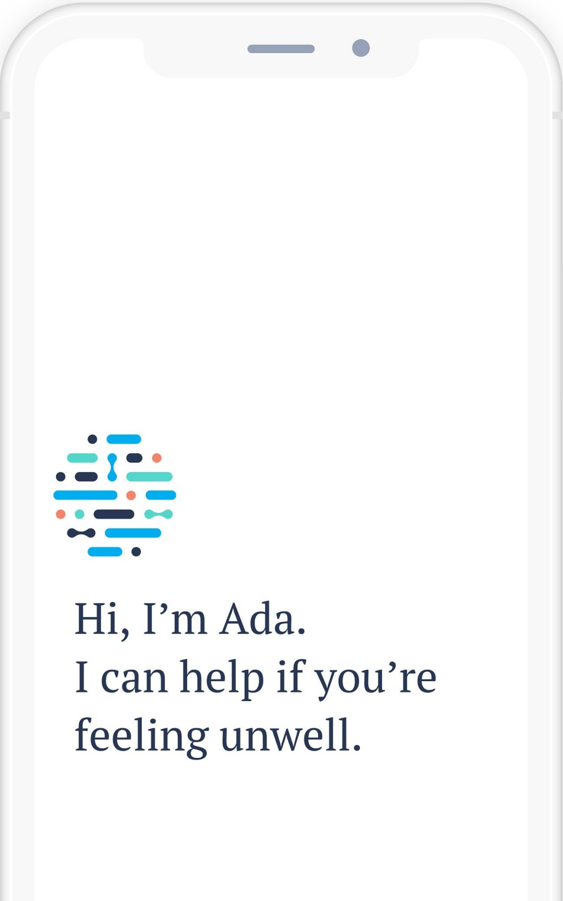
Ada Health

Our approach to assess Ada's diagnostic performance.

ITU Workshop on Artificial Intelligence for Health Session 3: Data Availability and Benchmarking Geneva, Switzerland, 25 September 2018

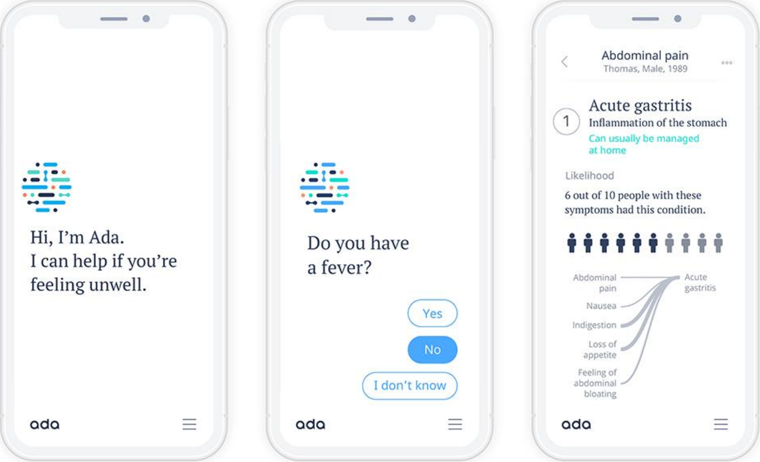


Hi, I'm Ada.
I can help if you're feeling unwell.

What is Ada?

Diagnostic decision support systems

- **Ada App**
  - End-user self-assessment app
  - AI based chatbot
  - Assessment report with possible next steps
- **Ada DX**
  - Professional diagnostic investigation tool
  - For doctors / experts



Hi, I'm Ada.
I can help if you're feeling unwell.

Do you have a fever?

Yes

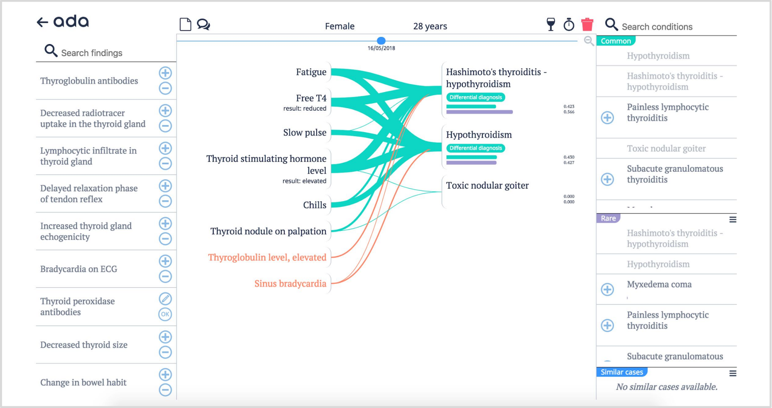
No

I don't know

Abdominal pain
Thomas, Male, 1989

1 Acute gastritis
Inflammation of the stomach
Can usually be managed at home

Likelihood
6 out of 10 people with these symptoms had this condition.



Search findings

Thyroglobulin antibodies
Decreased radiotracer uptake in the thyroid gland
Lymphocytic infiltrate in thyroid gland
Delayed relaxation phase of tendon reflex
Increased thyroid gland echogenicity
Bradycardia on ECG
Thyroid peroxidase antibodies
Decreased thyroid size
Change in bowel habit

Fatigue
Free T4
result: reduced
Slow pulse
Thyroid stimulating hormone level
result: elevated
Chills
Thyroid nodule on palpation
Thyroglobulin level, elevated
Sinus bradycardia

Hashimoto's thyroiditis - hypothyroidism
Differential diagnosis
0.423
0.566

Hypothyroidism
Differential diagnosis
0.430
0.427

Toxic nodular goiter
0.000
0.000

Search conditions

Common
Hypothyroidism
Hashimoto's thyroiditis - hypothyroidism
Painless lymphocytic thyroiditis
Toxic nodular goiter
Subacute granulomatous thyroiditis

Rare
Hashimoto's thyroiditis - hypothyroidism
Hypothyroidism
Myxedema coma
Painless lymphocytic thyroiditis
Subacute granulomatous

Similar cases
No similar cases available.

The Big Question

How can you be sure that your Health AI is good enough to give advice to real users?

It's about Health - not Pizza-Delivery

Symptoms & Finding
Total number about 8000-150000

Conditions & Diseases
Strongly depends on definition but about 15000-120000.

Closed Testing not Feasible
Assuming cases with 20 findings there are 8000^20 possible cases.

Ada’s QA System

You need a quality measurement system

- **Reference Cases (about 5k)**
  - Model-Cases (2-3 per disease)
  - Literature cases (if available)
  - Problematic user cases (on demand)
- **Full Evidence Measures**
  - P-M1, P-M3, P-M10
- **Question Flow Measures**
  - Only using questions asked by the AI
  - Checking for expected questions
- **Analysis and Visualization Tools**

Model Cases
A basic “Unit-Test” checking that the AI responds to the most obvious cases of the disease at all.

Literature Cases
Nicely prepared cases from books like “80 cases of Neurology”. Journals publishing cases.

User Cases
Confirmed user cases where the user told us a different diagnosis.

P-Mx-Measures
Probability to see the correct diagnosis within the top *x* results of the results ranked by probability

Sounds Simple, But …

How can we standardize something like this?

What do we need?

Symptom & Finding Ontology

Condition & Disease Ontology

Non-Clinical Triage Standard

Representative Cases

Agree on Metrics

Setup a Testing Framework

Agree on Ontologies

For speaking the same input/output language

- **Symptom & Finding & Factor Ontology**
  - Attribute support
  - Hierarchy support
  - No redundancy, no overlap
- **Condition Ontology**
  - Common conditions
  - Rare condition
  - Idiopathic conditions
- **Pre Clinical Triage Ontology**

Symptoms
The evidence provided by patients/users including their presenting complaints.

Findings
The evidence gathered by doctors, nurses, examinations, devices

Idiopathic
If symptoms occur without an underlying disease in a more or less healthy person

Pre Clinical Triage
There are standards for clinical triage but not for different shades of *“see your doctor soon”*

Reference Cases

Representative for all diseases

- **2-3 Cases Per Disease**
  - Simple case
  - Realistic & complex cases
  - Relevant stages & presentations
- **Covering All Relevant Features**
  - Negative evidence
  - Attributes & Factors
  - Multimorbidity?
- **Expected Output**
  - Peer reviewed
  - Carefully curated

Early Stage vs. Terminal
Many diseases change considerably over time, so you need several cases.

Exclusion Factor Testing
It’s a good idea to have tests for the impossible - it’s surprising how often you can provoke pregnant man.

Peer Review Needed!
Especially non experts show a high inter annotator variance - and sometimes even systematic bias.

Confirmed Diagnoses
You cannot expect a later confirmed diagnosis to rank #1 in an early stage.

Multimorbidity
Almost every user has more than one disease and they sometimes interfere.

Quality Measures

Measuring what a good result looks like

- **Rank Influence**
  - If the correct disease is not the top match
- **Prior scaling**
  - You want to see performance of rare diseases and the net performance in the real world
- **Define the role of emergency disease**
  - Maybe a less likely disease should be ranked higher if it requires immediate action
- **Define role of the questions flow details**
  - Number of questions asked, etc.

Ada - Learning
Doctors say they want disease ranked by probability but often they want as second place the disease that would win in case the top match would be wrong.

Expected values
If you have 2-3 cases per diseases the rare diseases are over represented in the average score, so you need to scale cases with the prior of their correct diagnosis to get the real-world-performance.

Multimorbidity
For systems supporting multimorbidity the scoring the results is much more complex.

Testing Framework

Define a fair but resilient testing methodology

- **Constraints**
  - Provider cannot give you the AI since it runs in a large cluster in a cloud and is their core IP
- **Real Time API Testing and Analysis**
  - Every provider can get real time quality test values at analysis at any time
  - E.g. on all data from former tests
- **Official Benchmarking**
  - Official testing on a regular base with new data
  - Publication of the results
  - ≤ every 6 month

Cases vs. Test Frequency
Six months is almost too long but getting 2-3 high quality cases per disease more often is difficult.

Published Benchmark Results
Can then provide the foundation to apply AI in certain scenarios e.g. WHO LMIC projects.

It’s challenging!
But it’s definitely worth the effort!

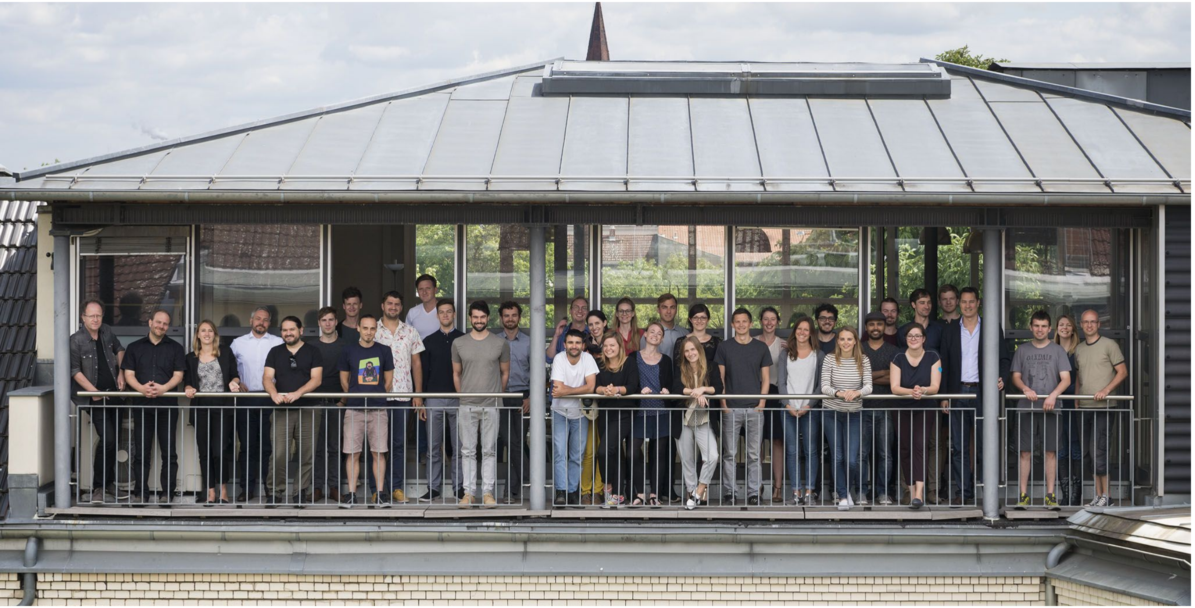
Thank you!

Henry Hoffmann
henry.hoffmann@ada.com

www.ada.com

Ada Health GmbH
Adalbert Straße 20
10997 Berlin
Germany

Ada today



> 7 years of research

> 100 employees

40 medical experts

5M users

> 7M assessments

> 100,000 ratings

#1 medical app in > 130 countries

CANNES LIONS 2017 SILVER WINNER

GLOMO Global Mobile Awards WINNER

Deutschland Land der Ideen

European Commission

bsi. ISO/IEC 27001 Information Security Management

CE

ada