CLEF 2018
CHS Task Report

(Task 3)

Dr Guido Zuccon
Queensland University of Technology
g.zuccon@qut.edu.au

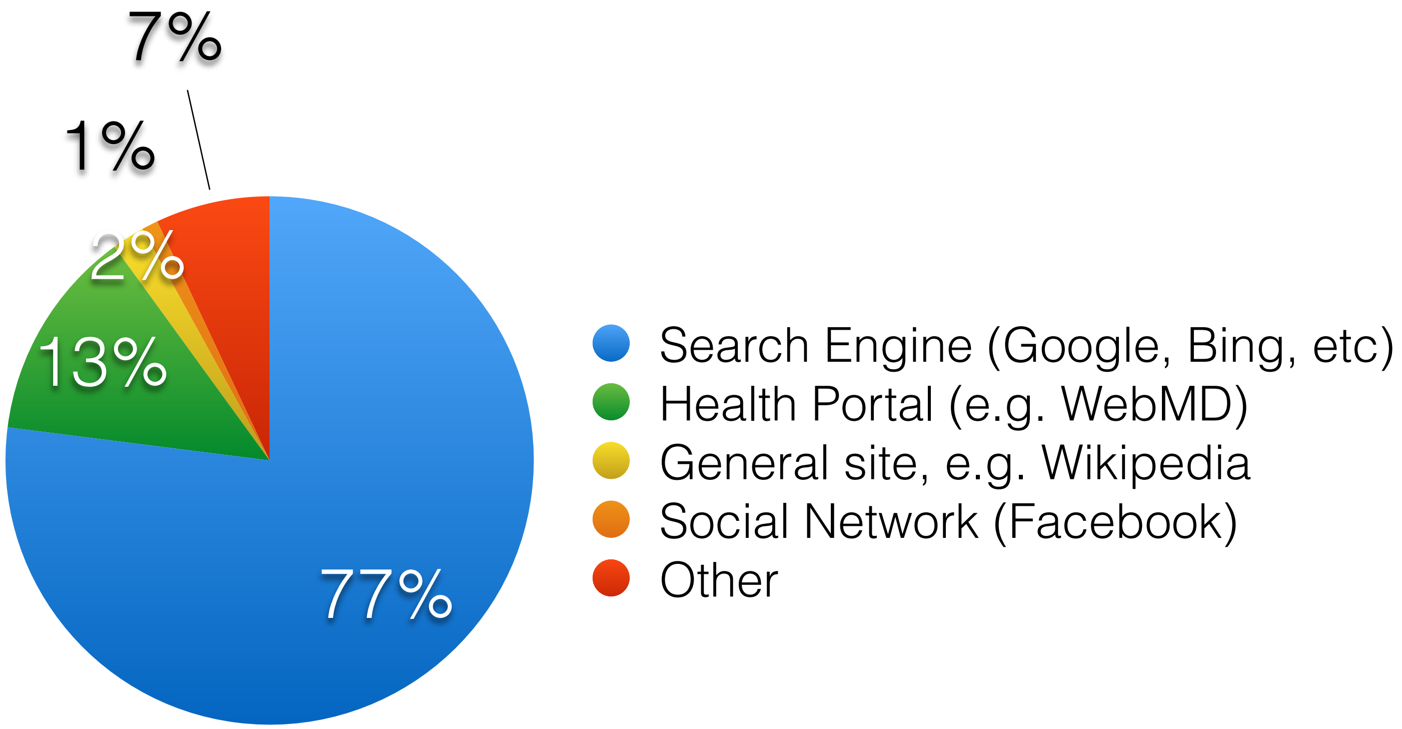
Task Organisers: Jimmy (QUT), Joao Palotti (QCRI, TUW), Guido Zuccon (QUT), Lorraine Goeuriot (UGA), Liadh Kelly (MU)

Why consumer health search?

Studies showing that a large majority of people seek health information online: e.g. 80% in Pew Research survey (2012)

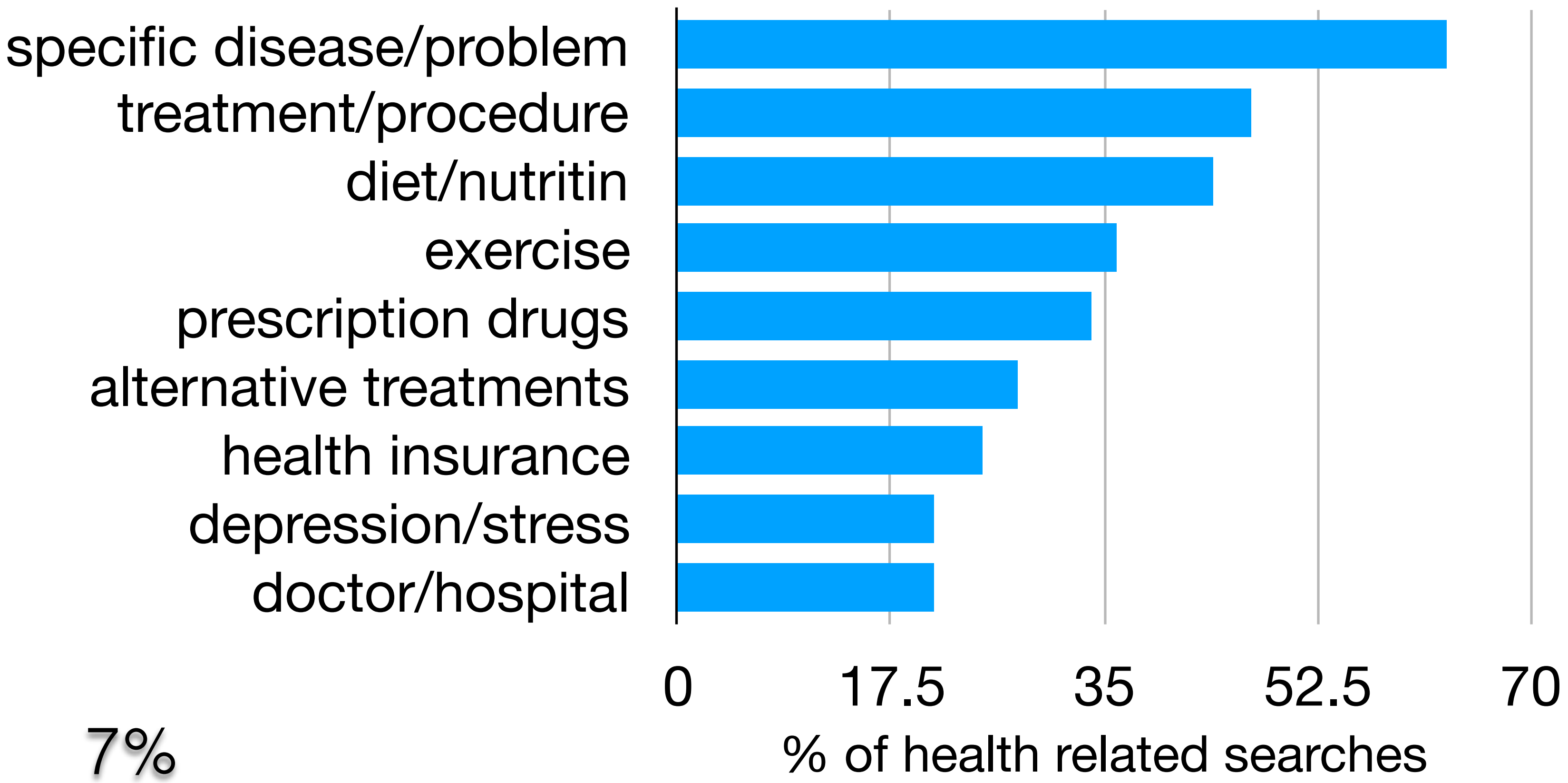
Why consumer health search?

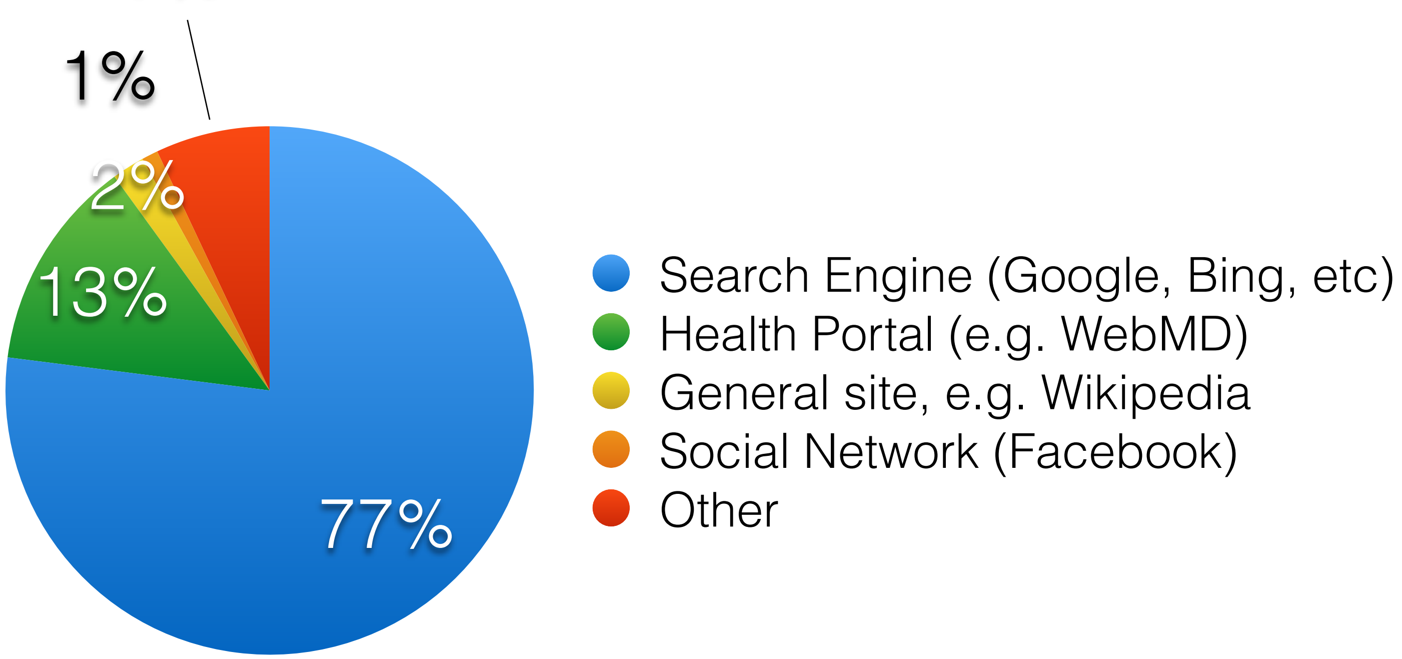
Studies showing that a large majority of people seek health information online: e.g. 80% in Pew Research survey (2012)



Why consumer health search?

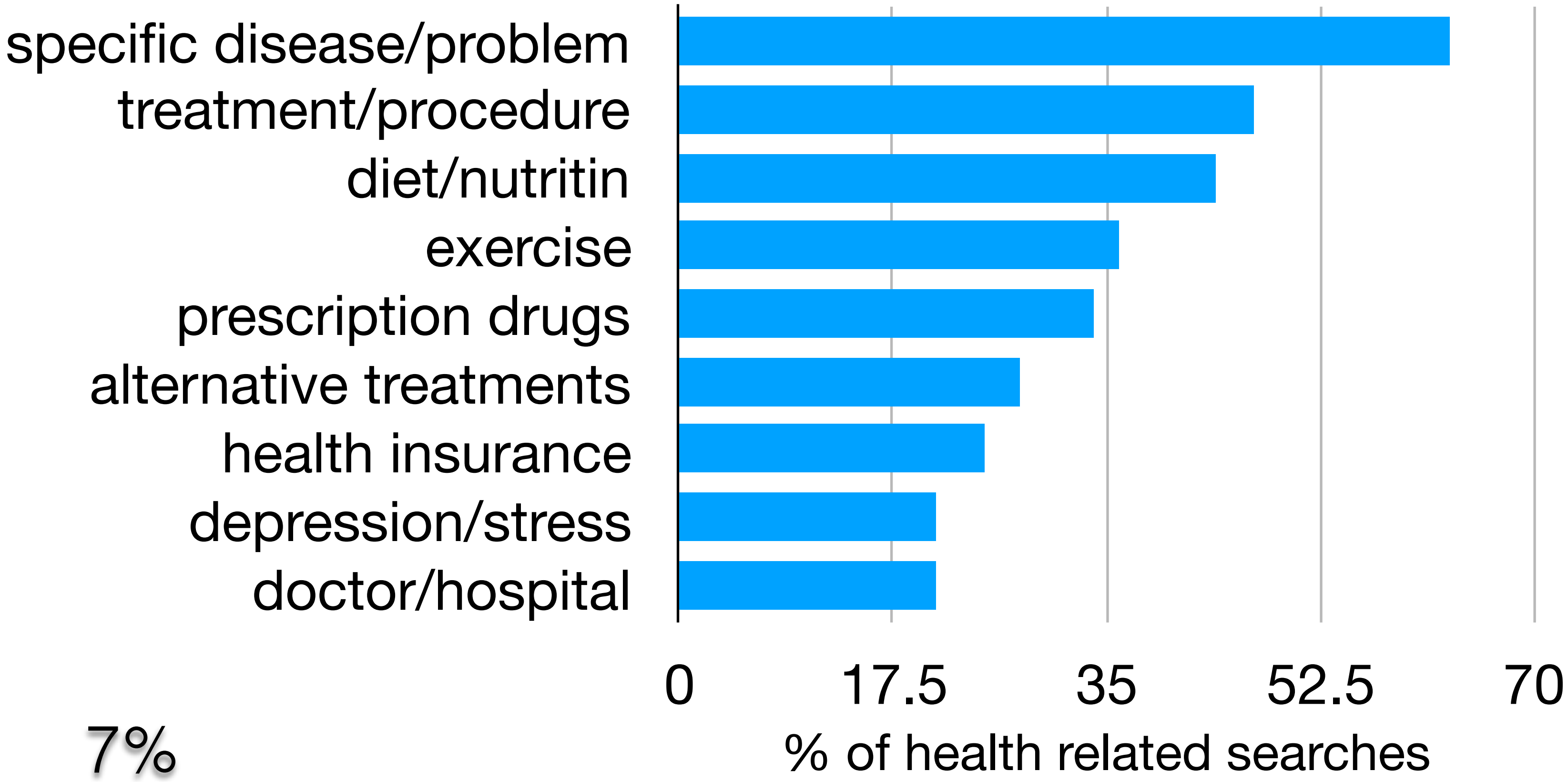
Studies showing that a large majority of people seek health information online: e.g. 80% in Pew Research survey (2012)

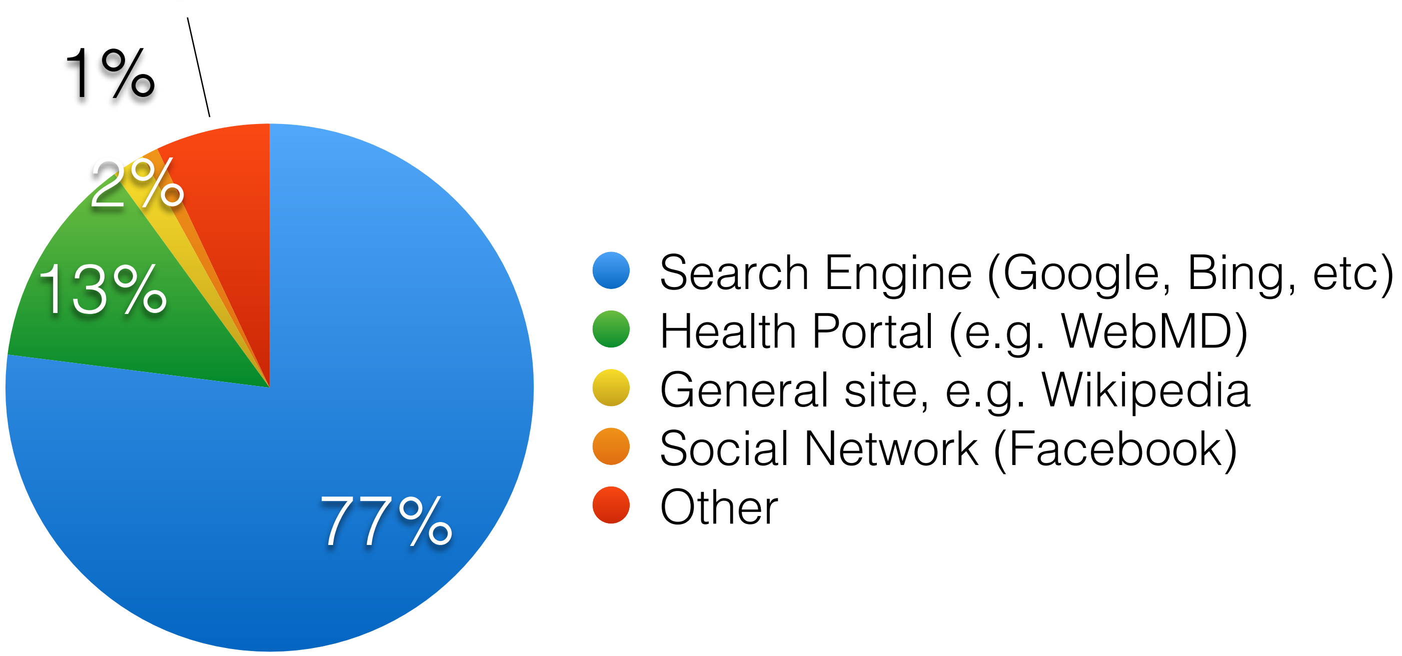


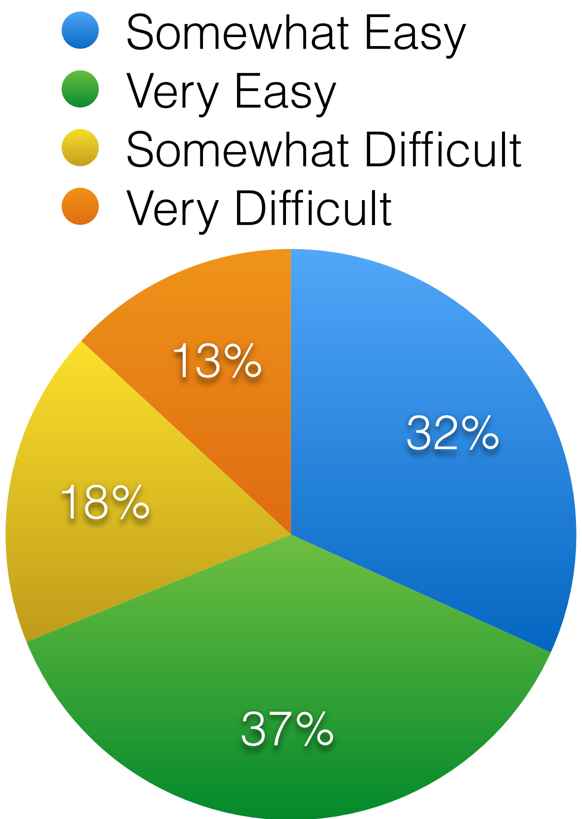


Why consumer health search?

Studies showing that a large majority of people seek health information online: e.g. 80% in Pew Research survey (2012)

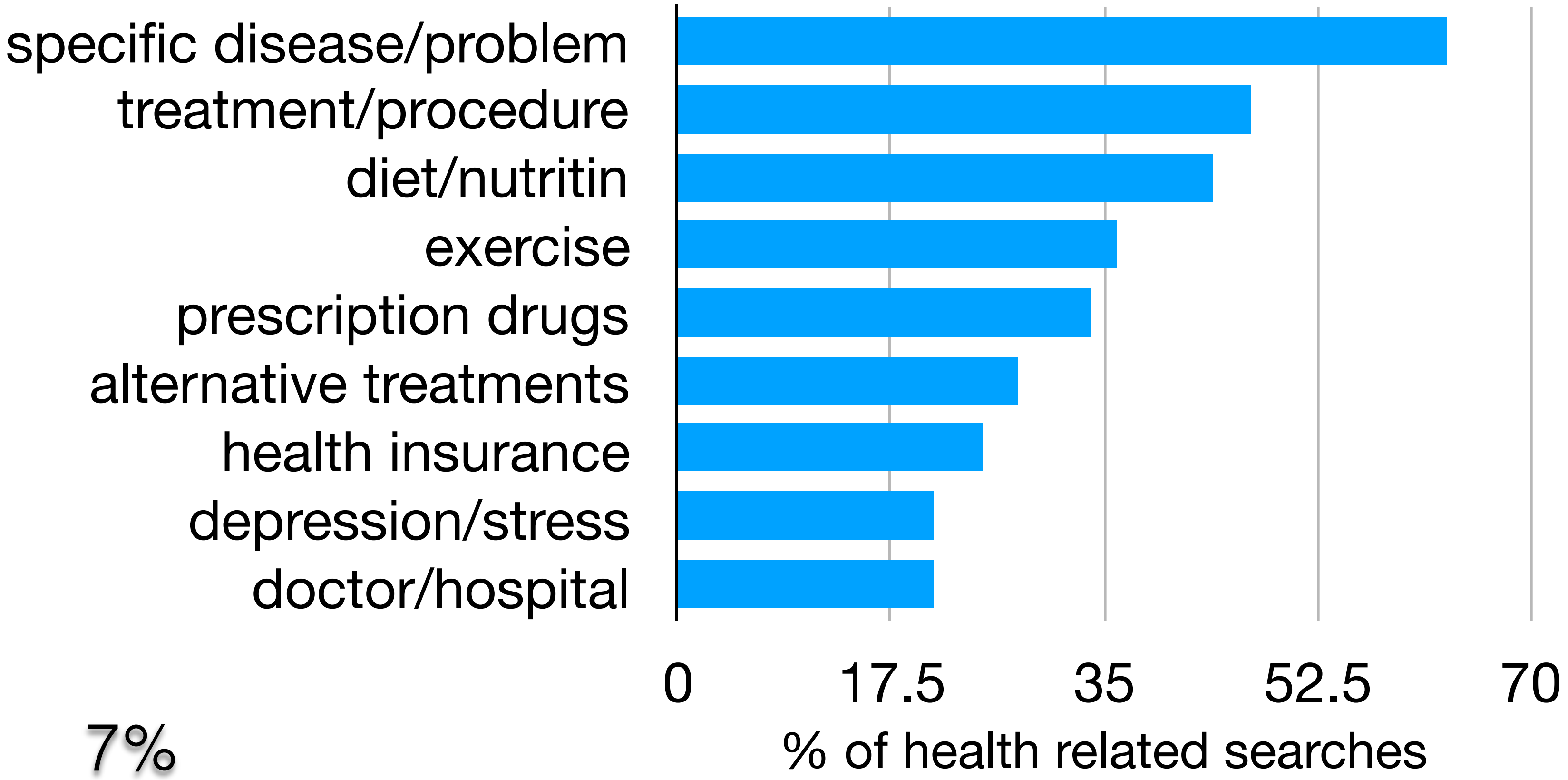


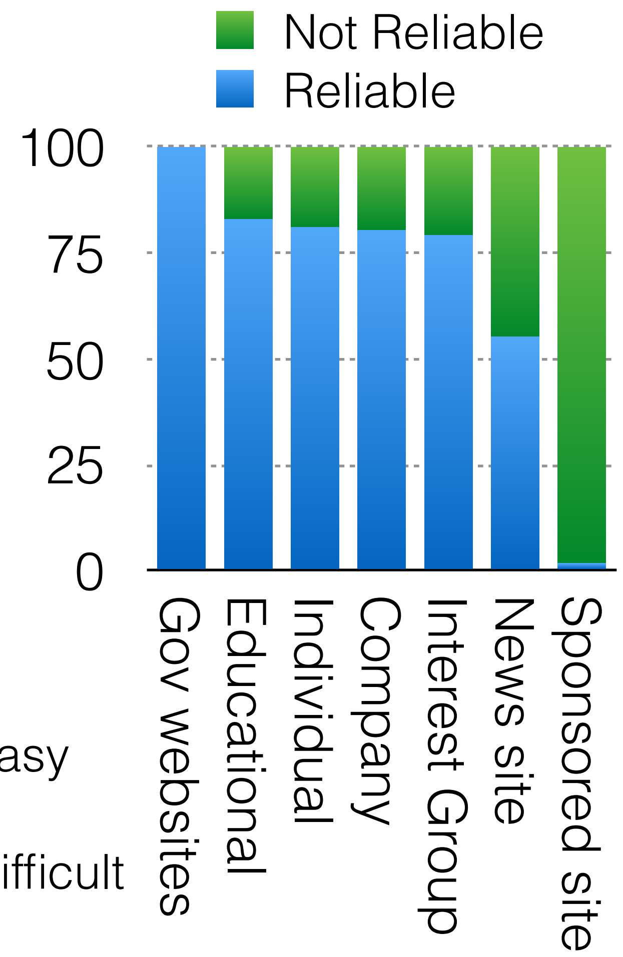


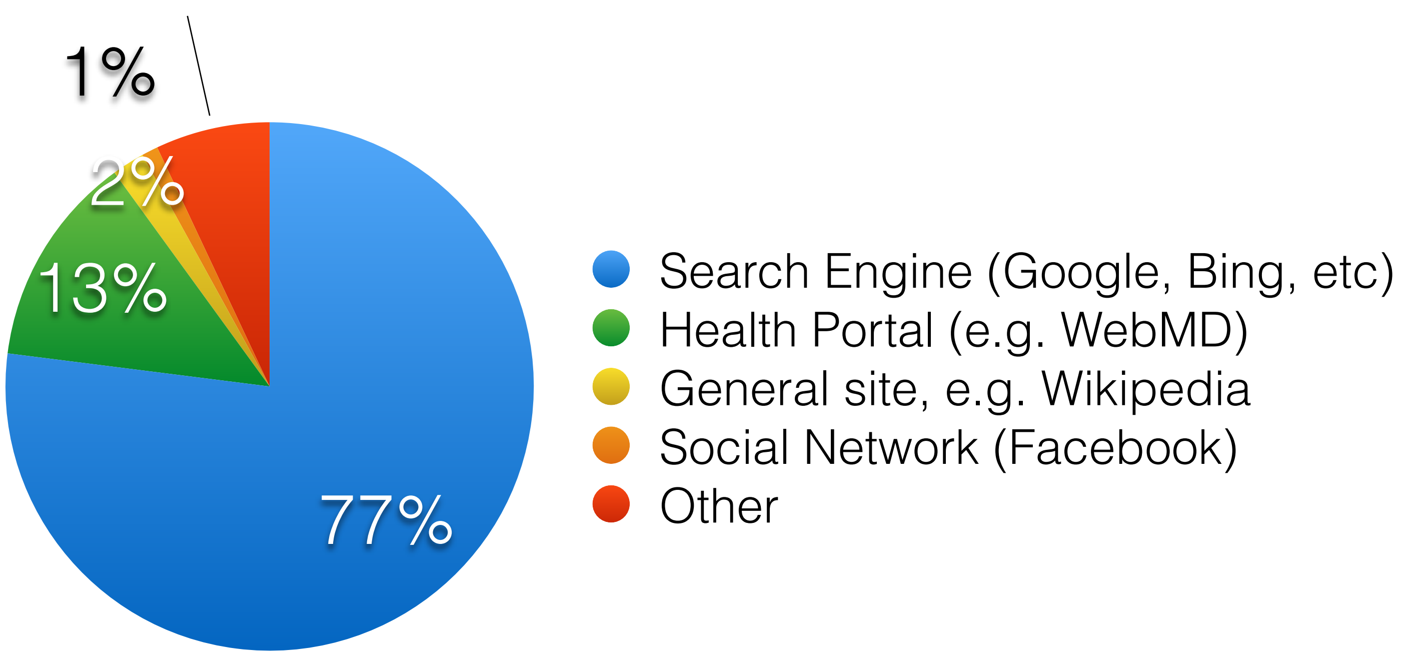


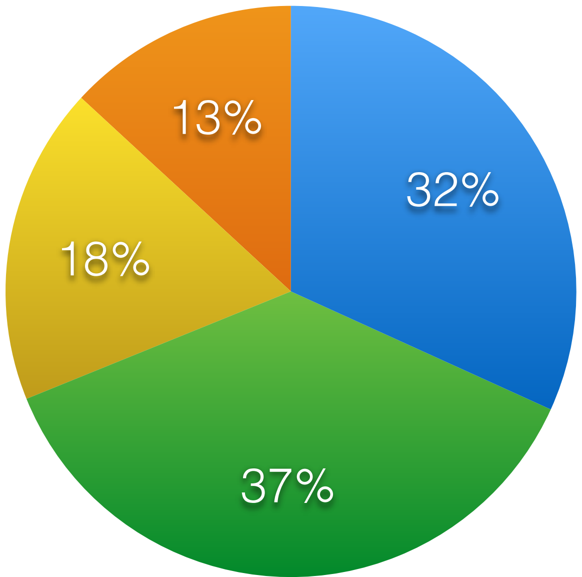
Why consumer health search?

Studies showing that a large majority of people seek health information online: e.g. 80% in Pew Research survey (2012)



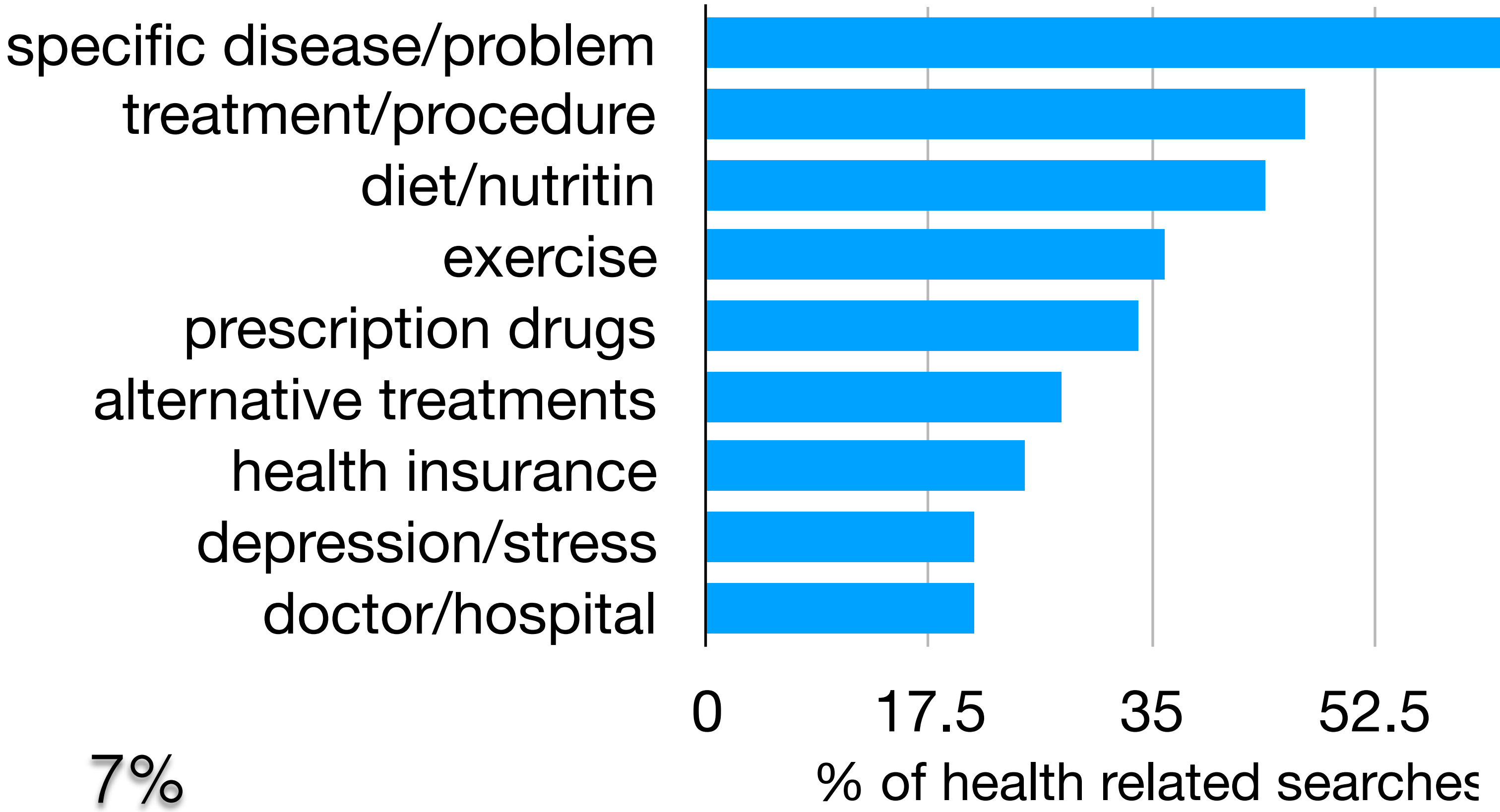


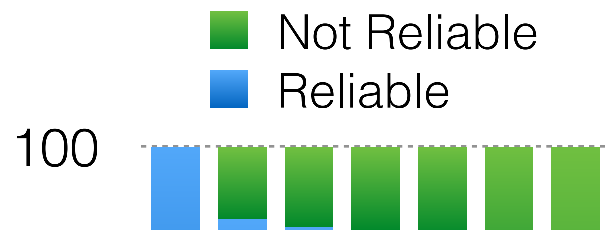


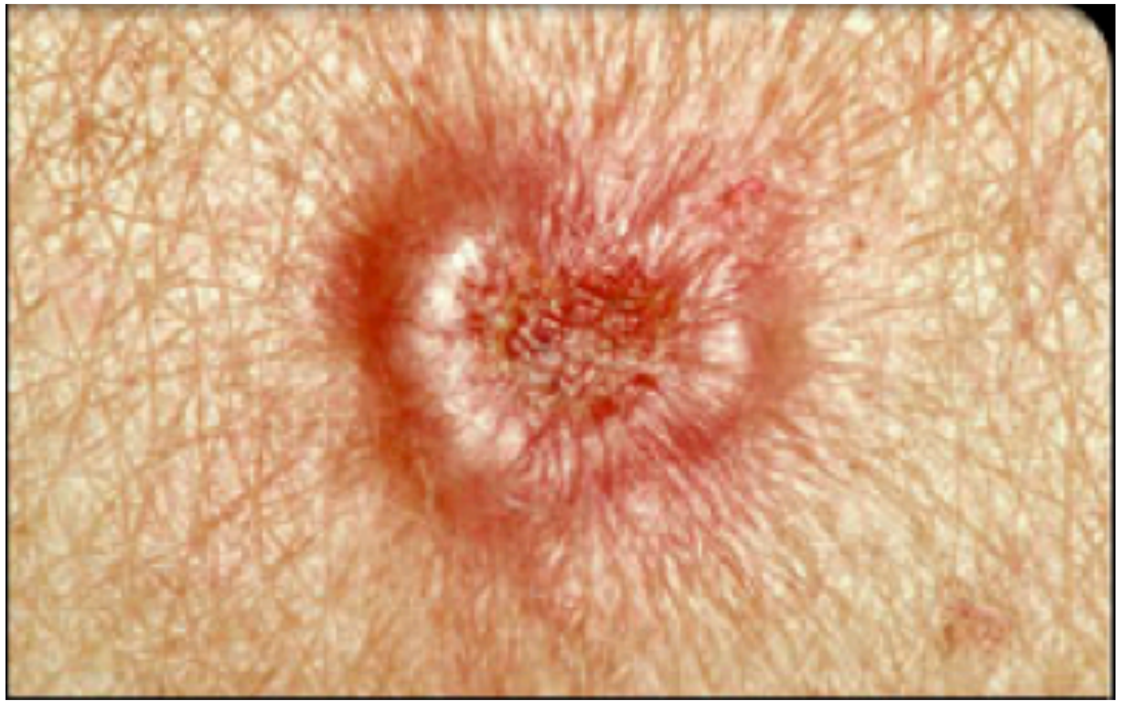


Why consumer health search?

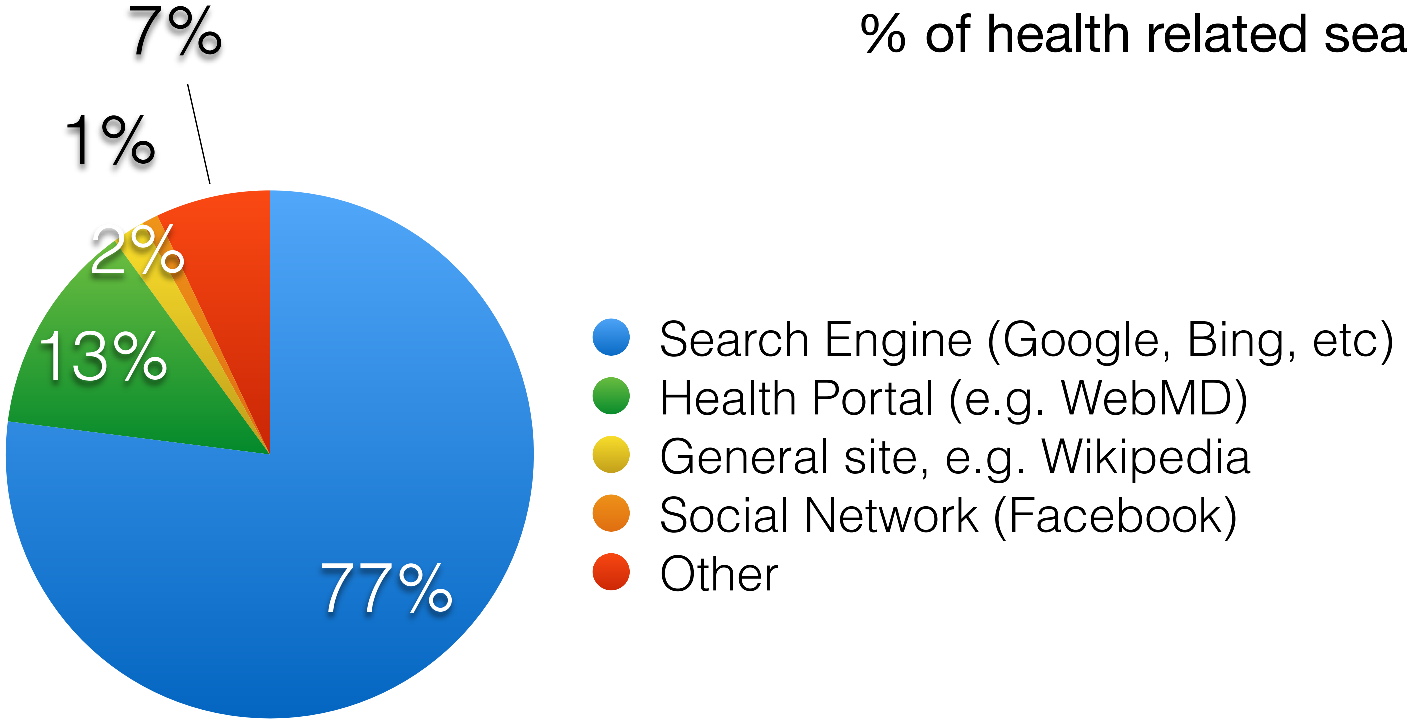
Studies showing that a large majority of people seek health information online: e.g. 80% in Pew Research survey (2012)







What would be your query to Google if you have this on your skin?



CHS Task @ CLEF 2016/2017

Topics

Generated by genuine user questions made on Reddit www.reddit.com/r/AskDocs/

6 query variations for each of the 50 information need

Corpus

50 million web pages from Clueweb 12B

Terrier, Indri and Elastic Search Indices were available on purpose Azure platform

Assessments

Assessments made by experts with respect to:

- Topical Relevance
- Understandability
- Trustworthiness

Assessments from 2016 and 2017 to be used together for deeper pool

Challenges for 2018:

- Participants found difficult to work with a collection as large (and not focused) as Clueweb 12B
- Budget reductions: no Azure anymore, less money for relevance assessments
- Need for new query set

The CLEF eHealth CHS 2018 Task

- New corpus, more focused on health topics than Clueweb12; also much smaller
- New set of queries, derived from EU Project Khresmoi, and augmented with query variations collected as part of the task
- Use of crowdsourcing (AMT) for relevance assessments
- 5 sub-tasks

The Clefehealth2018 corpus

- Web pages were acquired from the **CommonCrawl**, for a target domain
- Note, **we downloaded <u>all</u> pages available for a target domain**
- To **select target domains** to download:
    - We acquired the **top 50 domains retrieved by Bing** in answer to the queries used in the collection
        - Note, not all domains were in CommonCrawl
    - We compiled a list of **trustworthy health webpages** (e.g. webmd, etc), and got the associated data from CommonCrawl
    - We compiled a list of **not-trustworthy health webpages** and got the associated data from CommonCrawl
- This resulted in a corpus of **2,021 domains** (490 GB uncompressed).
- **Clefehealth2018_B**: smaller subset of the corpus (1,653 domains, 294 GB uncompressed)
    - Removed large domains, removed domains associated to newswire outlets, and more

The 2018 Queries

- Acquired **50 queries** issued by the general public to the **HON search service** (thanks to Khresmoi EU Project)
  - Each query came with a **classification into a taxonomy of health search intents** — more later
- For each query, we created a **user story (topic)**: we inferred this based on the query and the classification
- The topic was used to collect new **query variations** for the topic
  - 5 users recruited in lab; shown description, asked what they would query (and they entered the query to Google)
    - Note, there may be query duplicates among the variations
- Thus, the collection has **50 topics**. Each topic has
  - a **description**/narrative/user story
  - the **original query** issued to HON (queryid finishing with 1)
  - **5 additional query variations**, derived from the topic description

The 2018 Queries

```
Topic: 191
Your doctor said he wants to run some blood exams to test your liver
functions. You want to find out more information about how these tests could
be used to understand the correct behaviour of your liver functions.

<query> <id>191001</id>
        <en>liver function test uses</en>
</query>
<query> <id>191002</id>
        <en>blood test liver function</en>
</query>
<query> <id>191003</id>
        <en>how do blood tests inform liver function</en>
</query>
<query> <id>191004</id>
        <en>liver function blood test </en>
</query>
<query> <id>191005</id>
        <en>liver functions blood test understanding results and
treatments</en>
</query>
<query> <id>191006</id>
        <en>blood exam liver function</en>
</query>
```

The topic description

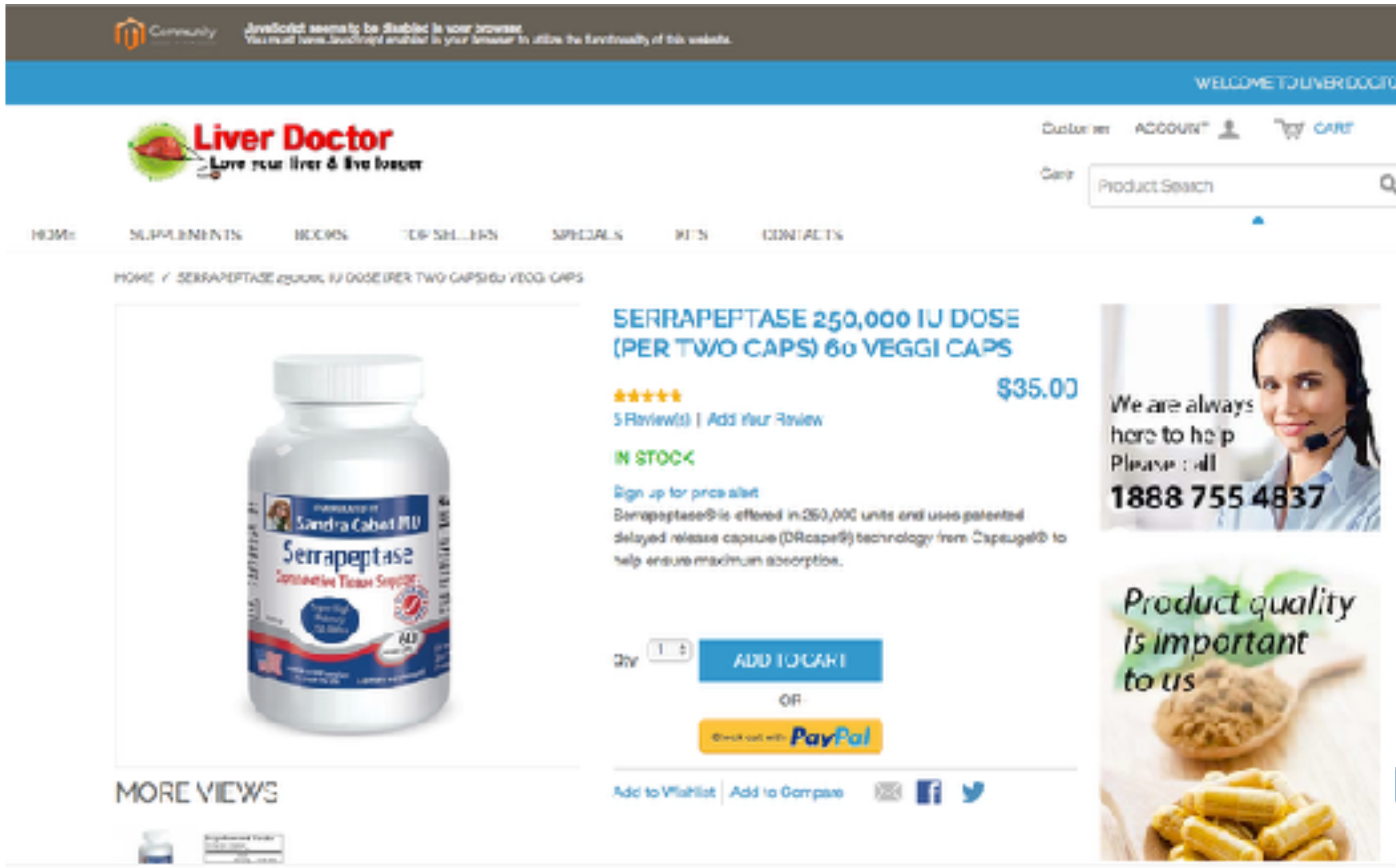
The original HON query

Sourced variations

The 2018 Assessments

- **Crowdsourcing** (AMT) was used to collect assessments. **Quality Control**:
    - Limit to US workers, 99% acceptance rate, at least 1000 HITs approved
    - In-lab, careful assessments of manually selected **honey-pots**
        - Examples of relevant + examples of irrelevant
    - For each topic, released HITs to AMT:
        - Each HIT batched **10 web pages** for assessment + 2 **honey-pots**
- AMT workers were shown topic description, and asked to assess: topicality, understandability, trustworthiness.
- **Pooling** using RBP Method A (Summing contributions): weight each document according to overall contribution to the effectiveness evaluation as provided by RBP
    - Selected the top 500 weighted documents per topic; Pool size: 25,000

The 2018 Assessments: Honey-pot Examples



Community JavaScript seems to be disabled in your browser. You must have JavaScript enabled in your browser to utilize the functionality of this website.

WELCOME TO LIVER DOCTOR!

Liver Doctor
Love your liver & live longer

Customer ACCOUNT CART

Cart Product Search

HOME SUPPLEMENTS BOOKS TOP SELLERS SPECIALS KITS CONTACTS

HOME / SERRAPEPTASE 250,000 IU DOSE (PER TWO CAPS) 60 VEGGI CAPS

SERRAPEPTASE 250,000 IU DOSE (PER TWO CAPS) 60 VEGGI CAPS

$35.00

★★★★★
5 Review(s) | Add Your Review

IN STOCK

Sign up for price alert

Serrapeptase® is offered in 250,000 units and uses patented delayed release capsule (DRcaps®) technology from Capsugel® to help ensure maximum absorption.

Qty 1 ADD TO CART

OR

Check out with PayPal

Add to Wishlist | Add to Compare

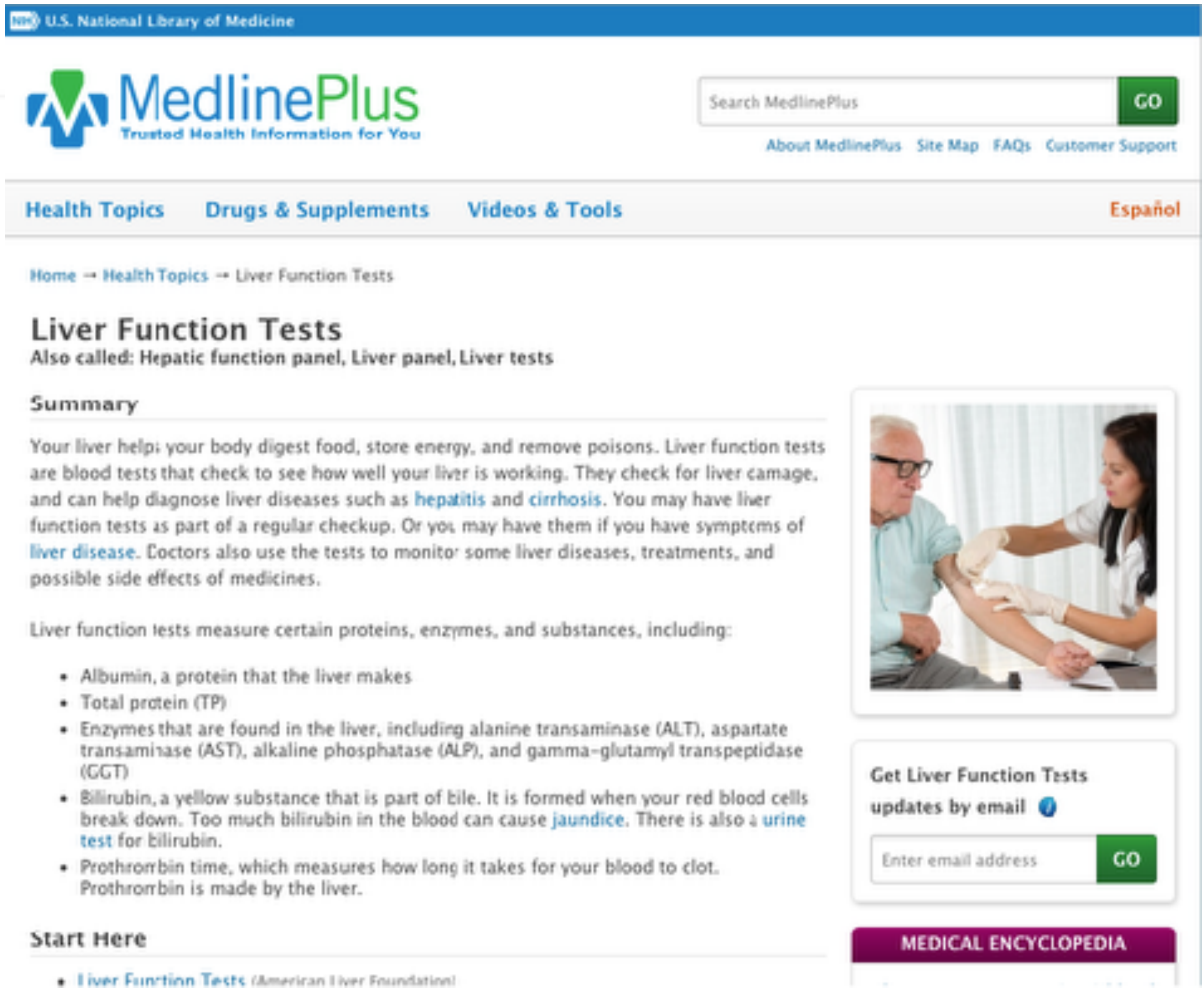
MORE VIEWS

We are always here to help
Please call
1888 755 4837

Product quality is important to us

Non-Relevant Honey-pot

Relevant Honey-pot



U.S. National Library of Medicine

MedlinePlus
Trusted Health Information for You

Search MedlinePlus GO

About MedlinePlus Site Map FAQs Customer Support

Health Topics Drugs & Supplements Videos & Tools Español

Home → Health Topics → Liver Function Tests

Liver Function Tests

Also called: Hepatic function panel, Liver panel, Liver tests

Summary

Your liver helps your body digest food, store energy, and remove poisons. Liver function tests are blood tests that check to see how well your liver is working. They check for liver damage, and can help diagnose liver diseases such as hepatitis and cirrhosis. You may have liver function tests as part of a regular checkup. Or you may have them if you have symptoms of liver disease. Doctors also use the tests to monitor some liver diseases, treatments, and possible side effects of medicines.

Liver function tests measure certain proteins, enzymes, and substances, including:

- Albumin, a protein that the liver makes
- Total protein (TP)
- Enzymes that are found in the liver, including alanine transaminase (ALT), aspartate transaminase (AST), alkaline phosphatase (ALP), and gamma-glutamyl transpeptidase (GGT)
- Bilirubin, a yellow substance that is part of bile. It is formed when your red blood cells break down. Too much bilirubin in the blood can cause jaundice. There is also a urine test for bilirubin.
- Prothrombin time, which measures how long it takes for your blood to clot. Prothrombin is made by the liver.

Start Here

- Liver Function Tests (American Liver Foundation)

Get Liver Function Tests updates by email

Enter email address GO

MEDICAL ENCYCLOPEDIA

THE SUB-TASKS

IRTask1: Ad-Hoc Retrieval

6 participants

- For each query, submit a ranking of documents, so as to maximise the relevance of the top results
- Evaluation:
  - User interested in only the first page of results:
    Precision at 10 (P@10)
  - User interested in only the first page of results, but cares about ranking:
    Normalized Discounted Cumulative Gain, depth 10 (NDCG@10)
  - User prefer relevant result at early ranks, but explore the ranks with medium-patience (i.e. beyond rank 10):
    Rank Biased Precision with μ=0.8 (RBP(0.8))

CHSTask2: Personalised Search

1 participant

- **Users have different abilities in understanding** content of health web pages
- A document that an health expert finds easy to understand, may be difficult for a common person
- If a user does **not understand** a document, does **not get gain (utility)**, regardless of the topical relevance of the document
- Task: systems need to retrieve&rank relevant documents that match the user understandability profile

CHSTask2: Evaluation

- Each user is characterised by an **understandability level parameter** (alpha); assessments wrt relevance, understandability and trustworthiness (not used)
- alpha parameter and understandability assessments expressed on the same scale: 0 - 1 (0-> easiest level; 1 -> hardest level). Example:
    - A user struggles to read web pages with medical terms/content: a small alpha, e.g., 0.20
    - A judge thinks that document D1 is quite easy to understand: score 0.15
    - document D2 is hard to understand : score 0.90
- Mapped fixed alpha to each query variation: thus each topic has a range of understandability pages required, e.g. variation 1: alpha = 0.2; variation 2: alpha 0.4;…
- Use understandability-biased measures for evaluation: uRBP + MM_RBP

CHSTask3: Query Variations

1 participant

- Each topic has 5 query variations
- The query variations capture the variability intrinsic in how people formulate queries when searching to answer the same information need
- Participants asked to exploit dependence between query variations, and submit one result ranking for each topic (i.e. a common one for all 6 variations)
- Evaluation: mean-variance evaluation framework (MVE)
  - Evaluation results for each query variation for a topic were **averaged**, and their **variance** also accounted for to compute final system performance

$$\mu - \alpha\sigma^2$$

CHSTask4: Multilingual Ad-hoc

1 participant

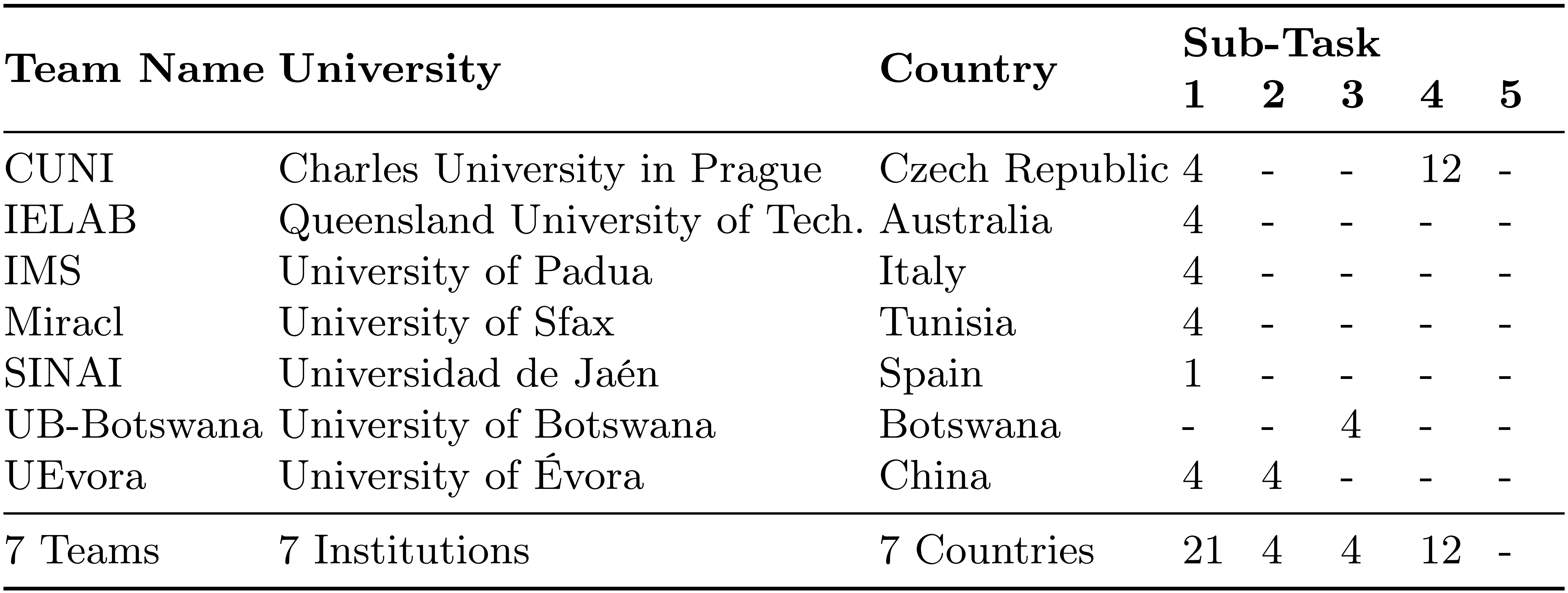
- Multilingual translations for each query
- Languages: Czech, French, and, German
- Evaluation: same as per IRTask1

CHSTask5: Multilingual Ad-hoc

0 participants

- Classify queries with respect to the underlying health intent
    - a query may have multiple intents
    - participants asked to submit top 3 intent predictions
- 8 high-level intents: (1) Disease/illness/syndrome/pathological condition, (2) Drugs and medicinal substances, (3) Healthcare, (4) Test & procedures, (5) First aid, (6) Healthy lifestyle, (7) Human anatomy, (8) Organ systems.
- For each high-level intent, up to 13 low-level sub-intents
- Evaluation: mean reciprocal rank (MRR) and NDCG@1,2,3.
    - Matches at high-levels intents are differentiated from matches at low-levels intents.

Participants Runs



Team Name	University	Country	Sub-Task				
			1	2	3	4	5
CUNI	Charles University in Prague	Czech Republic	4	-	-	12	-
*IELAB	Queensland University of Tech.	Australia	4	-	-	-	-
IMS	University of Padua	Italy	4	-	-	-	-
Miracl	University of Sfax	Tunisia	4	-	-	-	-
SINAI	Universidad de Jaén	Spain	1	-	-	-	-
UB-Botswana	University of Botswana	Botswana	-	-	4	-	-
UEvora	University of Évora	China	4	4	-	-	-
7 Teams	7 Institutions	7 Countries	21	4	4	12	-

*: Organisers contributed (-2 from 2017)

Although scarce participants to sub-tasks 2-5, relevance assessments can still be used, though may find systems have large number of unassessed for this task (thus unreliable)

NEXT STEPS

CHS Task @ CLEF 2019

- New queries? Or re-use existing queries, to enlarge assessment pool (more solid collection)? Will know after analysing the results

- Need to increase participation to ensure diversity in approaches/ reliability of collection

- Stabilise current sub-tasks (but no sub-task 5) + New sub-tasks?

    - Spoken Queries? Conversational Search?

    - Answer Card Generation?

    - Tell us your ideas/feedback!

- Come to see the CLEF 2018 participants submissions on

- We need you to help us building a reliable collection: Join us for the CLEF 2019 tasks!

CLEF
eHealth
2018

- GitHub repository for task: https://github.com/CLEFeHealth/CLEFeHealth2018IRtask
    - Evaluation scripts, queries, assessments, runs, etc.

- Slack team: https://goo.gl/ShnHR2