

mendelian

White Paper

Search
Performance

Only half the patients referred to a clinical geneticist for diagnosis of a genetic disorder will be diagnosed. Those patients may subsequently be referred to several physicians, undergo numerous clinical tests and if lucky be offered next-generation genome sequencing to search for causative genes.

Of those referred for genome sequencing, only about 25% will eventually get a confirmed diagnosis. On average, it takes 5.6 years in the UK for a rare disease patient to be diagnosed².

Background

Mendelian is an online rare disease search engine, built with the aim of increasing diagnostic hit rates. Rare diseases, their associated genes along with existing gene panels, are algorithmically matched to phenotypes.

The aim of this report, addressed to health professionals and clinicians, is to provide the simple metrics for the likelihood of finding a successful diagnosis through Mendelian.

How is Mendelian used?

Confronted with a rare disease case, patient's clinical features are first entered into www.mendelian.co, the tool outputs a focused shortlist of genes, diseases or disease clusters for consideration. For each result, Mendelian displays the underlying genetics, potential gene panels and a respective match score.

Evaluation

A large number of computer-generated cases were used to evaluate Mendelian's success rate. A successful event occurs when a patient's clinical history is entered and the correct diagnosis is listed in the search results. For the sake of this analysis a correct diagnosis occurs if in the first **25** results:

- The diagnosed and listed disease or disease clusters match
- Diagnosed disease falls into the listed disease cluster
- The patient's causative gene is listed in the result

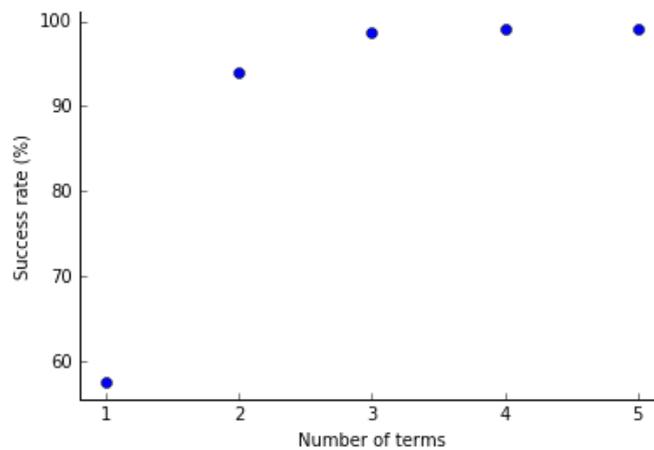
Computer generated cases

To improve the breadth of this evaluation, a far larger set of simulated cases was generated from the current disease database. We used four methods to simulate how real life patient may be input:

Incomplete data

Clinicians are unlikely to enter all the clinical features associated with a disease, many of these may not exist or have not yet been detected in the patient.

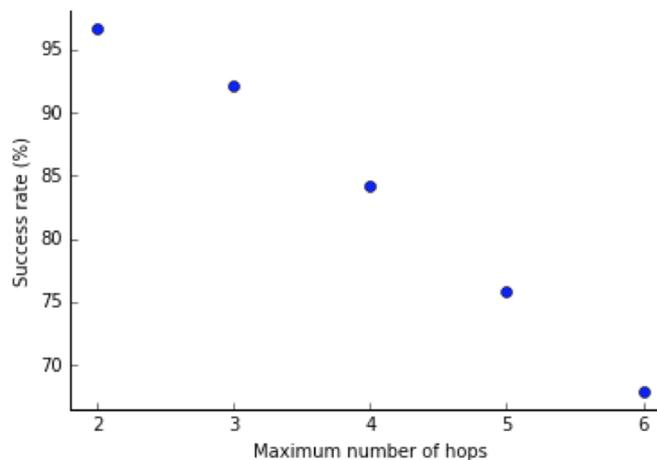
This was simulated by randomly selecting n clinical feature annotations for each disease. We perform r repeats of this test on each disease to ensure a good coverage over the disease's clinical features. Figure 1 shows that a favourable success rate of 94% is achieved at $n = 2$. At $n > 4$, the success rate stays at 99%. The average rank of successes at $n = 3$ was 1.48 -- the average rank is the average position in the results list.



Imprecise data

Clinical features are often recorded in subtly different ways, often due to the variability of the disease or the imprecision of a clinical test. We recreate this by running the previous benchmark with $n = 3$ and randomly replacing each query term with an imprecise term. Imprecise terms were generated using the Human Phenotype Ontology (HPO), a tree-like database of clinical features that are stratified, top-down, from general to more specific.

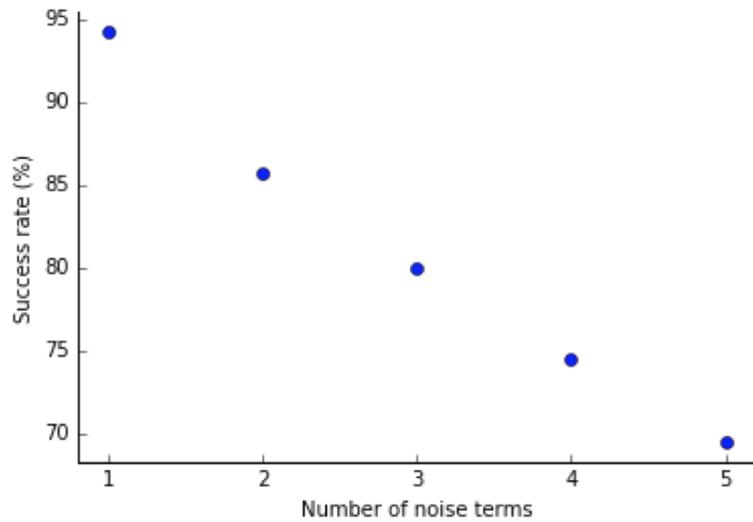
For each term we replace it with an ancestor that is h hops above, where h is drawn from $0, 1, \dots, h_{\max}$. At $h_{\max} = 4$, we manage a reasonable success rate of 84%. It later falls to under 68% when $h_{\max} > 5$, as shown in Figure 2. At $h_{\max} = 3$ the average rank of the successful document in our search was 2.52.



Note: Mendelian displays results only after 2 clinical terms or more are entered.

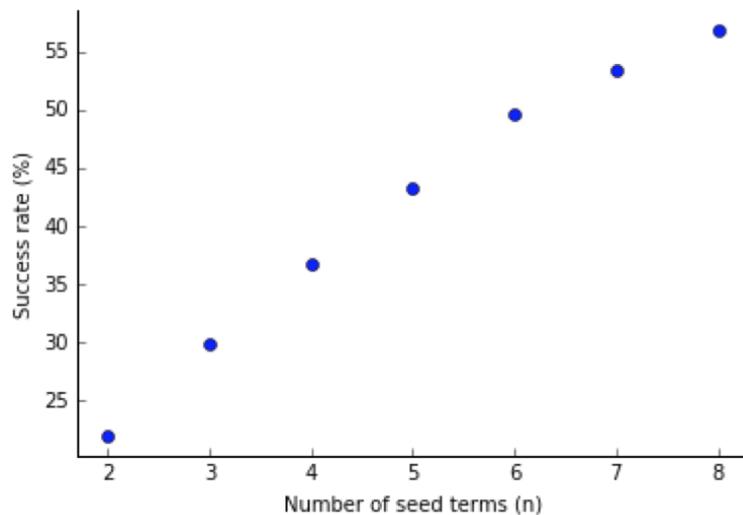
Noisy data

Most queries will introduce clinical features that are not known to be associated with the patient's disease. The disease's HPO annotations may be incomplete or the patient is exhibiting an unseen clinical feature. You can simulate noise by adding x randomly selected terms to the query. Figure 3, shows that the ideal case is where $x < 3$ we achieve a success rate of at least 80%. At $x = 3$ the average rank of the successful document in our search was 2.60.



Worst case data

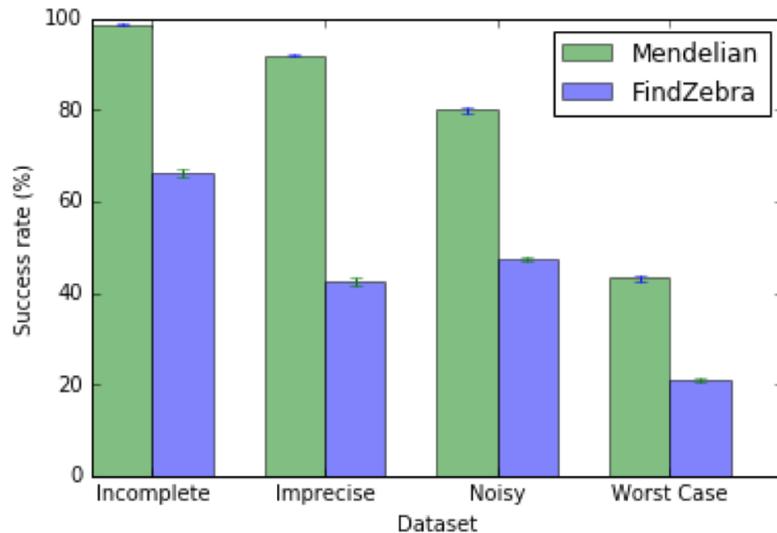
In real life a query will vary from the archetypal disease description in several ways. To estimate these extreme queries, we combine the imprecise and noisy simulations and see if the true disease is still shown. Figure 4 shows that at $n=8$, we expect a lower bound success rate of 56%.



Competitor Analysis

Below we compare these tests against FindZebra, an up to date rare disease search engine that covers several public disease databases. The FindZebra search requires symptoms to be entered as free text and returns documents linking to external data sources.

To evaluate it on our benchmark, we have converted our HPO query terms into a comma-separated list of terms. Successful searches occur when the correct document is referenced on the first page of results. We consistently perform better than FindZebra on all the tests described in this paper, as shown in Figure 5.



Conclusion

We estimate a success rate of between 56-99% for a well described case. The search algorithm degrades gracefully when confronted with imprecise or noisy queries. However, we restrict searches to have at least 2 clinical features and encourage users to be as specific as possible with their queries to get the best results.

In our comparison, we outperform FindZebra in all benchmarks. However, these tests assume that the patient exhibits clinical features found in public HPO disease annotations. To get the benefits of a semantic search in practice it is important to keep these annotations accurate and as comprehensive as possible.

Mendelian performs exceptionally well on a variety of cases to simulate the kind of queries found in the clinic. Given these assumptions, our search is shown to perform better than a leading alternative. The next step is to evaluate the search with real clinical data.

References

1. Vandana Shashi, et al. The utility of the traditional medical genetics diagnostic evaluation in the context of next-generation sequencing for undiagnosed genetic disorders. *Genetics in Medicine*, 16(2)
2. PA Engel, S Bagal, M Broback, and N Boice. Physician and patient perceptions regarding physician training in rare diseases: The need for stronger educational initiatives for physicians. *J. Rare Dis.* 2013.
3. Anna C Need and David B Goldstein. Neuropsychiatric genomics in precision medicine: diagnostics, gene discovery, and translation. *Dialogues in clinical neuroscience*, 18(3):237, 2016.
4. FindZebra: A search engine for rare diseases - 26 Feb. 2013,