Psychological Test and Assessment Modeling, Volume 63, 2021 (4), 542-560

# Simulation-Based Learning of Complex Skills: Predicting Performance With Theoretically Derived Process Features

*Laura Brandl[1]*, Constanze Richters[1]; Anika Radkowitsch[2], Andreas Obersteiner[3], Martin R. Fischer[4], Ralf Schmidmaier[5], Frank Fischer[1], Matthias Stadler[1]*

**Abstract**

Simulation-based learning is often used to facilitate complex problem-solving skills, such as collaborative diagnostic reasoning (CDR). Simulations can be especially effective if additional instructional support is provided. However, adapting instructional support to the learners' needs remains a challenge when performance is only assessed as the outcome after using the simulation. Researchers are, therefore, increasingly interested in whether process data analyses can predict outcomes of simulated learning tasks and whether such analyses allow early identification of the need for support. This study developed a random forest classification model based on theoretically derived process indicators to predict success in a simulated learning environment. The context of the simulated learning environment was medicine. Internists interacted with a simulated radiologist to identify possible causes of an illness. Participants' CDR was conceptualized via log-data, coded on a broad, domain-general level for better generalizability. Results showed a satisfactory prediction rate for CDR performance, indicated by diagnostic accuracy. The model predicted accurate and inaccurate diagnoses and was therefore suitable for making statements about the performance by only using process data of CDR. The findings

[1] Department Psychologie, Ludwig-Maximilians-Universität München, Munich, Germany

[2] IPN - Leibniz Institut für die Pädagogik der Naturwissenschaften und Mathematik, Abteilung Didaktik der Mathematik, Kiel, Germany

[3] Heinz Nixdorf-Stiftungslehrstuhl für Didaktik der Mathematik, TUM School of Social Sciences and Technology, Technische Unversität München, Munich, Germany

[4] Institut für Didaktik und Ausbildungsforschung in der Medizin, LMU Klinikum, Ludwig-Maximilians-Universität München, Munich, Germany

[5] LMU Klinikum, Medizinische Klinik und Poliklinik IV, Ludwig-Maximilians-Universität München, Munich, Germany

*Correspondence concerning this article should be addressed to Laura Brandl, Leopoldstr. 13, 80802 München, Germany. E-Mail: L.Brandl@psy.lmu.de

contribute to the development of more adaptive instructional support within simulation-based learning through being able to predict the individuals' learning outcomes already during the process.

Keywords: simulation-based learning, complex problem solving, learning analytics, process-based performance prediction, adaptive instructional support

Simulation-based learning is thought to facilitate complex problem-solving skills (Chernikova, Heitzmann, Fink, et al., 2020). Simulations represent relevant aspects of real-life problems (Grossman, 2021) and can be especially effective if they provide adaptive instructional support (Leutner, 1993). Adaptivity of instructional support is understood as the provision of support adjusted to individuals' specific needs. The aim of adaptive instructional support is twofold: Enhancing learning outcomes and enhancing self-regulation skills concerning learning processes. When a simulation can identify the needs of learners to better self-regulate their learning process and provide adaptive instructional support accordingly, this can allow learners to progress in their learning more efficiently than with non-adaptive support (Plass & Pawar, 2020).

Methods from the field of Learning Analytics seem to be helpful to enable adaptive instructional support because they focus on predicting future outcomes based on behavioral data during the assessment or training process, rather than solely observing the outcome of assessments (Baker & Siemens, 2014). One application of Learning Analytics is the prediction of learning performance using process data, thereby identifying learners at risk of showing inadequate performance (e.g., Gašević, Jovanovic, Pardo, & Dawson, 2017). The present study aims to apply Learning Analytics to the context of collaborative diagnostic reasoning (CDR) in simulation-based learning environments. CDR is an example of a complex problem-solving skill (Fiore et al., 2018) and refers to individual and collaborative skills that enable diagnosticians to diagnose problem states of specific systems (e.g., patients) while working together in teams, based on their conceptual and strategic knowledge (Radkowitsch, M.-R. Fischer, Schmidmaier, & F. Fischer, 2020).

Particularly, we predict CDR performance, indicated by diagnostic accuracy, based on the collaborative diagnostic process derived from existing theoretical models. In addressing this goal, the study serves as preparatory research for developing more adaptive instructional support within simulation-based learning.

## *Simulation-Based Learning of Complex Skills*

Although most complex tasks require intensive training to be performed expertly, many are not easily accessible as training situations as they may be scarce (e.g., natural disasters) or too critical to be approached by novices (e.g., some medical procedures). Simulation-based learning enables the deliberate practice of complex tasks that learners cannot solve (Ericsson, 2004), with the opportunity to provide additional

instructional support. It represents a promising instructional approach to facilitate the development of complex skills by providing authentic situations approximating real-life diagnostic problems (Cook, Brydges, Zendejas, Hamstra, & Hatala, 2013; Heitzmann et al., 2019). As Chernikova, Heitzmann, Stadler, et al. (2020) report in a recent meta-analysis, simulation-based learning significantly fosters complex problem-solving skills.

Complex problem solving is a multidimensional set of skills needed to solve complex problems (Dörner & Funke, 2017). Complex problems require active knowledge acquisition to create a mental representation of the problem (Stadler, Niepel & Greiff, 2019). If complex problems are solved with another person or simulated agent, this process is called collaborative problem solving (Fiore et al., 2018; Stadler, Herborn, Mustafić & Greiff, 2020). One example is CDR, which can be conceptualized as the set of skills to solve a problem, such as diagnosing a patient, "by generating and evaluating evidences and hypotheses that can be shared with, elicited from, or negotiated among collaborators" based on their conceptual and strategic knowledge (Radkowitsch et al., 2020, p. 2). The first entails declarative knowledge about constructs (e.g., diagnoses and symptoms) and their relation, the second is about knowledge of how to apply strategic knowledge through problem-solving (Stark, Kopp & M.-R. Fischer, 2011). The goal of CDR is to reduce the uncertainty of decision-making by diagnosing a phenomenon, such as a patient's symptoms, in a collaborative effort. As such, CDR requires individual diagnostic as well as collaborative processes. To successfully solve a diagnostic problem, diagnosticians draw inferences from latent or hidden patterns of a phenomenon based on their current knowledge (Heitzmann et al., 2019). Heitzmann et al. (2019) described the process of *individual* diagnosing using the scientific reasoning and argumentation framework by F. Fischer et al. (2014), stating that, similar to scientific reasoning, diagnosing can be described with eight epistemic activities (e.g., evidence evaluation, evidence generation, hypothesis generation). In an attempt to extend these considerations to *collaborative* diagnostic processes, Radkowitsch et al. (2020) proposed the CDR model. The CDR model is based on the scientific discovery as dual search model by Klahr and Dunbar (1988) and its further development by Van Joolingen and De Jong (1997) and describes how individual diagnostic processes (F. Fischer et al. 2014) and collaborative activities (Liu et al., 2015) interact with each other. Liu and colleagues (2015) suggest four social skills (sharing ideas, negotiating ideas, regulating problem-solving, and maintaining communication) to describe collaborative activities. One of the main functions of the collaborative activities is to construct a shared problem representation (Roschelle & Teasley, 1995) through sharing and eliciting relevant information, as information might not be distributed equally between all collaborators. Hence, it is crucial to accurately share all relevant information to diagnose the patient‘s illness. These activities seem particularly relevant in a field such as medicine in which physicians from different fields of expertise collaborate frequently. In such situations, it is crucial for an accurate diagnosis of the patient's problem that all relevant evidence and hypotheses for the specific collaborators are shared (Kiesewetter, F. Fischer , & M.-R. Fischer, 2017).

The CDR model specifies such collaborative diagnostic processes by suggesting collaborative diagnostic activities (CDAs). CDAs combine individual and collaborative diagnostic activities such as evidence elicitation, evidence sharing, and hypotheses sharing. Evidence and hypotheses, which are results of individual diagnostic processes and stored in an individual's cognitive storage (see Klahr & Dunbahr, 1988), can become part of *collaborative* cognitive processes by, for instance, sharing or eliciting them. In the medical context, evidence is, for example, patient information about symptoms and other parameters which are identified as relevant for a diagnosis. A hypothesis is a suspected diagnosis that refers to an underlying illness that could explain the patient´s symptoms. Evidence elicitation is, then, the activity of collaboratively generating new information, for example, by conducting medical examinations like radiological tests (Radkowitsch et al., 2020). Adequate performance of CDR in the context of medicine is defined as performing those activities with high quality resulting in an accurate diagnosis (Tschan et al., 2009). However, there is currently no assumption about the linearity and sequence of the performance of CDAs required to reach an accurate diagnosis, and not all CDAs might be necessary for all collaborative diagnostic scenarios.

In summary, simulation-based learning offers a promising approach for the training of complex problem-solving skills, such as CDR, by providing authentic diagnostic situations for learners to engage in (Chernikova, Heitzmann, Stadler, et al., 2020; H. G. Schmidt & Rikers, 2007) while allowing to provide adequate instructional support. However, adapting these support measures (such as prompts or worked-out examples; Belland, 2017) to the learners' needs remains a challenge because it requires assessing the learner's current knowledge during the simulation rather than after using the simulation. Analyzing data stemming from the CDR process to inform a learner model (Ding, Zhu, & Guo, 2018) while the learner is still working on the simulation might lead to more timely support when necessary.

## *Learning Analytics and Process Data in Simulation-Based Learning*

Using technologically-enhanced simulations that store data on the learning process immediately in log-files allows analyzing process data without the need for additional assessments with dedicated tests. Analyzing process data instead of only product data (the assessment's outcome) allows insights into the process leading to the eventual outcome (e.g., Goldhammer, Naumann, Rölke, Stelter, & Tóth, 2017). Widely used process data is often not at all straightforward to interpret. For example, more time spent on a task may indicate cognitive factors (i.e., the tasks are challenging) or motivational factors (i.e., tedious tasks). Nevertheless, process data analyses can increase understanding of the analyzed process (Greiff, Niepel, Scherer, & Martin, 2016). The results can be used to improve the theoretical understanding of the processes involved and approaches to assessing and facilitating them (Goldhammer, Naumann, Stelter, Tóth, Rölke & Klieme, 2014).

Using process data allows for the prediction of performance, enabling researchers to identify learners at risk to show inadequate performance, such as to benefit little from engaging in a learning activity (e.g., Leitner, Khalil, & Ebner, 2017), and to provide them with additional instructional support (e.g., scaffolding, Tabak & Kyza, 2018). Such support is ideally timed and adapted to the learners' needs (Plass & Pawar, 2020). Previous research has shown that the number of clicks and the time on task can be predictive for task success (Goldhammer et al., 2017). Stadler, Hofer, and Greiff (2020) analyzed differences between the time-on-task and the number of clicks of participants having the same outcome in a simulation of complex problem-solving. Despite having equal scores, participants differed in both time-on-task and number of clicks. The results indicate that process indicators depict individual differences in the ability not depicted in product data. This illustrates the need to take process data into account to assess learners' abilities. This is also in line with the assumption that complex problem-solving is not only about a task's outcome but also about the process to get there (Dörner & Funke, 2017).

However, it is difficult to deduct information on specific problems a learner might have with a task or what instructional support might be beneficial using process data. Therefore, researchers have called for a more robust link from process data to learning theories to understand better and facilitate learning (Gašević, Dawson, & Siemens, 2015). The identification of suitable features for the prediction of learning outcomes within process data should always be supported with theoretical models (Tomasevic, Gvozdenovic, & Vranes, 2020) in order to make findings replicable and generalizable beyond idiosyncratic learning environments.

## *Goal and Research Question*

The current study uses activities theoretical derived from the CDR model (Radkowitsch et al., 2020) and constructed from process data to predict the performance of complex problem-solving skills, such as CDR, in simulation-based learning. It addresses the research question to what extent theoretically derived process indicators are suitable to predict learners' diagnostic accuracy in the context of simulation-based learning of CDR. Since CDR frequently occurs in medical settings and has been identified to be a significant challenge for physicians (e.g., Tschan et al., 2009; Brady et al., 2012), the simulation was embedded in the context of medical education and developed based on the CDR model. Three CDAs proposed in the CDR model are particularly relevant in the simulated situation: evidence elicitation, evidence sharing, and hypotheses sharing. Hence, the current study investigates to what extent diagnostic accuracy can be predicted using the CDAs constructed from process data of a simulated learning environment in medical education. In addressing this research question, the current study contributes to developing more adaptive instructional support within simulation-based learning through showing the possibilities of learning analytics methods, being able to predict the outcome already in the process.

## Method

### *Simulation and Learner Task*

The simulation was integrated into CASUS (https://www.instruct.eu/; M. R. Fischer, Aulinger, & Baehring, 1999), a case-based learning platform, where learners worked on five different patient cases within the simulation. Medical experts from internal medicine, radiology, and general medicine constructed the patient cases. In the simulation, the learners' task was to interact with an agent-based (i.e., simulated) radiologist to diagnose fictitious patient cases suffering from unknown fever. To that end, learners requested further information about the patient from the radiologist who conducted radiological examinations. This required learners to engage in the CDA evidence elicitation, evidence sharing, and hypotheses sharing. Medical experts who supported the development of the situation considered these CDAs as particularly important for the specific collaborative diagnostic situation. The collaboration took place after the learners studied a health record (containing all current information about the patient). The collaboration consisted of filling out a radiological request form and receiving the requested results from the simulated radiologist only if the request form contained sufficient evidence and hypotheses relevant for the radiologist to conduct and interpret the radiologic test. Specifically, learners needed to elicit evidence by choosing an exam method to be performed by the radiologist, sharing evidence by choosing information from the health record relevant for the radiologist, and sharing suspected diagnoses as hypotheses. After the collaboration, learners were asked to indicate their final diagnosis individually. For a detailed description of the simulation's development and validation, see Radkowitsch et al. (2020).

### *Sample and Design*

Data for this study was taken from a more extensive experimental study conducted within the COSIMA Project. The study's design was an experimental setting with four groups investigating the effect of different kinds of instructional support. One group received an adaptive collaboration script; one was encouraged to have reflection phases, one both kinds of support, and the control group received none of them. In order to avoid confounding effects of the experimental conditions for the current study, only the control condition was used for the current analyses. Data was collected online from 9 male and 26 female intermediate learners from the $4^{th}$ – $6^{th}$ year of medical studies. In total, the study program includes six years of studying. Learners had an average age of $M = 25.43$ years ($SD = 2.54$ years) and studied medicine on average in their $M = 5^{th}$ year ($SD = 0.76$ years). Learners were recruited through an email distribution list and flyers. For full participation, learners received 10€ compensation per hour of testing. In line with the university's ethics requirements, participation was voluntary, and learners could terminate participation at any time. Given the focus of the study on the CDR process, the unit of analysis was the patient case and

not the participant. As learners worked on five patient cases, this led to a total of $n = 167$ after excluding missing data on diagnostic accuracy. The ethics committee of the medical faculty of LMU Munich declared ethical clearance prior to data collection (approval number 18-262).

## Measures

### *Diagnostic Accuracy*

Each patient case is assigned to one primary diagnosis, consented by experts. After working on the patient case and requesting a radiological examination, the learners indicated their final diagnosis using a free text field with suggested options out of a list of 249 diagnoses, based on the first letters entered, to shorten and standardize the input. Diagnostic accuracy was calculated by coding the final diagnosis's compliance with the expert solution. To that end, two independent coders each coded the complete data. Differences in the coding were discussed until all codes were identical. Accurate diagnoses were coded with 1, while inaccurate diagnoses were coded with 0. For example, when the patient suffers from hospital-acquired pneumonia, this diagnosis would be coded with 1, while only pneumonia or any other diagnosis would be coded with 0.

### *Process Data*

Every click in the simulation leading to an interaction with the system was stored with the corresponding timestamp in log file data allowing for analyzing process data. Based on the CDR model, the CDAs were coded depending on the learners' entries to a radiological request form during the collaboration with the simulated radiologist. Every activity where the learners selected a radiological examination by choosing a method and the body part to examine was coded as *evidence elicitation*. Every activity where the learners shared information from the health record to justify the radiological examination was coded as *evidence sharing*. Every activity where the learners indicated a potential diagnosis was coded as *hypotheses sharing*. Diagnoses were entered using a free text field with suggested options out of a list of 249 diagnoses, based on the first letters entered, to shorten and standardize the input. To illustrate this process, we will give an example of how a learner could have filled out the request form: The learner started to fill out the request form by choosing an x-ray of the chest as a radiological examination (evidence elicitation). This requires the learner to make two clicks in the simulation, one for selecting a method and another for selecting the respective body part. Next, the learner justified the decision for the examination method by ticking information presented in the health record (evidence sharing). In this example, the learner shared that the patient has decreased breathing sound, fever, is

male, and is a smoker. The learner identified and ticked the respective box to share that evidence, including this information. Lastly, the learner typed 'pneu' into the free text field on the bottom of the form. The system offered possible diagnoses starting with 'pneu' (e.g., pneumonia; community-acquired pneumonia; hospital-acquired pneumonia), the learner chose the share 'pneumonia' as a hypothesis with the simulated radiologist. Before sending the form, the learner decided to additionally share the evidence that the patient has an increased lymphocyte value.

First, the clicks in the simulation were coded automatically according to the CDAs using spreadsheet software. Then, each coded activity was decomposed into the number of seconds a participant spent on the activity. The activities coded in units of seconds were then summarized into behavioral strings that indicated, per learner and case, which CDA was performed, how long, and what activity followed. This information was stored in a string variable.

## *Analyses*

The proper selection of features is essential in prediction models. When process data depicts long sequences, exploratory approaches such as the *n-gram* method proposed by Damashek (1995) can be helpful. Here the process of activities is summarized as a sequence of n consecutive elements. This allows representing the sequence of activities as well as their frequency. For this study, we chose bigrams (n = 2) to ensure there are not too many different features in our prediction models. The bigrams represented either consistent activity (two instances of the same activity) or transitions from one behavior to another (two different activities). To apply the n-gram method, the string variable representing an individual's sequence of activities was separated in bigrams using the *n-gram* package in R (3.0.4; D. Schmidt & Heckendorf, 2017), leading to nine features constructed from the three theoretical derived activities, each summarizing how often this specific bigram occurred in the string variable.

Referring back to the previous example, the learner spent 60 seconds on evidence elicitation, which resulted in 59 instances of the EE.EE bigram. Further, the learner spent 200 seconds at the beginning and 6 seconds with evidence sharing when they returned to that activity after sharing the hypothesis resulting in 204 instances of the ES.ES bigram. Spending 150 seconds with hypotheses sharing results in 149 instances of the HS.HS bigram. Those three bigrams indicate consistent activity. Looking at transitions, the learner had a value of one on the bigrams EE.ES, ES.HS, HS.ES indicating changes between evidence elicitation and evidence sharing, evidence sharing and hypotheses sharing, as well as hypotheses sharing and evidence sharing, respectively.

For predicting diagnostic accuracy using bigrams of CDAs, the statistical software R (RStudio Team, 2020) was used. The essential packages were *ranger* (0.12.1; Wright & Ziegler, 2017) and *caret* (6.0-86; Kuhn, 2008). A random forest classification model (ranger algorithm; Wright & Ziegler, 2017) was developed to answer the research question. This model was chosen as it is highly accurate and able to deal with

relatively large numbers of features and few data points while considering complex interactions among the features. In contrast to more interpretable logistic regression models, random forest classification models are also less affected by multicollinearity issues (Breiman, 2001; Fernández-Delgado, Cernadas, Barro, & Amorim, 2014).

First, the data set was split into a training set (including 75 % of the data) and a testing set (including 25 % of the data). The training set was then used to fit the prediction model. To increase the model fit, hyperparameters were tuned automatically. A 10x3 cross-validation was applied to identify the hyperparameters to decrease the risk of overfitting. For the ranger algorithm, only the number of randomly selected predictors (mtry), the split rule (gini or extra trees), and the minimum node size needed to be determined (Kuhn, 2008). The prediction model was evaluated in the testing set and the training set using a confusion matrix (Buskirk, Kirchner, Eck, & Signorino, 2018). To assess classification quality of the prediction model classification accuracy (the total percentage of correct classifications), sensitivity (true positive classification relative to all positive classifications), and specificity (true negative classification relative to all negative specification), no-information rate (always predicting the most common class), and a one-sided significance test to see whether the developed model outperforms the no-information rate was evaluated (Alpaydin, 2010; Kuhn, 2008). Kappa, the agreement between predicted values and the actual data in relation to expected values by chance, is assessed, with a value of greater than .61 indicating sufficient strength of agreement (Landis & Koch, 1977).

Finally, a closer look into how each feature influenced the classification was done using feature importance. Due to complex interactions among different features, the interpretation of importance is not always straightforward and can only be done in relation to other features in the model, not by applying standardized cut off values (Kuhn, 2008; Liaw & Wiener, 2002; Strobl, Boulesteix, Zeileis, & Hothorn, 2007). The dataset and the code for the analyses are uploaded to the open science framework (OSF) repository and can be retrieved from https://osf.io/y6bfx/

## Results

Before looking at the predictability of diagnostic accuracy using process data, the used features are presented descriptively in Table 1. The bivariate correlation between the features and diagnostic accuracy is only minor, ranging from -.06 to .11.

**Table 1**

*Descriptive Results of the Features used for Prediction Diagnostic Accuracy*

| Feature | Accurate Diagnoses | | Inaccurate Diagnoses | | $r_p$ |
|---|---|---|---|---|---|
| | Median | Range | Median | Range | |
| EE.EE | 25.5 | 6 - 342 | 36.0 | 2 - 429 | -.05 |
| ES.ES | 146.0 | 0 - 587 | 135.5 | 0 - 581 | .11 |
| HS.HS | 79.0 | 0 - 568 | 67.5 | 0 - 520 | -.02 |
| EE.ES | 1.0 | 0 - 6 | 1.0 | 0 - 6 | .01 |
| EE.HS | 0.0 | 0 - 5 | 0.0 | 0 - 4 | .03 |
| ES.EE | 0.0 | 0 - 3 | 0.0 | 0 - 4 | -.04 |
| ES.HS | 0.0 | 0 - 4 | 0.0 | 0 - 5 | .05 |
| HS.EE | 0.0 | 0 - 4 | 1.0 | 0 - 4 | -.02 |
| HS.ES | 0.0 | 0 - 2 | 0.0 | 0 - 3 | .06 |

*Note.* EE = evidence elicitation, ES = evidence sharing, HS = hypotheses sharing

$r_p$ = Pearson correlation between feature and diagnostic accuracy

Investigating the predictability of diagnostic accuracy using process indicators, depicted through bigrams of CDAs, the identified random forest classification model (mtry = 2, splitrule = extra trees, min node size = 1) performed well. Classification accuracy of .98 (95 % CI [.93; 1.00]) was found for the training set, indicating strong predictive power. The results of the one-sided hypothesis test indicated that the developed model was significantly better than the no-information rate model (accuracy of .54, $p < .001$). The kappa for the model was .95, implying high agreement between the predicted values by the model and the actual data (Landis & Koch, 1977). Further evaluation revealed a sensitivity of .95 and a specificity of 1.00, indicating that the model could correctly predict accurate and inaccurate diagnoses in most cases.

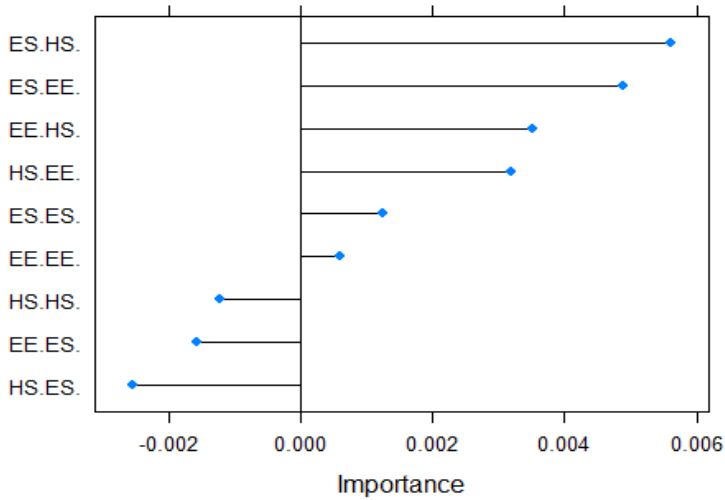
When using the testing set, the results supported the good ability of the model to predict diagnostic accuracy, with a predictive accuracy of .95 (95 % CI [.84;.99]) and a no-information rate of .73. The classification was also significantly better than the no-information model ($p <.001$) for the testing set. Results implied, again, a high agreement between the predicted values by the model that was trained based on the training

sample and the data of the testing sample with a kappa of .88. The additional evaluation metrics further indicated a sensitivity of .91 and specificity of .97, slightly worse than in the training set. Nevertheless, both measures indicated a high capacity of the model to predict accurate and inaccurate diagnoses in both the training and the testing data set.

Looking at the importance of the different features (see Figure 1), the most important one was the transition from evidence sharing to hypothesis sharing. This is followed by the transition from evidence sharing to evidence elicitation and the transition from evidence elicitation to hypotheses sharing. The fourth most important feature is the transition from hypothesis sharing to evidence elicitation. All those transitions are entailed in the process of CDR.

**Figure 1**

*Importance of Features Predicting Diagnostic Accuracy Using Process Data*




Note. EE = evidence elicitation, ES = evidence sharing, HS = hypotheses sharing

## Discussion

The current study aimed at investigating to what extent theoretically derived process indicators are suitable to predict learners' diagnostic accuracy (performance measure) in the context of simulation-based learning of CDR. A random forest algorithm

classified accurate and inaccurate diagnoses correctly based on bigrams of CDAs. The model predicted a large percentage of accurate and inaccurate diagnoses and is, therefore, suitable to support statements about the performance only using process data. This is in line with former research (e.g., Mahboob, Irfan, & Karamat, 2016), indicating that algorithms from the field of Learning Analytics are suitable for performance prediction.

Learning performance and its enhancement are widely investigated in Learning Analytics (Leitner et al., 2017). However, most of the studies lacked a theoretical grounding of their approach (Gašević et al., 2015). The present study used features for the prediction of diagnostic accuracy that were derived from the CDR model by Radkowitsch et al. (2020), which is theoretically rooted in well established theoretical frameworks (e.g., Klahr & Dunbahr, 1988; F. Fischer et al., 2014; Liu et al., 2015). The current results underline the relevance of epistemic activities, such as CDAs, and their sequences for diagnostic processes. However, so far, the CDR model does not consider predictions about the relation between the CDAs and diagnostic accuracy. It is only conceptualizing CDAs as part of the CDR process, which needs to be performed with high quality to draw an accurate final decision. Using Learning Analytics, we showed that the CDAs are relevant for diagnostic accuracy, being a performance indicator of CDR, even though the bivariate correlations between the bigrams and diagnostic accuracy were only minor.

The clear benefit of using machine-learning prediction models instead of traditional statistical models is the change of perspective. While the latter is concerned about explaining causal relationships and therefore has a retrospective view on the data, the former has the goal of predicting future data and therefore has a prospective view (Yarkoni & Westfall, 2017). Accordingly, predictive accuracy is the primary goal, and the ratio of bias and variance, which minimize the occurring error the best, should be chosen. In order to achieve this, one must be willing to allow for bias and nonlinearity for the sake of accurate prediction (Molnar et al., 2020; Yarkoni & Westfall, 2017). This focus on predictive accuracy can make prediction models, especially ensemble methods such as random forests, highly complex, resulting in accurate predictions but lacking an explanation of how they were achieved, leading to less transparent models, also known as black boxes (Molnar et al., 2018; Yarkoni & Westfall, 2017). There is a need to investigate non-linear relations between process indicators to enhance theoretical models. The current results highlight the relevance of theoretically derived process indicators for the performance of CDR in simulation-based learning and can be used to predict the performance of complex problem-solving skills in simulation-based learning already in the process. Such predictions may help provide learners with inadequate performance with additional (adaptive) instructional support. From the feature importance plot, we can see that the consistent features (e.g., time spent with evidence elicitation) and transitions from evidence elicitation to evidence sharing and from hypotheses sharing to evidence sharing are relatively unimportant. Future analyses should therefore focus less on these processes and more on the transitions from evidence sharing to hypotheses sharing, from evidence sharing to evidence elicitation and from evidence elicitation to hypotheses sharing and hypotheses

sharing to evidence elicitation. The most important feature is the transition from evidence sharing to hypotheses sharing. However, the feature did not differ considerably regarding accurate and inaccurate diagnoses. Therefore, a non-linear relationship or a complex interaction with one or more other features is assumed, which needs to be further investigated. However, we currently do not know precisely what indicates inadequate performance and how to foster it accordingly, as the interpretation of black-box models and feature importance is not straightforward, and the prediction is not linear but a result of complex interactions.

Nevertheless, the current study was able to show that predicting the performance in complex simulation-based learning environments based on theoretically derived indicators of behavior is possible, even if there are no linear correlations between behavior and performance. Since we were able to demonstrate a relation between the theoretically derived process indicators and the performance of CDR in simulation-based learning, the next step should be to investigate sequences of activities in depth, e.g., with sequence clustering (Piccarreta, 2017), allowing not only to identify learners who need additional instructional support but also to provide this support.

The current results are not limited to learning of CDR in the medical context but likely generalize to related fields such as teacher education (Heitzmann et al., 2019) and complex problem-solving skills in different domains as the indicators of behavior were coded on a domain-general level.

## *Limitations and Future Research*

The current study is not without limitations, which must be kept in mind when interpreting the results. First, it must be considered that all patient cases were analyzed independently, regardless of the order in which they appeared in the simulation, thus ignoring potential learning effects between the cases. Statistically speaking, this approach risks ignoring non-negligible random effects due to the clustered nature of the data. Extensions of the random forest algorithm have been proposed that consider clustered data (Hajjem, Bellavance, & Larocque, 2014). However, since our model performed exceptionally well, the intra-class correlation among participants is likely very low even without this extension. Another limitation is that only data from learners with an intermediate level of expertise was collected, limiting the observation of full expert and novice behavior. However, data showed a balanced frequency of accurate and inaccurate diagnoses. Future research might investigate whether participants of different expertise levels employ different strategies for their collaborative diagnosing, which would likely require an algorithm capable of including this information as an additional level of data.

Another potential limitation lies in the decision to observe only bigrams rather than n-grams that are more complex. N-grams that are more complex might provide further insights into more advanced strategies and might be more interpretable towards necessary support. However, the number of features increases exponentially with the length of observed n-grams. Even trigrams might have resulted in too many ($3^3 = 27$)

features for our limited sample. Future research might investigate longer n-grams using larger samples. An alternative would be the theoretical definition of specific sequences as predictors to explicitly test hypotheses on strategic behavior in a simulation. In line with the current results, the focus should be on the transitions between activities rather than on consistent behavior.

To help learners who potentially show inadequate performance as early as possible, future research will also need to investigate how early it is possible to predict the performance of complex problem-solving skills using process data. In addition, future research may also investigate how additional instructional support could look like. For example, Azevedo, Moos, Cromley, and Greene (2011) demonstrated that a combination of content and process-oriented adaptive scaffolding is suitable to facilitate self-regulated learning.

Currently, there is only little known about sequences of CDAs and their relation with diagnostic accuracy. However, we could show in this study that there are non-linear relations between those process indicators and learning performance. Future research should deepen this by investigating the transitions between activities to make further claims on refining existing theoretical process models. This is in line with the call for *explanatory learner models* that focus on optimal predictions using black-box models but use more interpretable methods to gain deeper insights into learning (Rosé et al., 2019). One approach in this context is the use of different kinds of data, such as process (e.g., log-file data), product (e.g., the outcome of a task), and learner data (e.g., self-report measures) using dispositional learning analytics (Buckingham Shum & Crick, 2021). This combination of data sources allows improving the design of adaptive scaffolding and interventions as it provides more profound insights into the origins of underperforming (Gašević et al., 2017). For example, Tempelaar, Rienties, and Nguyen (2021) combined this approach with a person-oriented type of research (instead of the traditional variable-oriented type) to identify five different learning profiles based on only learner data at the beginning and then by including more process data in a stepwise manner. This allows providing instructional support not only for a group of learners or an average learner but also for a specific individual learner, that is, personalized learning support.

## *Conclusion*

This study aimed to predict CDR performance using process data, indicated by diagnostic accuracy. Results show that using a Learning Analytics approach, a random forest prediction model, is suitable for predicting performance using process indicators theoretically derived and constructed from process data. Using Learning Analytics enables researchers to provide practical solutions such as identifying learners at risk to show inadequate performance in need of adaptive instructional support. The findings contribute to the development of more adaptive instructional support within simulation-based learning through being able to predict the individuals' learning outcomes already during the process.

## *Acknowledgments*

This research was supported by a grant from the Deutsche Forschungsgemeinschaft DFG (COSIMA; DFG-Forschungsgruppe 2385).

## References

Alpaydin, E. (2010). Introduction to machine learning (2nd ed.). Adaptive computation and machine learning. Cambridge Mass.: MIT Press.

Azevedo, R., Moos, D. C., Cromley, J. G., & Greene, J. A. (2011). Adaptive Content and Process Scaffolding: A key to facilitating students' self-regulated learning with hypermedia. Psychological Test and Assessment Modeling, 53(1), 106–140.

Baker, R., & Siemens, G. (2014). Educational Data Mining and Learning Analytics. In R. K. Sawyer (Ed.), The Cambridge Handbook of the Learning Sciences (pp. 253–272). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139519526.016

Belland, B. R. (2017). Instructional Scaffolding in STEM Education. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-02565-0

Brady, A., Laoide, R. Ó., McCarthy, P., & McDermott, R. (2012). Discrepancy and error in radiology: Concepts, causes and consequences. Ulster Medical Journal, 81(1), 3.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

Buckingham Shum, S., & Deakin Crick, R. (2012). Learning dispositions and transferable competencies. In S. Dawson & C. Haythornthwaite (Eds.), Proceedings of the 2nd international conference on learning analytics and knowledge - lak '12 (pp. 1–10). ACM Press. https://doi.org/10.1145/2330601.2330629

Buskirk, T. D., Kirchner, A., Eck, A., & Signorino, C. S. (2018). An Introduction to Machine Learning Methods for Survey Researchers. Survey Practice, 11(1), 1–10. https://doi.org/10.29115/SP-2018-0004

Chernikova, O., Heitzmann, N., Fink, M. C., Timothy, V., Seidel, T., & Fischer, F. (2020). Facilitating Diagnostic Competencies in Higher Education—a Meta-Analysis in Medical and Teacher Education. Educational Psychology Review, 32(1), 157–196. https://doi.org/10.1007/s10648-019-09492-2

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

Cook, D. A., Brydges, R., Zendejas, B., Hamstra, S. J., & Hatala, R. M. (2013). Technology-enhanced simulation to assess health professionals: A systematic review of validity evidence, research methods, and reporting quality. Academic Medicine: Journal of the Association of American Medical Colleges, 88(6), 872–883. https://doi.org/10.1097/ACM.0b013e31828ffdcf

Damashek, M. (1995). Gauging Similarity with n-Grams: Language-Independent Categorization of Text. Science (New York, N.Y.), 267(5199), 843–848. https://doi.org/10.1126/science.267.5199.843

Ding, W., Zhu, Z., & Guo, Q. (2018). A New Learner Model in Adaptive Learning System. In 3rd International Conference on Computer and Communication Systems (ICCCS) (pp. 440–443). IEEE. https://doi.org/10.1109/CCOMS.2018.8463316

Dörner, D., & Funke, J. (2017). Complex Problem Solving: What It Is and What It Is Not. Frontiers in Psychology, 8(1153). https://doi.org/10.3389/fpsyg.2017.01153

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, 3133–3181.

Fiore, S. M., Graesser, A., & Greiff, S. (2018). Collaborative problem-solving education for the twenty-first-century workforce. Nature Human Behaviour, 2(6), 367–369. https://doi.org/10.1038/s41562-018-0363-y

Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., . . . 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

Fischer, M. R., Aulinger, B., & Baehring, T. (1999). Computer-based-Training (CBT). Fallorientiertes Lernen am PC mit dem CASUS/ProMediWeb-System [Computer-based training (CBT). Case-oriented learning on the PC with CASUS/ProMediWeb System]. Deutsche medizinische Wochenschrift (1946), 124(46), 1401. https://doi.org/10.1055/s-2007-1024550

Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. TechTrends, 59(1), 64–71. https://doi.org/10.1007/s11528-014-0822-x

Gašević, D., Jovanovic, J., Pardo, A., & Dawson, S. (2017). Detecting Learning Strategies with Analytics: Links with Self-reported Measures and Academic Performance. Journal of Learning Analytics, 4(2), 113–128. https://doi.org/10.18608/jla.2017.42.10

Goldhammer, F., Naumann, J., Rölke, H., Stelter, A., & Tóth, K. (2017). Relating Product Data to Process Data from Computer-Based Competency Assessment. In D. Leutner, J. Fleischer, J. Grünkorn, & E. Klieme (Eds.), Methodology of Educational Measurement and Assessment. Competency Assessment in Education (pp. 407–425). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-50030-0_24

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

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

Grossman, P. (2021). Teaching core practices in teacher education. Harvard Education Press.

Hajjem, A., Bellavance, F., & Larocque, D. (2014). Mixed-effects random forest for clustered data. Journal of Statistical Computation and Simulation, 84(6), 1313–1328. https://doi.org/10.1080/00949655.2012.741599

Heitzmann, N., Seidel, T., Hetmanek, A., Wecker, C., Fischer, M. R., Ufer, S., . . . Opitz, A. (2019). Facilitating Diagnostic Competences in Simulations in Higher Education A Framework and a Research Agenda. Frontline Learning Research, 1–24. https://doi.org/10.14786/flr.v7i4.384

Kiesewetter, J., Fischer, F., & Fischer, M. R. (2017). Collaborative clinical reasoning-a systematic review of empirical studies. The Journal of Continuing Education in the Health Professions, 37(2), 123−128. https://doi.org/10.1097/CEH.0000000000000158

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. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software,, 28(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. Biometrics. (33), 159–174.

Leitner, P., Khalil, M., & Ebner, M. (2017). Learning Analytics in Higher Education—A Literature Review. In A. Peña-Ayala (Ed.), Studies in Systems, Decision and Control. Learning Analytics: Fundaments, Applications, and Trends (Vol. 94, pp. 1–23). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-52977-6_1

Leutner, D. (1993). Guided discovery learning with computer-based simulation games: Effects of adaptive and non-adaptive instructional support. Learning and Instruction, 3(2), 113–132. https://doi.org/10.1016/0959-4752(93)90011-N

Liaw, A., & Wiener, M. (2002). Classification and Regression by RandomForest. R News. 2(3), 18–22.

Liu, L., Hao, J., Davier, A. von, Kyllonen, P., & Zapata-Rivera, D. (2015). A tough nut to crack: Measuring collaborative problem solving. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), Handbook of research on technology tools for real-world skill development (pp. 344–359). Hershey, PA: IGI Global. https://doi.org/10.4018/978-1-4666-9441-5.c

Mahboob, T., Irfan, S., & Karamat, A. (2016). A machine learning approach for student assessment in E-learning using Quinlan's C4.5, Naive Bayes and Random Forest algorithms, 2016 19th International Multi-Topic Conference (INMIC), pp. 1-8. https://doi.org/10.1109/INMIC.2016.7840094.

Molnar, C., Casalicchio, G., & Bischl, B. (2020). Interpretable machine learning − a brief history, state-of-the-art and challenges. In I. Koprinska, M. Kamp, A. Appice, C. Loglisci, L. Antonie, A. Zimmermann, R. Guidotti, Ö. Özgöbek, R. P. Ribeiro, R. Gavaldà, J. Gama, L. Adilova, Y. Krishnamurthy, P. M. Ferreira, D. Malerba, I. Medeiros, M. Ceci, G. Manco, E. Masciari, . . . J. A. Gulla (Eds.), Communications in Computer and Information Science. ECML PKDD 2020 Workshops (Vol. 1323, pp. 417−431). Springer International Publishing. https://doi.org/10.1007/978-3-030-65965-3_28

Piccarreta (2017). Joint Sequence Analysis: Association and Clustering. Sociological Methods and Research, 46(2), 252–287. doi: 10.1177/0049124115591013.

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

Radkowitsch, A., Fischer, M. R., Schmidmaier, R., & Fischer, F. (2020). Learning to diagnose collaboratively: validating a simulation for medical students. GMS Journal for Medical Education, 37(5). https://doi.org/10.3205/zma001344

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 Berlin Heidelberg. https://doi.org/10.1007/978-3-642-85098-1_5

Rosé, C. P., McLaughlin, E. A., Liu, R., and Koedinger, K. R. (2019). Explanatory Learner Models: Why Machine Learning (Alone) Is Not the Answer. Br. J. Educ. Technol. 50, 2943–2958. https://doi.org/10.1111/bjet.1285

RStudio Team (2020). RStudio: Integrated Development for R. RStudio. Bosten, MA: PBC: PBC. Retrieved from http://www.rstudio.com/

Schmidt, D., & Heckendorf, C. (2017). ngram: Fast n-Gram Tokenization. Retrieved from https://cran.r-project.org/package=ngram

Schmidt, H. G., & Rikers, R. M. J. P. (2007). How expertise develops in medicine: knowledge encapsulation and illness script formation. Medical Education, 41(12), 1133–1139. https://doi.org/10.1111/j.1365-2923.2007.02915.x

Stadler, M., Herborn, K., Mustafić, M., & Greiff, S. (2020). The assessment of collaborative problem solving in PISA 2015: An investigation of the validity of the PISA 2015 CPS tasks. Computers & Education, 157, 103964.

Stadler, M., Hofer, S., & Greiff, S. (2020). First among equals: Log data indicates ability differences despite equal scores. Computers in Human Behavior, 111, 106442. https://doi.org/10.1016/j.chb.2020.106442

Stadler, M., Niepel, C., & Greiff, S. (2019). Differentiating between static and complex problems: A theoretical framework and its empirical validation. Intelligence, 72, 1-12. https://doi.org/10.1016/j.intell.2018.11.003

Stark, R., Kopp, V., & Fischer, M. R. (2011). Case-based learning with worked examples in complex domains: Two experimental studies in undergraduate medical education. Learning and Instruction, 21(1), 22−33.https://doi.org/10.1016/j.learninstruc.2009.10.001

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics, 8. https://doi.org/10.1186/1471-2105-8-25

Tabak, I., & Kyza, E. A. (2018). Research on Scaffolding in the Learning Sciences. In F. Fischer, C. E. Hmelo-Silver, S. R. Goldman, & P. Reimann (Eds.), International Handbook of the Learning Sciences (pp. 191–200). New York, NY : Routledge, 2018.: Routledge. https://doi.org/10.4324/9781315617572-19

Tempelaar, D., Rienties, B. & Nguyen, Q. (2021). Dispositional Learning Analytics for Supporting Individualized Learning Feedback. Frontiers in Education, 6. https://doi.org/10.3389/feduc.2021.703773

Tomasevic, N., Gvozdenovic, N., & Vranes, S. (2020). An overview and comparison of supervised data mining techniques for student exam performance prediction. Computers & Education, 143, 103676. 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

Van Joolingen, W. R., & De Jong, T. (1997). An extended dual search space model of scientific discovery learning. Instructional Science, 25(5), 307-346. https://doi.org/10.1023/A:1002993406499

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: A Journal of the Association for Psychological Science, 12(6), 1–23. https://doi.org/10.1177/1745691617693393