while interacting with trusted colleagues[8]. Hence, physicians tend to have similar views from the ones of their trusted colleagues on what is the right thing to do. As a consequence, we expect that physicians coordinate in judging the appropriateness of each of the actions proposed in this experiment, as possible solution to a given clinical case. In particular, when physicians work at the same hospital,

[6] Notice that, in the demographic survey, we asked participants whether or not the hospital which they work for adopt his internal guidelines.
[7] For instance, this is the current situation in the Italian region of Emilia Romagna.
[8] According to the same authors that is where coffee-room chat comes into play: physicians share their stories and their experiences so as to help each other solve clinical problems.

they face the same hospital internal guidelines. This is the case of our sample of physicians. In fact, in our experimental setting, it is possible to elicit physicians' beliefs about their peers' evaluations. Doing so, we can identify the actual 'norm' regulating them, either internal or national guidelines[9]. Even though participants are not able to exactly predict how each of their colleagues is going to answer to each of the proposed vignettes, they may be able to guess the most likely modal answer and stick with it. Therefore, we expect that physicians' valuations of each action do not differ, regardless of their medical specialties.

Behavioural hypothesis 1: *Physicians coordinate in judging the degree of appropriateness of each action*.

Second, we have already mentioned the role of guidelines into physicians' decision-making process. Even if they are not taken as directives, being unable to consider the multiple factors which play a part in clinical decisions, the guidelines generally identify recommended courses of action under certain circumstances (Institute of Medicine, 1990). For this reason, we expect the frequency of reporting such actions to be 'very appropriate' to be higher than the same frequency for any other action not prescribed by guidelines. Additionally, having been trained into the same national context, we can assume that all the physicians in our sample should be able to recognize the presence of the national guidelines among the proposed actions in the vignettes[10]. Hence, we expect that physicians coordinate in giving the same appropriateness judgment to the actions corresponding to the national guidelines.

Behavioural hypothesis 2a: *The average frequency of reporting an action corresponding to national guidelines to be 'very appropriate' is higher than the frequency for any other available action*.

Behavioural hypothesis 2b: *The average frequency of coordination on actions corresponding to national guidelines is higher than in any other case.*

[9] Although we acknowledge that internal norms may vary across physicians, in our experiment it cannot happen because all the physicians joining the experiment belong to the same Hospital.
[10] In each vignette, one out of the four proposed actions to be taken represented the national guideline to be followed in that specific situation.

Finally, we investigate whether physicians, in the Information treatment, switch to the action corresponding to guidelines, once they are made aware of guidelines content. Usually, physicians, in face of a specific clinical circumstance, form their own diagnostic hypotheses based on their own experience (i.e. mind-lines) and on internal guidelines, if present. However, when a physician asks for guidelines and notices that they differ from her previous chosen action, she may decide to change her answer (Gabbay and le May, 2004). In a nutshell, physicians may choose to conform to either internal or national guidelines on what is the best practice to adopt, overruling their mind-lines. This may be due to both following an action shared with colleagues and, by adhering to guidelines, reducing the risk of being sued for malpractice (Elstad et al., 2010).

Behavioural hypothesis 3: *Once knowing the guidelines contents, physicians change their choices according to guideline's suggestion.*

4. Basic setup and experimental protocol

4.1 Basic setup

To assess physicians' level of coordination and to investigate the role of national and internal guidelines, we have used the technique introduced by recent experimental literature to elicit norms (Gaechter et al. 2013; Barr et al. 2018; Burks and Krupka 2012). Such a technique combines the description of hypothetical situations through *ad hoc* vignettes with the coordination game structure. Physicians are randomly assigned to either the Coordination treatment or the Information treatment. Each treatment is composed by two stages. Before starting the first stage of any treatment, physicians have gone through the well-known Holt and Laury (2002) test to assess the attitude towards risk. Once, they have completed this test, the first stage starts.

In the first stage, which is common to both treatments, physicians face three different vignettes, each referred to a different specialty, depicting a patient suffering from a particular disease with a given diagnosis.[11] For each vignette, a set of four action (i.e., one of them has been set according to national guidelines) in response to the disease is proposed. Thus, physicians have to rank each of the

[11] To depict vignettes we asked for some specialists' availability. They suggested the three scenarios, confirming they were realistic and easy understandable to any physician regardless of his medical specialty.

alternatives, on a scale of one to four, according to their perceived degree of appropriateness, where 4 stands for 'very appropriate', while 1 corresponds to 'very inappropriate'. The instructions clarified that an action is appropriate when a physician believes that it is the correct thing to do given the specific situation. Participants are also made aware that, before moving to the next stage of the experiment, are going to be asked to indicate how certain they are about the answers given to vignettes, evaluating their confidence on a five-point scale, where 5 stands for most certain (following Kovacs et al., 2020 and Baumann et al., 1991).

In the second stage of both CT and IT, physicians are asked whether they would like to have access to guidelines content, alternatively they can move to the next scenario. Those who have asked for guidelines' contents received the requested information privately and then, they move to the next scenario. Once all participants have made their choices, an experimenter reads aloud guidelines content. Publicly announcing guideline's contents regarding previous scenario does not affect physicians' choices in the following one, belonging the scenarios to different and unrelated medical specialty. Differently, in the IT only, physicians who have decided to ask for guidelines can change their appropriateness judgments, after receiving the information on which action resembles guidelines contents on that specific disease.

Likewise Krupka and Weber (2008), once all the participants had completed the tasks, one of the 12 possible actions[12] was randomly selected and after having compared all the physicians' responses with the modal judgement, subjects were privately paid for correct matches. In particular, whereas in the CT, the modal answer has been computed considering the evaluations given in the first stage only, in the IT, modal answers are calculated considering the answers given in the first stage, together with new answers given by physicians who changed appropriateness judgments. This monetary award mechanism has been fully reported in the instructions and it has been clearly explained to participants at the beginning of each experimental session.

4.2 Experimental protocol

The experiment was conducted at the main hospital of Reggio Calabria, thanks to an agreement signed by the same hospital and the Mediterranean University of Reggio Calabria[13]. Thus, 52 physicians

[12] Notice that for each vignette, there were 4 possible actions.
[13] According to such agreement, the parties involved are committed to participating in the research project 'Experiments

took part in the experiment: 24 joined the Coordination Treatment, whereas the remaining were assigned to the Information Treatment. Although our vignettes specifically refer to three specialties (i.e. orthopaedics, paediatrics and oncology), we recruited physicians from any available specialty (including also those strictly related to the vignettes)[14]. The experimental sessions have been advertised by sending emails to hospital physicians' mailing list and by head physicians of each specialty department involved. To join the experiment, during their coffee-breaks, physicians came, individually or in group, to the meeting room, to avoid any interference with the working schedule[15].

As already mentioned, before starting the experiment, we have assessed physicians' attitude towards risk. Given that coordination games require that each player tries to guess the other players' behaviour (Heinemann et al., 2009), subjects' risk attitude can affect the outcome of the coordination game, like what happens in a lottery (Neumann and Vogt, 2009). To evaluate physicians' degree of risk aversion, we have implemented the well-known questionnaire taken from Holt and Laury (2002) with hypothetical rewards (Galizzi et al., 2016). The results of the questionnaire are slightly different from that obtained by Holt and Laury (2002): 44% of the subjects resulted risk averse; approximately 14% of the subjects showed extreme risk aversion; 29% of the subjects turned out to be risk-loving; and 13% of the subjects cannot be classified in any of the above category. The remarkably high number of risk-loving subjects can be due to the presence of many surgeons in the pool (roughly 70%), who usually show greater risk seeking, as reported by Pikkel et al. (2016).

After completing the risk assessment questionnaire, participants played the two stages of the experiment by pen and paper, which lasted approximately 15 minutes per subject. At the end of the experiment, while payoffs were calculated, subjects were asked to complete a demographic questionnaire, asking about physicians' years of experience, job satisfaction and their relationship

in health economics' and to sharing results which could be useful to the hospital for management policy.

[14] Although we control for this effect in the regression analysis provided in the following section, we acknowledge that this could be a problematic issue due to the fact that some physicians may not feel confident enough when dealing with disease referring to unfamiliar specialties. Also, physicians' willingness to know guidelines and then change decisions could vary with their field of specialization.

[15] To avoid any interaction between physicians who already joined the experiment and physicians who were about to participate, physicians enter in the experimental room through one door and then exit trough another door (i.e., to avoid behavioural spillovers).

with colleagues. Finally, physicians matching the modal answer[16] received their 10-euro meal ticket from the cafeteria[17]. The average reward for participants was €5.19[18].

5. Results

5.1 Descriptive analysis and non-parametric tests

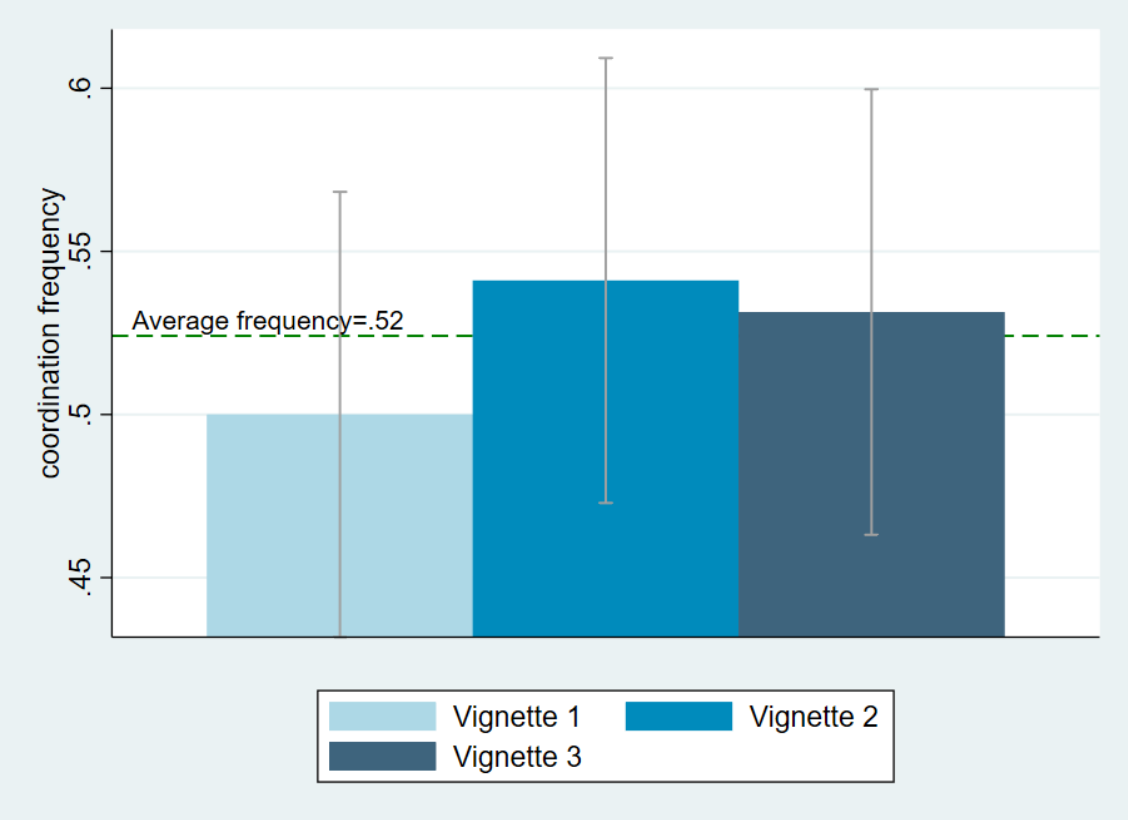
According to our experiment, coordination is achieved when physician's appropriateness judgment matches the modal valuation for the specific action considered, which, in turn, corresponds to the judgment given by most participants. In other words, we assess whether subjects equally evaluated any action. Such a choice contributes the external validity of the experiment, resembling the case of colleagues who generally work together (i.e., same hospital ward) and have to make an agreed decision (i.e., when prescribing a therapy to the patient who is about to be discharged). Figure 1 reports the average frequency of coordination achieved by participants, in each vignette.

Result 1: The overall average frequency coordination is 0.52[19]. In particular, average frequency of coordination is more or less constant across vignettes, being 0.5 in the meniscus injury scenario, 0.53 in the oral cavity cancer scenario, and 0.54 in the woman breast fissures one[20]. The highest level of coordination has been achieved by gynaecologists or neonatologists, that accounts for 36% of all physicians in the sample[21]. Comparing the average levels of coordination by vignettes, differences are not significant according to the Median test (p=0.23). However, although not all physicians have

[16] Physicians know that the modal answer was calculated considering all the answers given in the same day, though through different sessions.
[17] The above-mentioned incentive is reasonably salient for two reasons. First, cafeteria is the only hospital internal alternative available to physicians. Although there are some external cafés, walking distance from the hospital, their opportunity cost may be high (physicians would have to push out and walk for 15 minutes). Additionally, according to the regulation, the internal cafeteria must charge discounted rates (20% less than standard prices) to the hospital's employees. Also, see https://research.utoronto.ca/compensation-reimbursement-research-participants
[18] Although the financial incentive could be retained relatively low given the average income of the subject pool, their intrinsic motivation should be already enough to incentivize their performance in the experiment. In fact, according to Gneezy and Rustichini (2000) when an activity has a motivation its own, in this case contributing to medical research, introducing a monetary reward can have a detrimental effect on performance. As a result, under certain conditions, introducing a compensation contingent on performance may crowd out any endogenous incentive which instead the experimenter wishes to elicit (McKeganey, 2001).
[19] Notice that physicians, at this stage, neither receive information on which actions mirror guidelines content nor know that they will be given the opportunity to ask for them.
[20] In appendix B, looking at one vignette at a time, we check the average frequency of coordination for each action.
[21] Appendix D provides a table of the specialities' distribution across the sample.

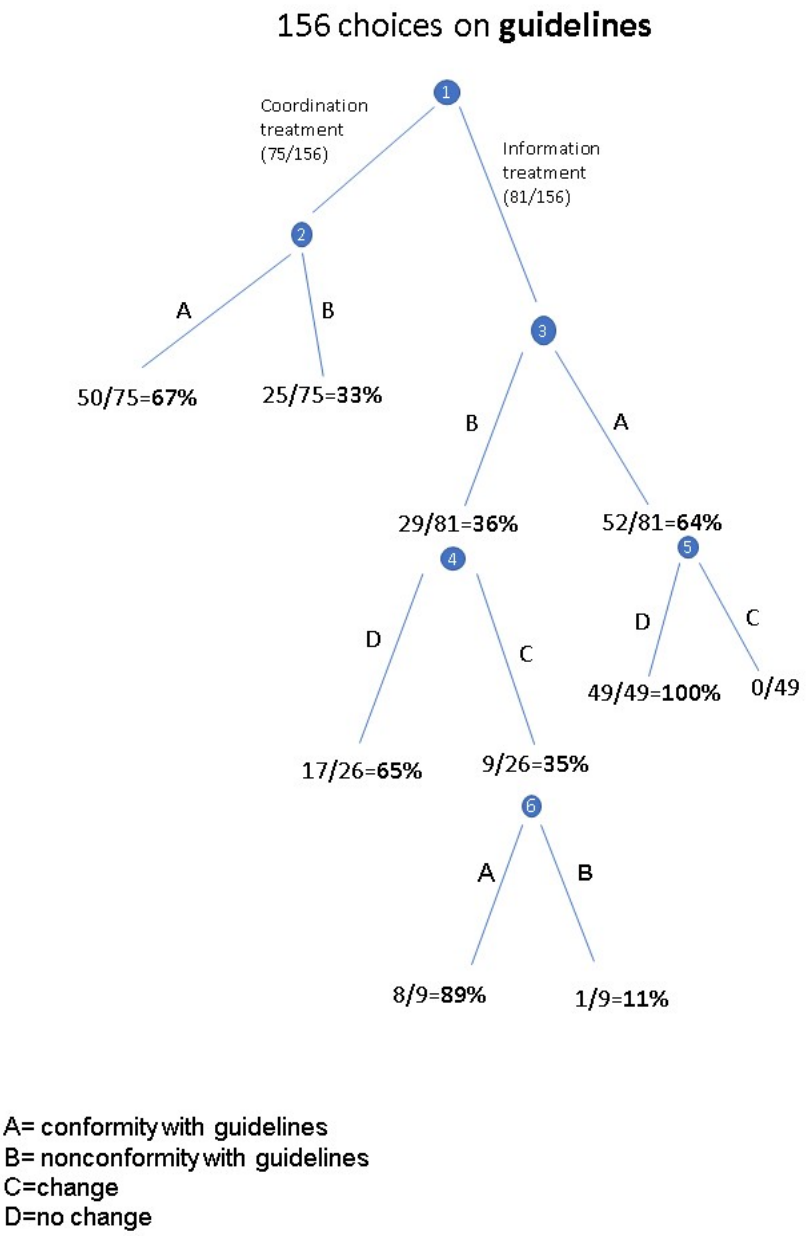
managed to coordinate on a given action, the average levels of coordination reached in the sessions are quite high.

Figure 1: Average frequency of coordination across the experiment



Besides coordination levels, it may be relevant to focus on physicians' attitude towards national guidelines. According to Connelly et al., (1990) accessibility and searchability (i.e., '*easy way to find relevant knowledge in the available source*'), are the two main factors affecting physicians' use of information (Connelly et al., 1990). Also, in our design, physicians do not have any opportunity cost (e.g. time and energy required to obtain information, Timpka et al., 1989) in knowing guidelines, since they are easily and free of charge accessible just checking the option 'yes'. As a consequence, only 8% of the physicians did not want to know national guidelines content, which corresponds to four physicians, equally divided between the two treatments. In order to depict more clearly the sequence of physicians' decisions and their respective average frequency, we report, in Figure 2, a game-tree style structure starting from the second stage of both treatments (i.e., when participants are asked whether to find out which action corresponds to the national guideline).

Figure 2: Structure of individual decisions on guidelines and average frequencies



All the numbers reported in the decision tree indicate the frequency of choices. Each of the 52 physicians made one guidelines-related choice in each of the three vignettes[22], leading to 156 choices. In 67% of the times, physicians have conformed with guidelines from the very first stage, assigning

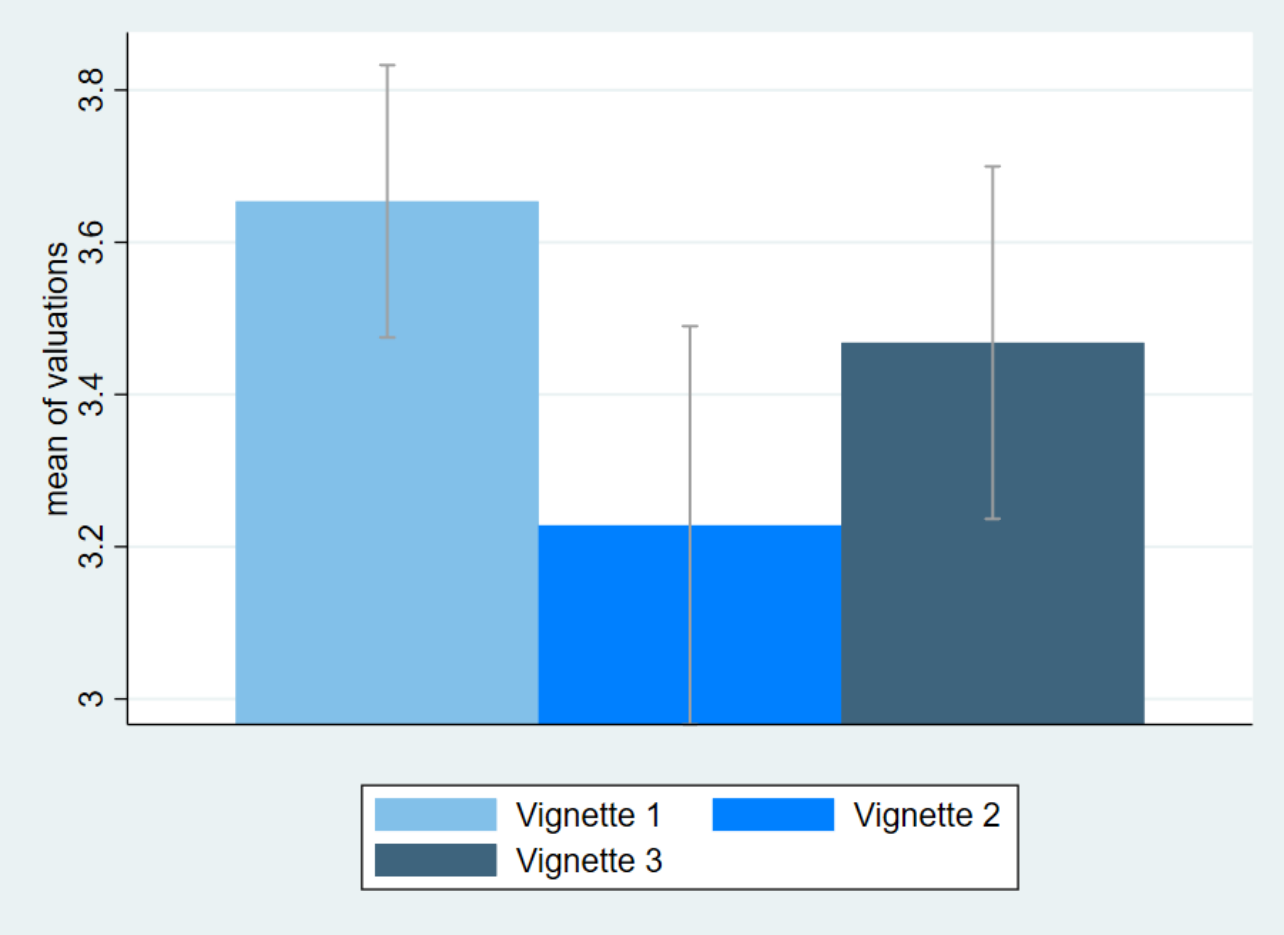
[22] Appendix C provides figure 2 broken down by vignette.

the highest appropriateness level to the action reflecting guidelines, regardless of treatment. Looking at the two treatments, the average frequency levels are roughly the same: 48% in the CT (left branch of node 2), and 52% in the IT (right branch of node 3). Furthermore, 33% of the times, subjects playing in the Coordination Treatment did not conform with guidelines at the first stage. In this case, 67% of the times physicians, although being told the guidelines and having the opportunity to change judgement, have decided to stick to their decisions[23]. On the contrary, 35% of the times physicians have decided to change their previous decisions (right branch of node 4), giving the maximum appropriateness judgements to the action corresponding to national guidelines in 89% of the times (left branch of node 6).

Result 2a: Figure 3 shows that, for each proposed scenario, the actions corresponding to national guidelines achieved very high average appropriateness ratings, though physicians have not yet received information on guidelines content. In particular, by comparing the distributions of the statement 'very appropriate' for such actions with those of other statements, the Wilcoxon rank-sum test reports significant differences (p-value<0.001). Also, on average, guidelines' evaluations have been higher than any other action suggested in the experiment according to the Median test (p-value<0.001). Hence, our data confirm hypothesis 2a.

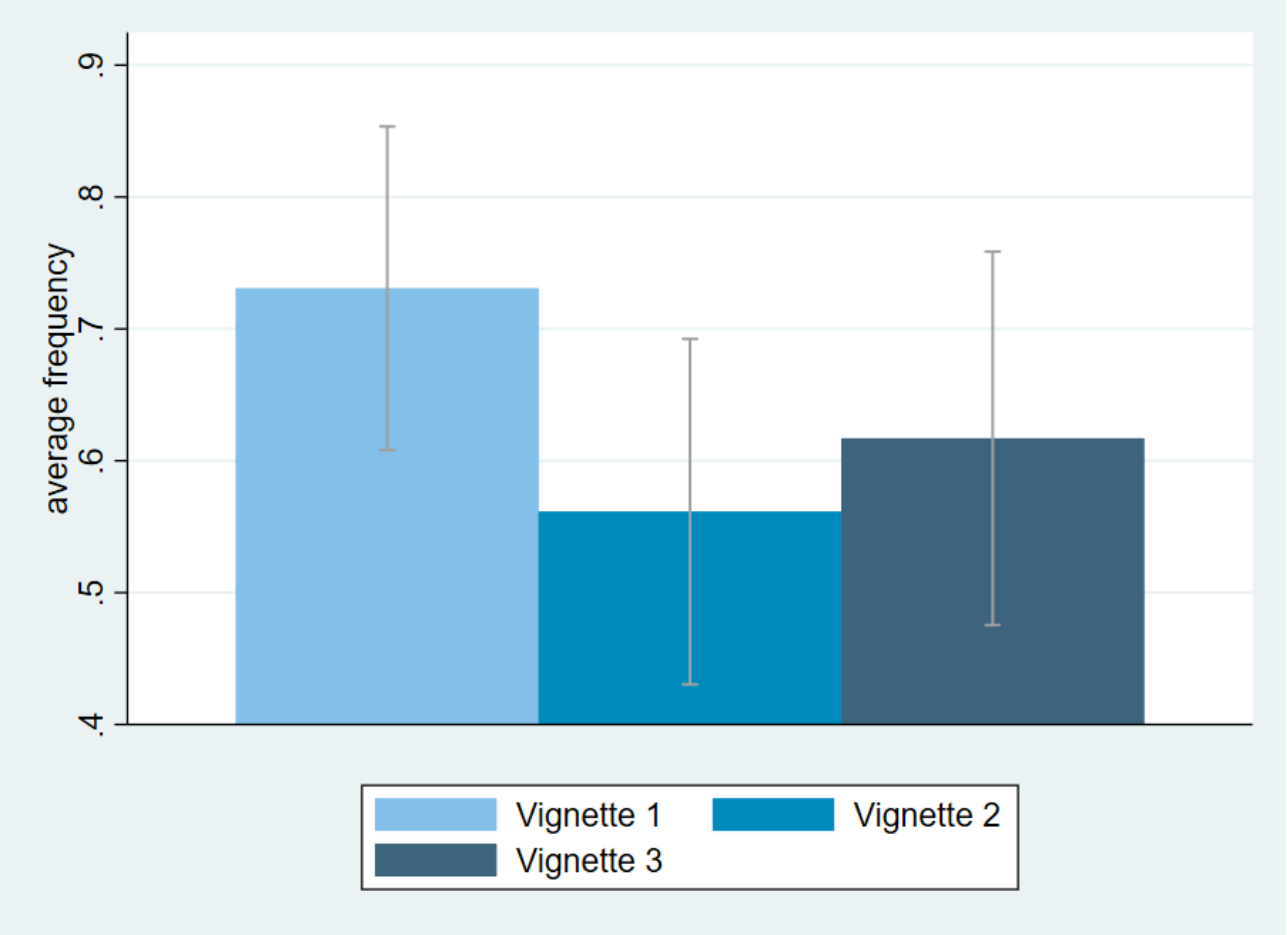
[23] Notice that from the 33 choices, the number decreases to 26 in node 4, removing the cases where physicians did not access the third stage, having refused to know guidelines.

Figure 3: Average evaluations of national guidelines



Result 2b: Figure 4 reports that the average frequency of coordination on the actions corresponding to national guidelines accounts for more than 60% of the cases. This shows that most of participants either followed national guidelines or choose actions corresponding to the national guidelines and expect other physicians to do the same. Thus, the role of national guidelines as a tool to achieve higher coordination among physicians cannot be neglected. Also, Figure 4 illustrates the differences among vignettes in term of average coordination levels which are not significant according to the Wilcoxon rank-sum test (p-value=0.32). However, by comparing coordination levels for guidelines with those of all the other actions, differences are significant according to the Kolmogorv-Smirnov test (p-value<0.05). Hence, our data provide support to hypothesis 2b.

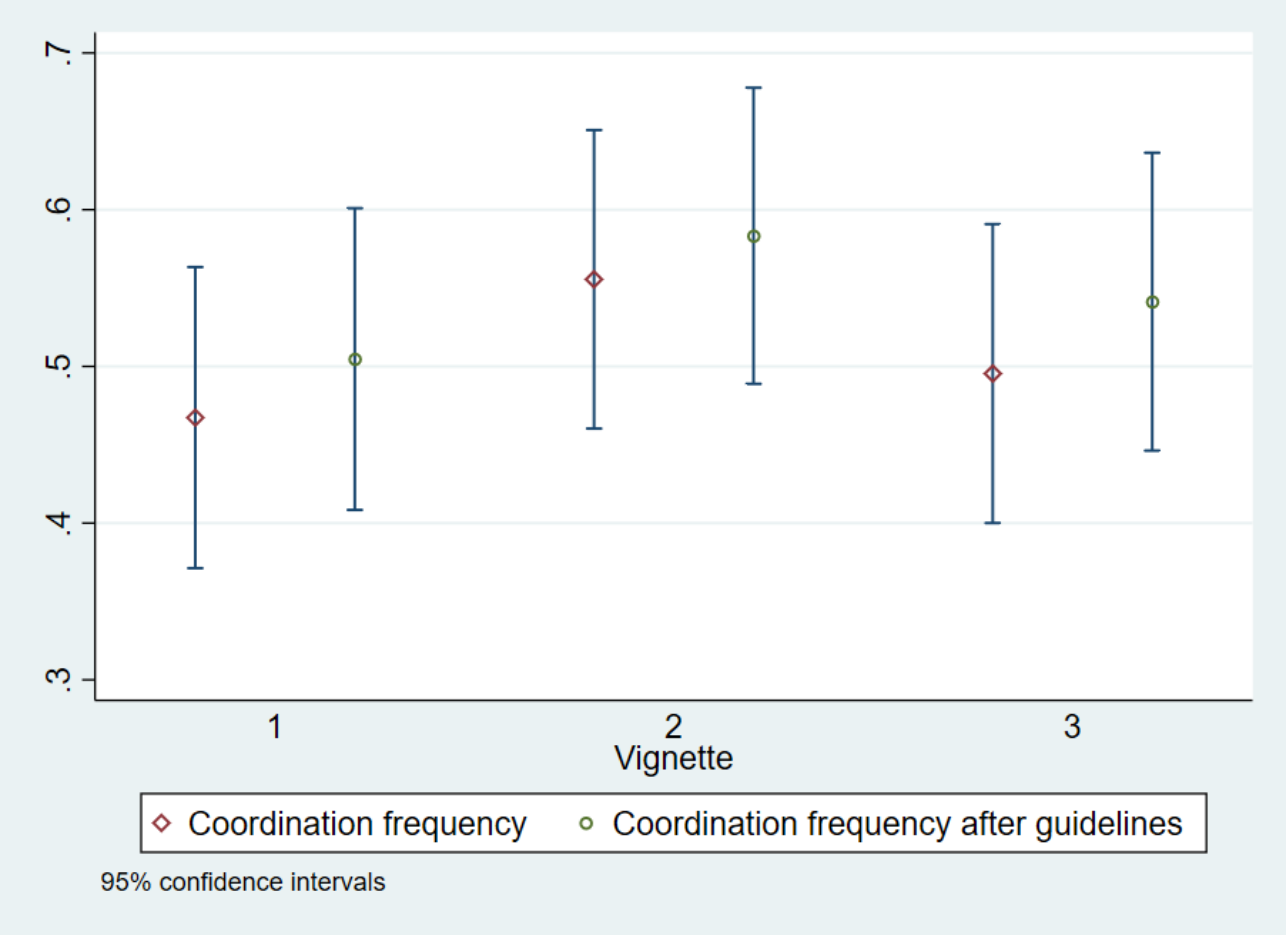
Figure 4: Average frequency of coordination on national guidelines



Result 3: Focusing on Information Treatment, Figure 5 reports the average coordination levels reached before and after having the chance of modifying own judgement for each vignette. In particular, physicians increase average frequency of coordination when changing their appropriateness judgements, once they have obtained information on guidelines. Overall, the average frequency of coordination in the first stage of IT is 0.51, which rises to 0.54 in the third stage of the treatment. Despite differences being not significant according to the Wilcoxon rank-sum test (p-value=0.35), and confidence interval bars overlapping, the greatest benefit of guidelines introduction can be seen in the third vignette (even though values are very close from one vignette to another)[24]. Also in this case, the experimental data support hypothesis 3.

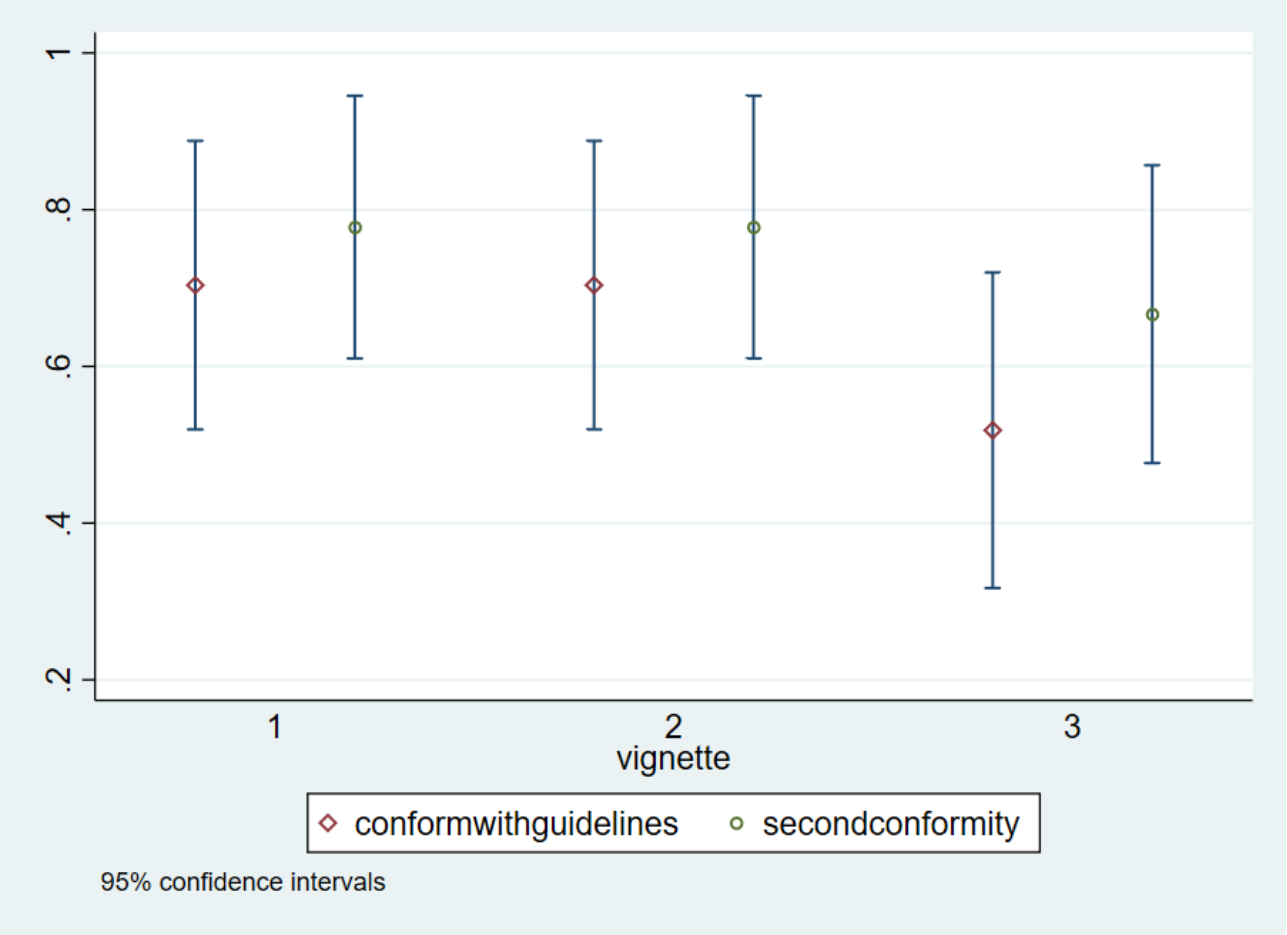
[24] By comparing paired vignettes (e.g. vignette 1vs vignette 2) differences are not significant as well.

Figure 5: Coordination frequency in the Information Treatment



Finally, considering the Information treatments again, it may be worth checking the average frequency of awarding the highest judgment rank (4 out of 4) to actions corresponding to national guidelines. In the remaining of the paper, we refer to the choice of attributing the highest appropriateness level to the action reflecting guidelines as "conformity with national guidelines". As shown by Figure 6, conformity average frequency goes from 0.64 (i.e., prior to introducing guidelines) to 0.74 (i.e., once physicians have the possibility of changing their judgments). Although, conformity with guidelines slightly increases with the possibility of changing judgments, the differences are not significant according to the Wilcoxon test (p-value=0.13).

Figure 6: Conformity with guidelines in the Information Treatment



5.2 Parametric analysis

Based on the above-mentioned results, we now investigate which factors may affect coordination levels and the choices regarding national guidelines. Before moving to the empirical models and results, we report the descriptive statistics of all the variables being considered in the regression analysis in Table 2. These variables have been obtained from physicians' answers to a questionnaire submitted to participants at the end of the experiment (see, Appendix A). Additionally, dummy variables for each vignette are introduced to control for specific effects.

Table 2: Descriptive statistics

Variables	Description	Mean	St. Dev.
male	Dummy for gender	0.60	0.49
age	Age	51.50	8.76
years of service	Years of employment	12.58	10.48
confidence	Individual perceived correctness on a five-point scale	3.82	0.86
specialty	Dummy for specialization in the specialty of the vignette	0.19	0.39
research	Dummy for carrying out scientific research	0.46	0.50
updating	Frequency of consulting scientific journals on a f.p.s[25]	3.25	0.51
colleagues' advice	Frequency of asking for colleagues' advice on a f.p.s	2.77	0.61
negative influence	Influence of colleagues' divergent view on a f.p.s	2.83	0.64
positive influence	Influence of colleagues' concordant view on a f.p.s	2.46	0.75
team	Frequency of taking decisions in team on a f.p.s	3.08	0.70
leader	Dummy for the presence of a leader in the team	0.29	0.45
whatsapp	Dummy for joining a whatsapp group with ward colleagues	0.65	0.48
risk seeking	Dummy for risk seeking	0.21	0.40
guidelines	Dummy for action prescribed by guidelines	0.25	0.43
coordination	Dummy for coordination	0.52	0.50
conformity with guidelines	Dummy for compliance with guidelines	0.65	0.48
request guidelines	Dummy for asking for guidelins	0.92	0.27
changing decisions	Dummy for changing decisions	0.11	0.31
Vignette 1	Dummy for vignette	0.33	0.47
Vignette 2	Dummy for vignette	0.33	0.47
Vignette 3	Dummy for vignette	0.33	0.47

To investigate which factors may affect coordination levels prior to receive information on national guidelines, we have employed a logit regression on a dataset of 624 observations (52 subjects x 3 vignettes x 4 judgements). Table 3 reports the result of 3 empirical models. The dependent variable, 'coordination', is a dummy variable which assumes the value 1 whether one subject coordinates with the modal answer. Cluster robust standard errors have been used to account for data being obtained from multiple observations per physician (Cameron et al., 2008). We start with the most parsimonious

[25] Four-point scale.

model employing just key variables (i.e., confidence, speciality and leader) and, then gradually we add on controls. Due to some missing answers to the final questionnaire, the number of observations is reduced moving from model (1) to the others, when further variables of interest are included.

Table 3: Logit for coordination

VARIABLES	(1)	(2)	(3)
confidence	-0.133	-0.0901	-0.0594
	(0.118)	(0.103)	(0.121)
speciality	0.852***	0.858***	0.910***
	(0.264)	(0.251)	(0.257)
leader	0.131	0.364***	0.375***
	(0.199)	(0.140)	(0.139)
male		-0.00685	0.000203
		(0.230)	(0.229)
age		-0.0293*	-0.0292*
		(0.0155)	(0.0152)
years_of_service		0.00241	0.00255
		(0.0133)	(0.0131)
updating		-0.183	-0.180
		(0.152)	(0.154)
whatsapp		-0.154	-0.159
		(0.243)	(0.241)
risk_seeking		-0.554***	-0.557***
		(0.158)	(0.159)
positive_influence		0.323**	0.324**
		(0.129)	(0.129)
negative_influence		-0.389***	-0.387***
		(0.136)	(0.135)

vignette1			-0.181
			(0.297)
vignette2			-0.229
			(0.285)
constant	0.412	2.448**	2.435**
	(0.474)	(1.155)	(1.167)
observations	604	584	584

Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Table 3 shows that the probability of coordinating on the action to take in a given vignette is positively affected by the fact that a physician holds the same medical specialization as the one recalled by the vignette. Specifically, the probability of coordinating increases by 21 % on average (marginal effect). As a result, it is more likely that a physician coordinates with another colleague specialized in the same field than with a colleague from a different medical field.

Then, the presence of a team leader increases the probability that physicians coordinate on a shared result by 9%. This result is empirically confirmed also by other works showing that the presence of a leader inside an organization facilitates coordination, overcoming possible misalignments between employers' objectives (Bolton et al., 2012). Additionally, leadership skills are found to benefit patients and improve healthcare organization (Rotenstein et al., 2018). Notice that leader turns out to be significant in model (2) and (3) where more controls have been added. Intuitively, when a team is led by a leader the exchange of ideas and thus of positive feedbacks between colleagues are stimulated (Carnevale et al., 2017), which in turn boost coordination. In a nutshell, a common intent between physicians is more likely in the presence of a leader.

Also, risk seeking turns out negatively affects coordination. The rationale stems from physicians' preference to accept higher risk levels when following their own ideas instead of coordinating with others, conforming with the popular opinion[26]. Physicians generally exhibit high risk propensity (see e.g. McCulloch et al., 2020), as confirmed by results from the Holt and Laury (2002) questionnaire, where almost one third of the sample shows risk seeking behaviour, differently from the original results of Holt and Laury (2002) reporting roughly 15% of risk loving subjects.

[26] As confirmed by McKibbon et al. (2007) risk loving physicians less often use resources in search for answers to clinical questions, than their risk avoiding colleagues.

Finally, positive influence increases physicians' probability to coordinate. Intuitively, receiving positive feedbacks promotes shared understanding and contributes to group cohesion and coordination (Slof et al., 2010). Specifically, according to Janssen and Bodemer (2013), consensus knowledge triggers successful collaboration.

However, although sharing opinions is essential for coordination (Bromme et al., 2005), problems could arise when ideas do not match. This could explain why negative influence reduces the likelihood of coordination. In particular, if a physician takes a colleague's view into proper consideration when they initially do not agree, this may lead to divergence of interpretation of the clinical case which may even worsen coordination.

As a next step, we have looked at coordination on the three choices corresponding to national guidelines (one for each vignette). Table 4 shows that the longer a physician works at the same hospital, the higher the probability of coordinating on judging guidelines is. By estimating the marginal effect of the variable of interest, on average, one additional year of service would predict a 2% increase in the probability of coordination. In fact, experience is one of the many contributing factors to coordination (Dezso et al., 2012). More specifically, working for many years with the same people creates team familiarity which boosts communication and coordination skills, improving team performance (Faraj and Sproull, 2000). Differently, one year increase in age reduces physicians' likelihood to coordinate by 3%. As it is well-known in the literature, physicians experience (regardless of own specialty and of the hospital they work for) results in cognitive rigidity (Choudhry et al., 2005; Clark, 1998). Consequently, more experienced physicians may be less willing to conform with guidelines, when they differ from their mind-lines. This, in turn, explains the negative sign.

Table 4 also shows that men are less likely to coordinate on guidelines than women. There is a wide literature on gender differences under several points of view. In particular, women are usually found to be more prone to follow social norms and shared code of conduct (Jhangiani et al., 2014; Okten et al., 2020). This result can be interpreted looking at subjects' responses to the final questionnaire. In addition, men declared to be less willing to share results of diagnostic tests or outcomes of surgical procedures with their colleagues and, they reported to seldom participate in training courses and brainstorming sessions which could jeopardize coordination.

Table 4: Logit model – Dependent variable: Coordination on guidelines

VARIABLES	(1)	(2)	(3)
confidence	0.202	0.374*	0.341
	(0.184)	(0.221)	(0.247)
speciality	1.013**	1.460**	1.696**
	(0.496)	(0.603)	(0.766)
leader	-0.263	0.525	0.501
	(0.418)	(0.417)	(0.406)
male		-0.943	-0.836
		(0.612)	(0.640)
age		-0.138***	-0.134***
		(0.0402)	(0.0394)
years_of_service		0.0933***	0.0894***
		(0.0296)	(0.0290)
updating		-0.770*	-0.796**
		(0.407)	(0.406)
whatsapp		0.133	0.0159
		(0.579)	(0.611)
risk_seeking		-1.341**	-1.365**
		(0.544)	(0.560)
positive_influence		0.852**	0.864**
		(0.371)	(0.373)
negative_influence		-0.480	-0.508
		(0.352)	(0.369)
vignette1			0.407
			(0.587)
vignette2			-0.623
			(0.589)
constant	-0.369	6.597**	6.817**
	(0.806)	(2.698)	(2.734)

observations	151	146	146

Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Finally, still considering first stage data only, we assessed physicians' likelihood to judge the action prescribed by guidelines as 'very appropriate' one (i.e. 4 out of 4). As before, we have restricted the dataset to the three courses of action suggested by national guidelines. Table 5 displays the output of the logit model where the dependent variable 'Conformity with guidelines' takes the value 1 when a physician reports the action corresponding to the guidelines to be very appropriate.

Table 5: Logit model – Dependent variable: Conformity with guidelines

VARIABLES	(1)	(2)	(3)
confidence	-0.0248	-0.0682	-0.177
	(0.214)	(0.253)	(0.276)
speciality	0.00840	0.311	0.471
	(0.375)	(0.350)	(0.409)
leader		0.162	0.157
		(0.394)	(0.414)
male		-1.009***	-1.086***
		(0.373)	(0.413)
age		-0.0908***	-0.0938***
		(0.0295)	(0.0322)
years_of_service		0.0637**	0.0660**
		(0.0251)	(0.0271)
risk_seeking	-0.110	-0.403	-0.489
	(0.448)	(0.454)	(0.468)
positive_influence		0.775***	0.855***
		(0.269)	(0.294)
negative_influence		-0.311	-0.310
		(0.346)	(0.369)
vignette1			1.343**
			(0.554)
vignette2			0.287
			(0.473)
team		-0.487*	-0.541**
		(0.261)	(0.276)

constant	0.695	5.455***	5.493***
	(0.882)	(1.739)	(1.906)
observations	150	150	150

Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

In Table 5, we added the variable Team to some of the others already employed in previous tables. In particular, the variable Team has been built from physicians' answers to the question 'How often decisions about a patient are taken as a result of a team's valuation?' on a four-point scale, where 4 stands for the highest frequency, namely. Results show that working in team decreases the probability of evaluating the actions corresponding to national guidelines as the most appropriate, which might appear counterintuitive. In fact, we would expect that teamwork facilitates the diffusion and the adoption of national guidelines. However, it could also be the case that members of the team have different ideas (Hollenbeck et al., 2011) and share them with their colleagues, making actions differing from guidelines equally acceptable as the right solution to adopt, given their success in similar previous clinical cases (similar to Clark et al., 2002). In fact, whereas a wide range of experimental literature (e.g. Kocher et al., 2004; Blinder and Morgan, 2005; Feri et al., 2008) finds that teams perform better than individuals in decision making, De Dreu and Weingart (2003) and van Woerkom and Sanders (2010) show that teamwork success depends on group cohesion. Specifically, disagreement within the team could undermine cooperation and negatively affect openness for sharing opinions (i.e., in this case guidelines). Regarding the remaining significant variables, the discussion provided for table 4 still holds.

Notwithstanding the above-mentioned evidence, the role of guidelines as coordination device can be investigate more deeply looking at the behaviour of physicians joining the Information treatment. In particular, we are interested in assessing whether, after knowing guidelines contents, physicians change their appropriateness judgements according to the guidelines.

It has to be acknowledged that not all 28 subjects, joining the Information treatment, decided to change previous judgment after being made aware of guidelines contents. In fact, some participants have chosen not to know the guidelines, not reaching the last stage of the treatment. Since physicians' choices determine whether or not they are observed in the subsample, we face the problem of selection bias, when the sample is non-randomly selected (Lennox et al., 2012). In this case, relying on standard regression technique produces inaccurate estimates '*of the effects of the independent variables conditional on a case of being into the sample'* (Sartori, 2003). Hence, we have to employ an econometric technique that accounts for the presence of censored observations for changing

decisions. When physicians do not access the last stage of the game, not requiring guidelines, a two-stage limited dependent variable model is needed. Among all the other options, the Tobit test[27] suggests the Cragg's model as the most suitable one for this dataset (Blundell and Meghir, 1987). The Cragg's model, which allows to distinguish between the variable effects on selection and outcome, consists of two different estimations: a logit for the selection choice which is the probability that the latent variable takes a nonnegative value[28] (changing decisions can take a value when physicians opt for guidelines) and a truncated regression for the outcome of changing decisions. As done in previous regressions, we implement robust cluster errors. However, given the small number of clusters, bootstrap methods, with 50 replications, originating pseudo-samples from the initial pool are used (Roodman et al., 2019). The first estimation of Cragg's model is reported in Table 6, where the dependent dummy variable is 'Request guidelines', taking the value 1 if physicians ask for guidelines.

[27] Tobit test=-42.34

[28] To distinguish physicians who were not allowed to change decisions given that they did not opt for guidelines from those who, made aware of guidelines, decided not to change their decisions after guidelines, 'change decisions' was set equal to -1 in the former case, while 0 in the latter.

Table 6: Logit model – Dependent variable: Asking for national guidelines in IT

VARIABLES	(1)	(2)
years_of_service	0.407***	0.407**
	(0.146)	(0.171)
vignette 1		0.00001
		(0.001)
vignette 2		0.0000001
		(0.0001)
research	-1.626**	-1.626
	(0.768)	(1.034)
collegues_advice	-0.888	-0.888
	(2.382)	(2.177)
constant	3.887	3.887
	(7.160)	(6.382)
observations	81	81

Robust errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Table 6 shows that being involved in scientific research decreases the probability of requesting guidelines. According to Katzka (2017), physicians who do clinical research explore a number of possibilities in clinical care, trying to respond to unanswered questions. Researchers sometimes experiment new solutions, never used before, that is the reason why they may not rely upon guidelines but rather improvise and think critically.

Also, physicians' years of experience in hospital increases the probability of asking for guidelines. Intuitively, physicians who have longer been in practice perceive the need to stay abreast of new medicine evidence. For instance, Alper et al. (2004) estimate that virologists for example need 627.5 hours per month to keep current, given the volume of monthly published literature. As a result, the earlier physicians' training, and so the greater their years of services, the more pressing the need for updating.

Finally, to check whether knowing guidelines might change physicians' attitude, inducing them to change their previous decisions we resort to a truncated regression, which excluded 24 observations, as the second estimation of Cragg's model. The dependent dummy variable 'changing decisions'

takes the value 1 whether physicians choose to change their selection in the vignette[29], 0 otherwise. Table 7 shows that specialty often results in practitioners' narrow-mindedness. As a consequence, physicians specialized in one specific vignette are often less likely to change their decisions. In fact, physicians could have detected the action matching guidelines from the very beginning (presumably with a high probability given their specialization) and, thus, they confirm their previous option.

Table 7: Truncated regression – Dependent variable: Changing decisions

VARIABLES	(1)	(2)
speciality	-0.145***	-0.0847
	(0.0512)	(0.0666)
age		0.0103
		(0.00668)
years_of_service		-0.0119*
		(0.00659)
updating	0.0882	0.124
	(0.0854)	(0.0821)
risk_seeking	0.223	0.167
	(0.143)	(0.173)
vignette 1		-0.127
		(0.0947)
vignette 2		-0.103*
		(0.0573)
constant	-0.190	-0.617*
	(0.272)	(0.342)
sigma	0.301***[30]	0.285***
	(0.0422)	(0.0461)

[29] Notice that physicians have the chance of changing decisions only if they ask for guidelines and that only 2 physicians over 28 do not require guidelines support.
[30] Sigma value is the estimated standard error of the regression. The value is comparable to the root mean squared error that would be obtained in an OLS regression.

observations	75	75

Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

6. Conclusions

This framed field experiment assesses the level of coordination achieved among physicians and investigates physicians' adherence to national guidelines and their potential use as a policy instrument. By designing three vignettes describing different clinical cases and a set of four possible prescriptions among which a physician can choose, we asked, in the Condition treatment, 52 physicians to evaluate the appropriateness of each of the possible prescriptions on a scale of one to four in order to match the modal evaluation. As a second step, physicians could immediately and freely of charge being informed on national guidelines contents for the specific medical case. Differently, in the Information treatment, participants, after receiving information on which actions correspond to national guidelines, were allowed to change their previous responses to the vignette.

Results show that the overall average level of coordination across the experiment is 52%. The empirical analysis reports that coordination increases when physicians do often exchange opinions and share positive feedbacks with colleagues. In addition, the presence of a leader in a medical ward facilitates coordination. Furthermore, when physicians have been working since long time at the same hospital, the probability that physicians share guidelines and judge them as the most appropriate actions to adopt in the specific clinical case increases.

For instance, even though in recent years Italy distinguished itself for the initiatives aimed at promoting integrated care through the creation of networks and cooperatives between different health professionals for example in Emilia Romagna, Lombardy and Piemonte (European Union Report, 2017), there is still a lot to do. More generally, collaborations between general practitioners and specialists together with more formal integrated path among professionals belonging to different levels of care organizations should be encouraged to fight lack of coordination. Furthermore, since sharing ideas and feedback is said to increase the level of coordination, teamwork should be incentivized through brainstorming and planning sessions. In this regard, leaders must be trained to be able to manage a team, facilitating collaboration and communication between different members.

Focusing on the role of guidelines, our results show that only 8% of the physicians did not want to know national guidelines content. However, only 23% of the subjects belonging to the Information treatment decided to change their appropriateness judgment after realizing they were in contrast with

guidelines in either one or two vignettes. Finally, the more physicians consult scientific sources, the more willing to accept suggestions from guidelines and, to change their decisions when they are wrong.

Although many physicians comply with guidelines from the very beginning, a fair number of them not only do not recognize them but they show to be reluctant to change their decisions, once they realize they were wrong (i.e., being informed on guidelines). This result supports the assumption that guidelines have hardly changed physicians' behaviour (Hayward, 1997), either due to their overconfidence or to guidelines inflexibility and inability to include all the multiple factors which come into play in medical decisions.

Notwithstanding this, since guidelines dissemination is shown to increase the level of coordination between physicians, hospitals should implement guidelines dissemination programmes. To do this, hospitals should take into account physicians' willingness to update, reflected in the high percentage of them eager to know guidelines (92%), and encouraged in this experiment by guidelines ease of accessibility. For instance, introducing a newsletter program and providing an alternative learning option to the standard education courses could be a solution (Strasser, 1978). More than other instruments, newsletter would have the features of accessibility and searchability required by physicians. Additionally, since physicians' flexibility and open mindedness increases in their frequency of updating, which positively correlates with physicians involvement in research, the latter should be incentivized. Since 46% of the physicians joining the experiment reported to conduct scientific research, introducing possible incentive mechanisms could allow to increase such percentage. Finally, any policy intervention should consider the subjects' age and training experience.

However, we are aware of some limitations of our work. For instance, even though vignettes seem to be very realistic, they can hardly resemble the way physicians are asked to make decisions in real life situations. For, results should be read as the coordination reached on hypothetical scenarios, taken from real diagnoses, in a simplified and controlled environment, though outside the lab. However, if coordination is not reached in such a friendly environment, it seems to be very implausible that coordination could be reached in a hospital where the environmental variables become more pressing. As a result, caution should be taken when reporting the not satisfactory level of coordination reached in our framed field experiment. Also, we acknowledge that it may have been useful to include more and specialty-wise diversified vignettes to better capture physicians' peculiarities. However, this would have increased the length of experimental sessions, possibly biasing physicians' answers.

As future development, we plan to include additional vignettes, representing scenarios where subjects in different hierarchical positions (e.g. physicians and nurses) have to coordinate on providing care

to patients. This would allow to compare coordination levels of subjects belonging to different positions in the organizational structure of an hospital. In addition, by varying the matching answer (i.e., by distinguishing modal answers for different categories of participants), similarly to Burks and Krupka (2012), the coordination game would allow to distinguish between norms held by different groups inside the same organization.

References

Allery, L.A., Own, P.A., Robling, M.R., (1997), Why general practitioners and consultants change their clinical practice: a critical incident study, *BMJ*, 314, 870

Alper, B. S., Hand, J. A., Elliott, S. G., Kinkade, S., Hauan, M. J., Onion, D. K., & Sklar, B. M., (2004), How much effort is needed to keep up with the literature relevant for primary care?, *Journal of the Medical Library Association*, 92(4), 429–437

Argote, L., (1982), Input uncertainty and organizational coordination in hospital emergency units, *Administrative Science quarterly*, 27, 420-434

Baggs, J.G., Ryan, S.A., Phelps, C.E., Richeson, J.F., & Johnson, J.E., (1992), The association between interdisciplinary collaboration and patient outcomes in a medical intensive care unit, *Heart lung: the journal of critical care*, 21(1), 18-24

Barr, A.,& Lane, T., & Nosenzo D., 2018, On the social appropriateness of discrimination, *Journal of Public Economics*, Volume 164, 153-164

Bauchner, H., Simpson L., Chessare, J., (2001), Changing physician behaviour, *Archives of Disease in Childhood*, 84 (6), 459-462

Baumann, A. O., Deber, R. B., & Thompson, G. G., (1991), Overconfidence among physicians and nurses: the 'micro-certainty, macro-uncertainty'phenomenon, *Social science & medicine*, *32*(2), 167-174

Beilby, J. J., & Silagy, C. A., (1997), Trials of providing costing information to general practitioners: a systematic review, *The Medical journal of Australia*, 167(2), 89-92

Berner, E. S., & Graber, M. L., (2008), Overconfidence as a cause of diagnostic error in medicine, *The American journal of medicine*, 121(5), S2-S23

Beuselinck, E., Verhoest, K., & Bouckaert, G., (2007), Reforms of Central Government Coordination in OECD-Countries for Cross-National Unifying Processes? In: *Cultural Aspects of Public Management Reform*, 77–109, Eds. Kuno Schedler & Isabella Proeller, Amsterdam: JAI Press

Blinder, A. S., & Morgan, J., (2005), Are two heads better than one? Monetary policy by committee, *Journal of Money, Credit and Banking*, 789-811

Blundell, R., & Meghir, C., (1987), Bivariate alternatives to the Tobit model, *Journal of Econometrics*, 34(1-2), 179-200

Bolton, P., Brunnermeier, M. K., & Veldkamp, L., (2012), Leadership, coordination, and corporate culture, *Review of Economic Studies*, 80(2), 512-537

Borra, V., De Buck, E., & Vandekerckhove, P., (2015), RICE or ice: what does the evidence say? The evidence base for first aid treatment of sprains and strains. In: *Filtering the information overload for better decision*, Ab- stracts of the 23rd Cochrane Colloquium, 2015 3-7 Oct, Vienna, Austria, John Wiley & Sons

Braddock III, C. H., Edwards, K. A., Hasenberg, N. M., Laidley, T. L., & Levinson, W., (1999), Informed decision making in outpatient practice: time to get back to basics, *Jama*, 282(24), 2313-2320

Bromme, R., Hesse, F. W., & Spada, H., (2005), Barriers, biases and opportunities of communication and cooperation with computers: Introduction and overview, In Barriers and biases in computer-mediated knowledge communication, 1-14, Springer, Boston, MA

Brosig-Koch, J., Hennig-Schmidt, H., Kairies-Schwarz, N., & Wiesen, D., (2016), Using artefactual field and lab experiments to investigate how fee-for-service and capitation affect medical service provision, *Journal of Economic Behavior & Organization*, 131, 17-23

Burks, S., & Krupka E., (2012), A Multi-Method Approach to Identifying Norms and Normative Expectations within a Corporate Hierarchy: Evidence from the Financial Services Industry, *MANAGEMENT SCIENCE*, Volume 58, No. 1, 203–217

Bursztyn, L., González, A. L., & Yanagizawa-Drott, D., (2018), Misperceived social norms: Female labor force participation in Saudi Arabia (No. w24736), *National Bureau of Economic Research*

Camera civile di Firenze, (2017), Le Corti Fiorentine, *Rivista di diritto e procedura civile*, n.3/17

Cameron, A. C., Gelbach, J. B., & Miller, D. L., (2008), Bootstrap-based improvements for inference with clustered errors, *The Review of Economics and Statistics*, 90(3), 414-427

Carnevale, J. B., Huang, L., Crede, M., Harms, P., & Uhl-Bien, M. (2017). Leading to stimulate employees' ideas: A quantitative review of leader–member exchange, employee voice, creativity, and innovative behavior. *Applied Psychology*, 66(4), 517-552.

Carrier, E., Dowling, M. K., & Pham, H. H., (2012), Care coordination agreements: barriers, facilitators, and lessons learned, *The American journal of managed care*, 18(11), e398-404

Cartabellotta, N., Cottafava, E., Luceri, R., & Mosti, M., (2019), 4 Rapporto GIMBE sulla sostenibilità del Servizio Sanitario Nazionale, Fondazione GIMBE, Available at : www.rapportogimbe.it

Choudhry, N.K., Fletcher, R.H., & Soumerai, S.B., (2005), Systematic re- view: the relationship between clinical experience and quality of health care, *Ann Intern Med*, 142(4), 260–273

Clark, A. J., (1998), Reframing: A therapeutic technique in group counselling, *Journal for Specialists in Group Work*, 23(1), 66-73

Clark, M. A., Amundsen, S. D.,&Cardy, R. L., (2002), Cross-functional team decision-making and learning outcomes: A qualitative illustration, *Journal of Business and Management*, 8(3), 217–236

Connelly, D. P., Rich, E. C., Curley, S. P., & Kelly, J. T., (1990), Knowledge resource preferences of family physicians, *J Fam Pract*, *30*(3), 353-9

Cragg, J. G., (1971), Some statistical models for limited dependent variables with application to the demand for durable goods, *Econometrica*: *Journal of the Econometric Society*, 829-844

Currie, J., & Macleod, W.B., (2006), First do no harm? Tort reform and birth outcomes, *National Bureau of Economic Research*, Nber Working paper series

Das, J., & Hammer, J., (2005), Which doctor? Combining vignettes and item response to measure clinical competence, *Journal of Development Economics*, 78(2), 348-383

Das, J., Hammer, J., & Leonard, K., (2008), The quality of medical advice in low-income countries, *Journal of Economic perspectives*, 22(2), 93-114

Das, J., Kwan, A., Daniels, B., Satyanarayana, S., Subbaraman, R., Bergkvist, S., ... & Pai, M., (2015), Use of standardised patients to assess quality of tuberculosis care: a pilot, cross-sectional study, *The Lancet Infectious Diseases*, 15(11), 1305-1313

De Dreu, C. K. W., & Weingart, L. R., (2003), Task versus relationship conflict, team performance, and team member satisfaction: A meta-analysis. *Journal of Applied Psychology*, 88, 741–749

Dezso, C. L., Grohsjean, T., & Kretschmer, T., (2012), Coordination experience and team performance: Evidence from the electronic games industry, *Robert H. Smith School Research Paper No*, RHS, 06-136

Duflo, E., & Saez, E., (2003), The role of information and social interactions in retirement plan decisions: Evidence from a randomized experiment, *The Quarterly journal of economics*, 118(3), 815-842

Elstad, E. A., Lutfey, K. E., Marceau, L. D., Campbell, S. M., von dem Kne- sebeck, O,. & McKinlay, J. B, (2010), What do physicians gain (and lose) with experience? Qualitative results from a cross-national study of diabetes, *Social science & medicine*, 70(11), 1728-1736

European Union, (2017), Tools and methodologies to assess integrated care in Europe, *Report by the Expert Group on Health Systems Performance Assessment*

Faraj, S., & Sproull, L., (2000), Coordinating expertise in software development teams, *Management science*, *46*(12), 1554-1568

Fargason, C.A. Jr., & Haddock, C.C., (1992) Cross-functional, integrative team decision making: Essential for effective QI in health care, *QRB Qual Rev Bull*, 18, 157-163

Feri, F., Irlenbusch, B., & Sutter, M., (2010), Efficiency gains from team-based coordination—large-scale experimental evidence, *American Economic Review*, 100(4), 1892-1912

Field, M. J., & Lohr, K. N., (1990), Clinical practice guidelines, Directions for a new program. Washington: National Academy Pr

Friedman, C., Gatti, G., Elstein, A., Franz, T., Murphy, G., & Wolf, F., (2001), Are clinicians correct when they believe they are correct? Implications for medical decision support, *Studies in health technology and informatics*, (1), 454-458

Gabbay, J., & le May, A., (2004), Evidence based guidelines or collectively constructed "mindlines?" Ethnographic study of knowledge management in primary care, *BMJ*, 329(7473):1013

Gachter, S., Nosenzo, D., & Sefton, M., (2013), Peer effects in prosocial behaviour: Social norms or social preferences?, *Journal of the European Economic Association*, 2013, vol. 11, issue 3, 548-57

Galizzi, M. M., Miraldo, M., & Stavropoulou, C., (2016), In sickness but not in wealth: field evidence on patients' risk preferences in financial and health domains. *Medical Decision Making*, 36(4), 503-517

Gittell, J.H., (2000), Organizing work to support relational coordination, *International Journal of Human Resource Management*, 11(3), 517-539

Glassman, P. A., Luck, J., O'Gara, E. M., & Peabody, J. W., (2000), Using standardized patients to measure quality: evidence from the literature and a prospective study, *The Joint Commission journal on quality improvement*, 26(11), 644-653

Gneezy, U., & Rustichini, A., (2000), Pay enough or don't pay at all, *The Quarterly Journal of Economics*, 115(3), 791-810

Gorter, K., de Poel, S., de Melker, R., & Kuyvenhoven, M., (2001), Variation in diagnosis and management of common foot problems by GPs, *Family practice*, 18(6), 569-573

Grimshaw, J., Shirran, L., Thomas, R., Mowatt, G., Fraser, C., Bero, L., . . . O'Brien, M., (2001), Changing Provider Behavior: An Overview of Systematic Reviews of Interventions, *Medical Care*, 39(8), II2-II45

Havighurst, C.C., (1991), Practice guidelines as legal standards governing physician liability, *Law and Contemporary Problems*, 54(2), 87-117

Hayward, R.S.A., (1997), Clinical practice guidelines on trial, *CMAJ*, 156, 1725-1727

Heinemann, F., Nagel, R., & Ockenfels, P., (2009), Measuring strategic uncertainty in coordination games, *The Review of Economic Studies*, 76(1), 181-221

Hollenbeck, J. R., Ellis, A. P., Humphrey, S. E., Garza, A. S., & Ilgen, D. R., (2011), Asymmetry in structural adaptation: The differential impact of centralizing versus decentralizing team decision-making structures, *Organizational Behavior and Human Decision Processes*, 114(1), 64-74

Holt, C. A., & Laury, S. K., (2002), Risk aversion and incentive effects, *American Economic Review*, 92(5), 1644-1655

Institute of Medicine (US), (1990), *Clinical Practice Guidelines: Directions for a New Program*, Committee to Advise the Public Health Service on Clinical Practice Guidelines, Field MJ, Lohr KN, editors, Source Washington (DC): National Academies Press (US)

Janssen, J., & Bodemer, B., (2013), Coordinated Computer-Supported Collaborative Learning: Awareness and Awareness Tools, *Educational Psychologist*, 48:1, 40-55, DOI: 10.1080/00461520.2012.749153

Jhangiani, R., Tarry, H., & Stangor, C., (2014), Principles of social psychology–1st international edition, Adapted from principles of social psychology from Charles Stangor. BC Open Textbook Project

Katzka, D. A., (2017), How to balance clinical work and research in the current era of academic medicine, *Gastroenterology*, 153(5), 1177-1180

King, A. S., (1974), Expectation effects in organizational change. *Administrative science quarterly*, 221-230

Kocher, M., Strauß, S., & Sutter, M., (2006), Individual or team decision-making—causes and consequences of self-selection, *Games and Economic Behavior*, 56(2), 259-270

Kovacs, R. J., Lagarde, M., & Cairns, J., (2020), Overconfident health workers provide lower quality healthcare, *Journal of Economic Psychology*, 76, 102213

Krupka, E. L., Leider, S., & Jiang, M, (2016), A meeting of the minds: in- formal agreements and social norms, *Management Science*, 63(6), 1708-1729

Krupka, E.L., & Weber, R. A., (2008), *Identifying social norms using coordination games: why does dictator game sharing vary?*, (No. 3860), IZA discussion papers

Lennox, C. S., Francis, J. R., & Wang, Z., (2012), Selection models in accounting research, *The accounting review*, 87(2), 589-616

Leonard, K. L., & Masatu, M. C., (2005), The use of direct clinician observation and vignettes for health services quality evaluation in developing countries, *Social science & medicine*, 61(9), 1944-1951

Lewis, R.D., (2006), *When Cultures Collide: Leading Across Cultures*, Boston: Nicholas Brealey International

Luxembourg: Healthcare & Long-Term Care Systems, (2016), The Joint Re- port on Healthcare and Long-Term Care Systems & Fiscal Sustainability, Paper 37 Volume 2 - *Country Documents*

Mapelli, V., & Lucioni C., (2003), Spending on pharmaceuticals in Italy: macro constraints with local autonomy, *Value Health*, 6(S1), S31–S45

McCulloch, M. R., Howgate, M. D., Pearce, K. G., & Flynn, M., (2020), Risk Behaviour and Attitudes in Medical Professionals, *International Journal of Health Sciences*, 8(2), 1-6

McKeganey, N., (2001), To pay or not to pay: respondents' motivation for participating in research, *Addiction*, 96(9), 1237-1238

McKibbon, K. A., Fridsma, D. B., & Crowley, R. S., (2007), How primary care physicians' attitudes toward risk and uncertainty affect their use of electronic information resources, *Journal of the Medical Library Association*, 95(2), 138

Meoli, B., Sica, S., & Stanzione, P., (2018), Commentario alla legge 8 Marzo 2017 n.24, "Disposizioni in materia di sicurezza delle cure e della persona assistita, nonché in materia di responsabilità professionale degli esercenti le professioni sanitarie", Edizioni Scientifiche Italiane, 52-59

Ministero della salute, (2019), *Allattare al seno: un investimento per la vita*, available at www.salute.gov.it

Mohanan, M., Vera-Hernandez, M., Das, V., et al., (2015), The know-do gap in quality of health care for childhood diarrhea and pneumonia in rural India, *JAMA Pediatr*, 169, 349–57

National Center for Complementary and Integrative Health, (2017), National Institutes of Health: *Clinical Practice Guidelines*, available at http://nccih.nih.gov/health/providers/clinicalpractice.htm

Neumann, T., & Vogt, B., (2009), Do Players' beliefs or risk attitudes determine the equilibrium selections in 2x2 coordination games?, Univ., FEMM

OECD (2015), OECD Reviews of Health Care Quality: Italy 2014: Raising Standards, OECD Reviews of Health Care Quality, OECD Publishing, Paris, available at https://doi.org/10.1787/9789264225428-en

Okten, I. O., Gollwitzer, A., & Oettingen, G., (2020), Gender differences in preventing the spread of coronavirus, *Behavioral Science & Policy*

Pikkel, D., Igal, Y. S. P., Sharabi-Nov, A., & Pikkel, J., (2016), Are doctors risk takers?, *Risk management and healthcare policy*, 9, 129

Roodman, D., Nielsen, M. Ø., MacKinnon, J. G., & Webb, M. D., (2019), Fast and wild: Bootstrap inference in Stata using boottest. *The Stata Journal*, 19(1), 4-60

Rotenstein, L.S., Sadun, R, Jena, A.B., (2018), Why Doctors Need Leadership Training, *Harward Business Review*

Samuels, M. A., Ropper, A. H., Poses, R. M., & Diaz, J. A, (2005), Clinical Experience and Quality of Health Care/IN RESPONSE/IN RESPONSE, *Annals of internal medicine*, 143(1), 84

Sandvik, H., (1995), Criterion validity of responses to patient vignettes: an analysis based on management of female urinary incontinence, *Family Medicine*, 27(6):388-392

Sartori, A. E., (2003), An estimator for some binary-outcome selection models without exclusion restrictions, *Political Analysis*, 11(2), 111-138

Shekelle, P. G., Woolf, S. H., Eccles, M., & Grimshaw, J., (1999), Developing guidelines, *Bmj*, 318(7183), 593-596

Shortell, S.M., Zimmerman, J.E., Rousseau, D.M., Gillies, R.R., Wagner, D.P., Draper, E.A., et al., (1994), The performance of intensive care units: Does good management make a difference?, *Medical Care*, 32, 508-525

Slof, B., Erkens, G., Kirschner, P. A., Jaspers, J. G., & Janssen, J., (2010), Guiding students' online complex learning-task behavior through representational scripting, *Computers in Human Behavior*, 26(5), 927-939

Strasser, T. C., (1978), The information needs of practicing physicians in northeastern, New York State, *Bulletin of the Medical Library Association*, *66*(2), 200

Switzer, G. E., Halm, E. A., Chang, C. C. H., Mittman, B. S., Walsh, M. B., & Fine, M. J., (2003), Physician awareness and self-reported use of local and national guidelines for community-acquired pneumonia, *Journal of general internal medicine*, 18(10), 816-823

Tannenbaum, S.J., (1994), Knowing and acting in medical practice: the epis- temological politics of outcomes research, *Journal of Health Politics*, Policy and Law, 19, 27–44

Timpka, T., Ekström, M., & Bjurulf, P., (1989), Information needs and information seeking behaviour in primary health care, *Scandinavian journal of primary health care*, 7(2), 105-109

van Woerkom, M., & Sanders, K., (2010), The romance of learning from disagreement, The effect of cohesiveness and disagreement on knowledge sharing behavior and individual performance within teams, *Journal of business and psychology*, 25(1), 139-149

Veloski, J., Tai, S., Evans, A. S., & Nash, D. B., (2005), Clinical vignette-based surveys: a tool for assessing physician practice variation, *American Journal of Medical Quality*, 20(3), 151-157

Wang, J., Iversen, T., Hennig-Schmidt, H., & Godager, G., (2020), Are patient-regarding preferences stable? Evidence from a laboratory experiment with physicians and medical students from different countries, *European Economic Review*, 125, 103411

Webb, A., (1991), Coordination: A problem in public sector management, *Policy Politics*, Volume 19, Number 4, 229-242(14)

Yager, J., Linn, L. S., Leake, B., Gastaldo, G., & Palkowski, C., (1986), Initial clinical judgments by internists, family physicians, and psychiatrists in response to patient vignettes: I. Assessment of problems and diagnostic possibilities, *General hospital psychiatry*, 8(3), 145-151

Appendix A: Instructions

Experimental instructions

Welcome to this experiment.

In this experiment you will be asked to perform one task and you will re- ceive a payoff related to it. During the experiment, we request that you remain quiet and do not attempt to communicate with other participants. Participants not following this request may be asked to leave without receiv- ing payment. If you have any questions, please raise your hand and one of us will come to you.

At the end of the experiment you could receive money based on your choices and the choices of others in the task described below.

There will be one task for all participants to perform. You will not receive feedback on the outcome of the task, and you will not be paid until the end of the experiment.

Introductory questionnaire (taken from Holt and Laury 2002)

Before starting the experiment, please fill in the following questionnaire in all its parts. Your answers will not affect your future earnings.

Choose which of the lotteries you would play between the two lotteries proposed.

Lottery A	**Lottery B**	**Your choice**

2€ with probability 1/10 1,60€ with probability 9/10	3,85€ with probability 1/10 0,10€ with probability 9/10	
2€ with probability 2/10 1,60€with probability 8/10	3,85€with probability 2/10 0,10€ with probability 8/10	
2e with probability 3/10 1,60€ with probability 7/10	3,85€ with probability 3/10 0,10€ with probability 7/10	
2€ with probability 4/10 1,60€ with probability 6/10	3,85€ with probability 4/10 0,10€ with probability 6/10	
2€ with probability 5/10 1,60€ with probability 5/10	3,85€ with probability 5/10 0,10€ with probability 5/10	
2€ with probability 6/10 1,60€ with probability 4/10	3,85€ with probability 6/10 0,10€ with probability 4/10	
2€ with probability 7/10 1,60€ with probability 3/10	3,85€ with probability 7/10 0,10€ with probability 3/10	
2€ with probability 8/10 1,60€ with probability 2/10	3,85€ with probability 8/10 0,10€ with probability 2/10	
2€ with probability 9/10 1,60€ with probability 1/10	3,85€ with probability 9/10 0,10€ with probability 1/10	
2€ with probability 10/10 1,60€ with probability 0/10	3,85€ with probability 10/10 0,10€ with probability 0/10	

Once you have completed the questionnaire, we will start the experiment.

<u>Task one</u> (partially adapted from Barr et al. 2018 and Krupka and Weber 2008)

You will receive a description of three situations. This description corre- sponds to situations in which one person, “Doctor X,” must decide how to act. You will be given a description of various possible actions Doctor X can choose to take in response to each situation.

After you receive the description of the situations, you will be asked to eval- uate each of the various possible actions Doctor X can choose to take for each situation. You must indicate, for each of the possible actions, whether taking that action would be “appropriate” or “inappropriate” on a scale of 1 to 4, where 1 means very inappropriate and 4 means very appropriate. We consider an action to be appropriate when you think is the “correct” thing to do in each specific situation.

In each of your responses, we would like you to answer as truthfully as possi- ble, based on your opinions of what constitutes appropriate or inappropriate action. To give you an idea of how the experiment will proceed, we will go through an example situation and show you how you will indicate your responses.

Example Situation

Doctor X is treating a patient who has been admitted in the hospital rehab block due to a compound fracture of his shoulder. Doctor X can choose 4 possible actions to take: referring the patient to a fifteen-minute magneto- therapy per day; referring the patient to a forty-minute magneto-therapy per day, referring the patient to a three-hour magneto-therapy per day; referring the patient to a seven-hour magneto-therapy per night.

The table below presents the list of the possible actions Doctor X can choose. For each of the actions, you would be asked to indicate whether you believe choosing that action is very inappropriate, somehow inappropriate, somehow appropriate, or very appropriate. To indicate your response, you would put a cross in the row corresponding to your belief about the degree of appropriateness.

	15-minute	40-minute	3-hour	7-hour

	magneto-therapy per day	magneto-therapy per day	magneto-therapy per day	magneto-therapy per night
1Very inappropriate				
2Somehow inappropriate				
3Somehow appropriate				
4Very appropriate				

If this was the situation for this study, you would consider each of the possible actions above and, for that action, indicate the extent to which you believe taking that action would be “appropriate” or “inappropriate”. Recall that by appropriate we mean action that you think is the “correct” thing to do in each specific situation.

For example, suppose you thought that referring the patient to a three-hour magneto-therapy per day was very inappropriate, referring the patient to a fifteen-minute magneto-therapy per day was somehow inappropriate, referring the patient to a three-hour magneto-therapy per day was somehow appropriate and referring the patient to a seven-hour magneto-therapy per night was very appropriate, then you would indicate your responses as follows:

	15-minute magneto-therapy per day	40-minute magneto-therapy per day	3-hour magneto-therapy per day	7-hour magneto-therapy per night
1 Very inappropriate			X	
2 Somehow inappropriate	X			
3 Somehow appropriate		X		
4 Very appropriate				X

After completing the table, you will be asked to indicate how much certainty do you feel about your choices on a five-point scale, where 5 stands for a great deal of certainty.

If you have any questions about this example situation or about how to indicate your responses, please raise your hand now and we will assist you privately.

You will next be given the description of three situations where Doctor X faces various possible actions. You will be given a paper with the description of the situations and a pen to write down your answer. After you read the description, you must consider the possible actions and indicate on the paper you receive how appropriate these are in a table similar to the one shown above in the example situation.

Once you have completed the task an experimenter will come to collect your paper.

Payment procedure

Once all the papers have been collected, a computer will randomly select one scenario and for that scenario one action Doctor X can choose. Your evaluation of this action will be compared with the response selected by the most people here today. If your evaluation coincides with the most frequently chosen option, you will receive 10 euros for this task, otherwise you will receive zero. For instance, imagine the example situation above was the actual situation and the possible action "Fifteen-minute magneto-therapy" was selected by the computer. If your evaluation had been "somehow inappropriate" then your task earnings would be 10 euros if this was the response selected by most other people in today's session and zero otherwise.

While the experimenters are calculating your total payoff, we ask you to complete a short, anonymous questionnaire. Please leave the questionnaire on your desk once you have completed it.

Scenarios

Now we present three different scenarios similar to the previous example. For each scenario, we propose four actions that Doctor X can take in response to that. For each of the actions, we ask you

to indicate whether you believe choosing that action is very inappropriate, somehow inappropriate, somehow appropriate, or very appropriate. Recall that by appropriate we mean action that you think is the "correct" thing to do in each specific situation.

First scenario

Doctor X deals with a 26-year-old patient who sustained a tear to the anterior cruciate ligament together with a meniscus injury. Doctor X has four alternate options: performing joint aspiration (arthrocentesis), prescribing cryotherapy and magneto-therapy applications, suggesting surgery treatment, recommending ice and rest. Here is the table reporting all the four available actions. For each of the actions you have to indicate whether you believe choosing that action is very inappropriate, somehow inappropriate, somehow appropriate, or very appropriate. Recall that by appropriate we mean action that you think is the "correct" thing to do in each specific situation. To indicate your response, you have to put a cross in the row corresponding to your belief about the degree of appropriateness for each of the actions which Doctor X can take. Remember that at the end of the experiment your evaluations will be compared with the most common answers provided today. You have a chance to earn additional money by matching the most frequently chosen option.

	Prescribing magneto -therapy and cryotherapy	Recommending ice and rest	Surgery treatment	Performing joint aspiration
1 Very inappropriate				
2 Somehow inappropriate				
3 Somehow appropriate				
4 Very appropriate				

How much certainty do you feel about your choice of optimum treatment? (on a scale 1 to five where 1 stands for no certainty and 5 for a great deal of certainty)

Do you want to know national guidelines content?

- Yes
- No

Do you want to change your previous answers?

- Yes
- No

Second scenario

Doctor X deals with a 28-year-old woman who suffers from breast fissures.

Doctor X has four alternate options: suggesting discontinuing nursing, suggesting discontinuing nursing and in the meanwhile prescribing protective creams, recommending correcting errors in latch, suggesting using pump. Here is the table reporting all the four available actions. For each of the actions you have to indicate whether you believe choosing that action is very inappropriate, somehow inappropriate, somehow appropriate, or very appropriate. Recall that by appropriate we mean action that you think is the ”correct” thing to do in each specific situation. To indicate your response, you have to put a cross in the row corresponding to your belief about the degree of appropriateness for each of the actions which Doctor X can take. Remember that at the end of the experiment your evaluations will be compared with the most common answers provided today. You have a chance to earn additional money by matching the most frequently chosen option.

	Suggesting discontinuing nursing	Suggesting discontinuing nursing and prescribing protective creams	Recommending correcting errors in latch	Suggesting using pump
1 Very inappropriate				
2 Somehow inappropriate				
3 Somehow appropriate				
4 Very appropriate				

How much certainty do you feel about your choice of optimum treatment? (on a scale 1 to five where 1 stands for no certainty and 5 for a great deal of certainty)

Do you want to know national guidelines content?

- Yes
- No

Do you want to change your previous answers?

- Yes
- No

Third scenario

Doctor X deals with a 54-year-old patient who has a cancer to the oral cavity

in the retromolar region which involves the pterygoid muscle. Doctor X has four alternate options: treating it with surgery and radiotherapy, treating it with chemotherapy, treating it with radiotherapy, treating it with surgery. Here is the table reporting all the four available actions. For each of the actions you have to indicate whether you believe choosing that action is very inappropriate, somehow inappropriate, somehow appropriate, or very appropriate. Recall that by appropriate we mean action that you think is the "correct" thing to do in each specific situation. To indicate your response, you have to put a cross in the row corresponding to your belief about the degree of appropriateness for each of the actions which Doctor X can take. Remember that at the end of the experiment your evaluations will be compared with the most common answers provided today. You have a chance to earn additional money by matching the most frequently chosen option.

	Treating it with surgery and radiotherapy	Treating it with chemotherapy	Treating it with radiotherapy	Treating it with surgery
1 Very inappropriate				
2 Somehow inappropriate				
3 Somehow appropriate				
4 Very appropriate				

How much certainty do you feel about your choice of optimum treatment? (on a scale 1 to five where 1 stands for no certainty and 5 for a great deal of certainty)

Do you want to know national guidelines content?

- Yes
- No

Do you want to change your previous answers?

- Yes
- No

Questionnaire

_____

Participant ID . . . . . .

Demographic Questionnaire

The following questions ask for some information about you. Please answer each question by placing a mark where appropriate or by writing a brief response.

When you finish a page you may go on to the next one.

In this booklet, unlike the others, if you have a question and you raise your hand, we can assist you privately in deciding what is the best answer for your situation.

1. **What is you age?**.....

2. **What is your gender?**

   - Male
   - Female

3. **What best describes your race or ethnicity (please select all that apply)?**

   - White
   - Black/African/America
   - Asian or Pacific Islander
   - Hispanic
   - Multiracial
   - Other

4. **What hospital/clinic do you work for?**.....

5. **How long have you been with that hospital/clinic?.....**

6. **In general, how satisfied are you with the hospital/clinic you work in?**

   - Very dissatisfied
   - Somehow dissatisfied
   - Somehow satisfied
   - Very satisfied

7. **Which kind of medical specialty do you have?**.....

8. **How many doctors work in your department?.....**

9. **How often do you ask for colleagues' advices to take decisions?**

- Never
- Seldom
- Often
- Very often

10. **How much does your colleague's opinion affect you if he agrees with you?**

- Not at all
- Slightly
- Somehow
- Very much

11. **How much does your colleague's opinion affect you if he does not agree with you?**

- Not at all
- Slightly
- Somehow
- Very much

12. **How often decisions about a patient are taken as a result of a team's valuation?**

- Never
- Seldom
- Often
- Very often

13. **How often do you share stuff with other colleagues (e.g. re- sults of diagnostic tests, outcomes of surgical procedures)?**

- Never
- Seldom
- Often
- Very often

14. When you work in team

- a. Any decision taken is the result of a shared opinion
- b. There is a member whose opinion weights the most

15. If you have answered a to question 13, please skip to question 16. **What is the role of the team member whose opinion weights the most?**.....

16. Do you have a whatsapp group with your colleagues where you discuss decisions to take?

- Yes
- No

17. How often do you take part in training courses?

- Never
- Seldom
- Often
- Very often

18. How often do you take part in brainstorming sessions?

- Never
- Seldom
- Often

- Very often

19. How often do you consult scientific journals?

- Never
- Seldom
- Often
- Very often

20. Do you carry out and public scientific research?

- Yes
- No

21. How important are national guidelines in your decisions?

- Very unimportant
- Somehow unimportant
- Somehow important
- Very important

22. Does your firm adopt internal guidelines?

- Yes
- No

23. If you have answered no to question 22, please skip to question 24.

How important are internal guidelines in your decisions?

- Very unimportant
- Somehow unimportant
- Somehow important

- Very important

24. Do you think that your colleagues follow guidelines while taking decisions

- Never
- Seldom
- Often
- Very often

25. What is your average opinion of your colleagues?

- Bad
- Mediocre
- Somehow good
- Very good

26. What is your average opinion of your colleagues with respect to yourself?

- Bad
- Mediocre
- Somehow good
- Very good

27. Do you have an insurance?

- Yes
- No

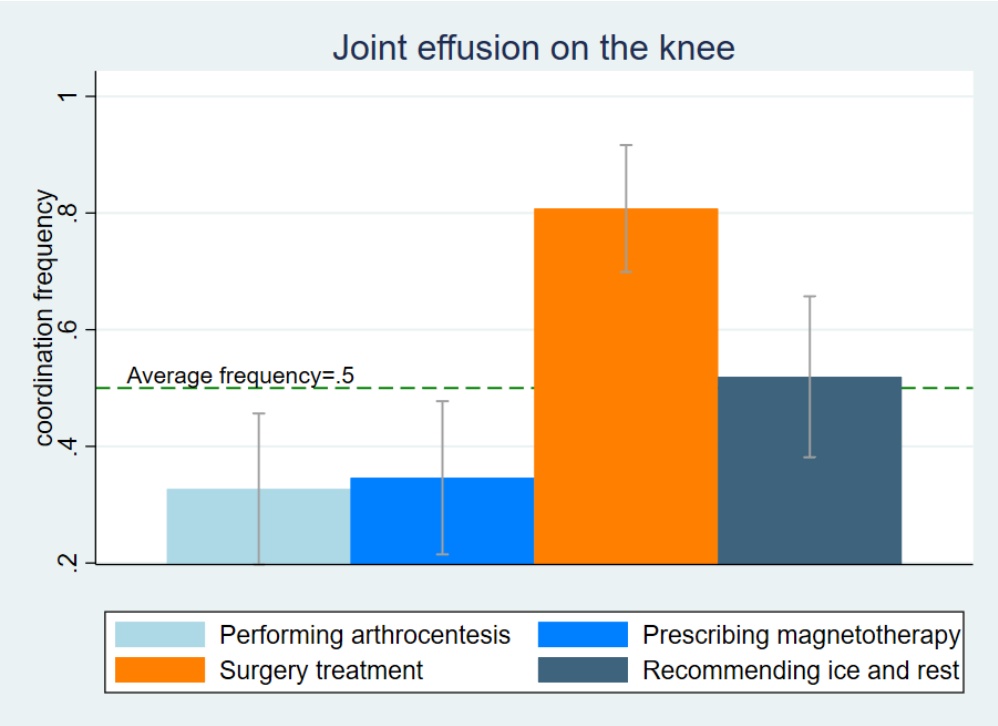
28. How do you judge your department in terms of coordination between colleagues?

- Bad
- Mediocre
- Somehow good
- Very good

Appendix B: Descriptive analysis by vignette

Figure 7 displays the average frequency of coordination for the meniscus injury scenario broken down by action. The orange bar represents the course of action suggested by national guidelines. The level of coordination for such action is significantly above the average of the vignette, which indicates that subjects recognize the guideline, despite the heterogeneity of the pool of physicians[1]. This assumption is also confirmed by the highest number of perfect appropriateness judgments reported for such action (81%). Ice and rest placed second for both coordination and level of appropriateness, being generally considered a more conservative therapy. In fact, any physician, regardless of his specialty, is likely to prescribe ice and rest which are said to be immediately effective after an injury (Borra et al., 2015).

Figure 1B: Average frequency of coordination in the first vignette with confidence intervals

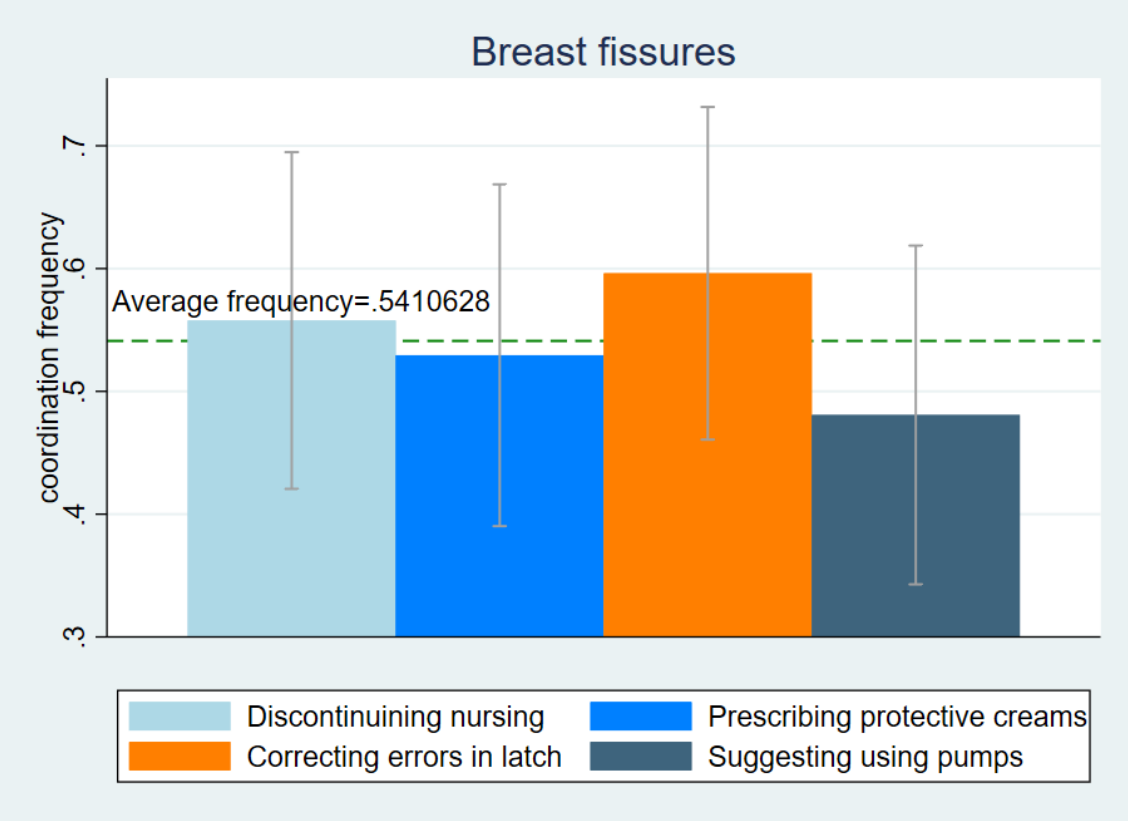


The highest level of coordination is observed in the second vignette whose average frequency is illustrated in Figure 8. Here physicians face a situation where a 28-year-old woman suffers from breast fissures. As usual, the orange bar corresponds to the treatment suggested by national guidelines

[1] Only 8 physicians were specialized in orthopaedics, which this vignette specifically refers to, while the other practitioners were specialized in various branches of medicine (i.e. pediatrics, anesthesia).

(Ministero della salute, 2019) which reconciles physicians, given the high level of coordination. On the other hand, the grey bar counterbalances the good level of coordination of the other actions.

Figure 2B: Average frequency of coordination in the second vignette with confidence intervals

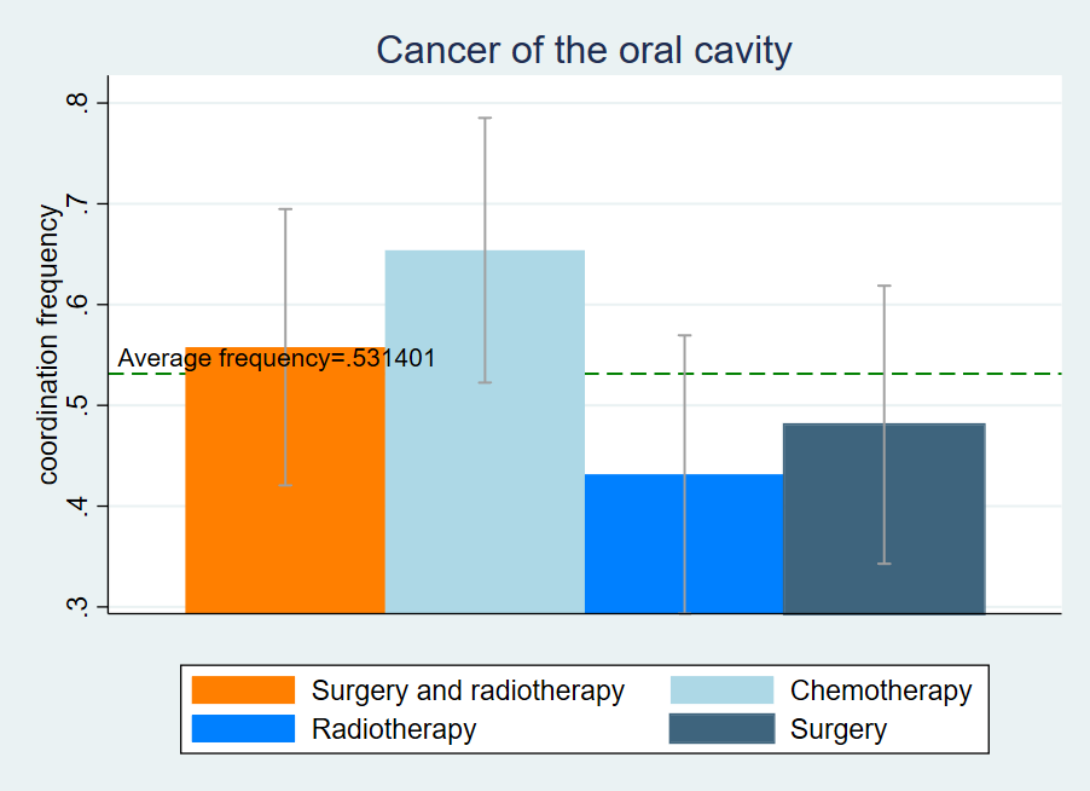


Focusing in detail on this vignette, a slight gender effect can be detected which may be responsible for the observed level of coordination. In fact, valuations of appropriateness vary between men and women across the actions as figure 9 reports, and this difference is significant according to the Mann-Whitney test (p-value<0.05). In such a scenario, it is very unlikely that men, not specialized in paediatrics or childcare and who did not witness similar episodes, understand the 28-year-old woman's pain. Thus, as the lower right box displays, men could assign the highest degree of appropriateness to a very unadvisable treatment such as 'suggesting using pump'.

Figure 3B: Gender difference in the second vignette

Finally, figure 10 shows the average frequency of coordination of the last vignette (i.e. portraying a patient suffering from an oral cavity cancer), where the highest level of coordination is reached in the second action which is ‘chemo- therapy’, reported to be ‘appropriate’ by 62% of the subjects.

Figure 4B: Average frequency of coordination in the third vignette

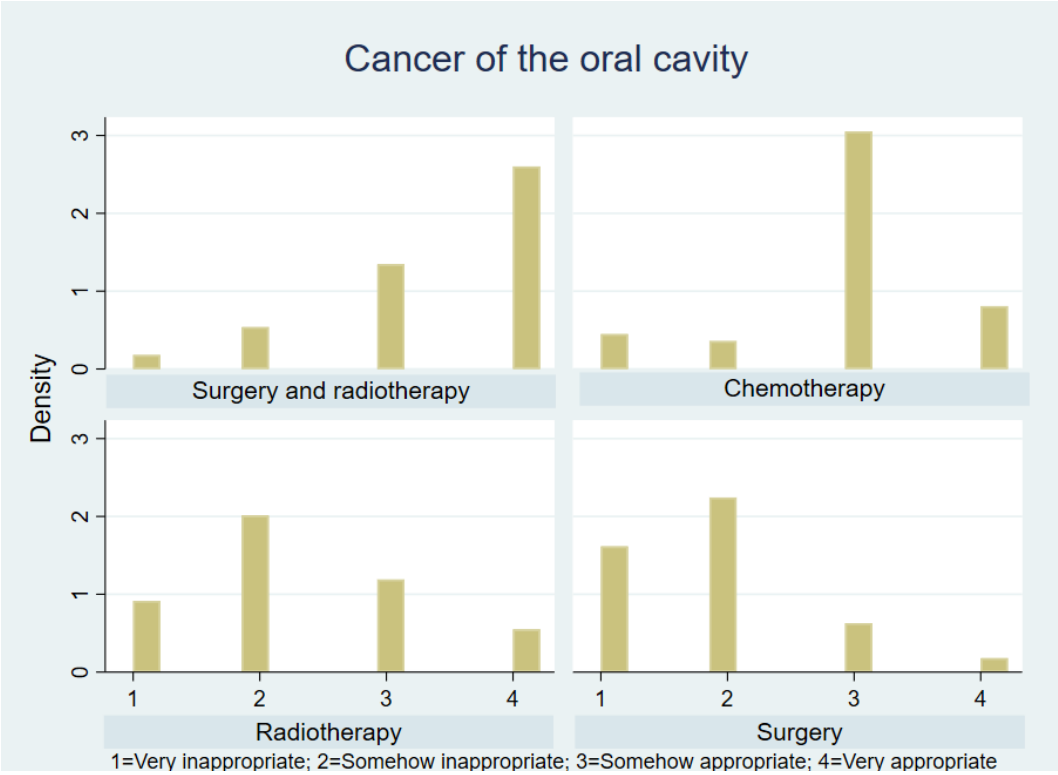


Such result has its foundation in recent literature on cancer treatments, where chemotherapy is expected to reduce the size of the tumor and to facilitate conservative treatment surgery[1]. Going further into details, figure 11[2] illustrates how each of the actions is assessed according to the appropriateness judgment. In particular, 'only surgery' stands out for the action with the highest likelihood of being judged as 'very inappropriate'. Similarly, the therapy corresponding to the guideline 'Surgery and radiotherapy' results in the highest probability of being judged as 'very appropriate'.

1 http://www.arquivosdeorl.org.br/additional/acervo_port.asp?id=228

[2] The same figures for the two remaining vignettes are provided by the end of this section.

Figure 5B: Density of the appropriateness judgment for action for the third vignette



Comparing the evaluations distribution across the actions, differences are significant according to the Kruskal Wallis test (p-value<0.001). The same results have been obtained for the two remaining vignettes (for the first vignette p-value< 0.001, for the second vignette p-value< 0.001[1]).

[1] Notice that in addition to the Kruskal Wallis test which compares all the actions at once, the ranksum test has been performed to compare the actions in twos for each vignette. Results are all significant at the 1% level.

Figure 6B: Density of the appropriateness judgment for action for the first vignette

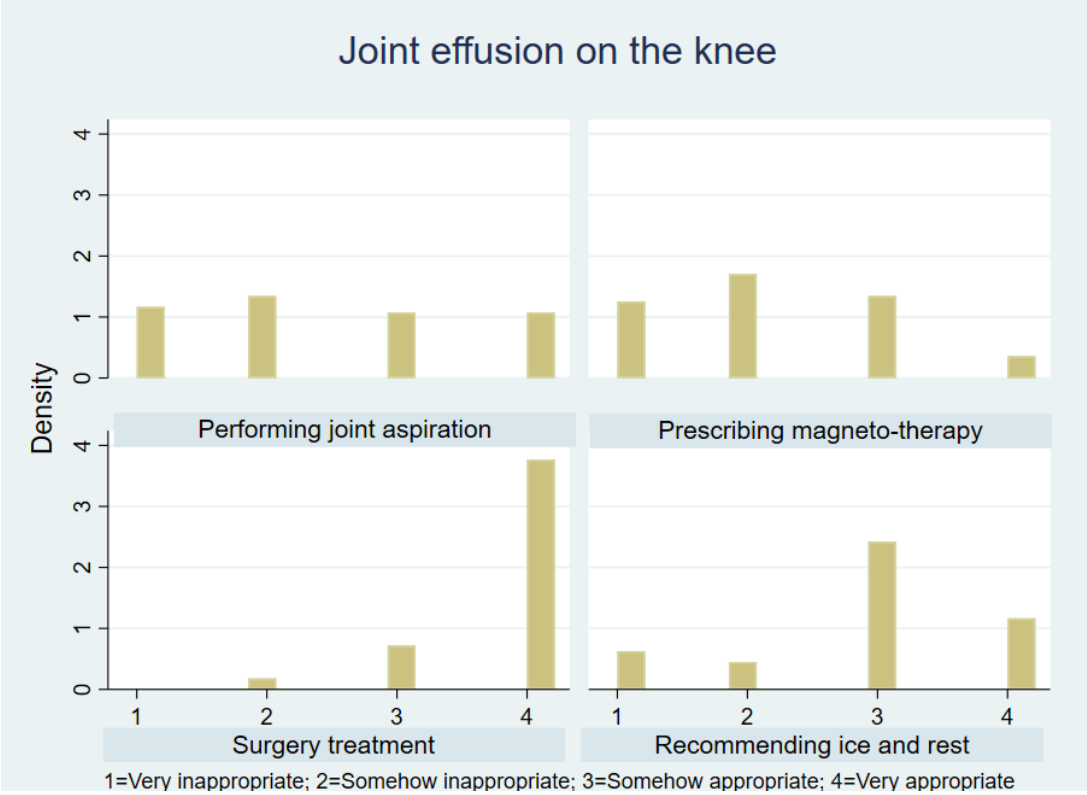
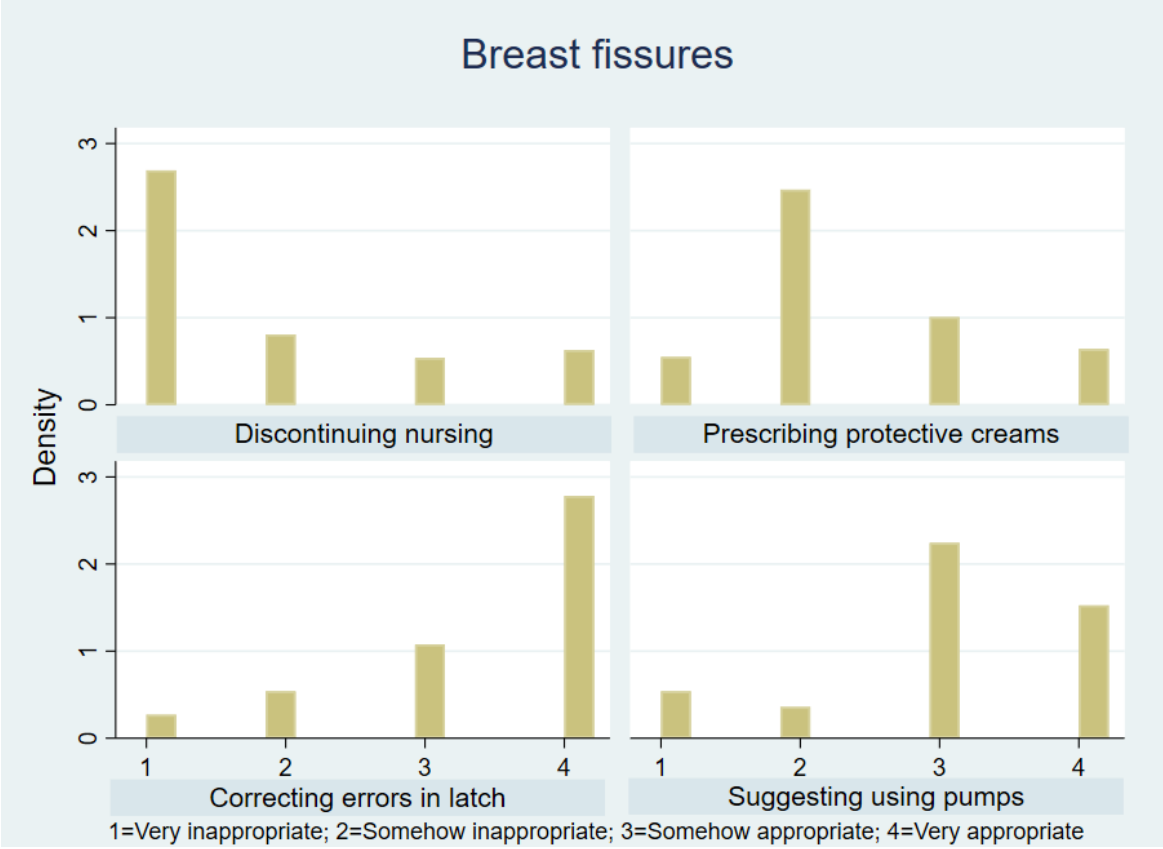
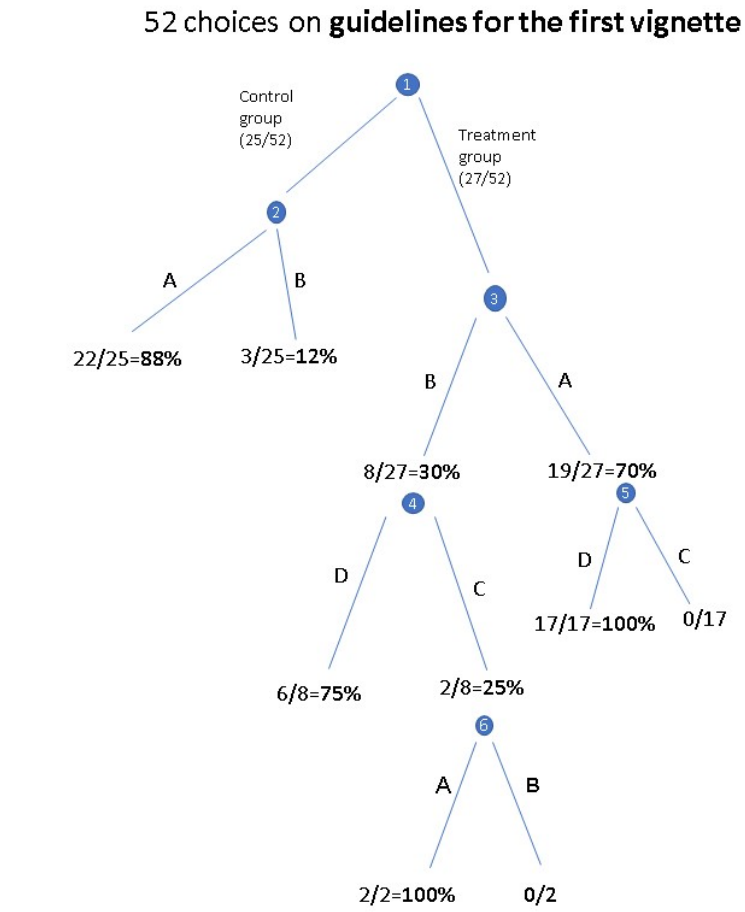


Figure 7B: Density of the appropriateness judgment for action for the second vignette



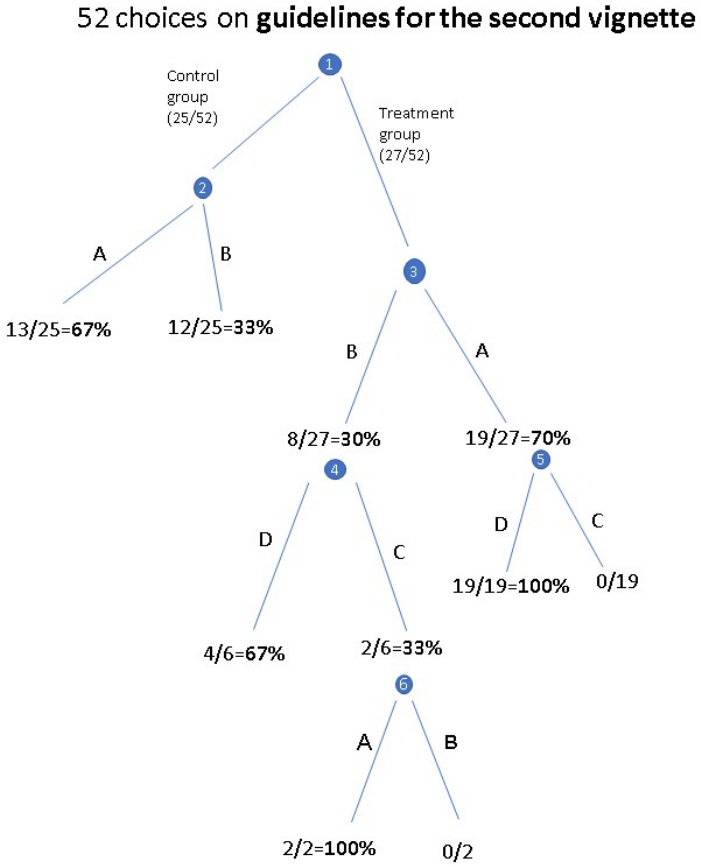
Appendix C

Figure 1C: Structure of the experiment for the first vignette



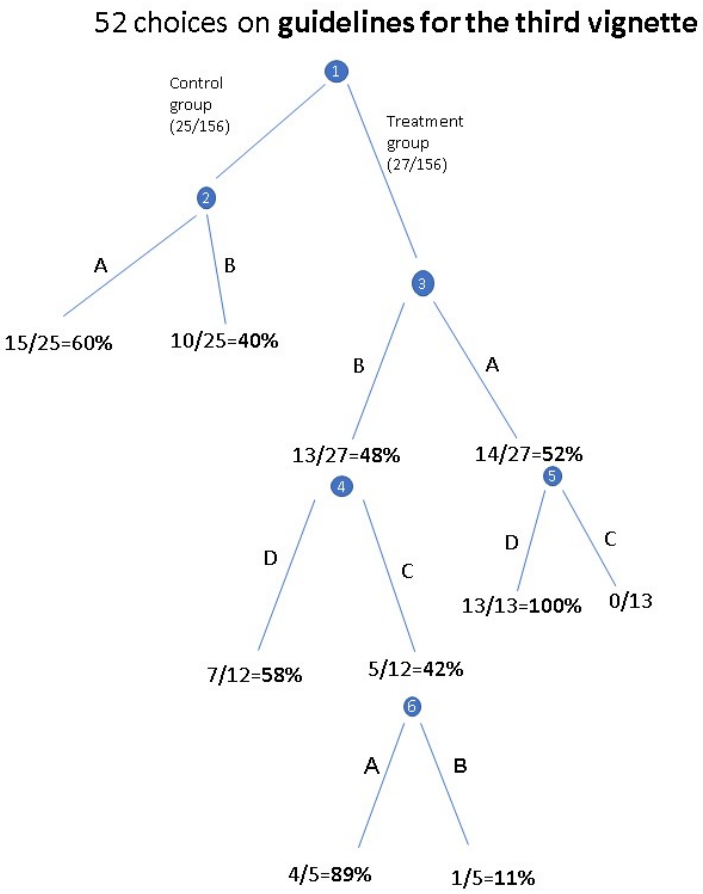
A= conformity with guidelines
B= nonconformity with guidelines
C=change
D=no change

Figure 2C: Structure of the experiment for the second vignette



A= conformity with guidelines
B= nonconformity with guidelines
C=change
D=no change

Figure 3C: Structure of the experiment for the third vignette



A= conformity with guidelines
B= nonconformity with guidelines
C=change
D=no change

Appendix D

Table 1D: Specialties' distribution across the sample

Specialty	**Percentage of the sample pool**
Gynaecologists	19%
Neonatologists	17%
Orthopedists	15%
Anaesthetists	15%
Surgeons	10%
Pathologists	6%
Otolaryngologists	4%
Hearth Surgeons	4%
Haematologists	4%
Diagnostic radiology	2%
Neurologists	2%
Internists	2%