Clustering of Medical Publications for Evidence Based Medicine Summarisation

Sara Faisal Shash Diego Mollá
Department of Computing, Macquarie University, Sydney, Australia

Abstract

We present a study of the clustering properties of medical publications for the aim of Evidence Based Medicine summarisation. Given an annotated dataset of documents that have been manually assigned to groups related to clinical answers, we apply K-Means clustering and verify that the documents can be clustered reasonably well. We advance the implications of such clustering for natural language processing tasks in Evidence Based Medicine.

Clustering for EBM

The ultimate goal is to build a query-based multi-document summarisation system for Evidence Based Medicine that groups (clusters) the input documents according to the answers, and generates (summarises) the answers.

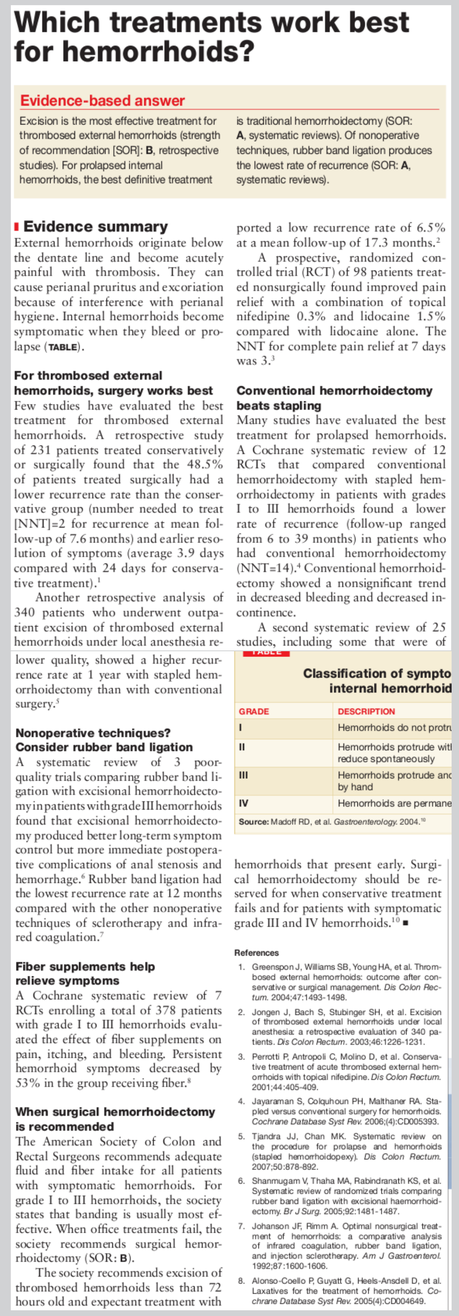
Input

QUESTION:
Which treatments work best for hemorrhoids?
DOCUMENTS:
[11289288], [12972967], [1442682], [15486746], [16235372], [16252313], [17054255], [17380367].

clustering ⟹ summarisation

Output

1. Excision is the most effective treatment for thrombosed external hemorrhoids. [11289288], [12972967], [15486746].
2. For prolapsed internal hemorrhoids, the best definitive treatment is traditional hemorrhoidectomy. [17054255], [17380367].
3. Of nonoperative techniques, rubber band ligation produces the lowest rate of recurrence. [1442682], [16252313], [16235372].



Which treatments work best for hemorrhoids?

Evidence-based answer

Excision is the most effective treatment for thrombosed external hemorrhoids (strength of recommendation [SOR]: **B**, retrospective studies). For prolapsed internal hemorrhoids, the best definitive treatment is traditional hemorrhoidectomy (SOR: **A**, systematic reviews). Of nonoperative techniques, rubber band ligation produces the lowest rate of recurrence (SOR: **A**, systematic reviews).

Evidence summary

External hemorrhoids originate below the dentate line and become acutely painful with thrombosis. They can cause perianal pruritus and excoriation because of interference with perianal hygiene. Internal hemorrhoids become symptomatic when they bleed or prolapse (TABLE).

For thrombosed external hemorrhoids, surgery works best

Few studies have evaluated the best treatment for thrombosed external hemorrhoids. A retrospective study of 231 patients treated conservatively or surgically found that the 48.5% of patients treated surgically had a lower recurrence rate than the conservative group (number needed to treat [NNT]=2 for recurrence at mean follow-up of 7.6 months) and earlier resolution of symptoms (average 3.9 days compared with 24 days for conservative treatment).[1]

Another retrospective analysis of 340 patients who underwent outpatient excision of thrombosed external hemorrhoids under local anesthesia reported a low recurrence rate of 6.5% at a mean follow-up of 17.3 months.[2]

A prospective, randomized controlled trial (RCT) of 98 patients treated nonsurgically found improved pain relief with a combination of topical nifedipine 0.3% and lidocaine 1.5% compared with lidocaine alone. The NNT for complete pain relief at 7 days was 3.[3]

Conventional hemorrhoidectomy beats stapling

Many studies have evaluated the best treatment for prolapsed hemorrhoids. A Cochrane systematic review of 12 RCTs that compared conventional hemorrhoidectomy with stapled hemorrhoidectomy in patients with grades I to III hemorrhoids found a lower rate of recurrence (follow-up ranged from 6 to 39 months) in patients who had conventional hemorrhoidectomy (NNT=14).[4] Conventional hemorrhoidectomy showed a nonsignificant trend in decreased bleeding and decreased incontinence.

A second systematic review of 25 studies, including some that were of lower quality, showed a higher recurrence rate at 1 year with stapled hemorrhoidectomy than with conventional surgery.[5]

Nonoperative techniques? Consider rubber band ligation

A systematic review of 3 poor-quality trials comparing rubber band ligation with excisional hemorrhoidectomy in patients with grade III hemorrhoids found that excisional hemorrhoidectomy produced better long-term symptom control but more immediate postoperative complications of anal stenosis and hemorrhage.[6] Rubber band ligation had the lowest recurrence rate at 12 months compared with the other nonoperative techniques of sclerotherapy and infrared coagulation.[7]

Fiber supplements help relieve symptoms

A Cochrane systematic review of 7 RCTs enrolling a total of 378 patients with grade I to III hemorrhoids evaluated the effect of fiber supplements on pain, itching, and bleeding. Persistent hemorrhoid symptoms decreased by 53% in the group receiving fiber.[8]

When surgical hemorrhoidectomy is recommended

The American Society of Colon and Rectal Surgeons recommends adequate fluid and fiber intake for all patients with symptomatic hemorrhoids. For grade I to III hemorrhoids, the society states that banding is usually most effective. When office treatments fail, the society recommends surgical hemorrhoidectomy (SOR: **B**).

The society recommends excision of thrombosed hemorrhoids less than 72 hours old and expectant treatment with hemorrhoids that present early. Surgical hemorrhoidectomy should be reserved for when conservative treatment fails and for patients with symptomatic grade III and IV hemorrhoids.[1,9] ■

TABLE

Classification of sympto internal hemorrhoid

GRADE	DESCRIPTION
I	Hemorrhoids do not protru
II	Hemorrhoids protrude with reduce spontaneously
III	Hemorrhoids protrude and by hand
IV	Hemorrhoids are permane

Source: Madoff RD, et al. *Gastroenterology.* 2004.[10]

References

1. Greenspon J, Williams SB, Young HA, et al. Thrombosed external hemorrhoids: outcome after conservative or surgical management. *Dis Colon Rectum.* 2004;47:1493-1498.
2. Jongen J, Bach S, Stubinger SH, et al. Excision of thrombosed external hemorrhoids under local anesthesia: a retrospective evaluation of 340 patients. *Dis Colon Rectum.* 2003;46:1226-1231.
3. Perrotti P, Antropoli C, Molino D, et al. Conservative treatment of acute thrombosed external hemorrhoids with topical nifedipine. *Dis Colon Rectum.* 2001;44:405-409.
4. Jayaraman S, Colquhoun PH, Malthaner RA. Stapled versus conventional surgery for hemorrhoids. *Cochrane Database Syst Rev.* 2006;(4):CD005393.
5. Tjandra JJ, Chan MK. Systematic review on the procedure for prolapse and hemorrhoids (stapled hemorrhoidopexy). *Dis Colon Rectum.* 2007;50:878-892.
6. Shanmugam V, Thaha MA, Rabindranath KS, et al. Systematic review of randomized trials comparing rubber band ligation with excisional haemorrhoidectomy. *Br J Surg.* 2005;92:1481-1487.
7. Johanson JF, Rimm A. Optimal nonsurgical treatment of hemorrhoids: a comparative analysis of infrared coagulation, rubber band ligation, and injection sclerotherapy. *Am J Gastroenterol.* 1992;87:1600-1606.
8. Alonso-Coello P, Guyatt G, Heels-Ansdell D, et al. Laxatives for the treatment of hemorrhoids. *Cochrane Database Syst Rev.* 2005(4):CD004649.

Dataset

- 456 questions from the "Clinical Inquiries" section of the *Journal of Family Practice*.
- Each question has several answer parts.
- Each answer part has its relevant documents.
- A document may be relevant to several answer parts (overlapping clustering).

Sample

```
Q    A  Documents in the answer
8269 1  17237298 16080084 12514443 15716561 16531939
8269 2  15716561
2095 2  11417373 9099150
2095 3  9099150
2095 4  12415081 1834190 9099150 7484689 11417373
```

Data For Clustering

- The source documents are encoded in PubMed XML.
- We performed original experiments using several kinds of information:
  1. Complete XML data.
  2. Abstract information only.
  3. Terms that have an Unified Medical Language System (UMLS) concept.
  4. UMLS semantic types.

Clustering Approach

- An independent clustering task for each question.
- Clustering approach was K-means.
- Words were lowercased, stop words removed, remaining words weighted using *tf.idf*.
- Final result is the average cluster entropy across all questions.

Clustering Results with Predefined K

Cluster Entropy

- Entropy of cluster i in question q:

$$Entropy(i) = -\sum_{j=1}^{K} p_{i,j} \log_2 p_{i,j} \qquad \text{where } p_{i,j} = \frac{\text{\# docs in } i \text{ and } j}{\text{\#docs in } i}$$

- Cluster entropy of question q: weighted average of $Entropy(i)$:

$$Entropy(q) = \sum_{i=1}^{K} Entropy(i) \frac{\text{\# docs in } i}{\text{\# docs in } q}$$

Table 1: Average entropy for optimal K clusters.

Measure	Whole XML	Abstract only	Concepts only	Semantic types
Euclidean	0.260	0.264	0.274	0.310
Correlation	0.348	0.362	0.349	0.347
Cosine	0.249	0.266	0.277	0.298
Dice	0.332	0.328	0.324	0.334
Jaccard	0.320	0.330	0.317	0.327
Manhattan	0.288	0.299	0.305	0.296

The entropy of pure random clustering is $-\log_2(1/K)$ for an average $K = 2.4$, giving 1.263, so simple k-means clustering gives good results.

Finding Best Number of Clusters K

User defined K: A constant value of K for each question.

Rule of Thumb: Based on the total number m of documents in a cluster. This provides a value of K that is distinct for each question.

$$K = \sqrt{m/2}$$

Cover Coefficient: Based on the number m of documents, the number n of terms, and the number t of non-zero entries in the matrix of bags of words.

$$K = \frac{m \times n}{t}$$

Table 2: Average entropy on full XML documents.

Measure	$K=2$	$K=3$	$K=4$	RoT	Cover	Original
Euclidean	0.489	0.309	0.205	0.163	0.235	0.260
Correlation	0.604	0.413	0.283	0.238	0.316	0.348
Cosine	0.479	0.298	0.213	0.154	0.224	0.249
Dice	0.572	0.368	0.250	0.204	0.290	0.332
Jaccard	0.562	0.360	0.252	0.191	0.293	0.320
Manhattan	0.522	0.327	0.226	0.174	0.281	0.288
Average K	2	3	4	3.8	2.8	2.4

Our system does not attempt to handle overlapping clustering. Considering that, the results are remarkably good. In future work we will look at overlapping clustering. We will also look at supervised clustering and incorporate information from the question.