INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS 84 (2015) 248–262

journal homepage: www.ijmijournal.com

A multiple-scenario assessment of the effect of a continuous-care, guideline-based decision support system on clinicians' compliance to clinical guidelines

Erez Shalom[a,*], *Yuval Shahar*[a], *Yisrael Parmet*[b], *Eitan Lunenfeld*[c]

[a] *Medical Informatics Research Center, Department of Information Systems Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel*
[b] *Department of Industrial Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel*
[c] *Department of Obstetrics and Gynecology, Soroka Medical Center, Faculty of Health Sciences, Ben-Gurion University of the Negev, Beer-Sheva, Israel*

ARTICLE INFO

Article history:
Received 3 September 2014
Received in revised form
8 January 2015
Accepted 9 January 2015

Keywords:
Medical informatics
Clinical decision support systems
Therapy
Computer-assisted
protocol-directed
Clinical guidelines
Quantitative evaluation
Preeclampsia/eclampsia

ABSTRACT

Objectives: To quantify the effect of a new continuous-care guideline (GL)-application engine, the Picard decision support system (DSS) engine, on the correctness and completeness of clinicians' decisions relative to an established clinical GL, and to assess the clinicians' attitudes towards a specific DSS.

Methods: Thirty-six clinicians, including residents at different training levels and board-certified specialists at an academic OB/GYN department that handles around 15,000 deliveries annually, agreed to evaluate our continuous-care guideline-based DSS and to perform a cross-over assessment of the effects of using our guideline-based DSS. We generated electronic patient records that realistically simulated the longitudinal course of six different clinical scenarios of the preeclampsia/eclampsia/toxemia (PET) GL, encompassing 60 different decision points in total. Each clinician managed three scenarios manually without the Picard DSS engine (Non-DSS mode) and three scenarios when assisted by the Picard DSS engine (DSS mode). The main measures in both modes were correctness and completeness of actions relative to the PET GL. Correctness was further decomposed into necessary and redundant actions, relative to the guideline and the actual patient data. At the end of the assessment, a questionnaire was administered to the clinicians to assess their perceptions regarding use of the DSS.

Results: With respect to completeness, the clinicians applied approximately 41% of the GL's recommended actions in the non-DSS mode. Completeness increased to the performance of approximately 93% of the guideline's recommended actions, when using the DSS mode. With respect to correctness, approximately 94.5% of the clinicians' decisions in the non-DSS mode were correct. However, these included 68% of the actions that were correct but redundant, given the patient's data (e.g., repeating tests that had been performed), and 27% of the actions, which were necessary in the context of the GL and of the given scenario. Only

* *Corresponding author*. Tel.: +97286477160.
E-mail address: erezsh@bgu.ac.il (E. Shalom).

http://dx.doi.org/10.1016/j.ijmedinf.2015.01.004

5.5% of the decisions were definite errors. In the DSS mode, 94% of the clinicians' decisions were correct, which included 3% that were correct but redundant, and 91% of the actions that were correct and necessary in the context of the GL and of the given scenario. Only 6% of the DSS-mode decisions were erroneous. The DSS was assessed by the clinicians as potentially useful.

Discussion: Support from the GL-based DSS led to uniformity in the quality of the decisions, regardless of the particular clinician, any particular clinical scenario, any particular decision point, or any decision type within the scenarios. Using the DSS dramatically enhances completeness (i.e., performance of guideline-based recommendations) and seems to prevent the performance of most of the redundant actions, but does not seem to affect the rate of performance of incorrect actions. The redundancy rate is enhanced by similar recent findings in recent studies. Clinicians mostly find this support to be potentially useful for their daily practice.

Conclusion: A continuous-care GL-based DSS, such as the Picard DSS engine, has the potential to prevent most errors of omission by ensuring uniformly high quality of clinical decision making (relative to a GL-based norm), due to the increased adherence (i.e., completeness) to the GL, and most of the errors of commission that increase therapy costs, by reducing the rate of redundant actions. However, to prevent clinical errors of commission, the DSS needs to be accompanied by additional modules, such as automated control of the quality of the physician's actual actions.

1. Introduction

1.1. *The need for an evaluation of the value of automated support to guideline-based care*

Clinical Guidelines (GLs) are one of the manifestations of the recent emphasis on *evidence-based medicine*, which tries to apply the best available evidence gained from scientific methods such as research studies, meta-analyses, and reviews, to support better clinical decision making [1]. Extensive evidence confirms that state-of-the-art GLs are a powerful method for standardization and uniform improvement of the quality of medical care and patient outcomes, often increasing patient survival rates while reducing the escalating costs of medical care [2–6].

Despite these useful findings, the level of adherence to GLs in daily practice is relatively low [2,3,7]. Therefore, care providers, health-care managers, and patients would benefit from automated support of GL-based care through the use of computerized GL-based decision support systems (DSSs). These systems include an electronic representation of GLs and support their automated dissemination and application at the point of care [8–11]. Over the past two decades, there have been a number of efforts to support the application of complex GLs in an automated fashion, typically providing static, one-time recommendations at several distinct points along the process of care [12–15]. However, none of these frameworks fully supports a continuous application of the GLs over significant stretches of time, providing recommendations when necessary, handling issues such as missing data in the electronic medical record (EMR), and also supporting a data-driven, asynchronous application (i.e., not just during a session with the care provider). Furthermore, there are very few large-scale assessments of the potential effect of using a GL-based DSS on the continuous application of a complex GL over time, especially assessments that use a meaningful number of clinicians. According to Isern [11] and others [15–17], there is a relative lack of research on the effects of GL application on the quality of clinical decisions by clinicians (i.e., their level of adherence to the GL's recommendations, and the percentage of their decisions that is correct according to the GL) and of "*in vivo*" evaluations in the area of GL application engines.

In a recent comprehensive methodological review summarizing the past decade's research regarding the life cycle of computerized GLs [18], Peleg noted that in general, only very few evaluations of GL-based DSSs have been made, since a full evaluation is often complicated by the fact that the DSS might allow clinicians to deviate from the GL's recommendations. Like others [11,15–17], Peleg concluded that additional research should be performed on the effect of GL-based DSSs on clinicians' behavior, in particular on improving their compliance to GLs.

1.2. *The objectives of this study*

The main objective of this study was to quantitatively evaluate the effects of a longitudinal GL-based DSS framework, designed for realistic, continuous use over multiple sessions, on the quality of medical decisions made by a group of physicians. We previously designed and implemented such a continuous-care framework (see Section 2.1). In the current study, we used a set of realistically simulated longitudinal medical records of patients, each presenting one of multiple clinical, GL-based scenarios that need management according to a known, well-established obstetrics GL for management of preeclampsia/eclampsia.

As a secondary objective, we wanted to assess the subjective perception of the clinicians regarding the GL-based decision-support framework.

2. Methods

2.1. The Picard framework for supporting guideline-based care

In order to support a longitudinal application of the GL (i.e., a continuous application of the GL over significant stretches of time, providing recommendations when necessary)., we developed and implemented a new GL-application architecture called the *Picard*[1] *framework* [20,21]. At the core of the Picard framework is a longitudinal, continuous-care, GL-application engine, the *Picard Decision Support* (*DSS*) *engine*. The Picard framework provides a set of application interfaces (APIs) to allow different types of client devices to interact with it, depending on the client's task type. For example, a GL-debugging task might include a client application that is specific for use by a knowledge engineer and that depicts the detailed state of the GL application engine at each step. A GL-simulation application might be a software component without any graphical interface, such as a simulation engine that simulates and tests over time the behavior of the GL-application engine on a simulated longitudinal patient database. GL-application tasks include, for example, a desktop decision-support application for the medical staff at the point of care, and an advisor for patients who are using their mobile phone as their computational infrastructure. The Picard DSS engine uses our *Digital Guideline Electronic Library* (DeGeL) architecture [22] to retrieve procedural GLs and specific knowledge items when applying a GL. Our IDAN temporal-abstraction mediator [23] is used to answer GL-specific queries regarding knowledge-based interpretations of the time-oriented clinical data. Currently, the Picard framework serves as the backbone of the MobiGuide European Union project [24], and is used to send alerts and personalized GL-based recommendations to chronic patients (based on data from bio sensors on the patients or at their home), through the patients' mobile devices, or to support the decisions of the patients' care providers at the ambulatory clinic or hospital.

2.2. An overview of the study design

Fig. 1 provides an overview of the study design; we present the details in Section 2.3. As the first step (Phase 1 in Fig. 1), we selected and then formalized the American College of Obstetrics and Gynecology (ACOG) preeclampsia/eclampsia/toxemia (PET) GLs [25]. Then, we defined six clinical GL-based scenarios and a total of 60 decision points across all scenarios (Phase 2 in Fig. 1). By using a new GL-specific simulation engine developed at our laboratory, we automatically generated a complete patient record for each scenario (Phase 3 in Fig. 1). We then performed a cross-over simulated clinical evaluation with the help of 36 clinicians, including both residents and board-certified specialists at an academic OB/GYN department that handles around 15,000 deliveries annually, who volunteered to assist us in the evaluation. Each clinician managed three scenarios without using the Picard DSS engine, i.e., in Non-DSS mode, and three scenarios using the Picard DSS engine, i.e., in DSS mode (Phase 4 in Fig. 1). After the experiment, a questionnaire was administered to the practitioners to assess their attitudes regarding the use of the Picard DSS engine, and their willingness to use it in the future. Finally, an expert assessed the correctness of the actions suggested by the physicians, relative to the ACOG PET GL, and their completeness, i.e., whether all of the ACOG GLs' relevant recommendations were followed (Phase 5 in Fig. 1).

2.3. Details of the experimental design

2.3.1. Choosing the clinical domain and guideline

We chose, together with the senior clinicians, a known, well-established, American College of Obstetrics and Gynecology (ACOG) PET GL [25]. Preeclampsia affects between 2% and 8% of pregnancies worldwide [26], and is a condition that occurs only during pregnancy. The diagnosis according to the GL is made through the combination of high blood pressure and protein in the urine, occurring after the 20th week of pregnancy. The GL manages both the mother and the fetus during different gestational periods, for several different scenarios of PET (e.g., mild preeclampsia, severe preeclampsia, and eclampsia). The target users of this GL are obstetricians. There was no official PET GL during the period of the experiment in the obstetrics ward in which we performed the experiment, and patients were treated according to textbook knowledge. The American College of Obstetrics and Gynecology (ACOG) PET GL was not formally defined at that time as the PET guideline adopted in this department, although all of the clinicians were familiar with it, and mentioned that they try to follow its spirit.

We found the PET GL to be appropriate for the current study because its semantics are reasonably complex, representing typical realistic GLs. In particular, the GL might be applied over several months, until a woman delivers her baby, and thus it is appropriate for examining the effect of a GL-based DSS on longitudinal GL application. Furthermore, we had previously specified several aspects of the PET GL as part of an experiment in GL specification [27].

We then formally and completely specified the PET GL in the ASBRU GL specification language [28] with the help of an obstetrics expert, and represented it within a GL repository that we had previously developed, the DeGeL digital GL library [22]. We used the multiple-step GL-specification methodology, as previously described by Shalom et al. [29].

2.3.2. Decomposing the guideline into clinical scenarios, steps, and decision points

In collaboration with a senior expert physician (different from the one who helped in the GL specification phase), we defined six mutually exclusive clinical *scenarios* that occur rather commonly when applying the PET GL. Table 1 shows the different GL scenarios chosen by the senior expert physician.

[1] Jean-Luc Picard was the captain of the USS Enterprise spaceship in the 1980s TV series *Star Trek: The Next Generation*, which was the sequel of the 1960s series *Star Trek*. In a metaphorical manner, the previously developed GL application framework, which was evaluated in this study, is called Picard because it is the next generation of "Spock"—the previous generation GL application tool of our lab [19]. The Picard guideline-based decision support engine assists whomever is the captain in charge of therapy—the physician, the nurse, or even the patient.

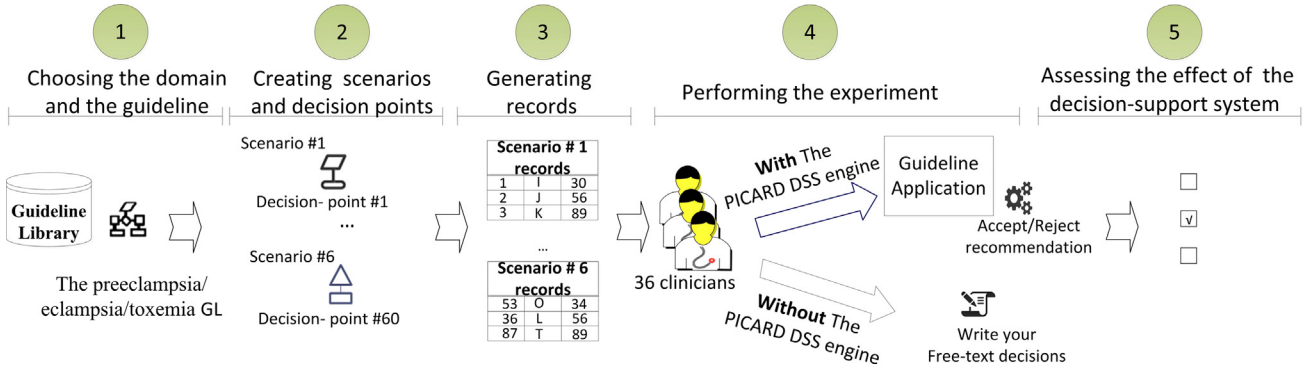

Fig. 1 – The overall study design.

To test the compliance of the clinicians, we further decomposed each scenario into several *steps*, each step taking place at a different point along the GL's timeline and being composed of several *decision-points*. Each scenario was presented to the clinicians as a series of single steps. As we shall see, in DSS mode the clinicians simply accepted or rejected the recommendations made by the DSS for each decision mode, possibly adding some free text; in the non-DSS mode, the clinicians had to provide their recommendations after each step in free text. (See Section 2.3.4 for how the scenario was re-aligned after each step, regardless of the clinician's responses.) Thus, each decision-point in our experiment represents an elementary decision unit in which, in the DSS-mode, the DSS can support a clinician's decision, or in non-DSS mode, we can test the clinician's compliance.

To classify the decision-point in a more refined fashion, we created six categories of decision-point types: "Order lab test", "Perform medical procedure", "Assert diagnosis", "Manage drug therapy", "Order a diagnostic imaging", and "Perform physical examination". Table 2 displays a partial example of the first scenario (numbered as 1), showing five of the decision-points into which it was decomposed. Each decision-point has a decision-point type. In total, there were 60 decision-points across all scenarios. The right column in Table 1 describes the number of decision-points in each scenario.

2.3.3. *Generating guideline-based simulated longitudinal medical records*

To simulate the required longitudinal data for each of the selected clinical scenarios, we generated for each of the scenarios one longitudinal medical record that included multiple transactions, using a specialized tool that we previously developed for GL-based generation of clinical scenarios and for scenario-based patient-record generation [30]. A set of transactions, representing multiple time points within a single longitudinal patient record, was generated for each GL scenario. In our simulation tool, this step can be performed using random generation of valid values, using the GL-based knowledge base (e.g., generating a systolic blood pressure that corresponds to the definition of "mild hypertension"), or can be performed manually by the knowledge engineer and the expert physician, to make sure that the records are clinically valid and consistent over time. We simulated on average 2300 transactions for each scenario.

2.3.4. *Performing the experiment: Presenting the scenarios to the clinicians in both modes*

Thirty-six physicians agreed to take part in the assessment of a simulated application of the GL: 24 residents and 12 board-certified specialists, all from an academic OB/GYN department. (We sent requests to all of 50 clinicians of the

Table 1 – Summary of the six clinical preeclapmsia/toxemia guideline scenarios used in the evaluation.

Scenario no.	Description	Gestational week of hospitalization	End of GL	Time duration	Number of decision points
1	Mild PET–delivery week 38	34.5	38.2	4 weeks	10
2	Mild PET–fetal heart rate monitoring non-reassuring	34.1	36.1	2 weeks	20
3	Severe PET–IUGR	33.1	35.1	2 weeks	5
4	Severe PET–magnesium toxicity	34.4	37.4	3 weeks	9
5	Severe PET–hypertension	35.1	35.3	2 days	5
6	Severe PET–eclampsia	36.1	36.3	2 days	11

Table 2 – Part of the decomposition of the first scenario (No. 1) into decision points of different types.

Step	Week	Time	Decision-point description	Decision-point type
1.A	34.5	7:00	1.1 Diagnosis is Mild PET	Assert a diagnosis
1.B	34.5	8:00	1.2 We suggest to hospitalize the patient	Perform a medical procedure
			1.3 We suggest to measure 24-h urine protein	Order a lab test
			1.4 We suggest to record fetal movements twice daily	Perform a medical procedure
	–	–	–	–
1.C	38	8:00	1.10 We suggest to deliver the baby	Perform a medical procedure

OB/GYN department; out of the 25 specialists, 12 agreed to assist in the experiment. Three additional specialists agreed to assist us in the different stages of the experimental evaluations. Out of the 25 residents, 24 agreed to assist us.)

We classified the clinicians into three groups, based on their level of training: (1) 1st year residents (9 clinicians); (2) Experienced residents—residents in the second or later residency year (but not board certified) (15 clinicians); (3) Specialists—all board certified (12 clinicians). All of the clinicians had at least textbook knowledge of handling PET, although they were not necessarily exposed to the specific GLs we used. All of them had previously used electronic medical record systems, but without DSS capabilities.

Since we had a limited number of clinicians, we did not want to have two separate groups—an experimental (DSS) group and a control (Non-DSS) group. In addition, due to the different levels of training among the clinicians, such a separation might have been non-homogenous. Thus, we decided to design the experiment as a cross-over study. Each of the clinicians, after a short introduction, was presented with three DSS scenarios and three Non-DSS scenarios (out of the six predefined clinical PET scenarios) within the same experimental session without any time limit.

For the experiment, we developed two versions of the same user interface for the two decision-support modes. The interfaces used by the two modes were colored differently: a yellow color signified the interface for the DSS mode, and a pink color signified the interface for the Non-DSS mode. In this fashion, we signaled to the clinicians that the two user interfaces represent decision-support modes that are different in nature. The colors also reminded the clinicians of which type of system they were using at each point.

At any point in time, the clinicians could open a simulated detailed longitudinal electronic medical record or a simulated fetal growth graph and examine them.

In the non-DSS mode, the text of the current scenario was shown to the clinician, and he was instructed to provide his free-text decisions (see Fig. 2).

In the DSS mode, the same text regarding the patients was presented, and the same electronic data were available, but in addition, several recommendations were presented to the user, and he was asked whether he agrees or disagrees. He could open an explanation dialog box that explains the recommendation. If he disagreed with the DSS recommendation, he had to provide an explanation. In addition, he could add additional free text if in his opinion some data were missing. (see Fig. 3).

(Note: Throughout this paper, the use of the terms "He" or "His" is only for convenience and does not specify any particular user gender).

Appendix A in the Supplementary materials includes a detailed example of the two user interfaces in the case of scenario 3.

Since we tried to simulate the real clinical settings as closely as possible, we developed a simulated electronic medical record (EMR) to show patient medical records for each of the scenarios of the experiment. This EMR included data such as history, lab tests, and risk factors. In addition, the EMR included a dynamic temporal representation of the fetal growth progress.

During the experiment, regardless of the clinician's answers (in either DSS or non-DSS mode), the next step in the scenario was presented, notifying the clinician that certain (in fact, the correct) decisions for the previous step were made. That is, for example, regardless of whether the clinician decided to hospitalize the patient or not, and assuming that hospitalization was the correct decision in the previous step, she was told: "*Mrs. Jones was hospitalized; her blood pressure on admission was . . .*". The scenario was aligned, if needed, with the identical one being presented to each clinician in both Non-DSS and DSS modes, and to make sure that each clinician was faced in each step of each scenario with precisely the same decision points.

To prevent a learning effect in our cross-over study, we made sure that the six scenarios were clinically quite different, and were essentially mutually exclusive with respect to their decision points. In this way, knowing the correct decisions regarding the previous steps (due to the alignment process explained above) did not materially affect making the decisions for the next steps, and no decision-point occurred twice within the same scenario or across different scenarios. (*Note:* The clinicians who performed the experiment were not aware of this conceptual decomposition into steps.)

To cater for a possible periodic effect, we defined 12 sequences of the six scenarios. Each sequence included three DSS scenarios and three Non-DSS scenarios, in the same order, except for a random order in the case of the first two, easier, scenarios, one of which was presented as a DSS scenario and the other as a non-DSS scenario. This arrangement made sure that each clinician experienced both the DSS mode and the non-DSS mode within the first two scenarios. Each of the 36 clinicians was assigned a single sequence. Each of the 12 sequences appeared, in total, three times. Altogether, each scenario was presented to the clinicians 36 times: 18 times in the DSS mode, and 18 times in the Non-DSS mode.

In addition, after the clinicians finished the experiment, they completed a questionnaire electronically in order to assess the perceived *usefulness* of the system by the clinicians based on a perceived usefulness and perceived ease of use questionnaire [31,32]. (The *usability* aspect was not tested in

Fig. 2 – An example of the Non-DSS mode user interface, colored pink in the original experiment. The textual scenario is presented at the top, and the clinician is instructed to provide his free-text management decisions in the bottom. On the right, the clinician can open the simulated EMR or the fetal growth chart. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

this study, since our intent in the current study was to ignore the particular interface design, and focus on the effect of the DSS on the clinicians' behavior.) The questionnaire was divided into two conceptual groups of questions, although they were presented as one set: Group A included questions about the general perception of the clinicians regarding the DSS. (e.g., "*Adding the computerized decision-support system to my current work environment would enable me to accomplish tasks more quickly*"). Group B included questions about the willingness of the clinician to use a DSS in the future. (e.g., "*In the future, I would like to use a similar computerized decision support in my work if it is offered to me* "). The clinicians were asked to grade each question between 1 (extremely likely) and 7 (extremely unlikely) (see Appendix B in the Supplementary material for the questionnaire questions and the user interface). Finally, an optional text-box was given to the clinicians to enter any comments, ideas, or suggestions about the experiment.

2.4. *The evaluation measures*

Since the current study did not measure any patient-oriented clinical outcomes, our evaluation used several process-oriented measures that assessed the compliance of the clinicians with respect to our gold standard GL.

Two measures were defined to assess the quality of a set of clinical decisions, with or without a DSS, relative to a given GL:

(1) *Completeness:* The percentage of relevant GL-based actions that was actually followed.
(2) *Correctness:* The percentage of the physician's actions that was correct, according to the relevant GL.

The Correctness and Completeness measures are relatively standard measures for quality assessment, such as when assessing compliance to certain policies. They are also similar to the Soundness and Completeness measures in algorithmic research; note that guidelines are essentially algorithms. Their semantics are also analogous to the standard Precision and Recall measures in the Information Retrieval area.

The completeness and correctness evaluation measures of a DSS mode scenario are calculated, in most cases, in a straightforward fashion: When the clinician adheres to the recommendation and accepts it, he receives a score of "1". If he disagrees, the expert working with us looked at his explanation (if provided) and his score was defined according to the explanation, if any, between 0 (not adhered to) and 1 (adhered to). The overall completeness and correctness are percentages, and thus range from 0 (no decision is correct, or no GL recommendation was followed) to 1 (all decisions were correct, or all GL recommendations were followed).

Table 3 shows an example of DSS completeness scoring: the Total DSS Completeness is $(1+1+0.5+0+0)/5 = 0.5$. That, is, the clinician followed half of the GL's recommendations. (For a formal mathematical definition of all evaluation measures, see the study's measures' complete specification [20].)

The Non-DSS evaluation was based on the text that the clinicians wrote. We evaluated the correctness and completeness of each portion of the text. *Correctness* refers to the percentage of the text the clinician wrote that is indeed correct according to the GL, and *completeness* refers to the percentage of the GL decision-points that is included by the text that the clinician wrote as his intention to follow. Table 4 shows an example of correctness scoring in a Non-DSS mode, in which the clinician wrote his proposed actions as text.

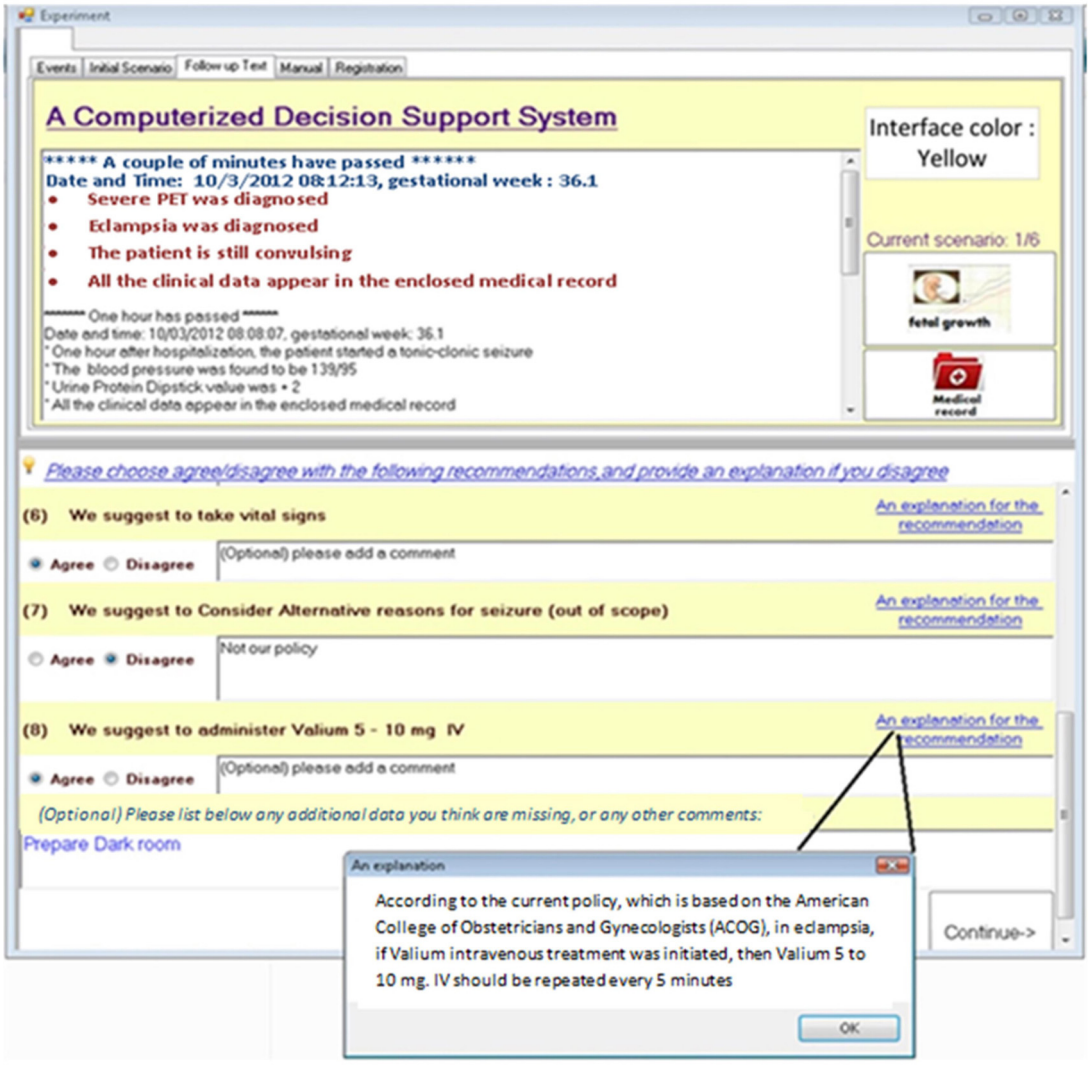

Fig. 3 – An example of the DSS mode user interface, colored yellow in the original experiment. The user is presented with several recommendations, and he is asked whether he agrees or disagrees with each of them. Here, the clinician added that a dark room should be prepared as well. See the text for a detailed description of the semantics of the interface. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3 – An example of scoring a guideline-based completness in the DSS mode.

The recommendation	Clinician's response	Clinician's comment	Score	Explanation for the score
Call for help	Agree	None	1	Clinician adheres
Consider alternative causes for eclampsia	Disagree	The etiology is preeclampsia unless other parameters or information is known such as epilepsy in the history	1	Clinician adheres
Administer 2 g of magnesium	Disagree	Elevated CCR. Consider 1 mg per hour	0.5	Gave 1 mg and not 2 mg
Administer valium 5 to 10 mg	Disagree	Should take magnesium	0	Incorrect explanation
Administer 1 g of magnesium	Disagree	In severe PET, no need to give loading dose. Tight follow up is enough	0	Incorrect explanation

Table 4 – An example of scoring the correctness of a series of actions in the Non-DSS mode.

Orders specified by users	Generally correct action	Redundantly correct action	Necessarily correct action
Physical examinations	1	1	0
Measure BP every hour	1	0	1
Measure proteinuria every 3 h	1	0	1
Measure CBC	1	1	0
Chemistry	1	1	0
PT	0	0	0
PTT	0	0	0
Fibrinogen every 6 h	0	0	0
Platelets	1	1	0
Creatinine	1	1	0
Urea	1	1	0
Liver enzymes	1	1	0
24 h Urinary protein collection	1	0	1

Going down to a finer resolution level, we differentiated between three types of correctness, all relative to a given GL: (1) *Generally* (in a context-free manner) *correct actions*, namely, the actions proposed by the clinician were in principle correct according to the GL, without considering the specific patient data that were presented to the clinician. General correctness can be further decomposed into two types of correctness: (1.A) *Correct* but *redundant* actions, namely, actions that were in principle correct according to the GL, but that were assessed as redundant in the current given context (either because the test should not be performed at this specific point in the GL's timeline, or because it should be performed, but its results were already provided in the scenario's text or in the EMR); and (1.B) *Correct* (context-sensitive) and *necessary* actions, i.e., the actions were justified and necessary in the context of the current GL scenario and of the actual patient data that were provided to the clinician (including the scenario's text and the EMR). For example, recent results of many laboratory tests were provided as part of the scenario's text or the EMR that we provided to the clinicians; thus, although it was in principle correct to perform these tests for a general patient according to the GL, it was not necessary (i.e., it was redundant) to do so in the particular scenario and data instance provided.

Calculating the correctness of the actions suggested by the clinician in Table 4, we find that the total number of items is 13, of which 3 were judged as correct by our expert, given this particular GL and scenario. Thus, the general GL-based correctness (without considering the context of this particular scenario) is 10/13, but the necessary (GL and scenario-based) correctness is only 3/13. Note that 3 of the 13 suggested actions, the PT, PTT, and Fibrinogen tests, were simply *incorrect*, in the sense that they were not even assessed as correct in general but redundant relative to the scenario and data supplied.

2.5. *The statistical analysis methods*

We performed several statistical tests to evaluate the contribution of using the DSS. We assumed that total completeness and correctness are two dependent variables, as they are bounded variables between zero and 1. Thus, we used a beta regression model with a logit link function for the mean response model; a log link function was fitted for the precision model [33,34]. This model is based on the assumption that the dependent variable is beta-distributed and that its mean is related to a set of regressors through a linear predictor with unknown coefficients and a link function; more details can found elsewhere [33].

Our full model included three factors: (1) DSS mode (DSS and Non-DSS), (2) Level of training (1st year resident, Experienced resident, specialist), and (3) Scenario (six different scenarios). We used the pseudo R^2 value (squared correlation of linear predictor and link-transformed response) to measure the overall goodness of fit of the model. To assess what are the important factors, we used a backward elimination algorithm; the final model is presented.

Unlike the total completeness measure, which was computed for the overall set of decision-points per scenario per clinician, when we analyzed the completeness measure for each decision-point type (e.g., "order lab test"), we increased our resolution and looked at the clinician's GL-based action per decision; therefore, our measured variable is binary, and a logistic regression with a logit link function [35] was used. As in the previous case, our full model included three factors; a backward elimination algorithm was used to reach the final model. (The completeness for each decision-point was also measured, as we show in the results, but did not go down to that low level of resolution in our statistical analysis, due to the small numbers involved.)

Finally, to measure that each question in the questionnaire was scored higher than 4, we used one sample *t*-test. In addition, we used Cronbach's α [36], which is a coefficient of internal consistency, to assess the reliability of the subjective-perceptions questionnaire. Cronbach's α was calculated for each of the two groups of questions, perception of the clinicians regarding the DSS, willingness of the clinician to use the DSS in the future, and for the overall questionnaire. The beta regression model was fitted using the DrichiletReg package in the R Language; the rest of the statistical analyses were performed using SPSS 20. In all of the statistical analyses, significance was set at $p < 0.05$.

3. Results

3.1. *Overall results*

The average time for a clinician to complete the experiment was 48 min: each Non-DSS scenario required an average of

9 ± 4.91 min, range 1.7 to 32.7 min, to be completed by the clinicians. A DSS scenario required 7.31 ± 4.4 min, range 2.1 to 21.9 min, to be completed. There were a total of 323 occurrences of opening the EMR in the Non-DSS mode and 215 occurrences of opening the EMR within the DSS mode, across all 36 clinicians and six scenarios. The mean number of occurrences of asking for an explanation per decision point, across all scenarios and all clinicians, was 4.6.

Table 5 summarizes the overall results across all scenarios and clinicians: when using the DSS mode, the clinicians made a total of 1069 decisions across all of the 60 decision points, including decisions found in the free text comments that they added. The clinicians added explanatory text to a total of 96 decision-points, across all clinicians and decisions; additional textual comments to the scenario were provided a total of 27 times, across all of the clinicians. All 1069 decisions were evaluated and tagged as being either necessarily correct relative to the PET guideline and the relevant scenario, redundantly correct, or incorrect (see Section 2).

All of the 1035 decisions accepted or rejected by the clinicians when using the DSS mode were assessed for completeness relative to the PET GL.

Note: There were in theory 36 × 60 = 1080 decision points; however, due to technical problems in the presentation of the 5th scenario, we discarded without any selection all of the 45 decisions (five decision points that were already presented to nine participants in the study) made within that scenario; otherwise, all scenarios were run as planned. (In the beginning of the experiment, we noted that we were using a slightly corrupt version of the knowledge base, with a typo in one of the items, and thus discarded all of the 5 decisions that were already made by 9 clinicians using the corrupt version, and fixed the typo in that item for the rest of the experiment.)

In the Non-DSS mode, the clinicians made 1643 decisions. All 1643 decisions were evaluated and tagged for necessary correctness, redundant correctness, or incorrectness.

All of the 1012 decisions potentially made or not made by the clinicians when using the Non-DSS mode were assessed for completeness relative to the PET GL by carefully examining the free text of their answers to each scenario. (*Note:* again, in theory, we could have had 1080 potential decision points; however, we had to discard the 5 decisions that were not made in DSS mode. This was relevant, according to the sequences design, to 10 clinicians; we therefore removed these 50 decisions from the evaluation. Furthermore, due to another technical problem, one decision was not presented at the end of one of the scenarios in the non-DSS mode to the 18 clinicians encountering it, removing another 18 decisions from the assessment.)In the case of judging correctness, we distinguished, as explained in the Methods, necessarily correct actions from correct, but redundant, actions, such as performing laboratory tests whose values appeared in the original case given to the clinicians, or within the longitudinal EMR. We shall now examine these results at a finer level of resolution.

3.2. *Mean completeness and correctness across all scenarios and clinicians*

Table 6 shows the mean completeness and general (context-free) correctness in both of the study's modes, and the range

Table 5 – Overall correctness and completeness of the clinicians' decisions, in the DSS and non-DSS modes, across all six scenarios and all 60 decision points.

DSS mode				Non-DSS mode			
Completeness	Necessary Correctness	Redundant correctness	Incorrect	Completeness	Necessary Correctness	Redundant correctness	Incorrect
93.71% (970/1035)	90.45% (967/1069)	3.18% (34/1069)	6.31% (68/1069)	49.80% (504/1012)	26.23% (431/1643)	68.22% (1121/1643)	5.53% (91/1643)

Table 6 – Mean completeness and general correctness across different scenarios and clinicians.

Measure	Mode	Dimension	Mean [of the means] ± Std (%)	Minimal value of the mean (%)	Maximal value of the mean (%)
General Completeness	DSS	All 6 scenarios	92.6 ± 4.5	88.88	100
		All 36 clinicians	93.45 ± 7.92	66.67	100
	Non-DSS	All 6 scenarios	41.22 ± 19.35	25	72.5
		All 36 clinicians	49.03 ± 18.58	17.24	84
General Correctness	DSS	All 6 scenarios	93.77 ± 4.57	88.17	100
		All 36 clinicians	93.35 ± 7.65	66.67	100
	Non-DSS	All 6 scenarios	94.22 ± 4.01	87.72	94.22
		All 36 clinicians	94.48 ± 6.74	70.58	100

of these means, calculated across all six scenarios and 36 clinicians.

Note that the mean value across the 6 scenarios, versus the mean value across the 36 clinicians, in each mode, is similar, but not quite the same, because the means were calculated from different values. Thus, for example, the mean DSS-mode completeness for each *scenario* (first row in Table 6) was calculated over all of the clinicians that encountered that scenario in DSS mode, and the overall scenario-completeness mean was computed as a simple (non-weighted) mean of the six scenario means; while the mean DSS-mode completeness for each *clinician*)second row in Table 6) was computed across all of the scenarios that the clinician had encountered, and the overall mean completeness for the 36 clinicians was computed as a simple (non-weighted) mean of the 36 clinician means. (As explained in Section 3.1, for technical reasons, there were some small differences in the number of decisions per clinician, and different scenarios were viewed by a somewhat different number of clinicians.)

After fitting the beta regression for the completeness with backward elimination, the final model included only the DSS, the scenario, and their interaction (pseudo $R^2 = 0.614$, overall $\chi^2_{11} = 211.7$, $p < 0.00001$). In the DSS mode, the average completeness level estimated (by the final beta regression model) over all scenarios was 92%, while in the Non-DSS mode it was estimated as 39% ($z = 16.48$, $p < 0.00001$). The level of training effect and its interactions were not included in the final model; hence, the level of total completeness of the three levels of training of clinicians is the same across all combinations of scenario (1 to 6) and mode of DSS (with and without DSS).

Table 7 shows the mean completeness, and the range of that mean, in both of the study's modes, for different levels of training, in both modes. (As in Table 6, note the two different points of view with slightly different values).

After fitting the beta regression for the "general correctness" with backward elimination, none of the examined factors and their second order interactions was found to be statistically significant ($p > 0.05$). However, after fitting the beta regression for "Necessary correctness" with backward elimination the final model included only the DSS, scenario, and their interaction (pseudo $R^2 = 0.62101$, overall $\chi^2_{11} = 249.75$, $p < 0.00001$). In the DSS mode, the average rate of the "Necessarily correct" decisions over all scenarios was estimated, using the beta regression model, as 91%, while in the Non-DSS mode it was estimated as only 35% ($z = 18.73$, $p < 0.00001$).

Regarding the "redundant" decisions rate, after fitting the beta regression with backward elimination, the final model included the DSS mode, the scenario, "level of training", and the interactions between scenario and mode, and scenario and level of training (pseudo $R^2 = 0.77$, overall $\chi^2_{23} = 310.77$, $p < .00001$). In the DSS mode, the average estimated "redundant" decisions rate over all scenarios was only 4%, while in the Non-DSS (manual) mode it was 59% ($z = 21.5$, $p < 0.00001$). The differences in redundant decision rates among different training levels varied within scenarios, but in an inconsistent manner.

Note that the overall (general) correctness was similar in both the DSS and the non-DSS modes. It was composed, however, of very different rates of necessary correctness and redundant correctness rates, with respect to the guideline, the clinical scenario, and the specific patient data provided to the clinicians.

3.3. *Completeness across decision-point types*

Table 8 shows the mean completeness values for all decision point types, across all of the clinicians and across all of the scenarios, in DSS mode and in non-DSS mode.

For example, when examining the results for the six scenarios, the mean completeness for the *Non-DSS* mode across all decision-point types, considering the scenarios dimension, ranged from 13.43 ± 7.8% (for the four scenarios in which there were decisions of type "manage drug therapy") to 83.33 ± 0% (in the case of the one scenario in which there were decisions of the "perform physical examinations" type, computed across all clinicians encountering decisions of that type for that scenario). The mean completeness for the *DSS* mode across all decision-point types, considering the scenarios dimension, ranged from 85.65 ± 0.65% to 1 ± 0%.

The "Manage drug therapy" decision-point type showed the highest level of improvement, from 12.26% in the Non-DSS mode, to 84.3% in the DSS mode, in the clinicians' dimension, and from 13.43% in the Non-DSS mode, to 88.19% in the DSS mode, in the scenarios' dimension ($z = 3.98$, $p < 0.000001$); the "Perform medical procedure" decision-point type showed a rather high improvement, from 38.65% in the Non-DSS mode, to 97% in the DSS mode in the clinicians' dimension, and from 33.64% in the Non-DSS mode, to 97.53% in the DSS mode, in the scenarios' dimension ($z = 13.5$, $p < 0.00001$). The smallest improvement occurred in the case of the "Perform physical examination" decision-point type, from 82.35% in the

Table 7 – Mean completeness levels across different levels of training in both modes across the six scenarios.

Level of training	Mode	Dimension	Mean [of the means] ± Std (%)	Minimal value of the mean (%)	Maximal value of the mean (%)
1st Year residents	DSS	All 6 scenarios	93.68 ± 7.74	80	100
		All 9 clinicians	93.89 ± 10.7	66.67	100
	Non-DSS	All 6 scenarios	43.3 ± 20.8	20	72.5
		All 9 clinicians	46.74 ± 19.39	20	72.72
Experienced residents	DSS	All 6 scenarios	92.78 ± 5.32	86.67	100
		All 15 clinicians	93.54 ± 6.3	80	1000
	Non-DSS	All 6 scenarios	40.95 ± 17.6	25	71
		All 15 clinicians	47.74 ± 19.8	17.24	84
Specialists	DSS	All 6 scenarios	94.39 ± 4.7	88.88	100
		All 12 clinicians	93.02 ± 8	75	100
	Non-DSS	All 6 scenarios	43.79 ± 22.67	26.67	73.12
		All 12 clinicians	52.68 ± 18.3	42.85	82.14

Non-DSS mode, to 100% in the DSS mode in the clinicians' dimension, and from 83.33% in the Non-DSS mode, to 100% in the DSS mode in the scenarios' dimension ($z = 2.168$, $p = 0.03$). The "Order lab test" decision type also showed a relatively small improvement, from 79.27% in the Non-DSS mode, to 98% in the DSS mode, in the clinicians' dimension, and from 74.81% in the Non-DSS mode, to 98.52% in the DSS mode in the scenarios' dimension ($z = 3.9$, $p < 0.00001$).

After fitting the logistic regression with backward elimination, the level of training effect and its interactions were found to be non-significant ($p > 0.05$); thus, the final model included the main three effects: scenario, DSS mode, and decision-point type and their interactions (overall $\chi^2_{31} = 438.5$, $p < 0.000001$). In general, the DSS increased the completeness for all decision point types, although the amount of improvement from the non-DSS to the DSS mode was different according to the decision type $\left(\chi^2_5 = 19.075,\ p = 0.002\right)$.

3.4. *Completeness per decision point*

Fig. 4 shows the 60 decision-points and the completeness distribution in the Non-DSS and in the DSS modes. The mean completeness of the decision-points (each decision point being represented by the mean completeness across all of the clinicians making that decision) for Non-DSS mode was 48.63 ± 29.5%, ranging from 0% to 94.4%. The mean completeness in the DSS mode was 93.98 ± 10.09%, ranging from 44.44% to 100%.

In general, we found a high completeness rate with low variability in the DSS mode, in contrast to the low scores

Table 8 – The completeness for all decision point types, across all of the clinicians and all of the scenarios, in DSS mode and in non-DSS mode.

Decision Point Type	Dimension	Mode	Mean [of the means] ± Std (%)	Minimal value of the mean (%)	Maximal value of the mean (%)
Manage drug therapy	All 4 scenarios	DSS	88.19 ± 12.08	72.22	100
		Non-DSS	13.43 ± 7.8	8.33	25
	All 36 clinicians	DSS	84.3 ± 23.71	0	100
		Non-DSS	12.26 ± 16.49	0	50
Perform medical procedure	All 6 scenarios	DSS	97.53 ± 3.3	91.67	100
		Non-DSS	33.64 ± 22.97	0	55.56
	All 36 clinicians	DSS	97 ± 10.7	40	100
		Non-DSS	38.65 ± 22.32	0	100
Assert a diagnosis	All 5 scenarios	DSS	93.33 ± 4.65	88.89	100
		Non-DSS	55.28 ± 22.14	37.5	88.89
	All 36 clinicians	DSS	91.67 ± 20.41	0	100
		Non-DSS	55.75 ± 31.29	0	100
Order a diagnostic imaging	All 2 scenarios	DSS	85.65 ± 0.65	86.11	85.19
		Non-DSS	53.24 ± 24.22	36.11	70.37
	All 36 clinicians	DSS	85.64 ± 26.09	0	100
		Non-DSS	53. 24 ± 47.77	0	100
Order lab test	All 3 scenarios	DSS	98.52 ± 2.57	95.56	100
		Non-DSS	74.81 ± 20.41	53.7	94.44
	All 36 clinicians	DSS	98 ± 5.2	72	100
		Non-DSS	79.27 ± 29.63	0	100
Perform Physical examination	1 Scenario	DSS	1 ± 0	100	100
		Non-DSS	83.33 ± 0	83.33	83.33
	All 18 clinicians	DSS	1 ± 0	100	100
		Non-DSS	82.35 ± 39.29	0	100

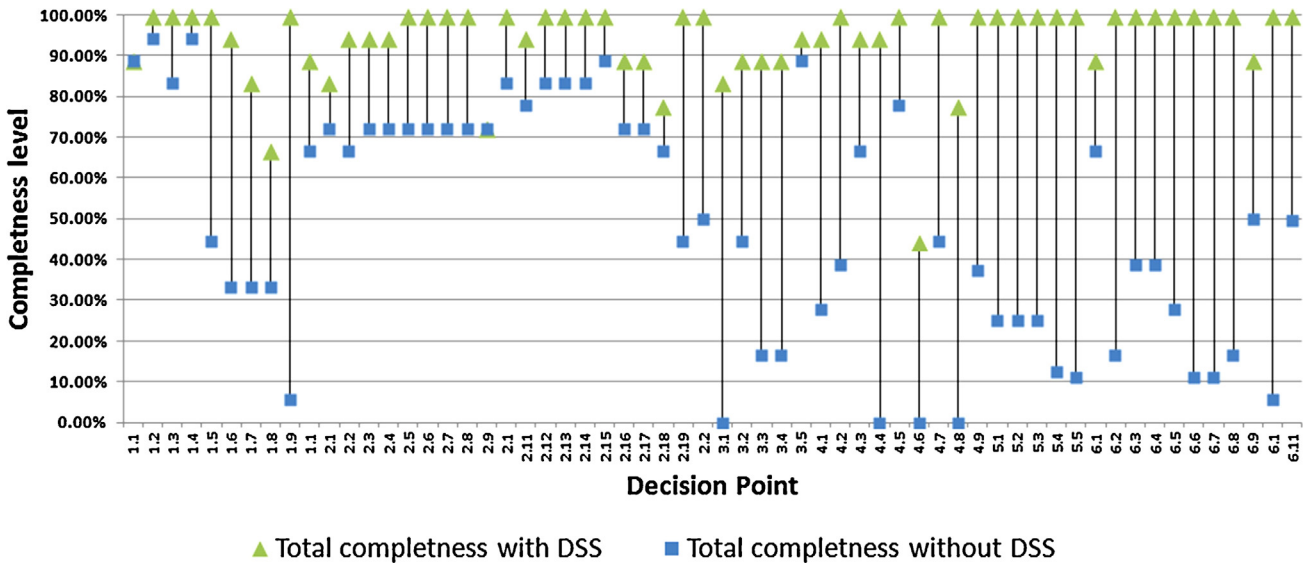

Fig. 4 – The mean completeness level of all decision points across all clinicians. Blue squares denote mean completeness for each decision point in the non-DSS mode; yellow triangles denote mean completeness for the same decision point in the DSS mode. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of completeness and high variability in the Non-DSS mode. Several decision-points had a low completeness rate in both modes. For example, decision-point 4.6 ("Administer 1 mg of magnesium") was a disagreement point: 0% (none) of the clinicians complied with the GL in the Non-DSS mode, but even in the DSS mode, only 44% were compliant, i.e., even after being provided with the DSS recommendation (and presumably, not agreeing to it in 56% of the cases). (See Section 4 for discussion about this result.)

3.5. *The subjective perspective*

For group A questions (general perception of the DSS), the mean score given by the clinicians was 5.42 ± 1.22 out of a maximal score of 7, and for group B (willingness to use or assess the DSS in the future), the mean score was 5.59 ± 1.39. All question scores were found to be significantly high ($p < 0.005$) in comparison with 4 (the expected value under the assumption of indifference) using a *t*-test. In addition, with respect to reliability, group A (general perception of DSS) had a Cronbach's α of 0.933, Group B (willingness to use DSS in the future) had a Cronbach's α of 0.88, and the overall questionnaire had a Cronbach's α of 0.95. We did not use the answers to questions 7 to 12 (scenario-specific assessment), since most of the clinicians, although answering the questions, mentioned informally that they were too detailed to answer meaningfully.

Finally, only 11 clinicians added free text comments about the overall experiment. Most of them pointed out that using the DSS is important to make better decisions; however, it is also important to leave sufficient room for the physicians to make the final decision, and to write their reservations, if any, to learn about frequent disagreements with the GL. Several suggestions were made regarding the need to deploy a more usable graphical interface in an actual clinical system.

4. Summary and discussion

In this study, we showed that using the DSS increased the clinicians' completeness (i.e., compliance with the guideline's recommendations) dramatically, *regardless of the level of training*.

Furthermore, it is interesting to note the completeness and correctness measures in the non-DSS mode (presumably, what might be considered as the baseline measures) were similar across different levels of training. A possible explanation is that the PET management is a common practice in an academic OB/GYN ward that handles approximately 15,000 deliveries annually. It is also possible that junior residents are trained by experienced ones, and thus display the same strengths and weaknesses.

In general, the total completeness across all decision-points for Non-DSS mode was $48.63 \pm 29.5\%$. The total completeness when using the DSS increased dramatically to $93.98 \pm 10.09\%$. Thus, the variance in completeness among different decisions decreased significantly, leading to a more uniform quality level, which is a major objective of quality-enhancement programs in any clinical organization.

The greatest improvements in completeness were in decisions of type "therapeutic procedures" (e.g., manage drug therapy or perform medical procedure). This result seems to imply that the DSS is mostly needed when the clinician needs to treat the patient, rather than for making a diagnosis or ordering a lab test.

These results seem quite meaningful for healthcare organizations. By ensuring more compliant decisions relative to evidence-based GLs, one might potentially improve the clinical outcomes in clinical settings and reduce direct and indirect costs (e.g., manpower and complications), as previously demonstrated in a neurological domain [37].

In spite of the high completeness scores in the DSS mode, some of the clinicians disagreed with several specific GL-based recommendations. We suggest, based on examining the affiliation of the clinicians, that this phenomenon was due, in our particular study, to a difference in the local clinical culture between two Obstetrics wards in the same medical center (for example, clinicians from one of the wards stopped administration of magnesium after toxicity to the drug was suggested, while clinicians from the other ward simply reduced the dose to 1 mg as a first step). This suggests that compliance levels to the DSS recommendations may identify local differences in GL application that would benefit from discussion and resolution between several local groups providing care within the same healthcare institution. In addition, a clinical manager can define a threshold for non-compliance to any decision-point, for example 90%. Thus, all extremely non-compliant decision-points can be reported, assessed, and addressed.

Regarding the actual clinical errors committed by clinicians, out of all decisions made by them, the 5.5% clinical error rate when not using the DSS was similar to the 6.3% error rate when using the DSS (note that an error in the DSS mode signifies a recommendation that was not accepted, and for which the free-text justification was not judged as plausible; or any incorrect recommendation added in the free text of any clinical step within a scenario). In other words, total (general) correctness was similar in both modes: around 94% to 95%.

This lack of difference between the DSS and non-DSS modes regarding clinical error rate is highly meaningful with respect to the quality of care when using a DSS. It means that although using the DSS can certainly and dramatically prevent errors of *omission*, which according to the U.S. Department of Health and Human Services report were responsible for $44B of the cost all adverse events [38], by reminding clinicians of what to do (manifested in the dramatic increase in completeness rate when using the DSS), it apparently cannot prevent errors of *commission*, i.e., additional (not mentioned by the DSS) incorrect actions that the clinicians decide to take. It might well be that to prevent errors of commission, additional, sophisticated mechanisms need to be put in place to monitor which actions were actually taken by the clinicians (even those that are seemingly outside of the guideline's immediate scope) and determine whether any of them contradict the spirit of the guideline.

However, not being able to prevent *clinical* errors of commission does not mean that we cannot significantly prevent other types of errors. In our study, although relative high correctness of approximately 94% in both modes, we found that 68% of the overall manual (Non-DSS) decisions were correct, but redundant. Although not clinically harmful, the main effect of redundancy when treating actual patients would be to increase expenses unnecessarily (for example, by ordering tests that were already ordered, or that are not justified by the GL at this particular point in time, or by suggesting unnecessary procedures), and possibly even inconvenience the patients or compromise their safety. Findings that are very similar to ours were discovered recently in a Johns Hopkins study, in which a 66% redundancy was discovered in cardiac biomarker tests involving troponin; an effective policy including training, education, and changes in the patient order-entry system saved the hospital $1.25 M after a year [39]. In our study, only 3% of such redundant decisions were found in the free text added by the clinicians when using the DSS mode, suggesting that when clinicians are faced with accepting or declining a set of specific recommendations, they are not likely to come up with additional, redundant actions. Nevertheless, it might well be argued that a higher rate of redundancy would have been found if we more strongly encouraged clinicians to list all of the actions that they would take in the real world, in addition to following (or not) the DSS recommendations. We suspect, however, that the number of correct but redundant actions will always be much lower when using a DSS, but that hypothesis remains to be verified "in vivo", in an actual clinical trial.

With respect to the subjective assessments, the results of the study suggest that the clinicians seemed to find the DSS to be potentially useful for their daily practice, particularly when managing the more complex cases. However, as studies by Bindels et al. had shown, such an enthusiasm does not guarantee real-world adoption of the DSS by the clinicians [40,41].

5. Limitations

The current study was performed in a simulated environment, using a set of realistically simulated longitudinal clinical records, but a clinical study at the point of care is still necessary. The study also included only 36 clinicians as assessors of the Picard framework's effects on their potential actions in the simulated environment, and a crossover evaluation. Although a larger number of clinicians than the numbers mentioned in most previous studies, we would like in the future to considerably increase the number of clinicians assisting us in assessing the system.

Designing the user interface was out of the scope of the current study; however, instead of a placeholder for textual content in the Non-DSS mode, in the future one might use structured lists, for, e.g., drugs or lab tests, or the interface of a real EMR software as used in the relevant clinical setting, to mimic the natural working environment of the clinicians.

6. Conclusions

Given our rather generic methodology, the results suggest that the Picard DSS engine ensures a high quality of clinical decision making (relative to a guideline-based norm), and is potentially beneficial for chronic patient monitoring using other GLs, and in particular, for care within a continuous, longitudinal time frame.

Altogether, the results imply that using a DSS dramatically increases the completeness across *all* scenarios and *all* decision types, especially in the therapeutic decisions, i.e., when managing medication therapies or performing a

Summary points

What was already known on this topic

- The level of adherence to clinical guidelines in daily practice is relatively low; therefore automated support of guideline-based care is beneficial.
- Existing guideline-based automated application frameworks do not fully support a continuous application of the GLs over significant stretches of time.
- There is need for a rigorous evaluation of the specific effect of automated support to guideline-based care on the quality of clinical decisions, and on clinicians' behavior, in particular with respect to improving their compliance to GLs.

What this study added to our knowledge

- Guideline-based decision support systems such as the Picard DSS engine have the potential to prevent errors of *omission* by ensuring a uniformly high quality of *completeness* in clinical decision making (relative to a GL-based gold standard), due to the increased adherence to the GL's recommendations, and reduction of redundant actions.
- Guideline-based decision support systems such as the Picard DSS engine also have the potential for preventing certain errors of *commission*, in the sense of significantly reducing the performance of *correct*, but *redundant*, actions.
- A GL-based DSS does not necessarily reduce the rate of actual clinical errors, i.e., *incorrect* actions. In order to prevent *clinical* errors of *commission*, a GL-based DSS needs to be companioned with additional modules, such as a quality-control system that monitors the clinician's actual actions, even those that are outside of the guideline's scope.

medical procedure, and less in the case of the diagnostic procedures. Correctness, however, is affected only in the sense of considerably reducing the rate of redundant actions, but not through the reduction in incorrect clinical actions introduced by the clinicians. Clinicians find the DSS to be potentially useful for their daily practice.

Conflict of interest statement

None.

Author contributions

Erez Shalom planned and conducted the research, developed the methods used in this study, performed the data collection and analyses, wrote the first draft of the manuscript and rewrote new drafts based on input from co-authors. Yuval Shahar planned the research, suggested computational methodologies, monitored the research progress as an advisor, and provided critical revisions of the manuscript. Yisrael Parmet designed part of the study, performed most of the statistical analyses, and provided input on manuscript drafts. Eitan Lunenefeld planned the research, assisted in the medical knowledge engineering process, recruited the physicians and provided critical revisions of the manuscript. All authors read and approved the final manuscript.

Acknowledgments

We would like to acknowledge all of the clinicians at the Obstetrics and Gynecology Division of the Soroka Medical Center, Beer-Sheva, who assisted us in evaluating the Picard framework. The Israel Ministry of Science Office, and the Israel National Institute for Health Policy and Health Services Research, provided part of the funding for Dr. Shalom's research. Dr. Shalom and Prof. Shahar were partially supported by the EU MobiGuide 7th Framework project (FP7-287811).

Appendices A and B. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ijmedinf.2015.01.004.

REFERENCES

[1] S. Timmermans, A. Mauck, Health Aff. (Millwood) 24 (2005) 18–28.
[2] T. Rotter, L. Kinsman, E. James, et al., The effects of clinical pathways on professional practice, patient outcomes, length of stay, and hospital costs: Cochrane systematic review and meta-analysis, Eval. Health Prof. 14 (2012) 3–27.
[3] J.M. Grimshaw, I.T. Russell, Effect of clinical guidelines on medical practice: a systematic review of rigorous evaluations, Lancet 15 (1993) 1317–1322.
[4] J. Grimshaw, M. Eccles, Is evidence-based implementation of evidence-based care possible? Med. J. Aust. 180 (2004) S50–S51.
[5] G. Micieli, A. Cavallini, S. Quaglini, Guideline Application for Decision Making in Ischemic Stroke (GLADIS) Study Group, Guideline compliance improves stroke outcome: a preliminary study in 4 districts in the Italian region of Lombardia, Stroke 33 (2002) 1341–1347.
[6] S. Quaglini, P. Ciccarese, G. Micieli, et al., Guideline Application for Decision Making in Ischemic Stroke (GLADIS) Study Group, Stud. Health Technol. Inf. 101 (2004) 75–87.
[7] M.D. Cabana, C.S. Rand, N.R. Powe, et al., Why don't physicians follow clinical practice guidelines? A framework for improvement, JAMA 282 (1999) 1458–1465.
[8] P. Elkin, M. Peleg, R. Lacson, et al., Toward the standardization of electronic guidelines, MD Comput. 17 (2000) 39–44.
[9] F. Rutten, W. Brouwer, L. Niessen, Practice guidelines based on clinical and economic evidence, indispensable tools in future market oriented health care, Eur. J. Health Econ. 6 (2005) 91–93.
[10] D. Isern, D. Sánchez, A. Moreno, HeCaSe2: a multi-agent ontology-driven guideline enactment engine, in: Proceedings of Multi-agent Systems and Applications V. Proceedings of Fifth International Central and Eastern

European Conference on Multi-agent Systems (CEEMAS 2007), Vol. 4696 of Lecture Notes on Artificial Intelligence, Springer, Berlin, Heidelberg, 2007, pp. 322–324.
[11] D. Isern, A. Moreno, Computer-based execution of clinical guidelines: a review, Int. J. Med. Inf. 99 (2008) 787–808.
[12] P.A. De Clercq, A. Hasman, J.A. Blom, et al., Design and implementation of a framework to support the development of clinical guidelines, Int. J. Med. Inf. 64 (2001) 285–318.
[13] M. Peleg, S.W. Tu, J. Bury, et al., Comparing computer-interpretable guideline models: a case-study approach, J. Am. Med. Inf. Assoc. 10 (2003) 52–68.
[14] D. Wang, M. Peleg, S.W. Tu, et al., Representation primitives process models and patient data in computer-interpretable clinical practice guidelines: a literature review of guideline representation models, Int. J. Med. Inf. 68 (2002) 59–70.
[15] A. Latoszek-Berendsen, H. Tange, H.J. Herik, van den, et al., From clinical practice guidelines to computer-interpretable guidelines. A literature overview, Methods Inf. Med. 49 (2010) 550–570.
[16] P. Gooch, A. Roudsari, Computerization of workflows, guidelines, and care pathways: a review of implementation challenges for process-oriented health information systems, J. Am. Med. Inf. Assoc. 18 (2011) 738–748.
[17] V. Patkar, D. Acosta, T. Davidson, et al., Using computerised decision support to improve compliance of cancer multidisciplinary meetings with evidence-based guidance, BMJ Open 2 (2012) e000439, http://dx.doi.org/10.1136/bmjopen-2011-000439.
[18] M. Peleg, Computer-interpretable clinical guidelines: a methodological review, J. Biomed. Inf. 46 (4) (2013) 744–763.
[19] O. Young, Y. Shahar, Y. Liel, et al., Runtime application of hybrid-asbru clinical guidelines, J. Biomed. Inf. 40 (2007) 507–526.
[20] E. Shalom, A Multi-Dimensional Framework for Realistic Application of Clinical Guidelines, Ben-Gurion University of the Negev, Beer-Sheva, Israel, 2013 (PhD Diss.).
[21] E. Shalom, I. Friedman, Y Shahar, et al., Towards a realistic clinical-guidelines application framework: desiderata, applications, and lessons learned, in: LNCS 7738: Process Support and Knowledge Representation in Healthcare, 2013, pp. 56–70.
[22] Y. Shahar, O. Young, E. Shalom, et al., A hybrid, multiple-ontology framework for specification and retrieval of clinical guidelines, J. Biomed. Inf. 37 (2004) 325–344.
[23] D. Boaz, Y. Shahar, A framework for distributed mediation of temporal-abstraction queries to clinical databases, Artif. Intell. Med. 34 (2005) 3–24.
[24] M. Peleg, Y. Shahar, S. Quaglini, Making healthcare more accessible, better, faster, and cheaper: the MobiGuide Project, Eur. J. ePractice: Issue Mobile eHealth 20 (2014) 5–20.
[25] ACOG Committee on Practice Bulletins-Obstetrics, ACOG practice bulletin: diagnosis and management of pre-eclampsia and eclampsia: number 33, January 2002, Obstet. Gynecol. 99 (2002) 159–167.
[26] World Health Organization, WHO Recommendations for Prevention and Treatment of Pre-eclampsia and Eclampsia, WHO, Geneva, 2011.
[27] A. Hatsek, Y. Shahar, M. Taieb-Maimon, et al., A scalable architecture for incremental specification and maintenance of procedural decision-support knowledge, Open Med. Inform. J. 4 (2010) 255–277.
[28] Y. Shahar, S. Miksch, P. Johnson, The Asgaard project: a task-specific framework for the application and critiquing of time-oriented clinical guidelines, Artif. Intell. Med. 14 (1998) 29–51.
[29] E. Shalom, Y. Shahar, M. Taieb-Maimon, et al., A quantitative evaluation of a methodology for collaborative specification of clinical guidelines at multiple representation levels, J. BioMed. Inform. 41 (2008) 889–903.
[30] I. Friedman, E. Shalom, Y. Shahar, Evaluation of guideline-application engines by longitudinal simulation: a position paper, in: Proceedings of the Third International Workshop on Knowledge Representation for Health Care 2011 (KR4HC'11), Bled, Slovenia, 2011.
[31] F.D. Davis, R.P. Bagozzi, P.R. Warshaw, User acceptance of computer technology: a comparison of two theoretical models, Manage. Sci. 35 (1989) 982–1003.
[32] F.D. Davis, Perceived usefulness, perceived ease of use, and user acceptance of information technology, MIS Q. 13 (1989) 319–340.
[33] S.L.P. Ferrari, F. Cribari-Neto, Beta regression for modeling rates and proportions, J. Appl. Stat. 31 (2004) 799–815.
[34] F. Cribari-Neto, A. Zeileis, Beta regression in R, J. Stat. Softw. 34 (2012) 1–24.
[35] A. Agresti, Categorical Data Analysis, third ed., John Wiley & Sons, New York, NY, 2013.
[36] L.J. Cronbach, Coefficient alpha and the internal structure of tests, Psychometrika 16 (1951) 297–334.
[37] S. Qualigini, P. Ciccarese, G. Micieli, et al., Non-compliance with guidelines: motivations and consequences in a case study, in: Proceedings of the Symposium on Computerized Guidelines and Protocols, 2004, pp. 75–87.
[38] D.R. Levinson, I. General, Adverse Events in Hospitals: National Incidence among Medicare Beneficiaries, Department of Health and Human Services Office of the Inspector General, Washington, DC, 2010.
[39] M.R. Larochelle, A.M. Knight, H. Pantle, S. Riedel, J.C. Trost, Reducing excess cardiac biomarker testing at an Academic Medical Center, J. Gen. Int. Med. 29 (11) (2014) 1468–1474.
[40] R. Bindels, A. Hasman, J.W.J. Van Wersch, J. Talmon, R.A.G. Winkens, Evaluation of an automated test ordering and feedback system for general practitioners in daily practice, Int. J. Med. Inf. 73 (2004) 705–712.
[41] R1 Bindels, A. Hasman, A.D. Kester, J.L. Talmon, P.A. De Clercq, R.A. Winkens, The efficacy of an automated feedback system for general practitioners, Inf. Prim. Care 11 (2) (2003) 69–74.