Thimo Marcin, Ailin Lüthi, Ronny R. Graf, Gert Krummrey, Stefan K. Schauber, Neal Breakey, Wolf E. Hautz and Stefanie C. Hautz*

# Is language an issue? Accuracy of the German computerized diagnostic decision support system ISABEL and cross-validation with the English counterpart

https://doi.org/10.1515/dx-2023-0047
Received April 25, 2023; accepted June 16, 2023;
published online July 24, 2023

**Abstract**

**Objectives:** Existing computerized diagnostic decision support tools (CDDS) accurately return possible differential diagnoses (DDx) based on the clinical information provided. The German versions of the CDDS tools for clinicians (Isabel Pro) and patients (Isabel Symptom Checker) from ISABEL Healthcare have not been validated yet.
**Methods:** We entered clinical features of 50 patient vignettes taken from an emergency medical text book and 50 real cases with a confirmed diagnosis derived from the electronic health record (EHR) of a large academic Swiss emergency room into the German versions of Isabel Pro and Isabel Symptom Checker. We analysed the proportion of DDx lists that included the correct diagnosis.
**Results:** Isabel Pro and Symptom Checker provided the correct diagnosis in 82 and 71 % of the cases, respectively. Overall, the correct diagnosis was ranked in 71, 61 and 37 % of the cases within the top 20, 10 and 3 of the provided DDx when using Isabel Pro. In general, accuracy was higher with vignettes than ED cases, i.e. listed the correct diagnosis more often (non-significant) and ranked the diagnosis significantly more often within the top 20, 10 and 3. On average, 38 ± 4.5 DDx were provided by Isabel Pro and Symptom Checker.
**Conclusions:** The German versions of Isabel achieved a somewhat lower accuracy compared to previous studies of the English version. The accuracy decreases substantially when the position in the suggested DDx list is taken into account. Whether Isabel Pro is accurate enough to improve diagnostic quality in clinical ED routine needs further investigation.

**Keywords:** clinical decision support; diagnostic accuracy; diagnostic decision support; diagnostic error; differential diagnosis generator; emergency medicine

## Introduction

Getting the right diagnosis is a key aspect of health care. It provides an explanation of a patient's health problem and informs subsequent health care and treatment. Misdiagnosis occurs in about 5–10 % of emergency room (ED) patients, sometimes with devastating medical and economic consequences [1–3]. The causes for diagnostic error can be diverse, but one major cause is human error. This includes the consideration of incomplete patient histories, failure to consider alternatives, lack of knowledge and lack of recognition of clinical findings by physicians [4–8]. Moreover, diagnostic hypothesis generation in the ED is often based on physician intuition and experience [9]. Because an accurate diagnosis is the basis for all treatment and care, improving the diagnostic process and accuracy is key to improving patient safety and outcomes.

Because healthcare is becoming increasingly digitalized, one possibility to reduce human error in the diagnostic process is the use of computerized diagnostic decision support systems (CDDS). Such programs can prompt physicians

Ailin Lüthi and Ronny R. Graf contributed equally to this work.

***Corresponding author: Dr. rer. medic. Stefanie C. Hautz**, Department of Emergency Medicine, Inselspital University Hospital Bern, Freiburgstrasse 10, 3010 Bern, Switzerland, E-mail: stefanie.hautz@extern.insel.ch. https://orcid.org/0000-0003-4715-8465
**Thimo Marcin, Gert Krummrey and Wolf E. Hautz,** Department of Emergency Medicine, Inselspital University Hospital Bern, Bern, Switzerland. https://orcid.org/0000-0001-7229-5985 (T. Marcin). https://orcid.org/0000-0002-8397-2336 (G. Krummrey). https://orcid.org/0000-0002-2445-984X (W.E. Hautz)
**Ailin Lüthi and Ronny R. Graf,** Department of Emergency Medicine, Inselspital University Hospital Bern, Bern, Switzerland; and Faculty of Medicine, University of Bern, Bern, Switzerland
**Stefan K. Schauber,** Centre for Educational Measurement, Faculty of Educational Sciences, University of Oslo, Oslo, Norway; and Centre for Health Sciences Education, Faculty of Medicine, University of Oslo, Oslo, Norway. https://orcid.org/0000-0002-1832-2732
**Neal Breakey,** Department of Medicine, Spital Emmental, Burgdorf, Switzerland. https://orcid.org/0000-0002-5809-8552

to ask relevant questions, can close knowledge gaps and suggest diagnostic steps and differential diagnoses (DDx) and thus have potential to improve the diagnoses and outcomes of patients [10, 11]. CDDS collect data on patient characteristics, suggest potential causes, and propose next steps in the diagnostic workup. While CDDS originally only supported physicians (Dr-CDDS), many today offer an interface to be used by patients (Pat-CDDS), and propose DDx and next steps to patients directly, such as doctor visits or self-care. In a recent review and meta-analysis of CDDS providing DDx, the best performing system was ISABEL, which was associated with the highest rates of accurate diagnosis retrieval compared to all other types of CDDS tools (pooled rate=0.89; 95 % CI=0.83–0.94) [11]. It should be noted, however, that the vast majority of studies evaluating CDDS used vignette cases (i.e. cases that authors constructed) as opposed to real-world cases (ED cases).

ISABEL Healthcare Ltd (UK) has recently developed a German version of both the Dr-CDDS and the Pat-CDDS. Neither of these have been validated so far. In contrast to other CDDS using chatbots and rule-based algorithms to generate DDx, ISABEL is based on natural language processing. More specifically, the users enter symptoms as free text either in medical terminology when using the Dr-CDDS (ISABEL Pro) or in lay terminology when using the Pat-CDDS (ISABEL Symptom Checker). These terms are then compared to a reference library using a search algorithm and matched with natural language processing to those DDx with the highest degree of matching symptoms. Each language version of Isabel is based on the same database, which is maintained in English. Each query runs first through a professionally translated synonym file. The database search is then expanded with all matched synonyms (e.g. “upper abdominal pain” with “upper stomach pain”). German terms not found in the synonym file are translated to English by the Google API and matched again with the synonym file. If a term or phrase cannot be found in Isabel’s synonym file, an NLP algorithm first removes ‘stop words’ such as ‘and’, ‘is’, ‘no’ and tries to match the remaining text with the synonym file. If still no match is found, the NLP algorithm searches the synonym file by using combinations of 2 or 3 words and lastly by searching single words. The natural language processing approach employed by ISABEL might be more vulnerable to errors in translations compared to systems using rule based algorithms (where there usually is one storyline for each of the predictable user request). Indeed, users can enter terms in many ways and combinations, which highlights the importance of taking language issues into consideration. Hence, it is critical to investigate the performance of the German versions of ISABEL. Consequently, we aimed to assess the accuracy of diagnoses suggested by the German version of ISABEL Pro and the ISABEL symptom checker on typical ED vignettes and real ED cases. In addition, we aimed to compare the provided DDx from the German and English Isabel Pro on the same cases for direct comparison between language versions. Further, we aimed to explore the potential of the Dr-CDDS to increase diagnostic quality, namely to evaluate whether the Dr-CDDS provides the correct diagnosis for a sample of ED patients who initially received a wrong ED discharge diagnosis.

## Subjects and methods

### Design

The present validation study was designed to examine the accuracy of correct diagnoses suggested by ISABELS Dr-CDDS (ISABEL pro) and the Pat-CDDS (ISABEL symptom checker) on ED vignettes and retrospectively on ED cases.

### Subjects

**Vignette cases:** The clinical vignettes were derived from a German text book containing 100 cases occurring in Emergency Medicine including cases from various medical disciplines such as cardiology, pulmonology, gastroenterology and others [12]. Of those 100 cases, 50 were purposefully selected by two independent investigators (AL, RG) in order to sample each discipline. Selection criteria were the presence of a correct diagnosis in the book and information regarding medical history and physical examination. Cases with an unspecific final diagnosis were excluded as well as resuscitation, trauma or psychiatric emergency cases, because the need for and value of CDDS for such cases is considered low.

**Real ED cases:** Real ED data were used from hospitalized patients who participated in a recent prospective observational study on diagnostic errors conducted in the ED of the university hospital of Bern [13, 14]. That former study assessed discrepancies between the ED diagnosis at hospital admission and the hospital discharge diagnosis, whereas the latter was considered as the correct diagnosis. To assure that the final diagnosis can be detected given the symptoms at ED presentation, only patient cases with a confirmed ED diagnosis (i.e. no discrepancy between ED and hospital discharge diagnosis) were considered eligible for the *validation analysis*. Patients with a diagnostic discrepancy were considered for analysing Isabel’s *diagnostic potential*. The original study excluded patients admitted to any internal medicine ward for reasons of age, comorbidities, palliative care, social reasons, surgical ward crowding and patients with an acute traumatic injury. For the present study, patients were further excluded in case of ED revisit, withdrawal of the hospital’s general informed consent, triaged with an acute life-threatening condition, missing information in EHR history and unspecific discharge diagnosis indicated by an ICD-10 code starting with R from Chapter 18 – *Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified within ICD-10*, because we did not expect ISABEL to provide unspecific DDx. In addition, cases were included where the hospital discharge diagnosis was neither completely identical nor different from the ED discharge diagnosis. From the

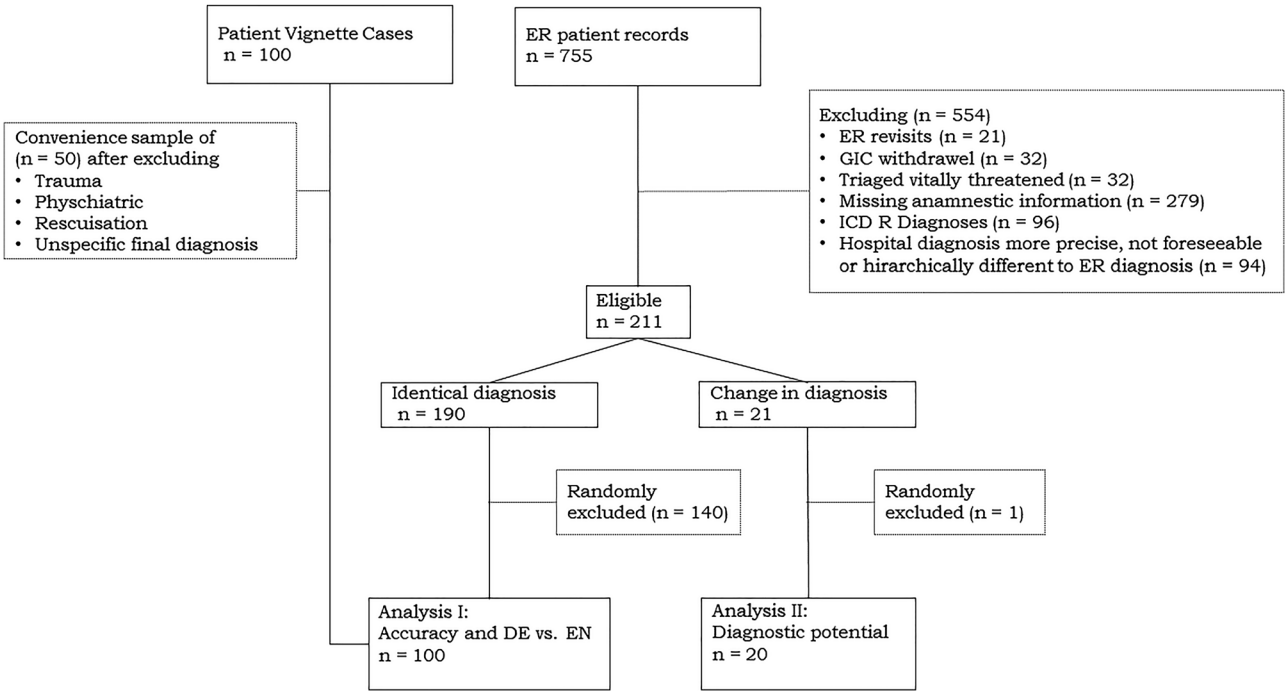


**Figure 1:** Flowchart.

eligible cases, 50 cases were randomly selected for the validation analysis and 20 cases for the diagnostic potential analysis using the statistical software R. The full flow chart is shown in Figure 1.

The study was approved by the Ethics Committee of the Canton of Bern, Switzerland (ID 2021-01449).

## Outcomes

Primary Outcome for this validation study was defined as the proportion of cases, where the correct diagnosis appeared in the DDx list provided by the Dr-CDDS. The correct diagnosis was defined as the vignettes text book diagnosis in case of the vignettes or as the hospitals discharge diagnosis for ED cases.

Secondary outcomes were the proportion of cases where the correct diagnosis appeared in the DDx list provided by the Pat-CDDS. Additional outcomes were the proportion of cases where the correct diagnosis appeared within the first 20, 10 and 3 DDx, the average number of DDx provided and the average position of the correct diagnosis on the DDx list provided by the Dr-CDDS and Pat-CDDS. An additional outcome was the level of agreement of the DDx provided between the English and German version when querying both language versions of the Dr-CDDS with the same (translated) clinical features. The diagnostic potential was assessed by the proportion of cases where the correct diagnosis (i.e. hospital discharge diagnosis) appeared within the list provided by the Dr-CDDS.

## Data collection

To collect ISABELS' DDx lists, two independent investigators prepared a file with age, sex, pregnancy status, travel history and key symptoms for each vignette and each ED case. One investigator (AL) prepared the query text for the Dr-CDDS, i. e. determined the key symptoms in medical terminology while another investigator (RG) determined the key symptoms in lay language for the Pat-CDDS. For 10 vignettes and 10 ED cases, key symptoms were derived from both investigators (AL, RG) in medical and lay terminology to assess an inter-rater agreement.

The information for the ED cases was derived from the EDs electronic health record (EHR) system from data collected and documented during clinical routine.

The query text prepared was sent to ISABEL using an application programming interface (API) that returned ISABELS' DDx lists.

To assess whether and at which position the correct diagnosis occurs on ISABELs' DDx list, two investigators (AL, RG) independently screened the DDx lists provided and determined the position of the first DDx on the list that matches the correct diagnosis. In case of different ratings, an agreement between the two raters was achieved by discussion and consensus.

Clinical features entered into Isabel Pro were translated from German into English by a native English speaking senior ED physician who is fluent in German and has extensive medical experience in German speaking ED settings.

Because the suggestions from Isabel change over time due to system maintenance, we have queried the English and German Version of the Dr-CDDS within few minutes using the API.

## Blinding

The investigators were aware of the correct diagnosis and therefore not blinded when deriving the key symptoms from the vignette cases or the ED cases.

## Statistical analysis

We used descriptive statistics such as counts and proportions or means and standard deviations as appropriate. In addition, we calculated the

margin of error for sample proportions as a Wilson confidence interval of 95 % for the primary outcome. Furthermore, we compared the primary and secondary outcomes achieved based on vignette cases and ED cases with exploratory methods as appropriate. For a random subsample of 10 vignettes and 10 ED cases, key symptoms were independently derived and entered into the online web version of Isabel Pro (Dr-CDDS) by two investigators (RG, AL) independently. Subsequently, Cohen's Kappa was calculated for inter-rater-agreement. For reasons of feasibility, the number of cases totally included for the primary analysis was pragmatically set at a total of 100 cases. Assuming that the correct diagnosis is provided on the DDx list in 89 % of the cases, we have expected a 95 % Wilson confidence interval width of 12 percent points, which we deemed as precisely enough [11].

To assess the level of agreement between the German and English Isabel Pro version, we analysed the number of DDx provided by both language versions and the number of DDx provided only by either one or the other version. ICD-10 Codes (three digits) of the output list were automatically compared for each case.

All statistical analyses were performed using statistical software R version 4.0.3.

# Results

Overall, 100 cases were used to analyse the accuracy for both, the German Dr-CDDS and the Pat-CDDS. Age of the included cases ranged from 12 to 90 years with a mean (standard deviation; SD) of 54.4 (19.5) years. Half of all discharge diagnoses could be categorized into diseases of the circulatory, digestive or respiratory system (Table 1).

## Accuracy – German Dr-CDDS

Overall, the German Isabel Pro provided the correct diagnosis in 82 % (95 % CI 0.73–0.88) of the cases. The accuracy was non-significantly (p=0.193) better for the text book vignette cases (88 %) compared to the retrospective ED cases (76 %) overall. The correct diagnosis also ranked significantly more often within the top 20, 10 and 3 DDx for vignette cases as compared to ED cases (Table 2). The Dr-CDDS provided on average 38 (SD 4.5) differential diagnoses (Table 2).

Table 3 shows the results of the post-hoc analysis regarding overall accuracy of the Dr-CDDS according the disease categorization (ICD-10-GM chapters). Accuracy was

**Table 1:** Patient characteristics.

| | | n=100 |
|---|---|---|
| Age | | 54.4 (19.5) |
| Female sex | | 45 (45 %) |
| Discharge diseases according to ICD-10 chapter GM | | |
| IX | Diseases of the circulatory system | 23 |
| XI | Diseases of the digestive system | 17 |
| X | Diseases of the respiratory system | 13 |
| XIII | Diseases of the musculoskeletal system and connective tissue | 9 |
| XVI | Certain infectious and parasitic diseases | 7 |
| IV | Endocrine, nutritional and metabolic diseases | 7 |
| III | Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism | 6 |
| XIV | Diseases of the genitourinary system | 5 |
| – | Diseases in other chapters | 13 |

**Table 2:** Accuracy of the Dr-CDDS in diagnosis retrieval.

| | Overall n=100 | Vignettes n=50 | EHR n=50 | p-Value[a] |
|---|---|---|---|---|
| True diagnosis in Isabel© DDx list | | | | |
| All DDx | 82 (82 %) | 44 (88 %) | 38 (76 %) | 0.193 |
| Top 20 | 71 (71 %) | 42 (84 %) | 29 (58 %) | 0.008 |
| Top 10 | 61 (61 %) | 38 (76 %) | 23 (46 %) | 0.004 |
| Top 3 | 37 (37 %) | 27 (54 %) | 10 (20 %) | <0.001 |
| Number of DDx | 38.33 (4.45) | 38.00 (4.66) | 38.66 (4.25) | |
| Position of similar DDx | 7.76 (8.75) | 4.82 (5.85) | 11.16 (10.28) | |

[a]p-value for difference in proportion between vignettes and EHR cases. Counts (percentage) or mean (standard deviation) as appropriate. DDx, differential diagnoses; EHR, electronic health record.

**Table 3:** Accuracy of the Dr-CDDS in diagnosis retrieval according to ICD-10 chapters.

| ICD-10 GM chapter | | Overall n=82/100 (82) | Vignettes n=44/50 | EHR n=38/50 |
|---|---|---|---|---|
| IX | Diseases of the circulatory system | 21/23 (91.3) | 12/12 | 9/11 |
| XI | Diseases of the digestive system | 13/17 (76.5) | 7/9 | 6/8 |
| X | Diseases of the respiratory system | 9/13 (62.2) | 2/3 | 7/10 |
| XIII | Diseases of the musculoskeletal system and connective tissue | 5/9 (55.5) | 2/3 | 3/6 |
| XVI | Certain infectious and parasitic diseases | 6/7 (85.7) | 3/3 | 3/4 |
| IV | Endocrine, nutritional and metabolic diseases | 7/7 (100) | 6/6 | 1/1 |
| III | Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism | 3/6 (50) | 2/4 | 1/2 |
| XIV | Diseases of the genitourinary system | 5/5 (100) | 3/3 | 2/2 |
| – | Diseases in other chapters | 13/13 (100) | 7/7 | 6/6 |

Values are counts (percentages) and indicate the proportion of cases where the true diagnosis (ER discharge diagnosis or vignette diagnosis) was listed on the DDx list.

lowest for respiratory diseases, diseases of the musculoskeletal system and diseases of the blood and blood-forming organs, while the accuracy for diagnoses of all other chapters was above 75 %.

## Accuracy – German Pat-CDDS

Querying the Pat-CDDS using lay terminology for the same cases resulted in a moderately lower accuracy compared to the Dr-CDDS using medical terminology as shown in Table 4. The correct diagnosis was among the provided DDx in 71 % of the cases, and in 58 , 48 and 24 % listed within the top 20, top 10 and top 3 respectively (Table 4).

## Unspecific diagnoses

As post hoc analyses, one of the investigators (RG) rated all correct diagnosis as either specific or non-specific. Although we already aimed to exclude non-specific diagnoses by excluding all ICD-10 diagnoses staring with R, 10 out of 100 diagnoses were categorized as non-specific. From those non-specific diagnoses, 6 (60 %) were not provided by the Dr-CDDS. In contrast, only 12 of 90 (13 %) specific discharge diagnoses were not retrieved from the Dr-CDDS.

## Inter-rater-agreement

The Dr-CDDS provided the correct diagnosis in 16 of 20 cases (queried by rater 1) and 17 of 20 cases (queried by rater 2) respectively resulting in a Cohens' Kappa of 0.32. The Pat-CDDS provided the correct diagnosis in 17 (rater 1) and 18 of 20 cases (rater 2) and resulted in a Cohens' Kappa of 0.57.

**Table 4:** Accuracy of the Pat-CDDS in diagnosis retrieval.

| | Overall n=100 | Vignettes n=50 | EHR n=50 | p-Value[a] |
|---|---|---|---|---|
| True diagnosis in Isabel© DDx list | | | | |
| All DDx | 71 (71 %) | 36 (73 %) | 35 (69 %) | 1 |
| Top 20 | 58 (58 %) | 31 (63 %) | 27 (53 %) | 0.543 |
| Top 10 | 48 (48 %) | 27 (55 %) | 21 (44 %) | 0.317 |
| Top 3 | 24 (24 %) | 13 (27 %) | 11 (22 %) | 0.815 |
| Number of DDx | 35.68 (7.75) | 35.22 (8.65) | 36.16 (6.12) | |
| Position of similar DDx | 9.48 (9.13) | 8.64 (9.23) | 10.41 (9.21) | |

[a]p-value for difference in proportion between vignettes and EHR cases. Counts (percentage) or mean (standard deviation) as appropriate. DDx, differential diagnoses; EHR, electronic health record.

## Agreement between German and English versions

When submitting the same queries translated to the English and German Isabel Pro version, 26.2 (SD 4.9) DDx were provided by both language versions (~75 %), while 8.6 (5.6) and 8.6 (5.4) were provided by either the German or the English version only. In one case, the hospital discharge diagnosis (I952 hypotension due to drugs) was provided by the German version (I959 hypotension unspecified), while the English version did not provide hypotension on the list. When considering only the top-10 DDX, 7.0 (1.7) were provided by both language versions while 2.8 (1.7) were provided only by either one of the systems. The German and English queries are provided together with the corresponding Isabel Output in Supplementary Table 3.

## Diagnostic potential

Of the 20 ED cases derived from a previous study, one case was erroneously tagged as having a discrepancy (i.e. having a different discharge diagnosis compared to the initial ED diagnosis), when in fact diagnoses did not differ. From the 19 ED remaining cases with discrepancies, the Dr-CDDS provided the hospital discharge diagnosis in 9 (47 %) of the cases, 6 of which ranked within the top 20 of the provided DDx and 4 ranked within the top 5. Table 5 shows the ED and hospital discharge diagnoses as well as the matching DDx provided by Isabel Pro (Dr-CDDS) for these 9 patients.

# Discussion

We aimed to assess the accuracy of differential diagnoses suggested by the German version of ISABEL Pro and ISABEL symptom checker on vignettes and ED cases. Our analysis shows robust results for both, vignettes and real ED cases.

Overall, we found an accuracy comparable to previous studies on DDx generators. A recent systematic review found an overall rate of accurate diagnosis retrieval of 0.70 (95 % CI 0.63–0.77). Isabel Pro was associated with the highest rate of accurate diagnoses among all of the investigated tools with a pooled rate of 0.89 (95 % CI 0.83–0.94), although a high heterogeneity was found between the 7 included studies [11]. One of the studies included queried Isabel Pro with data from 594 patients presenting to the ED and found a substantially greater accurate diagnosis retrieval rate of 95 % overall and 78 % in top 10 than we did [15]. The higher accuracy may be explained by the more strictly controlled data

**Table 5:** Diagnostic potential.

| ER diagnosis | Hospital discharge diagnosis | DDx provided by Isabel© |
|---|---|---|
| Atrial fibrillation and flutter, unspecified | Heart failure | Heart failure/CHF |
| Transient global amnesia [amnestic episode]. | Secondary malignant neoplasm of the brain and meninges | Brain tumors |
| Acute bronchitis caused by other specified agents | Pulmonary embolism without indication of acute cor pulmonale | Pulmonary embolism |
| Hypertensive heart disease with (congestive) heart failure: with indication of hypertensive crisis | Cerebral infarction | CVI/Stroke |
| Thrombosis, phlebitis and thrombophlebitis of unspecified location | Other infectious spondylopathies: Lumbar region | Infections of the spine |
| Disorders of vestibular function | Cerebral transient ischemia and related syndromes | Transient ischemic attack |
| Other aplastic anemias | Acute lymphoblastic leukemia [ALL]: without indication of complete remission | Leukemia |
| Gastritis, unspecified | Pneumonia due to other streptococci | Bacterial pneumonia |
| Pneumonia, unspecified | Acute rheumatic endocarditis | Endocarditis |

DDx, differential diagnoses; ER, emergency room.

collection and extraction in the prospective trial compared to our retrospective data analysis. However, as the authors did not provide details regarding discharge diagnoses, we may only speculate about this discrepancy.

Interestingly, we did not find differences in accuracy between the German and English language version, when sending the same (translated) queries to the Dr-CDDS. However, the provided DDX list did substantially differ between the language versions. Therefore, we assume that the algorithms in place are somewhat susceptible to minor changes in text entries.

Case vignettes, which are widely used to test Dr-CDDS, are often prototypical and illustrative. They may thus not be the best choice to test CDDS since ED case scenarios are hardly ever prototypical. Using both, ED cases and vignettes, we saw differences in the results of the DDx lists. The accuracy tended to be higher for vignette cases and if the correct diagnosis was retrieved, it was ranked substantially higher compared to ED cases. Therefore, we emphasize to use real-world cases for validation studies.

The authors of the aforementioned systematic review already pointed out the problem of lengthy DDx lists provided by CDDS. Although having an increased likelihood of retrieving the correct diagnosis, the value to clinicians may decrease, especially in a busy ED setting. Therefore, the importance of the rank where the correct diagnosis is listed should not be neglected.

Regarding the accuracy of the Pat-CDDS, Semigran et al. measured accuracy of self-diagnosis and triage advice provided by 23 symptom checkers by using 45 standardized patient vignettes. On average, the correct diagnosis was listed in 58 % of the cases among the first 20 DDx across all investigated CDDS. In this study, Isabel Symptomchecker provided the correct diagnosis in 69 % of the cases within the top 20 and thus performed better than in our study with 58 % [16].

In contrast to simulated vignette cases where usually a specific diagnosis can be defined, diagnoses in the real world are often rather vague and not definite, especially in emergency medicine. As we cannot expect a CDDS to come up with unspecific diagnoses, we excluded all diagnoses from analyses that had unspecific ICD-10 codes. Nevertheless, 10 patients included in this validation study still had some rather unspecific discharge diagnosis such as *unspecified infectious disease*, *lower back pain* or *disorder of muscle* (*unspecified*). For six of these 10 patients, Isabel could not provide any DDx that could have been assigned to the discharge diagnosis. Specific diagnoses that were not identified by the Dr-CDDS included colon diverticulum, immune thrombocytopenia or gastric perforation. The full list of failures including Isabel query and provided DDx list is shown in Supplementary Table 1. The full list with correctly identified diagnoses is provided in Supplementary Table 2 (in German).

The accuracy of differential diagnoses provided by the CDDS seems to differ regarding organs, respectively ICD disease category. Accuracy was higher for circulatory and endocrine conditions than for musculoskeletal and respiratory conditions. However, the sample size was rather low for many of the ICD-10 disease chapters.

Querying symptoms and clinical features available from the ED electronic health record system for real ED cases with known diagnostic errors (i.e. discrepancy between ED diagnosis and hospital discharge diagnosis), the Dr-CDDS provided the ‘correct’ hospital discharge diagnosis in half of cases. Of course, it can only be speculated whether the DDx list would finally have improved the diagnosis in ED.

However, it is likely that the Dr-CDDS would have at least broadened the DDx, as none of the ED diagnoses included the final hospital discharge diagnosis as differential diagnosis.

## Limitations

While the inclusion of both, vignette cases and real ED cases can be considered as a strength, inclusion of retrospective data for the real ED cases may be seen as a limitation. We randomly selected cases after excluding patients according to pre-specified exclusion criteria and ICD-10 coding of the discharge diagnosis. Therefore, it could not be avoided that some of the discharge diagnoses were still rather vague and not necessarily suitable to be detected by a DDx generator (i.e. lower back pain as a diagnosis). On the other hand, the risk of selection bias could be minimized.

As Isabel is based on free text, wording of the query text may influence the results. However, we found a good inter-rater agreement between two investigators (see results) querying the CDDS independently.

# Conclusions

We have found a somewhat lower accuracy using the German Version of Isabel Pro compared to previous studies on the English language version. Translating the German queries and querying the English version did not lead to improved accuracy, but a substantial part of the provided DDX differed between the language versions. CDDS have the potential to guide clinicians to the right diagnosis. However, further studies are needed to evaluate the effects when such tools are applied in a real world clinical setting (for example concerning an increase in diagnostic tests or increased length-of-stay due to additional investigations).

**Acknowledgments:** We would like to thank Isabel Healthcare to provide us access to Isabel Pro and Isabel Symptom Checker.

**Research funding:** The study was funded by the Swiss National Science Foundation under contract number 407740_187284/1. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Author contributions:** TM, WEH and SCH designed the study and provided oversight; AL and RG collected the data; GK implemented the API and developed a tool to allow for batch retrieval of DDx results; TM and SS analysed the data; SCH and TM drafted the manuscript; all authors read and approved the final manuscript. All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

**Competing interests:** WEH has received research funding from the European Union, Zoll foundation, Dräger Medical Germany, Mundipharma Research UK, MDI International Australia, Roche Diagnostics Germany, all outside the submitted work. WEH has provided paid consultancies to AO foundation Switzerland and MDI International Australia, all outside the submitted work. WEH has received financial support for a congress he chaired from EBSCO Germany, Isabel Healthcare UK, Mundipharma Medical Switzerland, VisualDx USA, all outside the submitted work. All other authors declare that they have no competing interests.

**Research ethics:** The study was approved by the Ethics Committee of the Canton of Bern, Switzerland (ID 2021–01449).

**Informed consent:** Informed consent was obtained from all individuals included in this study. Patients who withdrew the hospitals' general informed consent for further use of health related data for research were excluded. All methods were carried out in accordance with relevant guidelines and regulations.

# References

1. Singh H, Meyer AN, Thomas EJ. The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving US adult populations. BMJ Qual Saf 2014;23:727–31.
2. Newman-Toker DE, Wang Z, Zhu Y, Nassery N, Saber Tehrani AS, Schaffer AC, et al. Rate of diagnostic errors and serious misdiagnosis-related harms for major vascular events, infections, and cancers: toward a national incidence estimate using the "Big Three". Diagnosis 2021;8:67–84.
3. Graber ML. The incidence of diagnostic error in medicine. BMJ Qual Saf 2013;22:ii21–7.
4. Graber ML, Franklin N, Gordon R. Diagnostic error in internal medicine. Arch Intern Med 2005;165:1493.
5. Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. Acad Med 2003;78:775–80.
6. Norman GR, Monteiro SD, Sherbino J, Ilgen JS, Schmidt HG, Mamede S. The causes of errors in clinical reasoning: cognitive biases, knowledge deficits, and dual process thinking. Acad Med 2017;92:23–30.
7. Zwaan L, de Bruijne M, Wagner C, Thijs A, Smits M, van der Wal G, et al. Patient record review of the incidence, consequences, and causes of diagnostic adverse events. Arch Intern Med 2010;170:1015–21.
8. Zwaan L, Thijs A, Wagner C, van der Wal G, Timmermans DRM. Relating faults in diagnostic reasoning with diagnostic errors and patient harm. Acad Med 2012;87:149–56.
9. Pelaccia T, Tardif J, Triby E, Ammirati C, Bertrand C, Dory V, et al. How and when do expert emergency physicians generate and evaluate diagnostic hypotheses? A qualitative study using head-mounted video cued-recall interviews. Ann Emerg Med 2014;64:575–85.
10. Fraser H, Coiera E, Wong D. Safety of patient-facing digital symptom checkers. Lancet 2018;392:2263–4.

11. Riches N, Panagioti M, Alam R, Cheraghi-Sohi S, Campbell S, Esmail A, et al. The effectiveness of electronic differential diagnoses (DDX) generators: a systematic review and meta-analysis. Schmidt RL, herausgeber. PLoS One 2016;11:e0148991.
12. Fleischmann T. Fälle Klinische Notfallmedizin – Die 100 wichtigsten Diagnosen. München: Urban & Fischer in Elsevier; 2018:566 S p.
13. Hautz WE, Kämmer JE, Hautz SC, Sauter TC, Zwaan L, Exadaktylos AK, et al. Diagnostic error increases mortality and length of hospital stay in patients presenting through the emergency room. Scand J Trauma Resuscitation Emerg Med 2019;27:54.
14. Hautz SC, Schuler L, Kammer JE, Schauber SK, Ricklin ME, Sauter TC, et al. Factors predicting a change in diagnosis in patients hospitalised through the emergency room: a prospective observational study. BMJ Open 2016;6:e011585.
15. Ramnarayan P, Cronje N, Brown R, Negus R, Coode B, Moss P, et al. Validation of a diagnostic reminder system in emergency medicine: a multi-centre study. Emerg Med J 2007;24:619–24.
16. Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. BMJ 2015;351:h3480.

**Supplementary Material:** This article contains supplementary material (https://doi.org/10.1515/dx-2023-0047).