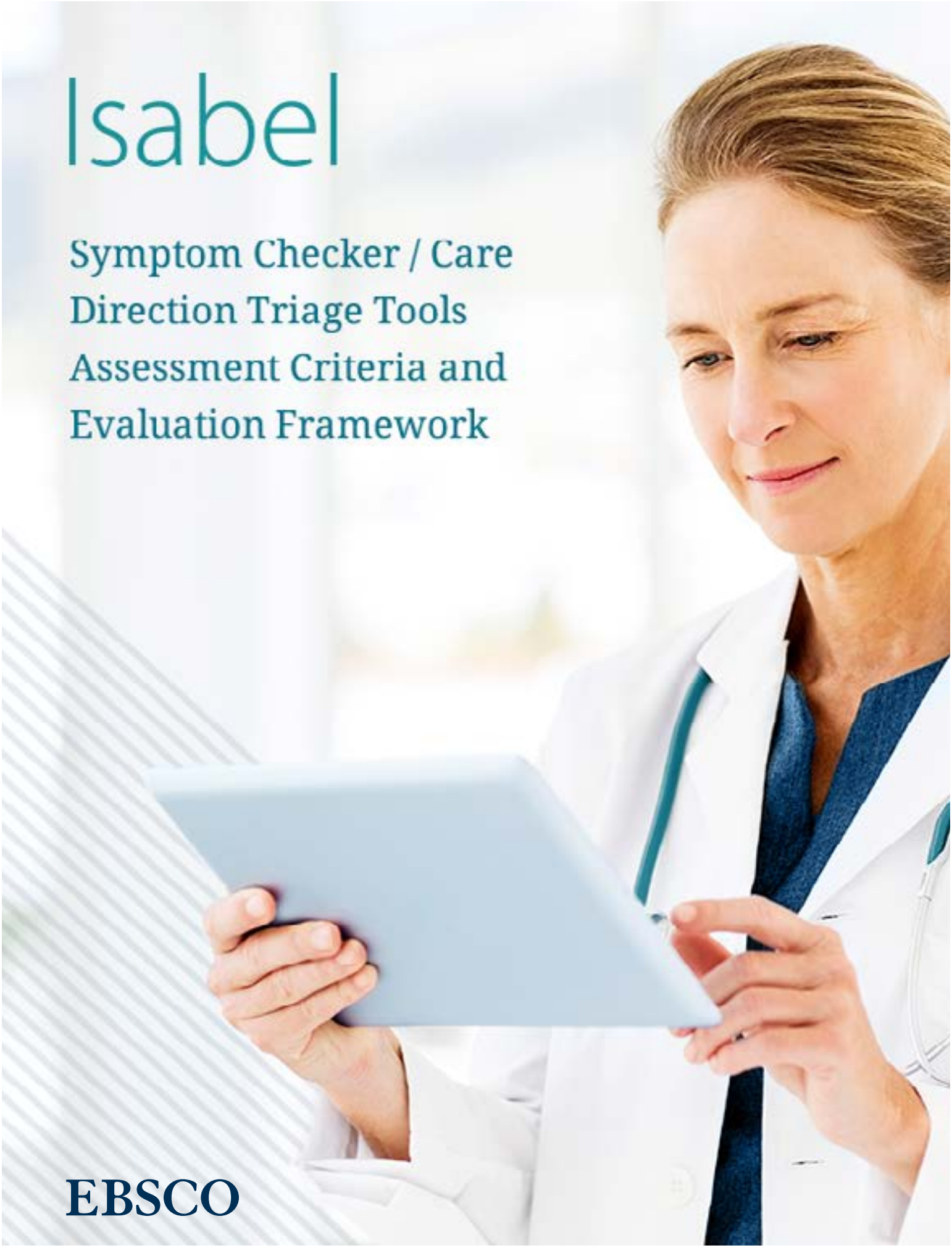
Isabel

Symptom Checker / Care Direction Triage Tools Assessment Criteria and Evaluation Framework



EBSCO

Introduction

The flood of patient symptom checkers/virtual triage tools that entered the market in the last two years is staggering, and it is making for a very confusing space. These tools are used in various patient workflows. They drive virtual visits for the providing company, supporting a number of patient workflows as the new "Digital Front Door" for health systems and payers, driving patients to the correct care setting to avoid unnecessary ED visits.

Often, these tools are evaluated on their 'look and feel' and some of the professed technical capabilities (bot enabled, use of AI (artificial intelligence), API support, etc.). Comparatively, not a lot of attention is focused on the clinical performance related to accuracy of their Clinical Engine foundations, which should be the most important aspect since they are used to direct patients to the most appropriate level of care. Other important criteria for consideration include the breadth of coverage (number of symptoms and conditions, are all age groups covered, etc.).

Expert Opinions

With this as a backdrop, it is being discovered that the tools often are not performing at the level they are marketed to be. Industry experts are weighing in and providing valuable insight into the emerging field. The following are just a few of the observations:

> "...What most folks don't realize is that the internal logic of which answers to provide are still hand coded decision trees (rules-based) in 95% of chatbots, not the result of some exotic AI/ML related search or automated intelligence..."

William Vorhies, Editorial Director for Data Science Central; President & Chief Data Scientist at Data-Magnum; has practiced as a data scientist since 2001 from ***Data Science Central article***

> "...a diagnostic (rules-based) engine, in a nutshell, is based on a complicated set of rules. These rules are decided by clinicians who type a range of probabilities for symptoms into their computers. As the number of rules grows, the software's path to making decisions becomes more complex and difficult to alter..."

John Taylor, the CEO of Action.AI stated in a ***Forbes report***

> "...Symptom checking apps gave conflicting results and advice when we presented them with the same set of symptoms...with the potential for incorrect or inadequate advice being given to patients... it can use the information you enter to provide triage advice, and that information on potential diagnoses... provides context for why it advises a particular course of action..."

Anna Studman, Author of Which?, the independent, charitable social enterprise in the United Kingdom from ***Can you trust AI symptom checkers? article***

Key Assessment Criteria

That said, what can be done to weave your way through the process of selecting a symptom checker/virtual triage tool for your organization? The following guideline tool consolidates key features and functionalities to help evaluate and score various tools in a simple and straight forward method. The questions target the most important capabilities of these tools: clinical accuracy and appropriateness of their results.

The following eight questions are designed to help your team debias the selection process and objectively evaluate tools you might be considering:

Key Criteria	Why is this Important?
Does the system force you to pick a chief complaint? Examples of how chief complaint is asked include: • *"What symptom is bothering you the most?"* • *"Which of these is your main problem?"*	If the system does, it is essentially forcing the patient to self-diagnose and biases the results given. Systems can give very different answers depending on the symptom the patient picks as the chief complaint and directs them to very different care settings. The order of symptom entry should not have an impact on the results!

Key Criteria	Why is this Important?
Does the system recognize and use all the symptoms entered by the patient?	If a patient has multiple symptoms, all should be considered when suggesting conditions, not just those the system recognizes or has built into their fixed and finite rules-based system. Patients can represent their symptoms in numerous ways and should be free to describe exactly how they are feeling. The patient's description should be used by the system to generate the list of possible conditions. If some of the symptoms they present with are not recognized, the results are skewed and biased.
Does the system have an age range limitation?	People in all age ranges should be covered by the tool, not just adults or just pediatrics. How would a mother or father find care for their child if the tool did not cover pediatrics?
Does the correct diagnosis appear in the top ten conditions listed by the system?	Clinical accuracy of the system should be the most important criteria. If the system does not come up with the correct condition in its list, especially in systems that force a patient to self-diagnose, how can it be relied upon to get the patient to the right care venue?
How many questions does the system ask the patient to get results?	Less is more. Many systems ask between 20-50 questions, sometimes repeating the same question or asking about information already entered, etc. It is critical to understand that the patient is not feeling well to start, and may be worried or scared, leading to high drop off rates and dissatisfaction.

Key Criteria

Is the patient asked any of the following or similar questions during the session before the list of possible conditions has been generated:

- *Which level of care are you considering?"*
- *"It would be helpful to know, based on your symptoms, what do you think you should do? Go to the ER; Go to Urgent Care; Go to a doctor; Nothing special; Don't know?"*
- *"Which (condition) do you think is the right answer?"*
- *"Do you feel your symptoms seem severe enough to require immediate medical help?"*
- *"Do you feel this looks like a life-threatening problem?"*

Why is this Important?

These are all forms of self-diagnosis and force an untrained patient to decide on their own treatment options.

Key Criteria	Why is this Important?
Is the patient forced to pick a condition from the generated list (self-diagnose) before the system provides a level of care recommendation?	When asking a patient to choose a condition to direct them to the appropriate level of care, the system forces them to self-diagnose. Published diagnosis error rates with trained physicians are 5% to 20%; should patients be put in this position? What if they pick the wrong condition (e.g., if three conditions are listed and the first suggests Emergency Room, the second Urgent Care walk-in and the third Primary Care Doctor, it is very confusing and self-defeating – which should they choose)? Basically, the patient is presented with a no-win dilemma. The level of care recommendation should be based on the patient's overall presentation, not variable based on a condition.
Does the system get the patient to the correct care venue based on their presentation?	This is a fundamental feature of these systems. What is the correct venue based on their actual clinical presentation? Getting this wrong can lead to treatment delays, possible increased cost, high drop off rates and patient dissatisfaction. Getting patients to the correct venue of care is critical for not only curtailing costs but also improving outcomes.

System Assessment Examples

To demonstrate how the questions from the table above come to light in evaluating systems, you can run cases through each system and evaluate their results in relation to each question. The example below takes two randomly selected real cases from independent sources published on July 26th, 2019 and runs them through ten systems, four of which are included in Appendix 2 as examples.

After the case summary section, the results of how each system performed in each of the different situations. The answers to the questions are tabulated and scored 0 for a positive response and 1 for a negative response, therefore a lower score is better.

As you will see, there were significant variances across the systems as mentioned by the industry insiders above.

Case One

Sourced from the **Society for Improvement in Diagnostic Medicine (SIDM) listserv (Posted: July 26, 2019)** or directly from the **Apple Podcast sponsored by the Kaiser Family Foundation**.

Key information starts at the 5:30 minute mark into the podcast for signs and symptoms and the physician's correct thyroid issue diagnostic path: 35-year old, female, constipation, weight gain, heavy menstrual periods. In this case, the doctor correctly recognized a thyroid problem.

Case Two

In the United Kingdom, the mother of a toddler who died four years ago from a twisted bowel has urged the NHS to make changes. This comes after an inquest heard that the 111 and Out of Hours Nurse services both missed chances, an incorrect diagnostic path of Gastritis, to save her daughter's life (Published: Friday, July 26, 2019).

Key information including signs and symptoms: 2-year old, female, abdominal pain, vomiting, rapid breathing; the correct diagnosis was Twisted Bowel (Volvulus).

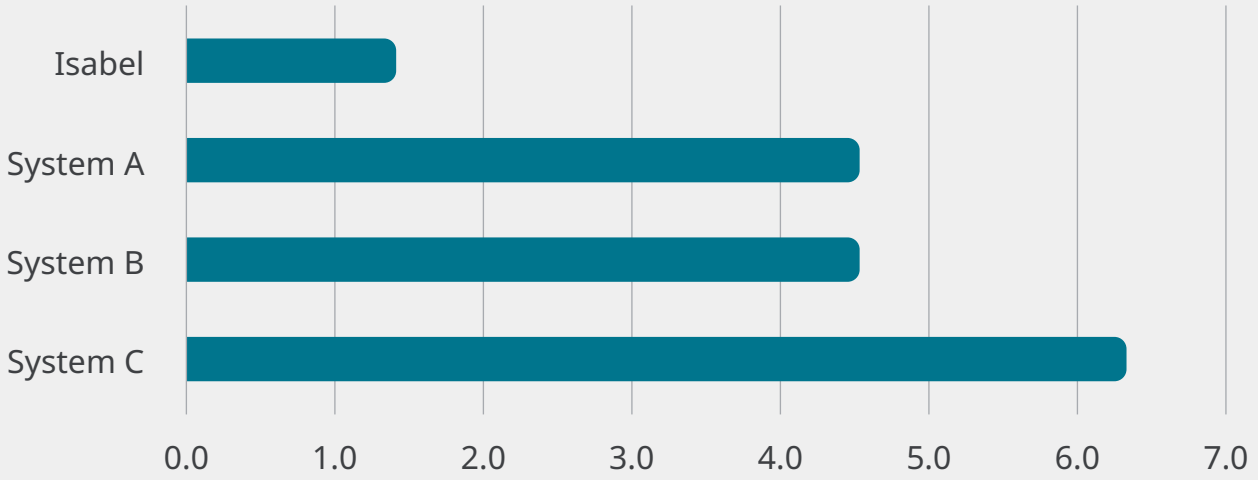
The symptoms are run through the system. Symptoms need to be run multiple times for systems that force the patient to pick the Chief Complaint, as depending on what symptom is selected the results may vary. The outputs are then reviewed for accuracy, (i.e., does the actual final diagnosis show up and is the care direction recommendation accurate).

System Average Scores

The table below represents the totals from four of the systems tested. It provides the total score for each system and the details (answers to the nine questions above and how the system performed) are provided in Appendix B.

Average from all Encounters

(Lower = More Accurate and Efficient for the Patient)



This simple, straight forward approach provides an objective review of the clinical performance of symptom checker/virtual triage systems. Ultimately, this is the most important capability of the system to consider when providing these tools to patients and consumers. While other criteria are important (e.g., the number of conditions covered, number of symptoms covered – this should always be infinite, how many different patient workflows can the tools support, has the tool been independently medically validated, etc.) the most important is the clinical accuracy and efficacy.

For more information or to receive a copy of a scoring template for your use, please contact us at **Don.bauman@isabelhealthcare.com**.

Appendix A

Expert Opinions

"...It is a striking thing that as we have this huge plethora of tools that have emerged...and yet we don't really know that basic question, 'What does it change?'"...How well they perform is still an open question, he said..."

Dr. Ateev Mehrotra, an associate professor at Harvard Medical School who has studied symptom checkers from ***Online symptom checker aims to provide care at the right time, place Modern Healthcare article***

"...Typically, you enter your symptoms and the app asks you follow-up questions and reacts to your answers...The importance of how you describe your symptoms, and the limitations of a check-box approach, became clear in our snapshot test...the question-based format didn't allow for important contextual information to come out..."

"... 'Usually it is considered good practice to ensure that the patient can talk freely...but there's no ability for the app to dissect free text. It's like playing "20 questions" at a party.' ..."

Dr. Margaret McCartney, GP from ***Can you trust AI symptom checkers? article***

"...Elizabeth Murray, Professor of eHealth and Primary Care at University College London, thinks it is unlikely that these symptom checkers will be able to make a safe diagnosis, because the apps haven't been developed on the basis of robust evidence, such as going through peer reviewing or clinical trials...These processes are at odds with how the tech industry likes to work: quickly, and with an emphasis on marketing...Dr Whitaker, GP and New Statesman columnist, puts it more bluntly. He thinks these algorithms are 'basically disasters'..."

Anna Studman, Author of Which?, the independent, charitable social enterprise in the United Kingdom from ***Can you trust AI symptom checkers? article***

> "...So buyers beware and be sure to satisfy yourself about the accuracy of any chatbot or similar AI/ML solutions before you put them in production..."
>
> *William Vorhies, Editorial Director for Data Science Central; President & Chief Data Scientist at Data-Magnum; has practiced as a data scientist since 2001 from* ***Data Science Central article***

Appendix B

Scoring Details Examples

Company / Version	Chief Complaint (CC) / Selected	Forced Chief Complaint? 0 = No 1 = Yes	All Symptoms Recognized by System? 0 = Yes 1 = No	Age Limitation for System? 0 = No 1 = Yes	Patient Forced to Self-Diagnose? 0 = No 1 = Yes	Patient Asked to Self-Diagnose? 0 = No 1 = Yes	# of Questions Asked? Avg = Can't Process	Correct Condition on List? 0 = Yes 1 = No	Care Direction Correct? 0 = Yes, 1 = No *or* Care Direction Variable by Condition? 0 = No, 1 = Yes
System C• Case 2	Vomiting	1	1	1	1	0	45	1	1
	Rapid Breathing	1	1	1	1	0	45	1	1
	Abdominal Pain	1	1	1	1	0	45	1	1
System C• Case 1	Weight Gain	1	0	1	1	0	47	0	1
	Heavy Menstrual Periods	1	0	1	1	0	43	0	1
	Constipation	1	0	1	1	0	48	0	1
System B• Case 2	Vomiting	1	1	0	1	1	30	1	1
	Rapid Breathing	1	1	0	1	1	31	1	1
	Abdominal Pain	1	1	0	1	1	28	1	1
System B• Case 1	Weight Gain	1	0	0	1	1	32	1	1
	Heavy Menstrual Periods	1	0	0	1	1	32	1	1
	Constipation	1	0	0	1	1	32	1	1
System A• Case 2	Vomiting	1	1	1	1	1	29	1	1
	Rapid Breathing	1	1	1	1	1	29	1	1
	Abdominal Pain	1	1	1	1	1	29	1	1
System A• Case 1	Weight Gain	1	1	1	1	1	29	1	1
	Heavy Menstrual Periods	1	0	1	1	1	29	1	1
	Constipation	1	0	1	1	1	29	1	1
Isabel• Case 2	Not Required	0	0	0	0	0	11	0	0
	Not Required	0	0	0	0	0	11	0	0
	Not Required	0	0	0	0	0	11	0	0
Isabel• Case 1	Not Required	0	0	0	0	0	11	0	0
	Not Required	0	0	0	0	0	11	0	0
	Not Required	0	0	0	0	0	11	0	0

About Isabel Healthcare

Isabel Healthcare was started in 1999 after the founder's daughter, Isabel, suffered a near fatal misdiagnosis. She was three years old and had the chicken pox. Her local family doctor and emergency department doctors all missed a secondary infection which turned out to be necrotizing fasciitis. Isabel spent three weeks in intensive care and four weeks in a high dependency unit. She survived and is a healthy young woman.

Isabel has been a proven diagnosis decision support system used by clinicians around the world. Over 30 articles have appeared in peer-reviewed articles covering various aspects of the system. The system was selected by the American Medical Association as the diagnosis tool for its portal. More recently, the *British Medical Journal* (BMJ) endorsed *Isabel* as a new joint product that was launched incorporating the BMJ's Best Practice tool.

Today, many high-profile health systems, family practices and individual physicians use *Isabel* to help improve the quality of care they provide. *Isabel* uses a database of over 10,000 diagnoses, of which 6,000 are diseases and 4,000 are drugs. This database has been manually built and populated over nearly two decades with knowledge about how each disease presents from a multitude of sources.

The *Isabel Symptom Checker*

The *Isabel Symptom Checker* is a unique and powerful tool designed to empower and engage healthcare consumers in your network. Adapted from Isabel's professional differential diagnosis decision support tool, *Isabel Symptom Checker* has been reengineered for patients and consumers. With the help of the highly sophisticated *Isabel Symptom Checker* and its seamlessly integrated, evidenced-based *Health Library* content, patients have instant access to a trusted resource for symptom and disease information. Symptoms can be entered using everyday language, providing patients with the information they need to engage in meaningful conversations about their personal wellness with a healthcare provider.

Learn More **Request a Demo**