Research Interest:
In recent years, we have been surrounded by a variety of diseases and medical challenges. With the rapid advancement of medical technology, identifying the current hot trends in medicine can provide a clearer picture of the prevailing research interests. This insight not only informs the public about key focus areas but also helps me determine where my data analysis skills could be most effectively applied.

Data source:
The first dataset is sourced from PubMed website: https://pubmed.ncbi.nlm.nih.gov/. Specifically, I choose to put the focus on the published "trend papers" from 2022 to 2024. Overall there are 907 articles.
The second dataset is sourced from Google Scholar. I type in the search keyword "medicine", and specify the time range from 2022 to 2024, collecting information from 200 web pages, with overall 2000 articles. The first web page can be accessed: https://scholar.google.com/scholar?start=50&q=medicine&hl=en&as_sdt=0,34&as_ylo=2022&as_yhi=2024.
I understand that the second data source should be tabular data, however, for my research interest, it is a bit hard for me to find a suitable tabular data for analysis. So I choose to collect information from a different website.

Method:
My initial intention is to scrape down the abstract description in each article, and provide text clustering and identify keywords information from the clusters. However, this method does not give me good results, even though I have preprocessed the text with many stop words, the keyword from clusters is not showing up as significant. So instead of scraping from "Abstract" section, I choose to scrape from another section "Keyword".

In the first data source PubMed, most of the articles contain a section called "Keyword" where the author of that article lists down the area covered, which I think reflects the research interest more precisely than the abstract. And then I provide frequency analysis to the keyword, sort them and gain insights.

In the second data source Google Scholar, the article does not have the same clear format as what is inside the PubMed. So I choose to scrape down the titles of each article, split them into words, categorize them into different groups, and identify each group's importance.

Result Analysis:

a. PubMed Data Source
Not all articles contain a “Keyword” section. Out of 907 articles, 853 contain keyword sections. And Figure1 and Figure 2 show the keyword distributions.

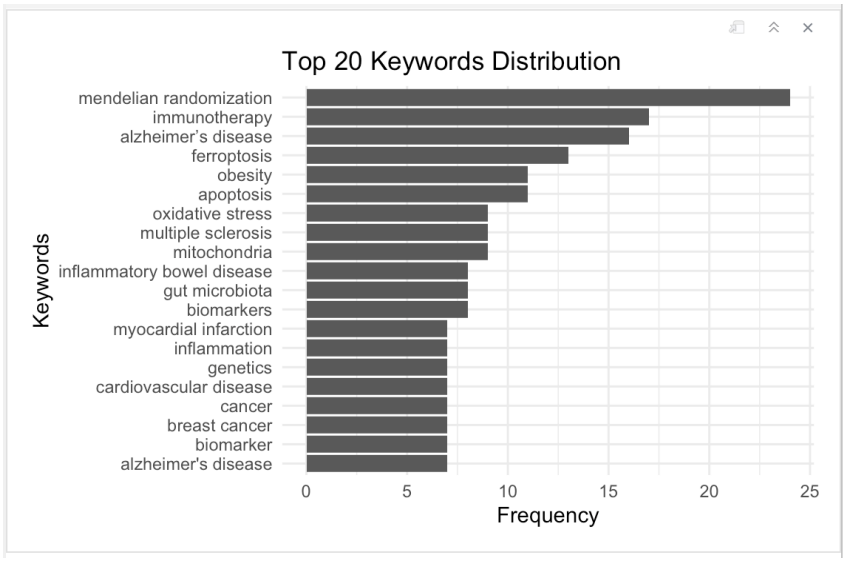
Figure 1

Description: df [2,260 × 2]

keywords <chr>	n <int>
mendelian randomization	24
immunotherapy	17
alzheimer’s disease	16
ferroptosis	13
apoptosis	11
obesity	11
mitochondria	9
multiple sclerosis	9
oxidative stress	9
biomarkers	8

1–10 of 2,260 rows Previous 1 2 3 4 5 6 … 100 Next

Figure 2



These keywords are medical-related, and precisely describe the research areas. So I don’t provide further processing to the keyword dataframe. From the keyword, we can see that several key trends are prominent in recent medical research. First: immunotherapy and cancer research. The consistent mention of terms such as immunotherapy, tumor microenvironment, and various forms of cancer indicates a strong focus on developing and refining immunotherapies. And also, with multiple terms related to Alzheimer’s disease, Parkinson’s disease, and neurodegeneration, there is a clear underscore on these diseases’s

pathophysiology. Topics like myocardial infarction, heart failure, and hypertension highlights the ongoing research into cardiovascular health, one of the leading causes of death worldwide. We can also identify the research interest in metabolic disorder from terms like “type 2 diabetes, insulin resistance.”, and with interest in genetic studies from terms “genetic testing, GWAS”. And one big area of focus is technology. Terms such as “artificial intelligence, machine learning, and data” appear a lot.

b. Google Scholar
   2000 article titles have been scrapped. I remove some usual meaningless words, identify them as “stop words” , remove them from the text dataframe, and analyze the occurrence of each word. The frequencies of words are shown in figure3.

Figure 3

word <chr>	count <int>
medicine	113
covid-19	32
cancer	18
clinical	17
precision	17
trial	16
disease	15
phase	15
review	15
study	14

1-10 of 1,144 rows Previous 1 2 3 4 5 6 … 100 Next

Since article titles contain more words and do not give as precise a description as keywords, I define five categories in the medical field: medicine and clinical (it involves surgery, anesthesia, medical treatment, therapies) ; health conditions (it refers to specific diseases: covid-19, cancer, diabetes, infections, Alzheimer’s) , technology and innovation (artificial intelligence, data), biomedical science (biological processes and mechanism), health policy. And then based on the similarity between words and each group, assign each word to the group. If the word is not similar to any one of the groups, assign it to “other”. It turns out that there are many distracting words in titles, and out of those meaningful words, medicine and clinical, health conditions, and technology occupy the largest area, which aligns with our finding in the first data source.

Figure 4

category <chr>	total_count <int>
Other	1626
MedicineAndClinical	213
HealthConditions	114
TechnologyInnovation	67
BiomedicalScience	40
HealthSystemsPolicy	40

6 rows

Discussion:
The analysis in the second dataset may not be accurate, because the categories and the words inside each category are self-defined and may miss many words that are supposed to be included. The real percentage of the distracting group “other” should be much smaller than in the current output. Dr.Evans, if you have any suggestions on how to deal with those distracting words, please comment and let me know.

And another big issue is that when I scrape down the title information from the second data source, it works fine for the first couple of times, but gives me http 429 error in the later attempts. I think it might be because I tried it too many times and google scholar identified this and blocked my ip address. I tried to run my script on another computer. It works fine.