# Avey vs. Infermedica: A Clinical Vignette Accuracy Study


## ABSTRACT

Medical self-diagnostic algorithms (or symptom checkers) are increasingly becoming an integral part of digital health and our daily lives. In this study, we conducted a comprehensive experiment that capitalizes on the standard clinical vignette approach to evaluate the accuracies of 2 symptom checkers, namely, Avey and Infermedica. We tested Avey and Infermedica on 150 gold-standard vignettes that were peer-reviewed by 7 external medical doctors with an average experience of 8.4 years. To establish a frame of reference, we further compared Avey's and Infermedica's accuracies against 3 highly seasoned primary care physicians with an average experience of 16.6 years. Results show that Avey significantly outperforms Infermedica across 5 standard accuracy metrics that are commonly used in the domain. For instance, Avey outpaced Infermedica by an average of 45.27% in producing the correct diagnosis at the top of their differential diagnosis lists.


## 1. Experimentation Methodology

### 1.1 Vignette Selection, Standardization, and Testing

Building on prior related work [1, 2, 3, 4, 5, 6], we adopted a clinical vignette approach to rigorously measure the performance of Avey and Infermedica. A seminal work at Harvard Medical School has established the value of this approach [3, 6] for testing symptom checkers, especially since it has also been a common method to test physicians on their diagnosis abilities [6].

To this end, we concretely defined our experimentation methodology in terms of 3 stages, namely, *vignette selection & standardization*, *vignette testing on symptom checkers*, and *vignette testing on doctors*.

In the *vignette selection & standardization* stage, our medical team checked out the conditions available in Infermedica's knowledge repository published on Infermedica's official website, updated on November 3, 2022. For a fair comparison, the team identified 150 conditions that exist in both Avey's and Infermedica's knowledge repositories. Subsequently, the team selected 150 gold-standard vignettes that include these conditions at the top of their differential diagnosis lists [7, 8].

The selected 150 vignettes involved 14 body systems and encompassed common and less-common conditions relevant to primary care practice (see Table 1). They fairly represent real-world cases in which patients might seek primary care or advice from a physician or a symptom checker. They were drawn and compiled by our medical team from reputable medical websites and training material for healthcare professionals [9, 10, 11, 12, 13, 14, 15, 16]. Afterwards, each vignette was reviewed by 7 external medical doctors from 4 different specialties, namely, Family Medicine, General Medicine, Emergency Medicine, and Internal Medicine, with an average experience of 8.4 years. None of these doctors had any involvement with Avey or Infermedica and they were all entirely unaware of them before they were asked to review the vignettes.

We designed and developed a full-fledged web portal to streamline the process of reviewing and standardizing the vignettes. To elaborate, the portal allowed our medical team to upload the vignettes to a web page that is shared across the 7 recruited doctors. Each doctor was able to access the vignettes and review them independently and opaquely (i.e., doctors could not see each other's work). After reviewing a vignette, a doctor can reject or accept it. Upon rejecting a vignette, a doctor can propose changes to improve its quality and/or clarity. Our medical team reviewed every suggested change of every vignette and made refinements accordingly, before re-uploading it to the portal for a new round of review. Multiple reviewing rounds can occur before a vignette is deemed gold-standard. A vignette was considered gold-standard only when there were no more suggested changes by any external doctor and at least 5 out of the 7 (i.e., super-majority) doctors accepted it.

In the *vignette testing on symptom checkers* stage, our medical team tested the gold-standard vignettes from stage 1 on Avey and Infermedica. To allow for external validation and the reproducibility of the results, we made all the 150 vignettes publicly available at [8].

Finally, to establish a frame of reference and interpret Avey's and Infermedica's results accordingly, we recruited 3 external primary care physicians with an average experience of 16.6 years. One of those physicians is a Family Medicine doctor with 30+ years of experience. The other two are also Family Medicine doctors, each with 10+ years of experience. None of these physicians had any involvement with the developments of Avey or Infermedica and were completely unaware of them before they were recruited. Furthermore, none of them were among the 7 doctors of stage 1 and were only recruited for "diagnosing" the gold-standard vignettes in what we refer to as the *vignette testing on doctors* stage.

For the purpose of this last stage and akin to [17], we concealed the main and differential diagnoses of the 150 gold-standard vignettes from the 3 recruited doctors and exposed the remaining information through our web portal. The doctors were granted access to the portal and asked to provide their main and differential diagnoses for each vignette without checking any references, mimicking as closely as possible real-world sessions where they typically diagnose patients on the spot without checking references. As an outcome, each vignette was "diagnosed" by each of the 3 doctors. We published the results of the doctors online at [8] to allow for external cross-validation.

**Table 1: The body systems and numbers of common and less-common diseases covered in our benchmark vignette suite.**

| Body System | # of Diseases | # of Common Diseases | # of Less Common Diseases | Common Diseases | Less Common Diseases |
|---|---|---|---|---|---|
| **Hematology** | 3 | 0 | 3 | 0.00% | 100.00% |
| **Cardiovascular** | 16 | 12 | 4 | 75.00% | 25.00% |
| **Neurology** | 7 | 3 | 4 | 42.86% | 57.14% |
| **Endocrine** | 9 | 7 | 2 | 77.78% | 22.22% |
| **ENT** | 8 | 7 | 1 | 87.50% | 12.50% |
| **GI** | 20 | 13 | 7 | 65.00% | 35.00% |
| **Obs/Gyn** | 16 | 15 | 1 | 93.75% | 6.25% |
| **Infectious** | 7 | 1 | 6 | 14.29% | 85.71% |
| **Respiratory** | 26 | 19 | 7 | 73.08% | 26.92% |
| **Orthopedics & Rheumatology** | 5 | 3 | 2 | 60.00% | 40.00% |
| **Ophthalmology** | 5 | 4 | 1 | 80.00% | 20.00% |
| **Dermatology** | 10 | 7 | 3 | 70.00% | 30.00% |
| **Urology** | 7 | 4 | 3 | 57.14% | 42.86% |
| **Nephrology** | 11 | 11 | 0 | 100.00% | 0.00% |

## 1.2 Accuracy Metrics

To evaluate the performance of Avey, Infermedica, and the doctors in stage 3, we utilized 5 standard accuracy metrics. In particular, we used the matching-1 ($M1$), matching-3 ($M3$), and matching-5 ($M5$) criteria to measure if Avey, Infermedica, or a doctor is able to output a gold-standard vignette's main diagnosis at the top (i.e., $M1$), among the first 3 diseases (i.e., $M3$), or among the first 5 diseases (i.e., $M5$) of their differential list. For Avey, Infermedica, and the doctors, we report the percentages of vignettes that fulfill $M1$, $M3$, and $M5$. The mathematical definitions of $M1$, $M3$, and $M5$ are given in Table 2.

Alongside, for each tested gold-standard vignette, we utilized *precision* as a measure of the percentage of diseases in Avey's, Infermedica's, or doctors' differential list(s) that are relevant. The average precision is defined mathematically in Table 2.

**Table 2: The descriptions and mathematical definitions of the 5 accuracy metrics used in our study.**

| Metric | Description | Mathematical Definition |
|---|---|---|
| **M1%** | The percentage of vignettes where the gold-standard main diagnosis is returned at the top of a symptom checker's or doctor's differential list | $\frac{\sum_{v=1}^{N} i_v}{N}$, where $N$ is the number of vignettes and $i_v$ is 1 if the symptom checker or doctor returns the gold-standard main diagnosis within $v$ at the top of their differential list; and 0 otherwise |
| **M3%** | The percentage of vignettes where the gold-standard main diagnosis is returned among the first 3 diseases of a symptom checker's or doctor's differential list | $\frac{\sum_{v=1}^{N} i_v}{N}$, where $N$ is the number of vignettes and $i_v$ is 1 if the symptom checker or doctor returns the gold-standard main diagnosis within $v$ among the top 3 diseases of their differential list; and 0 otherwise |
| **M5%** | The percentage of vignettes where the gold-standard main diagnosis is returned among the first 5 diseases of a symptom checker's or doctor's differential list | $\frac{\sum_{v=1}^{N} i_v}{N}$, where $N$ is the number of vignettes and $i_v$ is 1 if the symptom checker or doctor returns the gold-standard main diagnosis within $v$ among the top 5 diseases of their differential list; and 0 otherwise |
| **Average Precision** | Precision is the proportion of diseases in the symptom checker's or doctor's differential list that are also in the gold-standard differential list. The average precision is taken across all vignettes for each symptom checker and doctor | $\frac{\sum_{v=1}^{N} p_v}{N}$, where $N$ is the number of vignettes and $p_v = \frac{True\ Positive}{True\ Positive + False\ Positive}$ of the symptom checker or doctor for vignette $v$ |
| **Average NDCG** | Normalized Discounted Cumulative Gain (NDCG) is a measure of ranking quality. The average NDCG is taken across all vignettes for each symptom checker and doctor | $\frac{\sum_{v=1}^{N} \frac{DCG_v}{gold\ DCG_v}}{N}$, assuming $N$ vignettes, $n$ number of diseases in a gold-standard vignette, $v$, and $relevance_i$ for the disease at position $i$ in $v$'s differential list. $DCG_v = \sum_{i=1}^{n} \frac{2^{relevance_i - 1}}{\log_2(i+1)}$, which is computed over the differential list of a doctor for $v$. $gold\ DCG_v$ is defined exactly as $DCG_v$, but is computed over the gold-standard differential list of $v$ |

Finally, we measured the ranking qualities of Avey, Infermedica, and the doctors using the *Normalized Discounted Cumulative Gain* (NDCG) [18] metric that is widely used in practice [19]. To define NDCG, each disease at position $i$ in the differential list of a gold-standard vignette is assigned relevance$_i$. The higher the rank of a disease in the differential list, the higher the relevance of that disease to the correct diagnosis. To this end, Discounted Cumulative Gain (DCG) can be defined mathematically as $\sum_{i=1}^{n} \frac{2^{relevance_i - 1}}{\log_2(i+1)}$, assuming $n$ diseases in a vignette's differential list (see Table 2). As such, DCG penalizes a symptom checker or a doctor if they rank a disease lower in their output differential list than the gold-standard list. Capitalizing on DCG, Normalized DCG (NDCG) becomes the ratio of a symptom checker's or a doctor's DCG divided by the corresponding gold-standard DCG. Table 2 provides the complete mathematical definition of NDCG.

## 2. RESULTS

As indicated in Section 1.1, we tested the 150 gold-standard vignettes on Avey, Infermedica, and a panel of 3 highly seasoned physicians. Figure 1 demonstrates all the accuracy results. As shown, Avey resulted in averages of 92%, 95.33%, 95.33%, 76.24%, and 77.27% for *M*1, *M*3, *M*5, precision, and NDCG, respectively. On the flip side, Infermedica provided average *M*1, *M*3, *M*5, precision, and NDCG of 63.33%, 79.33%, 82.67%, 53.75%, and 65.81%, respectively. Consequently, Avey outperformed Infermedica by averages of 45.27%, 20.17%, 15.31%, 41.84%, and 17.41% using *M*1, *M*3, *M*5, precision, and NDCG, respectively. Interestingly, Avey outpaced Infermedica even when asking an average of 16.3% fewer questions. In particular, while Infermedica used an average of 26.9 questions per diagnostic session, Avey used an average of 22.5 questions.

Alongside Avey and Infermedica, Figure 1 depicts the accuracy results of MDs, which is the average performance of the three medical doctors presented in Section 1.1. As illustrated, the human doctors provided average $M1$, $M3$,

*M*5, precision, and NDCG of 64.22%, 75.33%, 75.55%, 72.39%, and 62.93%, respectively. To this end, Avey outperformed the three doctors by 43.26%, 26.55%, 26.18%, 41.84%, and 17.41% using *M*1, *M*3, *M*5, precision, and NDCG, on average, respectively.

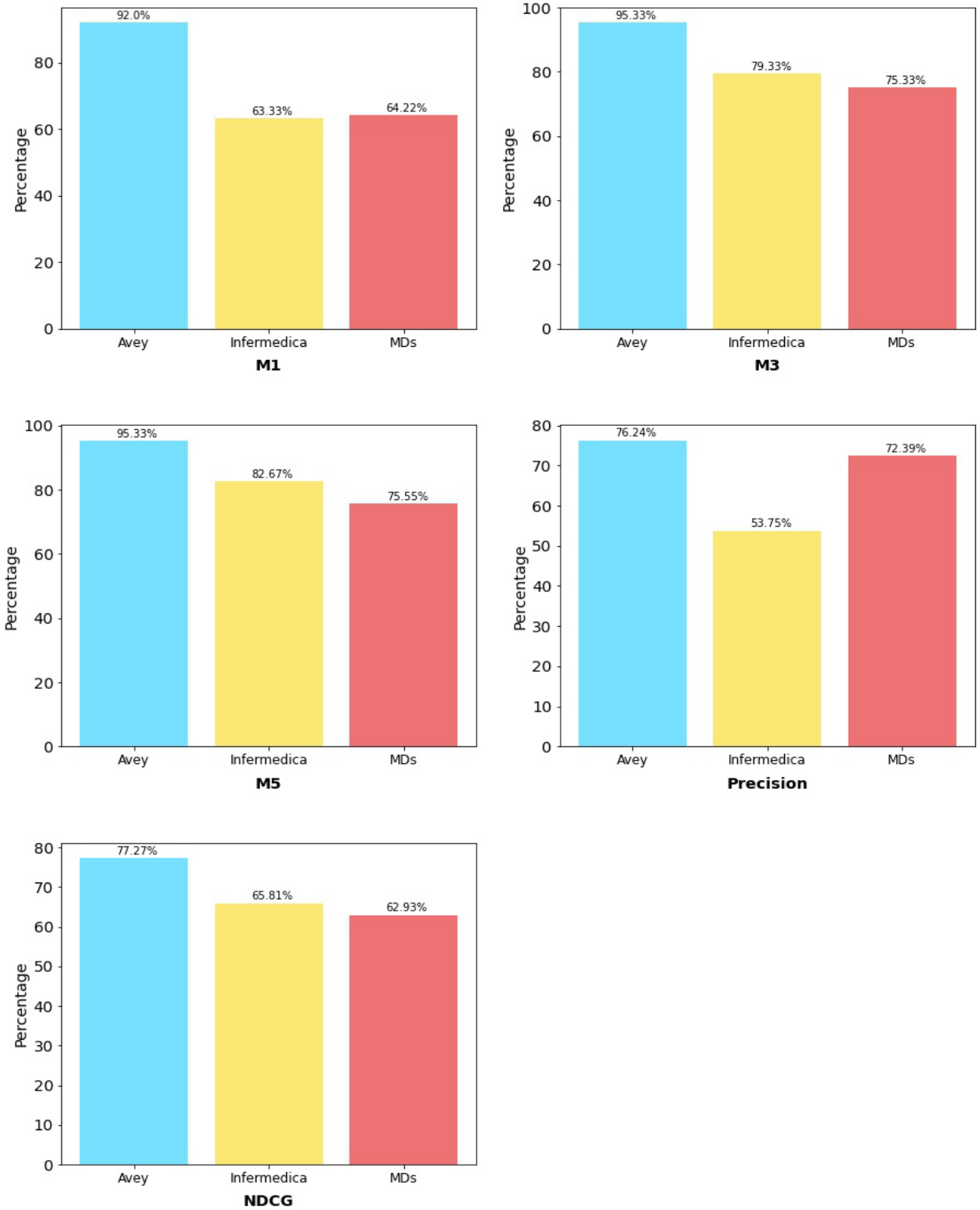


**Figure 1**: The accuracies of Avey, Infermedica, and MDs (i.e., a panel of external and highly seasoned medical doctors with an average clinical experience of 16.6 years) using the 5 standard accuracy metrics, *M*1, *M*3, *M*5, precision, and NDCG.

## 3. CONCLUDING REMARKS

Artificial Intelligence (AI) based symptom checkers that undergo rigorous research, development, and testing have the potential to become useful tools for timely, accurate, and instantly available self-diagnosis. In this study, we used the standard clinical vignette approach to compare the accuracy of Infermedica against that of Avey, a highly sophisticated and advanced AI-based symptom checker that took around 4 years of extensive research, design, development, and testing before it was launched. To put things in perspective and interpret the collected results accordingly, we further measured the accuracy of an external and independent panel of physicians with an average clinical experience of 16.6 years. Results show that Avey significantly outperforms Infermedica and the panel of physicians using the standard accuracy metrics in the field.

## REFERENCES


**[1]** Michella G Hill, Moira Sim, and Brennen Mills. 2020. The quality of diagnosis and triage advice provided by free online symptom checkers and apps in Australia. Medical Journal of Australia 212, 11 (2020), 514–519.

**[2]** David M Levine and Ateev Mehrotra. 2021. Assessment of Diagnosis and Triage in Validated Case Vignettes Among Nonphysicians Before and After Internet Search. JAMA network open 4, 3 (2021), e213287–e213287.

**[3]** Hannah L Semigran, David M Levine, Shantanu Nundy, and Ateev Mehrotra. 2016. Comparison of physician and computer diagnostic accuracy. JAMA internal medicine 176, 12 (2016), 1860–1861.

**[4]** Adam Ceney, Stephanie Tolond, Andrzej Glowinski, Ben Marks, Simon Swift, and Tom Palser. 2021. Accuracy of online symptom checkers and the potential impact on service utilisation. Plos one 16, 7 (2021), e0254088.

**[5]** Stephen Gilbert, Alicia Mehl, Adel Baluch, Caoimhe Cawley, Jean Challiner, Hamish Fraser, Elizabeth Millen, Maryam Montazeri, Jan Multmeier, Fiona Pick, et al. 2020. How accurate are digital symptom assessment apps for suggesting conditions and urgency advice? A clinical vignettes comparison to GPs. BMJ open 10, 12 (2020), e040269.

**[6]** Hannah L Semigran, Jeffrey A Linder, Courtney Gidengil, and Ateev Mehrotra. 2015. Evaluation of symptom checkers for self diagnosis and triage: audit study. bmj 351 (2015).

**[7]** Infermedica. 2022. Infermedica's knowledge base. https://developer.infermedica.com/docs/v3/available-conditions. [Accessed 22-Nov-2022].

**[8]** Rimads QSTP-LLC. 2022. Avey Benchmark Vignette Suite. https://avey.ai/research/avey-accurate-ai-algorithm/infermedica-vs-avey. [Accessed 22-Nov-2022].

**[9]** United States Medical Licensing Examination. 2019-2020. USMLE Step 2 CK. https://www.usmle.org/step-exams/step-2-ck. [Accessed 05-Feb-2022].

**[10]** John D Firth and Ian Gilmore. 2008. MRCP Part 1 Self-Assessment: Medical Masterclass Questions and Explanatory Answers. Radcliffe Publishing.

**[11]** Doug Knutson. 2018. Family Medicine PreTest Self-Assessment And Review. Mc- Graw Hill Professional.

**[12]** American Board of Family Medicine. 2018. In-Training Examination. https:

**[13]** American Academy of Pediatrics. 2020. 2021 PREP Self-Assessment. https://www.aap.org/. [Accessed 05-Feb-2022].

**[14]** CRC Press. 2011-2020. 100 Cases Book Series. https://www.routledge.com/100- Cases/book-series/CRCONEHUNCAS. [Accessed 05-Feb-2022].

**[15]** Alfred F Tallia, Joseph E Scherger, and Nancy Dickey. 2017. Swanson's Family Medicine Review E-Book. Elsevier Health Sciences.

**[16]** Ian Wilkinson, Ian Boden Wilkinson, Tim Raine, Kate Wiles, Anna Goodhart, Catriona Hall, and Harriet O'Neill. 2017. Oxford handbook of clinical medicine. Oxford university press.

**[17]** Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems (TOIS) 20, 4 (2002), 422–446.

**[18]** Xue Zhao. 2013. A Theoretical Analysis of NDCG Ranking Measures. (2013).