Richters *et al.*
Large-scale Assessments in Education (2023) 11:3
https://doi.org/10.1186/s40536-023-00151-1

Large-scale Assessments
in Education

RESEARCH **Open Access**

Who is on the right track? Behavior-based prediction of diagnostic success in a collaborative diagnostic reasoning simulation

Constanze Richters[1,2*], Matthias Stadler[1], Anika Radkowitsch[5], Ralf Schmidmaier[2,3], Martin R. Fischer[2,4] and Frank Fischer[1,2]

*Correspondence:
constanze.richters@psy.lmu.de

[1] Department of Psychology, Ludwig-Maximilians-Universität München, Munich, Germany
[2] Munich Center of the Learning Sciences (MCLS), Ludwig-Maximilians-Universität München, Munich, Germany
[3] Medizinische Klinik und Poliklinik IV, University Hospital, Ludwig-Maximilians-Universität München, Munich, Germany
[4] Institute of Medical Education, University Hospital, Ludwig-Maximilians-Universität München, Munich, Germany
[5] Leibniz Institute for Science and Mathematics Education, Kiel, Germany

Abstract

Background: Making accurate diagnoses in teams requires complex collaborative diagnostic reasoning skills, which require extensive training. In this study, we investigated broad content-independent behavioral indicators of diagnostic accuracy and checked whether and how quickly diagnostic accuracy could be predicted from these behavioral indicators when they were displayed in a collaborative diagnostic reasoning simulation.

Methods: A total of 73 medical students and 25 physicians were asked to diagnose patient cases in a medical training simulation with the help of an agent-based radiologist. Log files were automatically coded for collaborative diagnostic activities (CDAs; i.e., evidence generation, sharing and eliciting of evidence and hypotheses, drawing conclusions). These codes were transformed into bigrams that contained information about the time spent on and transitions between CDAs. Support vector machines with linear kernels, random forests, and gradient boosting machines were trained to classify whether a diagnostician could provide the correct diagnosis on the basis of the CDAs.

Results: All algorithms performed well in predicting diagnostic accuracy in the training and testing phases. Yet, the random forest was selected as the final model because of its better performance (kappa = .40) in the testing phase. The model predicted diagnostic success with higher precision than it predicted diagnostic failure (sensitivity = .90; specificity = .46). A reliable prediction of diagnostic success was possible after about two thirds of the median time spent on the diagnostic task. Most important for the prediction of diagnostic accuracy was the time spent on certain individual activities, such as evidence generation (typical for accurate diagnoses), and collaborative activities, such as sharing and eliciting evidence (typical for inaccurate diagnoses).

Conclusions: This study advances the understanding of differences in the collaborative diagnostic reasoning processes of successful and unsuccessful diagnosticians. Taking time to generate evidence at the beginning of the diagnostic task can help build an initial adequate representation of the diagnostic case that prestructures subsequent collaborative activities and is crucial for making accurate diagnoses. This information could be used to provide adaptive process-based feedback on whether learners are on the right diagnostic track. Moreover, early instructional support in a diagnostic training task might help diagnosticians improve such individual diagnostic activities and prepare for effective collaboration. In addition, the ability to identify successful diagnosticians even before task completion might help adjust task difficulty to learners in real time.

Keywords: Simulations, Collaborative diagnostic reasoning processes, Learning process analysis, Medical education, Logfile analysis, Supervised machine learning

Introduction

Training in collaborative diagnostic reasoning is important across various domains in higher education because, in practice, diagnosticians often work together in teams (e.g., in medical consultations, classrooms, scientific laboratories, therapeutical supervision, or industrial engineering). Previous research on collaborative problem solving (e.g., Graesser et al., 2018) has highlighted the need for training in collaboration skills, which form a key competence of the twenty-first century. For example, in order to assess a student's learning status or to diagnose a patient's health problem accurately, teachers or physicians, respectively, must be able to generate, elicit, and share evidence as well as come up with and share hypotheses and draw conclusions (so-called *collaborative diagnostic activities* [CDAs]; Fischer et al., 2014; Radkowitsch et al., 2022). The improvement of such complex skills is related to a constant increase in learners' current *zone of proximal development* (Vygotsky, 1978), which describes what learners are currently not able to solve on their own but could certainly solve with external help. Thus, for optimal learning outcomes, there is a need for learning environments that include problem-solving tasks that are slightly more difficult than what learners can already solve independently (Roosevelt, 2008).

Simulations are often used to train complex skills. They enable standardized repetitions of individual learning steps and deliberate practice (Ericsson, 2004) and training in rarely occurring or critical real-life situations (e.g., rare or deadly diseases). There is evidence that simulations are particularly effective when the embedded instructional support is adaptive (Chernikova et al., 2020). However, properly and immediately adjusting the appropriate instructional support to learners' individual needs represents a challenge for instructional designers and educators. Moreover, being able to identify at what point in time learners can already solve the task without additional support might also be helpful for removing or fading out (Pea, 2004) instructional support that might even hinder learning (Kalyuga et al., 2003). One starting point for such an adjustment involves using machine learning to analyze learners' behavior on the basis of process data that are recorded and stored by the computer system (e.g., log files). Previous studies have demonstrated that analyzing learners' behavior can help identify how learners approach certain problems (Griffin and Care, 2015) and can aid the understanding of specific misconceptions that arise in the learning process (e.g., Stadler et al., 2019). Earlier analyses showed that specific actions in the learning environment were associated with task completion success (Cirigliano et al., 2020). Thus, assessing behavioral indicators of diagnostic reasoning skills (e.g., CDAs) and relating them to the diagnostic outcome can provide insights into whether learners currently have adequate or inadequate representations of the diagnostic problem. For instance, such behavioral indictors may be beneficial for assessing whether a patient's relevant signs and symptoms are adequately interpreted (Charlin et al., 2012). If a learner's performance can be predicted before the diagnostic task is completed, instructors may be able to take early action to improve learning outcomes. The information obtained from the analysis of CDAs could provide a promising starting point for performance-based individualized instructional support and could make a positive contribution to effective diagnostic training.

Collaborative diagnostic reasoning as a complex skill

The process of diagnosing can be considered the "goal-oriented collection and interpretation of case-specific or problem-specific information to reduce uncertainty" (Heitzmann et al., 2019, p. 4) to be able to make professional decisions. Specific diagnostic situations require planned or initiated actions based on observations of and information about the problem to meet the diagnostic goal. Building on the conceptual framework of scientific reasoning and argumentation (Fischer et al., 2014), Heitzmann et al. (2019) defined such actions as *epistemic diagnostic activities*, which consist of, for example, evidence generation, evidence evaluation, hypothesis generation, and drawing conclusions (see also Klahr & Dunbar, 1988). These activities are grouped into a framework but cannot be placed in a fixed general sequence or order. According to Fischer et al. (2014), *evidence generation* refers to generating evidence in favor of or against a claim. Next, *evidence evaluation* is aimed at assessing "the degree to which a certain piece of evidence supports a claim or theory" (Fischer et al., 2014, p. 34). *Hypothesis generation* refers to the process by which students frame possible answers to the question, hereby deriving them from plausible models, available theoretical frameworks, or empirical evidence that they have access to. Finally, in *drawing conclusions*, students integrate different pieces of evidence "by weighing every single piece according to the method by which it was generated and by the rules and criteria of the discipline" (Fischer et al., 2014, p. 35).

To ensure high diagnostic quality, practicing scientists, physicians, psychologists, teachers, and engineers often need to diagnose in teams. Collaborative diagnostic reasoning (and, more generally, collaborative problem solving) has some advantages over individual reasoning, such as dividing labor according to individual professions, different perspectives, and knowledge bases (OECD, 2017), plus higher diagnostic accuracy (Tschan et al., 2009). However, existing research has demonstrated that students often lack collaborative skills (e.g., Hall & Buzzwell, 2012; O'Neill et al., 2013; Pauli et al., 2008) and that practitioners lack collaborative diagnostic reasoning skills (e.g., physicians; Tschan et al., 2009). By extending Fischer et al.'s (2014) framework of individual diagnostic activities to collaborative contexts, Radkowitsch and colleagues (2022) recently defined CDAs in their model of collaborative diagnostic reasoning. This model describes the diagnostic reasoning processes of two diagnosticians with different knowledge backgrounds. In doing so, Radkowitsch and colleagues (2022) distinguished individual activities from social or collaborative activities, namely, sharing, elicitation, negotiation, and coordination. The model can also be viewed as an integration and extension of Liu et al.'s (2015) collaborative problem-solving framework and Klahr and Dunbar's (1988) scientific discovery as dual search (SDDS) model. More precisely, the collaborative diagnostic reasoning model combines individual and collaborative activities and integrates them into CDAs referred to as *eliciting*, *sharing*, *negotiating*, and *coordinating evidence* as well as *hypotheses* (Radkowitsch et al., 2022). During the diagnostic reasoning process, these activities help diagnosticians construct and maintain a shared conception of a problem (Roschelle & Teasley, 1995). The quality of CDAs is assumed to be crucial for the success of the collaboration (Radkowitsch et al., 2022).

Using process data analysis for individualized learning support in the context of simulation-based complex skills training

To foster complex skills (e.g., collaborative diagnostic reasoning), simulations have been established in various domains in higher education. Flight simulators have been used in pilot training for many years (Landriscina, 2012) just as surgical simulations are common in the medical context (Al-Kadi & Donnon, 2013). Standardized training in simulations has different advantages over training in real-world scenarios. First, simulations can reduce the complexity of a situation while offering learners the opportunity to apply their knowledge to specific cases in standardized settings (Grossman et al., 2009). Second, simulations enable repetitive deliberate practice, which has been considered to be crucial for acquiring professional expertise (Ericsson, 2004). Third, unlike real-life scenarios, simulations enable training while ensuring ethical safety regarding mental or physical human conditions (Gegenfurtner et al., 2014; Grossman et al., 2009). Useful real-learning situations are often either rare (e.g., disruptive patient behavior) or too critical (e.g., amniotic fluid examination) to be used for training purposes. In real life, failure or complications would have serious unacceptable consequences (Ziv et al., 2003). A large number of primary studies and several meta-analyses have yielded positive effects of simulation-based learning and have provided recommendations for their implementation (e.g., Chernikova et al., 2020; Cook et al., 2013).

However, despite their potential, the effective use of simulations in training, especially in the field of collaborative diagnostic reasoning, remains challenging. To enhance highly effective learning that is based on complex and challenging problems, additional instructional support is often important (e.g., Hmelo-Silver et al., 2007). Instructional support is considered to be particularly effective when it is adapted to learners' individual needs (i.e., microlevel; e.g., Plass & Pawar, 2020). Dynamic assessment that can be realized by measuring learners' current performance in the problem-solving process (performance-based adaptation; e.g., VanLehn, 2011) can provide an adaptive basis for instructional support. One way to dynamically assess learners' performance is to analyze learners' behavior. This allows researchers to identify processes that are related to arriving at a successful solution to the problem (Griffin & Care, 2015) and to understand misconceptions in the learning process (e.g., Stadler et al., 2019). Compared with looking at only the summative outcome measure of a learning process, considering the learning process itself also offers the advantage of identifying subtler differences among learners that might not be reflected in the outcome measure (Stadler et al., 2020). To foster collaborative diagnostic reasoning skills, it might be useful to detect whether learners are currently leaning toward a correct or incorrect diagnosis—which is related to whether they have adequate or inadequate representations of the patient's problem—by predicting diagnostic accuracy. Following the hierarchical model of clinical reasoning processes (MOT; Charlin et al., 2012), which depicts the complex process of clinical reasoning as a network, these cognitive representations of the patient's problem evolve and change as the diagnostic reasoning process unfolds.

In recent years, interest in predicting learners' performance with machine learning has increased considerably (e.g., Baker & Inventado, 2014; Hilbert et al., 2021). For instance, previous studies have predicted learners' performance to identify those at risk of failing a course (e.g., Tomasevic et al., 2020) or to support an intervention (e.g., San Pedro et al., 2013). The data for such an assessment can be collected automatically in real time while the learners are exploring the learning content (e.g., stealth assessment; Shute, 2011). However, the analysis of learners' behavior—especially during collaborative diagnostic reasoning procedures for automated assessments—based on wide, general behavioral indicators has not yet been sufficiently investigated or implemented in practice. First, previous studies that have analyzed learners' behavior have tended to focus on problem-solving strategies (e.g., Stadler et al., 2019) rather than on diagnostic activities. Second, the chosen behavioral indicators have been highly specific to the problem context presented in the learning environment (e.g., necessary and unnecessary actions for fixing a water pump; Zhu et al., 2016). A more general and replicable approach may be found in relating successful learning to more generic behavioral indicators that can be found across a broader range of diagnostic contexts (O'Neil et al., 2003). Predictions of diagnostic success could inform learners and instructions in real time whether or not learners are currently in need of instructional support in the collaborative diagnostic reasoning process and can thus help to individually address learners' zone of proximal development (Vygotsky, 1978). Supporting learners with individual instructional support in single diagnostic cases enables dynamic diagnostic training, which is important for the learning of collaborative diagnostic reasoning skills. Research on complex problem solving has shown that learners use problems that have been solved as blueprints for similar new problems to find new solutions (Richter & Weber, 2013). The opportunity to use learners' learning behavior to readjust instructional support for each diagnostic case would offer the advantage of being able to take learning progress into account.

However, beyond the ability to predict diagnostic success or failure, in order to effectively adapt instructional support, it is necessary to better understand the behavior of successful and unsuccessful diagnosticians. We consider the CDAs to be broad process-based indicators of collaborative diagnostic reasoning skills that can be used in various collaborative diagnostic contexts—from diagnosing diseases to assessing a student's current learning status—to identify differences in successful and unsuccessful diagnostic reasoning processes.

This study

The goals of this study were twofold. First, to provide a general and replicable approach for analyzing diagnostic reasoning processes, we aimed to link diagnostic accuracy to broad behavioral indicators by analyzing the CDAs displayed in a medical training simulation using log files. We aimed to investigate differences in successful and unsuccessful diagnostic reasoning processes and to determine the extent to which CDAs could predict diagnostic accuracy. Second, we aimed to investigate how early diagnostic accuracy could be predicted from CDAs on the basis of behavior exhibited before, during, and after collaboration. In this way, we aimed to exploratively identify early starting points for effective ways to adapt instructional support.

We addressed the following research questions:

1. To what extent can CDAs predict diagnostic accuracy in a medical training simulation using machine learning classification models?
2. How early in the process of making a diagnosis can diagnostic accuracy be reliably predicted from CDAs in a medical training simulation using machine learning classification models?

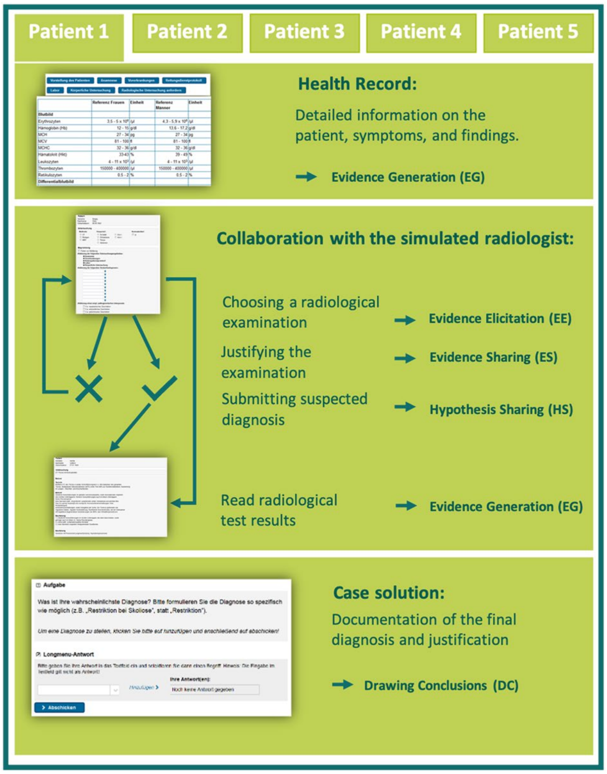
Methods

Sample, simulation, and procedure

To predict diagnostic accuracy, we selected a sample with sufficiently high variance in prior knowledge. Participants were 73 medical students ($N_{\text{female}} = 51$) in their clinical years from the 5th semester and higher ($M = 8.32$ semesters, $SD = 2.80$) of a 6-year study program and 25 physicians from internal medicine ($N_{female} = 11$) with a minimum of 3 years of clinical experience ($M = 13.6$ years of clinical work, $SD = 10.5$). Participation for medical students was limited to those in their clinical years because we assumed that, in principle, students in their preclinical years have not yet generated systematic prior knowledge of radiology and internal medicine. Participation was voluntary. The mean age of the participating medical students was $M = 24.9$ ($SD = 4.23$); for the participating physicians, it was $M = 42.0$ ($SD = 11.7$).

In the text-based simulation, participants acted in the role of an internal specialist in the emergency department of a hospital. Figure 1 presents an overview of the structure of the simulation. Five patient cases that all had the same structure had to be processed. Sequentially, participants received an electronic health record of five fictitious patients who all presented with a fever of unknown origin. The electronic health record was implemented as an electronic folder that contained information about the patients' admission, their medical history, findings from a physical examination, and laboratory results. Participants could navigate between these sections by clicking on representatively named buttons (e.g., medical history), which led to texts with the respective information. The health record could be accessed during the entire diagnostic procedure. After individually processing the information presented in the health record, participants were asked to collaboratively generate further evidence by requesting a radiological examination from an agent-based radiologist.

Participants filled out a request form by choosing a radiological examination and by sharing evidence of the suspected disease and hypotheses with the agent-based radiologist. The agent-based radiologist conducted the radiological examination only when the request was appropriately justified by the shared evidence and hypotheses. Participants then received a detailed document containing the radiological evidence they requested. Otherwise, participants were asked to revise their requests. After requesting the radiological examination, participants could request up to 10 additional radiological examinations. Finally, participants solved the patient case by indicating the diagnosis they thought was most likely. In sum, a participant's task was to collect evidence and generate


Patient 1
Patient 2
Patient 3
Patient 4
Patient 5
Health Record:
Detailed information on the patient, symptoms, and findings.
Evidence Generation (EG)
Collaboration with the simulated radiologist:
Choosing a radiological examination
Evidence Elicitation (EE)
Justifying the examination
Evidence Sharing (ES)
Submitting suspected diagnosis
Hypothesis Sharing (HS)
Read radiological test results
Evidence Generation (EG)
Case solution:
Documentation of the final diagnosis and justification
Drawing Conclusions (DC)
Aufgabe
Longmenu-Antwort
Ihre Antwort(en):
Hinzufügen ›
Noch keine Antwort gegeben
› Abschicken

Was ist Ihre wahrscheinlichste Diagnose? Bitte formulieren Sie die Diagnose so spezifisch wie möglich (z.B. „Restriktion bei Skoliose", statt „Restriktion").

Um eine Diagnose zu stellen, klicken Sie bitte auf hinzufügen und anschließend auf abschicken!

Bitte geben Sie Ihre Antwort in das Textfeld ein und selektieren Sie dann einen Begriff. Hinweis: Die Eingabe im Textfeld gilt nicht als Antwort!

Fig. 1 Overview of the Structure of the Simulation With Corresponding Assignment of Activities to the CDAs

hypotheses about a patient's illness to reduce uncertainty about the final diagnosis. The simulation was implemented in the learning platform CASUS (www.instruct.eu). For further information about the development, implementation, and validation of the simulation, see Radkowitsch et al. (2022). The study was conducted in a laboratory setting. Participants could work on the cases without time constraints but were asked to work efficiently. They were prompted to offer a solution to a case after 15 min. The total processing time per case was $Mdn_{\text{min}} = 15.26$. The minimum median processing time was 6.77 min, and the maximum was 26.03 min. Participants received 25€ as compensation for their participation.

Coding collaborative diagnostic activities and measuring diagnostic accuracy

Participants' activities (i.e., their clicks and text entries) during the diagnostic reasoning process were automatically recorded and assigned to the five abovementioned previously specified CDAs (Radkowitsch et al., 2022). Due to the implementation of the simulation, some activities were individual diagnostic activities (e.g., evidence generation), whereas other activities were collaborative diagnostic activities with the agent-based radiologist (e.g., evidence sharing). The overview of the structure of the simulation in Fig. 1 contains the corresponding assignment of activities to the CDAs within each section. The ways in which the activities were assigned to the activity categories is described below in more detail.

Evidence generation (EG)

Any individual activity by which learners directly received additional information about a patient's health status was coded as evidence generation. This included any clicking within the health record as well as reading the results from the radiological examination.

Evidence elicitation (EE)

An activity was coded as evidence elicitation whenever participants asked the agent-based radiologist to generate further evidence about a patient's health status. The specific activities included choosing a body part about whose status participants required further evidence as well as choosing a radiological examination (e.g., computer tomography [CT] scan) to examine the respective body part using the request form.

Evidence sharing (ES)

Anytime participants used the request form to share evidence about a patient's health status (e.g., main symptoms, course of the disease) with the agent-based radiologist to help them interpret the radiological evidence, an activity was coded as evidence sharing.

Hypothesis sharing (HS)

Anytime participants used the request form to share a differential diagnosis with the agent-based radiologist, an activity was coded as hypothesis sharing.

Drawing conclusions (DC)

Learners concluded a patient case by choosing a final diagnosis from a long menu containing over 200 entries. To do so, participants were asked to type in the initial letters of a diagnosis, after which matching entries popped up, and from which they could select a fitting diagnosis. In addition, participants were asked to justify their diagnosis using a free text field. This activity and the previous one were coded as drawing conclusions. The quality of the final diagnosis was used as an indicator of diagnostic accuracy.

Diagnostic accuracy

We used the final diagnoses proposed by the participants as indicators of diagnostic accuracy, which we used as an easy-to-interpret summative measure of diagnostic reasoning skills. The final diagnoses were coded by researchers from the learning sciences based on sample solutions developed by medical experts as either 1 (*correct*) or 0 (*incorrect*). Two trained raters independently coded 20% of the data set. They achieved perfect interrater agreement ($ICC = 1$). The remaining data set was split in half, and each half was coded by one of the trained raters.

Statistical analyses

All analyses described below were conducted in R 4.0.2 (R Core Team, 2020). The data sets, R script, and formulas are available from the open science framework (OSF) repository at https://osf.io/2ne3y/?view_only=13ae84318f164875a67b7919cf85fd21.

Feature extraction

To analyze participants' activities during the diagnostic reasoning process, the total time participants worked on the patient cases was split into seconds for each patient case. We logged the collaborative diagnostic activity that was being performed for each second. This procedure resulted in 490 individual strings of activities (98 participants with five patient cases each) with the length of the total time-on-task measured in seconds. Subsequently, 14 of the strings had to be removed due to missing values in the case solution, resulting in a final number of $N = 476$ strings. For the subsequent feature extraction, we opted to apply an exploratory approach.

An approach that was created for applying an exploratory search of repetitive patterns within long sequences is the *n*-gram method (Damashek, 1995). The *n*-gram method summarizes a long string of entries (e.g., individual diagnostic steps in a diagnostic reasoning process) as sequences of *n* consecutive elements. To limit the number of features, we split the strings of activities into *n*-grams of length 2 (bigrams), using the "ngram" R package (Schmidt & Heckendorf, 2017), resulting in 25 variables, each representing the frequency of the occurrence of a unique combination of activities (see He & von Davier, 2016). More precisely, the resulting bigrams included two types: bigrams consisting of one activity (e.g., EE.EE) and bigrams consisting of two activities (e.g., EE.ES). The more frequently bigrams of two identical activities occurred, the more time was spent on that activity. The more frequently bigrams of two different activities occurred, the more frequently the transition from the first to the second activity occurred. Bigrams that occurred in only a maximum of one participant's string of activities were not included in the following analyses.

To identify bigrams that led to correct or incorrect diagnoses, we employed the Chi-Square feature selection model proposed by He and von Davier (2016). Using this approach, we conducted a weighted Chi-Square test for each bigram to determine whether its occurrence and nonoccurrence were independent for participants who came up with the correct versus the incorrect diagnosis. We used the weighted frequencies of the bigrams in correct and incorrect diagnoses to calculate whether the bigrams were more typical of correct or incorrect diagnoses (more details can be found in Oakes et al., 2001).

Machine learning approaches

To investigate our research questions, we trained three different supervised machine learning models to classify whether a participant would provide the correct diagnosis for any specific patient on the basis of the bigrams. Specifically, we trained support vector machine (SVM) models with linear kernels, random forest (RF) models, and gradient boosting machine (GBM) models. We chose these models because they are widely used in educational data mining and are viewed, among others, as representatives of the state-of-the-art methods for predicting binary or categorical outcome variables inside and outside of educational assessment (e.g., Costa et al., 2017; Fernández-Delgado et al., 2014; Qiao & Jiao, 2018). Detailed insights into the calculation principles (including formulas) can be found in Bonaccorso (2017).

SVMs classify data into two classes by finding the hyperplane that captures the largest distance between the data points in one class and those in the other class. The maximum width of the slab parallel to the hyperplane, which has no inner data points, is called the margin (Cortes and Vapnik, 1995). The data points at the left and right sides of the margin closest to the hyperplane (support vectors) are used as the starting point for maximizing the margin. With the help of the so-called kernel function, which is applied to the predictor variables, SVMs raise the variable space to a higher dimension and can thus also identify nonlinear relationships (Hilbert et al., 2021). Previous studies have shown that SVMs achieve better performance than other algorithms such as RFs or naïve bayes (e.g., Costa et al., 2017). Moreover, SVMs offer the advantage of being suitable for smaller data sets (Hussain et al., 2019). For the application of SVMs to our data set, we chose linear kernels to map linear relationships in the data in addition to nonlinear relationships that we captured with RFs and GBMs. RFs are based on decision trees and are used in classification and regression problems.

RFs constructs a certain number of single decision trees using random parts of the data to be classified. The procedure uses the test data on all constructed trees and assigns the most frequently occurring outcomes as labels to the test data (Breiman, 2001). As ensembles of single decision trees, RFs have advantages over single trees in terms of predictive power (Fernández-Delgado et al., 2014). Due to the large number of trees (law of large numbers), RFs barely overfit compared with single decision trees or other tree-based ensemble methods, such as GBMs (Breiman, 2001). Moreover, RFs are easier to tune and less time-consuming than GBMs, as well as easier to interpret than other supervised machine learning models, such as SVMs (Hilbert et al., 2021).

In contrast to RF models, which train trees independently, GBMs construct decision trees sequentially so that each new tree can help compensate for errors in previous trees (gradient descent method). By limiting the maximum number of leaves and splits, each decision tree acts as a weak learner (a model that performs slightly better than a random classifier/regressor) and does not dominate the prediction. GBM models allow high flexibility (Natekin & Knoll, 2013) and often achieve better performance than RFs (e.g., Qiao & Jiao, 2018) due to various hyperparameter options. Moreover, a strength of GBM models is that they can easily handle plenty of features and unbalanced data sets (Schröders et al., 2022).

Model development and evaluation

To train the models, we used the R packages "caret" (Kuhn, 2020), "ranger" (Wright & Ziegler, 2017), and "gbm" (Greenwell et al., 2020).

For all methods, the same data were used to train and test the models. First, we randomly split the data set into a training set (70% of the data) and a testing set (30% of the data). This resampling strategy is also called the holdout estimator (Pargent et al., 2022). The training set was then used to fit the predictive models. Unlike more conventional statistical models (e.g., linear regression), machine learning algorithms involve hyperparameters that have to be set before they are run (Probst et al., 2019). For SVM models with linear kernels, only one hyperparameter (the cost value, which specifies how much the algorithm is "punished" for incorrect assignments) has to be tuned. The RF models were tuned to optimize minimal node size (the minimum number of data points required in any given node to split it), splitrule (gini or extra trees), and the number of predictors considered for splitting at each node (mtry). Important hyperparameters for GBM models include the basis of the number of trees (total number of trees in the ensemble), the interaction depth (maximum nodes per tree), the shrinkage (learning rate), and the minimal number of observations in a node (n.minobsinnode).

While training, the abovementioned hyperparameters were tuned automatically for each model on the basis of model performance using 10 × 3 cross-validation (Fushiki, 2011). The cross-validation resulted in 30 iterations (10 folds, three repetitions) of training for each model, thus allowing us to determine the optimal hyperparameters and estimate the stability of each model to avoid over- or underfitting.

The optimal model was selected automatically for each of the algorithms on the basis of the largest kappa value (degree of agreement between the classifications and the real data, taking into account the agreement that occurred by chance). To check whether the diagnostic accuracy could be predicted on the basis of unseen data (RQ1), the optimal model was evaluated in the testing data set. To evaluate the algorithms, the classification accuracy (proportion of correct classifications out of all classifications), sensitivity (proportion of true classified correct diagnoses), specificity (proportion of true classified incorrect diagnoses), positive predictive value (PPV; proportion of true classified correct diagnoses out of all diagnoses classified as correct), negative predictive value (NPV; proportion of true classified incorrect diagnoses out of all diagnoses classified as incorrect), and F1 value (weighted average of sensitivity and positive predictive value) were calculated in addition to kappa.

The algorithm with the best average kappa value resulting from the cross-validation (training phase) was selected for further analysis and interpretation. For this model, we estimated the relative importance (Chen et al., 2020) of each bigram with the R package "caret" (Kuhn, 2020), which indicates how each feature affected the model's performance (total classification accuracy). The higher the variable importance score, the more important the feature was for the overall prediction (Fisher et al., 2019). This provided some measure of how relevant any specific combination of activities was for the total prediction in relation to the others but could not be interpreted concerning size or direction. Machine learning models can become highly complex and are therefore sometimes referred to as black boxes (Yarkoni and Westfall, 2017), which make it difficult to interpret the individual contribution of each feature. However, for this study, we were mainly interested in the total prediction rather than in individual feature interpretation.

To address RQ2, the algorithm was then applied to 10 subsets of the original complete data, created by splitting the first 1200 s of the total processing time into time intervals of 120 s before extracting the features (bigrams). The data sets contained the behaviors (bigrams) that participants exhibited at the corresponding time points.

Results

Descriptive statistics

Table 1 presents the numbers of incorrect and correct diagnoses across the behavioral strings of physicians and medical students. Physicians and medical students came up

Table 1 Distributions of Incorrect and Correct Diagnoses Across Behavioral Strings of Physicians and Medical Students

	Number of behavioral strings		Total
	Incorrect diagnoses	Correct diagnoses	
Physicians	34	91	125
Medical students	128	223	351
Total	162	314	476

with correct diagnoses in 73% and 64% of the cases, respectively. However, this difference was not statistically significant, $X^2(1) = 3.52$, $p = .061$. Overall, there was a higher proportion of correct diagnoses.

Research question 1

To investigate whether diagnostic accuracy could be predicted from observed behavior (RQ1), we first took a closer look at differences in the CDAs between the incorrect and correct diagnoses.

Table 2 summarizes the numbers of strings of incorrect and correct diagnoses in which the bigrams occurred and the total frequencies in those strings. The three bigrams that occurred in only one string of activities in either correct or incorrect diagnoses (HS. DC, DC.ES, and DC.HS) were excluded from the following analyses, leaving a total of 22 bigrams. Further, Table 2 presents the results of the Chi-Square feature selection model, which shows the differences in the probabilities of the bigrams for participants who correctly diagnosed the patient case and those who did not. Bigrams with higher Chi-Square values were better at discriminating between the two groups.

When looking at the bigrams with only one activity (i.e., the bigrams that indicated how much time was spent on that activity), the bigram DC.DC (i.e., spending more time drawing conclusions) was by far the most discriminative bigram for participants who gave an incorrect diagnosis versus those who gave a correct diagnosis. Spending more time drawing conclusions occurred more often among participants who gave a correct diagnosis. Next was EE.EE (spending more time eliciting evidence), which was more typical of participants who gave an incorrect diagnosis, followed by HS.HS (spending more time sharing hypotheses) and EG.EG (spending more time generating evidence), both of which were more typical of participants who gave a correct diagnosis. For the bigrams with two activities (i.e., the bigrams that indicated more frequent transitions from the first to the second activity), EE.EG (switching back from the radiological request to the health record or to reading radiological test results), ES.EE, and HS.EE (both representing setbacks during the radiological request) were the most discriminative behaviors, all of which were more typical of participants who submitted an incorrect final diagnosis. Moreover, both switching between submitting the final diagnosis and requesting the agent-based radiologist (DC.EE, EE.DC, ES.DC) and studying the health record (DC. EG) were among the most discriminative behaviors, all of which were more typical of participants who gave an incorrect diagnosis. All of the described bigrams were statistically significantly able to discriminate between the two groups.

Table 2 Frequency of Occurrence of Bigrams in Incorrect and Correct Diagnoses

Bigram	Frequency in strings		Weight	Total frequency of bigrams				Chi-Square test		
	Incorrect diagnoses	Correct diagnoses		Incorrect diagnoses		Correct diagnoses		χ^2	p	Dir
				Raw	Wgt	Raw	Wgt			
EG.EG	162	314	0.03	71,931	410.50	110,662	631.53	144.17	< .001	+
EG EE	159	312	0.04	405	8.83	521	11.36	0.11	.735	–
EG ES	38	47	2.27	49	110.59	54	121.87	49.80	<.001	–
EG HS	14	21	3.04	18	54.56	29	87.91	31.62	<.001	+
EG DC	156	309	0.06	230	9.63	376	15.74	6.44	.011	+
EE EG	82	81	1.79	113	200.39	83	147.19	766.49	<.001	–
EE.EE	162	313	0.03	11,727	113.68	8801	85.32	403.93	<.001	–
EE ES	151	283	0.17	360	57.84	464	74.55	0.67	.414	–
EE HS	48	60	2.06	73	149.57	77	157.76	101.46	<.001	–
EE DC	9	3	3.33	9	29.93	3	9.98	410.76	<.001	–
ES EG	70	82	1.63	87	141.08	97	157.29	56.34	<.001	–
ES EE	54	39	2.22	74	163.26	54	119.14	633.63	<.001	–
ES.ES	157	298	0.14	29,944	3588.27	39,799	4769.22	0.43	.514	–
ES HS	146	280	0.20	301	57.41	464	88.49	19.51	<.001	+
ES DC	6	2	3.39	6	20.27	2	6.76	278.01	<.001	–
HS EG	147	270	0.24	265	59.37	423	94.77	31.10	<.001	+
HS EE	54	41	2.16	68	146.03	48	103.08	620.17	<.001	–
HS ES	52	78	1.81	58	104.47	97	174.71	88.15	<.001	+
HS.HS	159	311	0.06	14,990	537.64	23,578	845.66	257.95	<.001	+
HS DC	1	4	3.37	1	3.36	4	13.45	89.06	<.001	+
DC EG	46	60	2.12	83	174.87	85	179.08	149.63	<.001	–
DC EE	10	4	3.34	11	36.68	4	13.34	462.98	<.001	–
DC ES	1	2	3.26	1	3.25	2	6.50	9.10	.003	+
DC HS	0	2	3.10	0	0.00	2	6.19	114.52	<.001	+
DC.DC	160	313	0.04	20,863	441.81	40,842	864.89	1203.06	<.001	+

Note. EG = Evidence generation, *EE* = Evidence elicitation, *ES* = Evidence sharing. *HS* = Hypothesis sharing, *DC* = Drawing conclusions. Higher Chi-Square values indicate more discriminative bigrams. *Dir.* = Direction of the difference in the occurrence of bigrams between learners who diagnosed the case correctly and those who diagnosed the case incorrectly, "+" represents a more frequent occurrence of the bigram in the strings of learners who correctly diagnosed the case, "–" represents a more frequent occurrence of the bigram in the strings of learners who incorrectly diagnosed the case

Table 3 Mean Classification Accuracy and Kappa From the Cross-Validation for All Algorithms

Measures	SVM	RF	GBM
Mean accuracy	.73	.75	.74
$\mathrm{CI}_{\mathrm{Accuracy}}$	[.71–.76]	[.70–.79]	[.70–.76]
Mean kappa	.33	.37	.36
$\mathrm{CI}_{\mathrm{kappa}}$	[.24–.42]	[.31–.49]	[.30–.43]

Note. CI 95% confidence interval

Subsequently, we trained three different machine learning models to classify whether a participant would provide the correct diagnosis for any specific patient case on the basis of the 22 remaining bigrams. Table 3 summarizes the results for all models from the training phase (cross-validation) by presenting the average classification accuracy and kappa across all 30 repetitions. Generally, the different model iterations did not differ much, thus suggesting no substantial overfitting. All algorithms showed significantly higher average classification accuracy than the no information rate (NIR), which indicates how many observations out of all observations would have been correctly classified if only the label "correct diagnosis" (the larger class) would have been assigned. The NIR of .66 corresponds to the proportion of all correct diagnoses in all observations (see Table 1). Considering an ideal NIR of .50 (equally distributed classes; Batista et al., 2004), .66 deviates somewhat from this value but does not indicate a substantial skewness in favor of one of the classes. Beyond accuracy, the algorithms reached acceptable kappa values (Fleiss et al., 2003). Moreover, the models did not differ significantly in their average classification accuracy values, F(2, 87) = 0.56, $p = .559$, $\eta^2 = .01$, or in their average kappa values, F(2, 87) = 0.72, $p = .491$, $\eta^2 = .02$. However, since the RF showed descriptively a slightly better average kappa, it was selected to finally answer RQ1 and RQ2.

Table 4 presents the evaluation results of all algorithms in the testing data set. As can be seen, RF (final tuning parameters: min node size = 1, mtry = 2, and splitrule = gini), GBM (final tuning parameters: n.trees = 50, interaction.depth = 1, shrinkage = 0.1, and n.minobsinnode = 10), and SVM (final tuning parameter: cost value = 0.25) all achieved significantly higher classification accuracy than the NIR as well as acceptable kappa values (Fleiss et al., 2003). Strikingly, all models showed high sensitivity, and good PPV and F1 values but rather low specificity, indicating that correct diagnoses were substantially better predicted than incorrect diagnoses. However, all models reached acceptable NPV values, indicating precision in classifying incorrect diagnoses (many of the diagnoses classified as "incorrect" were indeed incorrect diagnoses). Overall, the algorithms did not differ greatly in their performance. The final selected algorithm, the RF model, achieved acceptable to good values on all measures (classification accuracy = .75, kappa = .40, sensitivity = .90, specificity = .46, PPV = .77, NPV = .71, and F1 = .83) and was therefore selected for further interpretation and analyses.

Figure 2 illustrates the bigrams' relative importance in the RF model. By far most important for the overall prediction was how much time was spent eliciting evidence (EE.EE) followed by the amount of time spent drawing conclusions (DC.DC), sharing evidence (ES.ES), generating evidence (EG.EG), and sharing hypotheses (HS.HS).

Table 4 Results of the Evaluation of the Algorithms in the Testing Data Set

Measures	SVM	RF	GBM
NIR (.66)			
Acc	.75	.75	.74
p-value [Acc > NIR]	.012	.011	.029
Kappa	.39	.40	.40
Sensitivity	.91	.90	.84
Specificity	.44	.46	.54
PPV	.76	.77	.78
NPV	.72	.71	.63
F1	.83	.83	.81

Note. NIR = Proportion of correct diagnoses in all observations, *Acc* = Classification accuracy, *PPV* = Positive predictive value, *NPV* = Negative predictive value

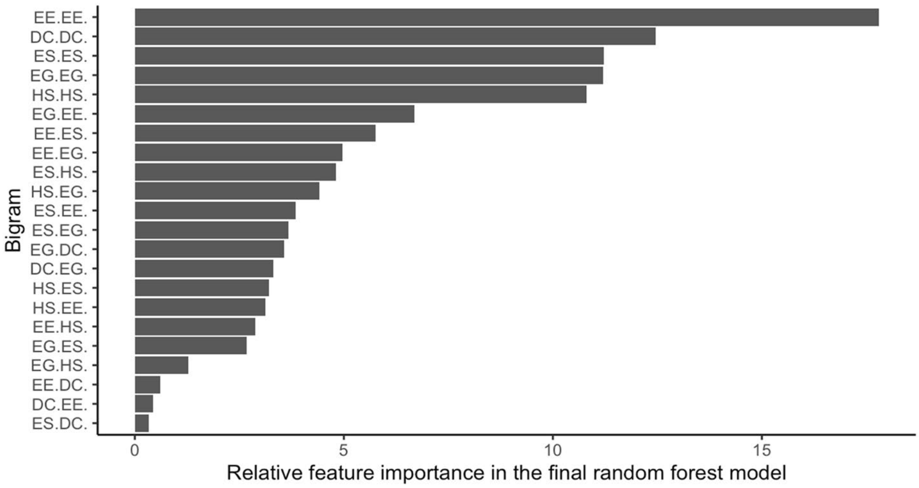


Fig. 2 Relative Importance of Each Bigram for the Final RF Model. *EG* Evidence generation, *EE* Evidence elicitation, *ES* Evidence sharing, *HS* Hypothesis sharing, *DC* Drawing conclusions

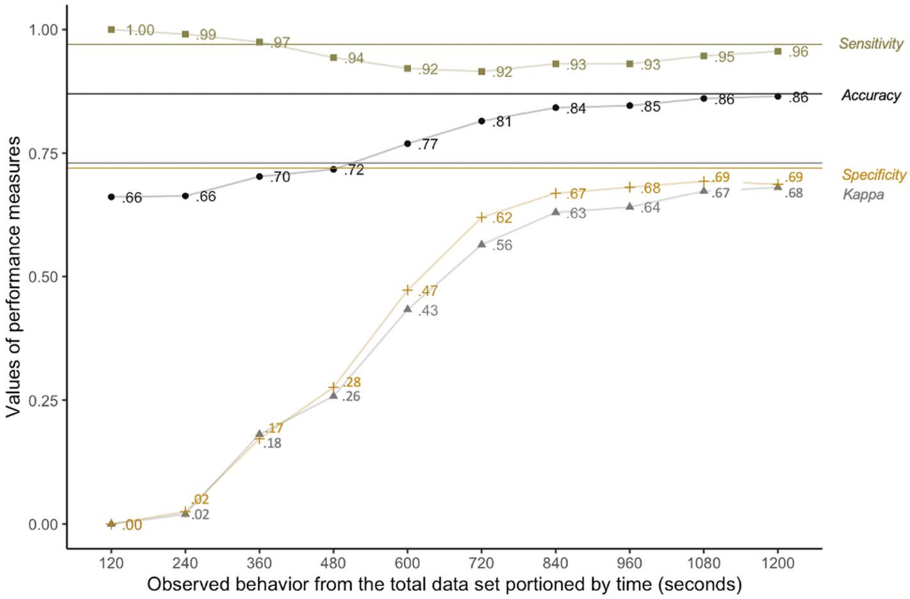


Fig. 3 Performance Measures for the Random Forest Model Applied to Increasing Amounts of Data. The horizontal lines represent the final values for the RF algorithm based on the original complete data set

Moreover, the analysis revealed that the most important bigrams with two activities were the frequency of switching between evidence generation and evidence elicitation (EG.EE; EE.EG) as well as the frequency of transitions from evidence elicitation to evidence sharing (EE.ES).

Research question 2

To investigate how early during diagnosing it is possible to reliably predict diagnostic accuracy on the basis of CDAs (RQ2), we applied the final RF model to a sequence of subsets of the complete data that included only the actions observed in the first 120 to 1200 s. As can be seen in Fig. 3, classification accuracy, kappa, sensitivity, and specificity approximated the values estimated for the complete data (horizontal lines) after 1200 s. In the first 120 s, the model did not perform better than the NIR of .66 (classification accuracy = .66, sensitivity = 1, kappa = 0, specificity = 0). From second 240, the performance slowly increased and asymptotically approached the final values in the complete data set. More precisely, in second 360, the accuracy exceeded the NIR until it reached approximately its final value in the complete data set at second 1200 with 0.86. Similarly, the kappa value increased over time (largest increase between seconds 600 and 840). At second 120, the RF began with a sensitivity (correct classification of correct diagnoses) of 1 (100%) because, in the beginning, the model classified all observations as "correct." Up to second 720, the sensitivity slowly decreased, while kappa and specificity increased, until sensitivity approximately reached its final value in the complete data set with .96 after 1200 s. By contrast, at second 120, the model began with a specificity (correct classification of incorrect diagnoses) of 0 (0%) but approximately approached the final value over time with .69. Overall, it can be seen from the graph that the model's performance took on acceptable predictive values from about second 840. Correct diagnoses could be predicted particularly well after 600 s (10 min) or after two thirds (66%) of the median time (15 min) had been spent on the patient case.

Discussion

This study examined the extent to which and how quickly diagnostic accuracy could be predicted from learners' engagement in CDAs based on log file data from a medical simulation with the help of machine learning. Three different classification algorithms (SVM, RF, GBM) reached acceptable overall prediction quality. Due to slightly better performance, the RF model was selected for further interpretation and analysis of how early it is possible to achieve a reliable prediction of diagnostic accuracy during diagnosing on the basis of CDAs. The results showed that after approximately two thirds of the median time learners spent on the diagnostic task, the RF algorithm was able to reliably predict diagnostic success. Moreover, the time spent on CDAs was especially important for predicting diagnostic accuracy and was the best at distinguishing between correct and incorrect diagnoses. While spending more time engaged in individual activities (e.g., generating evidence and drawing conclusions) was more typical of successful diagnosticians, spending more time engaged in collaborative activities (e.g., eliciting and sharing evidence; i.e., interaction with the agent-based radiologist) tended to be behavior that was more typical of unsuccessful diagnosticians. These findings are aligned with previous work that showed somewhat similar results in the context of complex problem solving. For example, Stadler et al. (2019) found that successful problem solvers spent more time reflecting on the task (i.e., they spent more time drawing conclusions), whereas unsuccessful problem solvers spent more time performing activities that involved gathering information. However, the equivalent results for unsuccessful diagnosticians in the context of our simulation apply only to collaborative engagement with the evidence (i.e., spending more time eliciting and sharing evidence as opposed to spending more time generating evidence).

Previous research found that time spent on tasks was moderated by prior knowledge level (e.g., Goldhammer et al., 2014). Our study adds to this line of research by qualifying the types of activities within the task. Considering the MOT model (Charlin et al., 2012), in contrast to unsuccessful diagnosticians, successful diagnosticians should be able to identify early case cues, have more specific initial representations, and be better able to determine the relevant objectives of the encounter. Applied to our simulation, when successful diagnosticians have a concrete suspected diagnosis, they are able to make a more specific radiological request that they know will help them find support for or falsify their diagnosis. As a consequence, they consult the radiologist less often and elicit less evidence. Instead, they spend more time carefully processing the information from the health record and radiological test results, and at the end, they spend more time drawing conclusions before settling on a final diagnosis. On the other hand, unsuccessful diagnosticians might have trouble identifying early cues in the patient case and determining the appropriate objectives of the patient encounter (Bowen, 2006). Compared with diagnosticians who have a proper initial patient representation, they urgently require further radiological information to be able to diagnose the case but might have trouble further processing this large amount of weakly organized information (Stadler et al., 2019), as they lack a proper initial representation. Thus, these diagnosticians have trouble making optimal use of collaboration as a source of information (Radkowitsch et al., 2022) because they have both no clue about what additional information to look for in the patient and problems with sharing *relevant* information with the collaboration partner (Tschan et al., 2009), leading to an increasing amount of time spent selecting appropriate examinations and sharing evidence from the health record. This interpretation would be supported by the frequent transitions and setbacks typically encountered by unsuccessful diagnosticians while working in the simulation. One reason for frequent transitions within the radiological request is that these diagnosticians request a larger number of examinations, supporting the assumption that they have a greater need for additional radiological evidence. Diagnosticians who displayed frequent switches from the radiological request form to the health record may have lacked a concrete idea about the patient's problem at that time, had several possible suspected diagnoses in mind, and were unable to retain information from the health record in their working memory while simultaneously implementing the requirements of collaboration. Further, switching back and forth between submitting the final diagnosis (drawing conclusions) and dealing with evidence by either requesting the radiologist and studying the health record (generating and eliciting evidence) or sharing patient information with the radiologist (sharing evidence) are typical behaviors of unsuccessful diagnosticians. This finding most likely indicates that these diagnosticians have problems using the evidence appropriately to validate or exclude a particular hypothesis from their set of suspected hypotheses (evidence evaluation).

Notably, the Chi-Square feature selection model revealed that the above described transitions from one CDA to another and switching between CDAs, both of which are related to incorrect diagnoses, better distinguish between successful and unsuccessful diagnosticians than the time spent on these activities. However, in the RF model, the time spent on CDAs was clearly most important for the overall prediction. Thus, we assume that beyond the Chi-Square test, the prediction of the RF model may have revealed additional nonlinear relationships between CDAs and diagnostic accuracy (black box problem; Yarkoni and Westfall, 2017).

Taken together, these findings on the differences between successful and unsuccessful diagnosticians suggest that, at least in the context of our simulation, an adequate initial representation of the case is crucial for diagnostic success. The information on adequate or inadequate initial representations of the case could be used to provide adaptive process-based feedback on whether learners are heading toward correct diagnoses. On the other hand, an inadequate representation can hardly be compensated for by subsequent collaboration with the agent-based radiologist. Thus, in the context of our simulation, the agent tended not to be helpful to diagnosticians who were on the wrong track. Further, deviations from the intended structure of the simulation were more likely to be indicators of misdiagnoses, thus applying to a wide range of expertise. However, referring to the high sensitivity but low specificity achieved by our model, we were able to reliably predict correct diagnoses better and earlier than incorrect ones. We assume that one reason for the low specificity compared with the high sensitivity is that in our sample successful diagnosticians may not differ in their behavior as much as unsuccessful diagnosticians. After reading the health record, successful diagnosticians enter the collaboration with an adequate mental representation, through which they can make targeted radiologic requests to reduce diagnostic uncertainty regarding suspected diagnoses, and solve the diagnostic case correctly. In contrast, the misdiagnoses of unsuccessful diagnosticians could be due to cognitive misbehavior of various causes, which manifests itself at the simulation level in different behavior. For example, recent analyses on the behavior after impasses in the context of the same simulation show that diagnosticians differ in their success in identifying and subsequently compensating for errors in the diagnostic reasoning process (Heitzmann et al., 2023). Future research may follow this line of research and examine the behaviors that lead to an incorrect diagnosis in more detail.

Our study represents a “proof of concept” for one way in which the prediction of successful and unsuccessful diagnosticians using the behavior displayed in the simulation could be used in microadaptive learning environments. Yet, further research will be necessary. Early predictions of learners heading toward a correct diagnosis can inform instructors and educators to remove instructional support in real time before it has negative effects on learning (Kalyuga et al., 2003). Our prediction of correct diagnoses was successful only after two thirds of the diagnostic reasoning process and thus cannot necessarily be considered an early prediction, for example, as shown by Ulitzsch et al. (2022), when they used only about one third of their examined clickstream data in the context of complex problem-solving. However, because we obtained the information on diagnostic success before learners completed the diagnostic task, it is still possible to adjust the task difficulty in real time or in the upcoming task (Roosevelt, 2008) to address learners’ zone of proximal development (Vygotsky, 1978). Moreover, to increase the likelihood of building a correct initial case representation that prepares and prestructures the individual diagnostic reasoning process for collaborating with the agent-based radiologist, learners could receive prompts that remind them to review the health record and radiological test results properly and help them integrate the information into hypotheses. Conceivable types of scaffolding may be reflection prompts (Mamede and Schmidt, 2017), which encourage learners to reflect on the evidence they generated in terms of potential hypotheses.

Limitations and further research

In interpreting these findings, there are some limitations to be considered. The first relates to the prediction of diagnostic success after beginning the diagnostic reasoning process, which was possible only after 10 min because the behavior in the earlier minutes was probably not diverse enough.

The reason for this finding can be seen in the rather coarse granulation level of the coded log files of the CDAs, which might not have been fine enough to identify early subtle differences in the behaviors of successful and unsuccessful diagnosticians. However, the use of broad diagnostic indicators is also one of the strengths of this study, as they can be applied to other diagnostic contexts for generalization at a low threshold. Nevertheless, future process analyses could investigate diagnostic behavior at finer coding levels to uncover further latent differences between successful and unsuccessful diagnosticians.

Second, at least to some extent, the use of bigrams limited the insights that could have been gained about the behavior of successful and unsuccessful diagnosticians if trigrams (e.g., EE.ES.HS), which would have included two transitions, had been used. Alternatively, unigrams (e.g., EE) might have been interpretable in a more straightforward way. However, trigrams would have extensively increased the number of possible features ($k = 125$), and unigrams would have indicated only the time spent on CDAs without considering transitions from one to another. To verify our choice of bigrams, we repeated the Chi-Square test with trigrams to control for possible significant sequences of two transitions. We found that the ranking of the most important indicators of diagnostic success and failure did not change such that, for each strong discriminative bigram (e.g., EE.EG.), both possible trigrams (EE.EE.EG; EE.EG.EG) discriminated equally well. Interested readers can find these analyses on the OSF. Moreover, our approach to feature extraction did not consider participants' pauses between activities, even though pausing behavior may provide a valuable source of information (e.g., Tenison and Arslan, 2020). Pausing behavior, for instance, may indicate reflective thinking about the diagnostic reasoning process or may be linked to behavioral responses following errors or impasses. Since the n-gram approach is not necessarily the best one to capture pausing behavior, approaches more appropriate for timing data may be considered in future research.

Third, we did not consider case difficulty, case typicality, or the prior knowledge or expertise level of diagnosticians in our prediction models. However, the fact that our algorithm was able to reliably predict diagnostic accuracy across different cases and expertise levels is a strong sign of robustness. Further, another study with the same tasks found that changes in difficulty across tasks led to changes in time on task regardless of participants' level of expertise (Stadler et al., 2021), further supporting their equivalence in typicality. However, our interpretations of the behavior of successful versus unsuccessful diagnosticians were mainly valid for cases in which early cues already pointed to the correct diagnosis (typical cases). The extent to which the algorithms can predict similar results exclusively for atypical cases needs to be investigated in further studies.

Moreover, the present analysis focused on diagnostic accuracy and not on learning as a change in knowledge and skills. It is possible that our participants who "gambled the radiologist" by sharing and requesting a lot of information may be among those who still failed to reach a correct conclusion but still learned a lot from the simulation. Exploring complex problem-solving tasks with the goal of finding out as much as possible, without the goal of establishing a well-supported solution or diagnosis may be an effective approach to learning, as it is connected to lower cognitive load (goal-free instruction; Sweller et al., 2019). Finally, the study participants in our setting interacted with an agent. A recent study found no differences between agents and human collaborators in the assessment of collaborative problem solving in PISA (Herborn et al., 2020), yet agent-based collaboration carries the risk of being a poor substitute for natural collaboration. However, we chose agent-based collaboration for one significant advantage: In contrast to human-to-human collaboration, it enabled the standardized measurement of collaborative diagnostic reasoning processes by holding the agent's behavior and knowledge level constant. In addition, the simulation's interface (request form) and its structure were carefully developed by learning scientists and medical experts on the basis of real clinical situations in which an internist collaborates with a radiologist, who serves as a potential additional source of evidence, to reduce further diagnostic uncertainty. Yet, future research should address the transfer to human-to-human collaboration in diagnostic settings.

Conclusion

Even though having the competence to provide a correct diagnosis collaboratively is relevant in many domains, the fostering of collaborative diagnostic reasoning has yet to be thoroughly investigated. Simulations with dynamic individual learning support are a promising approach for fostering such complex skills. The present study identified behavioral characteristics for successful and unsuccessful diagnosticians in a collaborative medical training simulation based on CDAs—broad theoretical indicators that can be found in various diagnostic contexts. We used these indicators to develop a model that enabled a reliable and robust prediction of diagnostic accuracy across diagnosticians with varying expertise levels and different diagnostic cases. The study provides preliminary evidence that (a) the individual diagnostic reasoning process controls the collaborative diagnostic reasoning process and is thus crucial for overall diagnostic success and that (b) diagnostic success can be predicted better than diagnostic failure, and after only 66% of the average time spent on the diagnostic case, which might be due to the fact that diagnostic failure underlies more heterogeneous behavior than diagnostic success.

Our study is an example of how log-file-based process data analyses could be further used in adaptive learning environments to individually foster collaborative diagnostic reasoning skills in a targeted manner. These insights can open up new ways to conduct collaborative diagnostic training both within and outside of higher education.

Abbreviations

CDAs	Collaborative diagnostic activities
DC	Drawing conclusions
EE	Evidence elicitation

EG	Evidence generation
ES	Evidence sharing
GBM	Gradient boosting machine
HS	Hypothesis sharing
OECD	Organization for Economic Co-operation and Development
OSF	Open Science Framework
RF	Random forest
SDDS	Scientific discovery as dual search
SVM	Support vector machine

Acknowledgements
Not applicable.

Author contributions
CR: writing—original draft, conceptualization, methodology, project administration. MS: methodology, formal analysis, writing—original draft, writing—review and editing, conceptualization. AR: resources, writing—original draft, writing—review and editing, conceptualization. RS: funding acquisition, conceptualization, writing—review and editing. MF: funding acquisition, conceptualization, writing—review and editing, FF: conceptualization, funding acquisition, supervision, writing—review and editing. All authors read and approved the final manuscript.

Funding
This work was supportet by a grant from the Deutsche Forschungsgemeinschaft (DFG, FOR 2385) for a subproject conducted by the CoSiMed research group (FI 792/11-1).

Availability of data and materials
The datasets analysed during the current study are available in the open science repository (OSF), https://osf.io/2ne3y/?view_only=13ae84318f164875a67b7919cf85fd21.

Declarations

Ethics approval and consent to participate
Ethical clearance was declared by the Ethics Committee at the Medical Faculty of LMU Munich prior to data collection.

Consent for publication
The authors consent to the publication of the manuscript in ***Large-scale Assessments in Education.***

Competing interests
The authors declare no competing interests.

Received: 13 January 2022 Accepted: 4 January 2023
Published online: 20 January 2023

References

Al-Kadi, A. S., & Donnon, T. (2013). Using simulation to improve the cognitive and psychomotor skills of novice students in advanced laparoscopic surgery: a meta-analysis. *Medical Teacher, 35*(sup1), S47–S55. https://doi.org/10.3109/0142159X.2013.765549

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. A. Larusson & B. White (Eds.), *Learning analytics* (pp. 61–75). Springer. https://doi.org/10.1007/978-1-4614-3305-7_4

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of severalmethods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter, 6*(1), 20–29. https://doi.org/10.1145/1007730.1007735

Bonaccorso, G. (2017). *Machine learning algorithms: A reference guide to popular algorithms for data science and machine learning*. Packt Publishing.

Bowen, J. L. (2006). Educational strategies to promote clinical diagnostic reasoning. *New England Journal of Medicine, 355*(21), 2217–2225. https://doi.org/10.1056/NEJMra054782

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Charlin, B., Lubarsky, S., Millette, B., Crevier, F., Audétat, M.-C., Charbonneau, A., Caire Fon, N., Hoff, L., & Bourdy, C. (2012). Clinical reasoning processes: Unravelling complexity through graphical representation. *Medical Education, 46*(5), 454–463. https://doi.org/10.1111/j.1365-2923.2012.04242.x

Chen, R.-C., Dewi, C., Huang, S.-W., & Caraka, R. E. (2020). Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*. https://doi.org/10.1186/s40537-020-00327-4

Chernikova, O., Heitzmann, N., Stadler, M., Holzberger, D., Seidel, T., & Fischer, F. (2020). Simulation-based learning in higher education: A meta-analysis. *Review of Educational Research, 90*(4), 499–541. https://doi.org/10.3102/0034654320933544

Cirigliano, M. M., Guthrie, C. D., & Pusic, M. V. (2020). Click-level learning analytics in an online medical education learning platform. *Teaching and Learning in Medicine, 32*(4), 410–421. https://doi.org/10.1080/10401334.2020.1754216

Cook, D. A., Hamstra, S. J., Brydges, R., Zendejas, B., Szostek, J. H., Wang, A. T., Erwin, P. J., & Hatala, R. (2013). Comparative effectiveness of instructional design features in simulation-based education: Systematic review and meta-analysis. *Medical Teacher, 35*(1), e867–e898. https://doi.org/10.3109/0142159X.2012.714886

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.

Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior, 73*, 247–256. https://doi.org/10.1016/j.chb.2017.01.047

Damashek, M. (1995). Gauging similarity with n-grams: Language-independent categorization of text. *Science, 267*(5199), 843–848. https://doi.org/10.1126/science.267.5199.843

Ericsson, K. A. (2004). Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Academic Medicine, 79*(10), S70–S81. https://doi.org/10.1097/00001888-200410001-00022

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research, 15*(1), 3133–3181.

Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., Neuhaus, B., Dorner, B., Pankofer, S., Fischer, M., Strijbos, J.-W., Heene, M., & Eberle, J. (2014). Scientific reasoning and argumentation: Advancing an interdisciplinary research agenda in education. *Frontline Learning Research., 2*(3), 28–45. https://doi.org/10.14786/flr.v2i2.96

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research, 20*(177), 1–81. http://jmlr.org/papers/v20/18-760.html

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). Wiley. https://doi.org/10.1002/0471445428

Fushiki, T. (2011). Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing, 21*(2), 137–146. https://doi.org/10.1007/s11222-009-9153-8

Gegenfurtner, A., Quesada-Pallarès, C., & Knogler, M. (2014). Digital simulation-based training: A meta-analysis. *British Journal of Educational Technology, 45*(6), 1097–1114. https://doi.org/10.1111/bjet.12188

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology, 106*(3), 608–626. https://doi.org/10.1037/a0034716

Graesser, A. C., Fiore, S. M., Greiff, S., Andrews-Todd, J., Foltz, P. W., & Hesse, F. W. (2018). Advancing the science of collaborative problem solving. *Psychological Science in the Public Interest, 19*(2), 59–92. https://doi.org/10.1177/1529100618808244

Greenwell, B., Boehmke, B., Cunningham, J., & GBM Developers. (2020). *Package 'gbm'* (Version 2.1.8) [Computer software]. https://cran.r-project.org/web/packages/gbm/gbm.pdf

Greiff, S., Niepel, C., Scherer, R., & Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior, 61*, 36–46. https://doi.org/10.1016/j.chb.2016.02.095

Greiff, S., Stadler, M., Sonnleitner, P., Wolff, C., & Martin, R. (2015). Sometimes less is more: Comparing the validity of complex problem solving measures. *Intelligence, 50*, 100–113. https://doi.org/10.1016/j.intell.2015.02.007

Griffin, P., & Care, E. (2015). *Assessment and teaching of 21st century skills*. Dordrecht: Springer. https://doi.org/10.1007/978-94-017-9395-7

Grossman, P., Compton, C., Igra, D., Ronfeldt, M., Shahan, E., & Williamson, P. W. (2009). Teaching practice: A cross-professional perspective. *Teachers College Record, 111*(9), 2055–2100.

Hall, D., & Buzwell, S. (2012). The problem of free-riding in group projects: Looking beyondsocial loafing as reason for non-contribution. *Active Learning in Higher Education, 14*(1), 37–49. https://doi.org/10.1177/1469787412467123

He, Q., & Von Davier, M. (2016). Analyzing process data from problem-solving items with n-grams. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development* (pp. 749–776). IGI Global. https://doi.org/10.4018/978-1-4666-9441-5

Heitzmann, N., Seidel, T., Opitz, A., Hetmanek, A., Wecker, C., Fischer, M. R., Ufer, S., Schmidmaier, R., Neuhaus, B., Siebeck, M., Stürmer, K., Obersteiner, A., Reiss, K., Girwidz, R., & Fischer, F. (2019). Facilitating diagnostic competences in simulations in higher education: A framework and a research agenda. *Frontline Learning Research., 7*(4), 1–24. https://doi.org/10.14786/flr.v7i4.384

Heitzmann, N., Stadler, M., Richters, C., Radkowitsch, A., Schmidmaier, R., Weidenbusch, M., & Fischer, M. R. (2023). Learners' adjustment strategies following impasses in simulations—effects of prior knowledge. *Learning and Instruction*. https://doi.org/10.1016/j.learninstruc.2022.101632

Herborn, K., Stadler, M., Mustafić, M., & Greiff, S. (2020). The assessment of collaborative problem solving in PISA 2015: Can computer agents replace humans? *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2018.07.035

Hilbert, S., Coors, S., Kraus, E. B., Bischl, B., Frei, M., Lindl, A., Wild, J., Krauss, S., Goretzko, D., & Stachl, C. (2021). Machine learning for the educational sciences. *Review of Education*. https://doi.org/10.1002/rev3.3310

Hmelo-Silver, C. E., Duncan, R. G., & Chinn, C. A. (2007). Scaffolding and achievement in problem-based and inquiry learning: A response to Kirschner, Sweller, and Clark (2006). *Educational Psychologist, 42*(2), 99–107. https://doi.org/10.1080/00461520701263368

Hussain, M., Zhu, W., Zhang, W., Abidi, S. M. R., & Ali, S. (2019). Using machine learning to predict student difficulties from learning session data. *Artificial Intelligence Review, 52*(1), 381–407. https://doi.org/10.1007/s10462-018-9620-8

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist, 38*(1), 23–31. https://doi.org/10.1207/S15326985EP3801_4

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12*(1), 1–48. https://doi.org/10.1207/s15516709cog1201_1

Kuhn, M. (2020). *caret: Classification and Regression Training* (Version 6.0–86) [Computer software]. https://CRAN.R-project.org/package=caret

Landriscina, F. (2012). Simulation and learning The role of mental models. In N. M. Seel (Ed.), *Encyclopedia of the sciences of learning.* Springer. https://doi.org/10.1007/978-1-4419-1428-6_1874

Liu, L., Hao, J., von Davier, A. A., Kyllonen, P., & Zapata-Rivera, J. D. (2015). A tough nut to crack. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Advances in higher education and professional development (AHEPD) book series. Handbook of research on technology tools for real-world skill development.* IGI Global. https://doi.org/10.4018/978-1-4666-9441-5.ch013

Mamede, S., & Schmidt, H. G. (2017). Reflection in medical diagnosis: A literature review. *Health Professions Education, 3*(1), 15–25. https://doi.org/10.1016/j.hpe.2017.01.003

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*. https://doi.org/10.3389/fnbot.2013.00021

Norman, G. (2005). Research in clinical reasoning: Past history and current trends. *Medical Education, 39*(4), 418–427. https://doi.org/10.1111/j.1365-2929.2005.02127.x

Oakes, M., Gaaizauskas, R., Fowkes, H., Jonsson, A., Wan, V., & Beaulieu, M. (2001). A method based on the chi-square test for document classificatioDn. In D. H. Kraft, W. B. Croft, D. J. Harper, & J. Zobel (Eds.), *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 440–441). ACM Press. https://doi.org/10.1145/383952.384080

OECD. (2017). PISA 2015 Assessment and analytical framework: Science, reading, mathematic, financial literacy and collaborative problem solving. *PISA, OECD Publishing*. https://doi.org/10.1787/9789264281820-en

O'Neil, H. F., Chuang, S.-H., & Chung, G. K. W. K. (2003). Issues in the computer-based assessment of collaborative problem solving. *National Center for Research on Evaluation, Standards, and Student Testing, 10*(3), 361–373. https://doi.org/10.1080/0969594032000148190

O'Neill, T. A., Allen, N. J., & Hastings, S. E. (2013). Examining the "Pros" and "Cons" of TeamConflict: A Team-Level Meta-Analysis of Task, Relationship, and Process Conflict. *Human Performance, 26*(3), 236–260. https://doi.org/10.1080/08959285.2013.795573

Pargent, F., Schoedel, R., & Stachl, C. (2022). An introduction to machine learning for psychologists in R. *PsyArXiv*. https://doi.org/10.31234/osf.io/89snd

Pauli, R., Mohiyeddini, C., Bray, D., Michie, F., & Street, B. (2008). Individual differences in negative group work experiences in collaborative student learning. *Educational Psychology, 28*(1), 47–58. https://doi.org/10.1080/01443410701413746

Pea, R. D. (2004). The social and technological dimensions of scaffolding and related theoretical concepts for learning, education, and human activity. *Journal of the Learning Sciences, 13*(3), 423–451. https://doi.org/10.1207/s15327809jls1303_6

Plass, J. L., & Pawar, S. (2020). Toward a taxonomy of adaptivity for learning. *Journal of Research on Technology in Education, 52*(3), 275–300. https://doi.org/10.1080/15391523.2020.1719943

Probst, P., Boulesteix, A.-L., & Bischl, B. (2019). Tunability: Importance of hyperparameters of machine learning algorithms. *Journal of Machine Learning Research, 20*(1), 1–32. https://www.jmlr.org/papers/volume20/18-444/18-444.pdf

Qiao, X., & Jiao, H. (2018). Data Mining Techniques in Analyzing Process Data: A Didactic. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2018.02231

Radkowitsch, A., Sailer, M., Fischer, M. R., Schmidmaier, R., & Fischer, F. (2022). Diagnosing collaboratively: A theoretical model and a simulation-based learning environment. In F. Fischer & A. Opitz (Eds.), *Learning to diagnose with simulations: Teacher education and medical education* (pp. 123–141). Springer Nature. https://doi.org/10.1007/978-3-030-89147-3

Richter, M. M., & Weber, R. O. (2013). Case-Based Reasoning. *Springer*. https://doi.org/10.1007/978-3-642-40167-1

R Core Team. (2020). *R: A Language and environment for statistical computing* (Version R4.0.2) [Computer software]. https://www.R-project.org/

Roosevelt, F. D. (2008). Zone of proximal development. In N. J. Salkind (Ed.), *Encyclopedia of educational psychology* (pp. 1017–1022). SAGE Publications. https://doi.org/10.4135/9781412963848.n282

Roschelle, J., & Teasley, S. D. (1995). The construction of shared knowledge in collaborative problem solving. In C. O. Malley (Ed.), *Computer supported collaborative learning* (pp. 69–97). Springer. https://doi.org/10.1007/978-3-642-85098-1_5

San Pedro, M., Baker, R. S., Bowers, A. J., & Heffernan, N. T. (2013). Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. In S. D'Mello R. Calvo, & A. Oldey (Eds.), *Proceedings of the 6th international conference on eduactional data mining* (pp. 177-184).

Schmidt, D., & Heckendorf, C. (2017). *Guide to the ngram package: Fast n-gram tokenization* (Version 3.0.4) [Computer software]. https://cran.r-project.org/package=ngram

Schröders, U., Schmidt, C., & Gnambs, T. (2022). Detecting careless responding in survey data using stochastic Gradient boosting. *Educational and Psychological Measurement, 82*(1), 29–56. https://doi.org/10.1177/00131644211004708

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. *Computer Games and Instruction, 55*(2), 503–524.

Stadler, M., Fischer, F., & Greiff, S. (2019). Taking a closer look: An exploratory analysis of successful and unsuccessful strategy use in complex problems. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2019.00777

Stadler, M., Hofer, S., & Greiff, S. (2020). First among equals: Log data indicates ability dif-ferences despite equal scores. *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2020.106442

Stadler, M., Radkowitsch, A., Schmidmaier, R., Fischer, M., & Fischer, F. (2021). Take your time: Invariante of time-on-task in problem-solving tasks across expertise levels. *Psychological Test and Assessment Modeling, 65*(4), 517–525.

Sweller, J., van Merriënboer, J. J., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review, 31*(2), 261–292.

Tenison, C., & Arslan, B. (2020). Characterizing pause behaviors in a science inquiry task. In T. C. Stewart (Ed.), *Proceedings of the 18th International Conference on Cognitive Modeling* (pp. 283–298). Applied Cognitive Science Lab.

Tomasevic, N., Gvozdenovic, N., & Vranes, S. (2020). An overview and comparison of supervised data mining techniques for student exam performance prediction. *Computers & Education*. https://doi.org/10.1016/j.compedu.2019.103676

Tschan, F., Semmer, N. K., Gurtner, A., Bizzari, L., Spychiger, M., Breuer, M., & Marsch, S. U. (2009). Explicit reasoning, confirmation bias, and illusory transactive memory: A simulation study of group medical decision making. *Small Group Research, 40*(3), 271–300. https://doi.org/10.1177/1046496409332928

Ulitzsch, E., Ulitzsch, V., He, Q., & Lüdtke, O. (2022). A machine learning-based procedure for leveraging clickstream data to investigate early predictability of failure on interactive tasks. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-01844-1

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*(4), 197–221. https://doi.org/10.1080/00461520.2011.611369

Vygotsky, L. S. (1978). Mind in society: Development of higher psychological processes. *Harvard University Press*. https://doi.org/10.2307/j.ctvjf9vz4

Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in *C*++ and *R*. *Journal of Statistical Software, 77*(1), 1–17. https://doi.org/10.18637/jss.v077.i01

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Zhu, M., Shu, Z., & von Davier, A. A. (2016). Using networks to visualize and analyze process data for educational assessment: Network analysis for process data. *Journal of Educational Measurement, 53*(2), 190–211. https://doi.org/10.1111/jedm.12107

Ziv, A., Wolpe, P. R., Small, S. D., & Glick, S. (2003). Simulation-based medical education: An ethical imperative. *Academic Medicine, 78*(8), 783–788. https://doi.org/10.1097/00001888-200308000-00006

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.