Applied Health Data Science Summative Assessment:

Health Data Science Mini Project Report

1. Data processing

A description of the steps you undertook to pre-process the data (e.g. re-formatting, data cleaning).

i. Fetching Data

The first step in our data pre-processing pipeline was to fetch relevant research articles in accordance with the search for a particular keyword. This is done using a shell script that interacts with the PubMed database through their API. Running this script creates a URL to query their database for articles matching the keyword. The PMIDs of those articles are fetched and stored in XML format. Each PMID is then utilized to fetch the complete metadata of the article that corresponds to it, and the data again gets stored in XML format in the specified directory. The approach will be quite organized and repeatable with regard to data acquisition.

ii. Processing Articles

The fetched article metadata is then processed further in R. This part of the workflow will carve out from the XML files the information and details and store the information in a tab-delimited file with three columns: PMID, year, and title. Data transformation and storage are performed here in a structured format that is more accessible by analysis in later steps.

iii. Clean Titles

The last step in the data pre-processing phase is the cleaning of article titles. First, tokenization is done into individual words, stop words removal is done since common words are of no use for any analytical features. Text is changed to lowercase, and then numeric characters are removed from the words, including embedded numbers to ensure uniformity in text data. The script also reduces the words by stemming, or to base or root form, so normalizing variations of a word helps in comparison analysis among words. The words are then regrouped after cleaning to reconstruct cleaned titles for each article.

2. Data visualisation with description and justification of the approach

One static visualisation produced in R and embedded in text describing the question you intended your plot to explore, the methods you used to visualise the data, the theoretical reasons you chose this approach, and your interpretation of what the plot shows.

Exploring Latent Topics in Research Data: An Analysis of Term Importance Using LDA

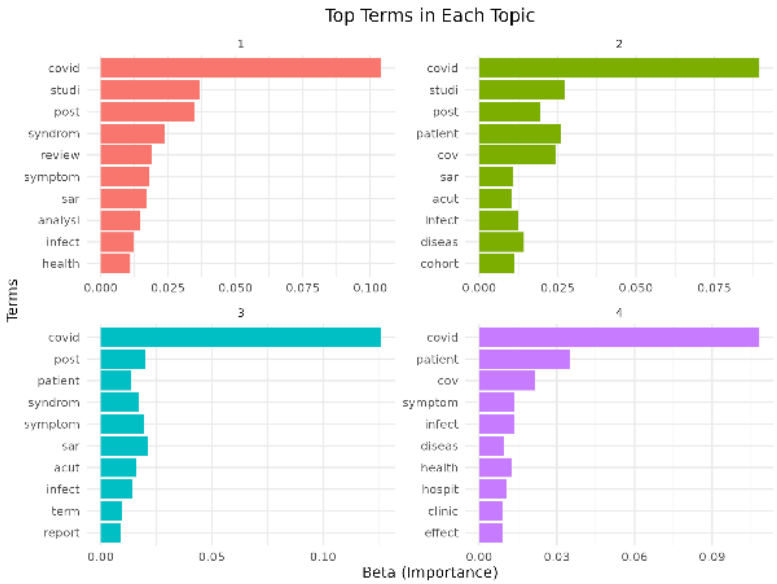

Figure1: Top Terms in Each Topic

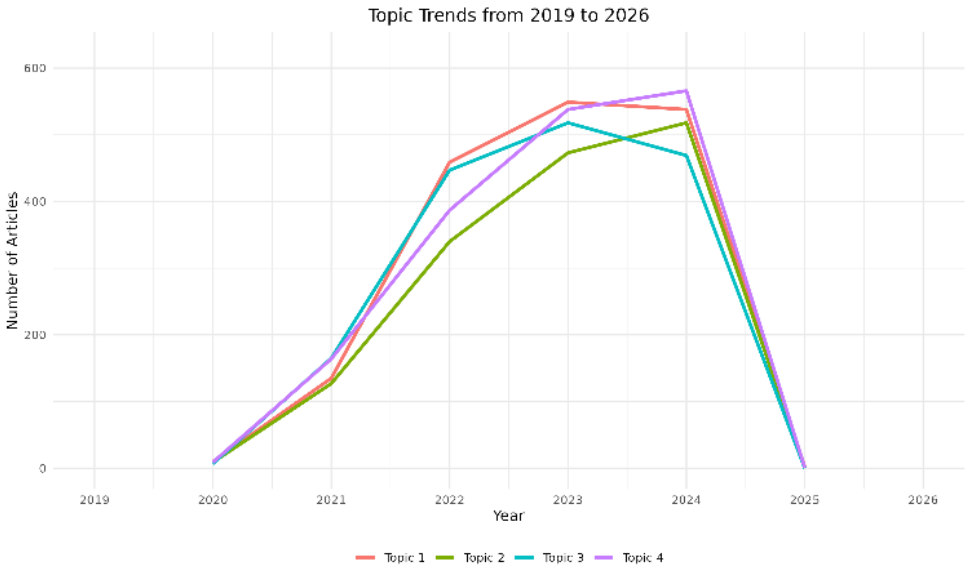

Figure2: Topic Trends from 2019 to 2026

i. Problem Statement

The problem of our analysis was to identify dominant themes in research articles over time. The study sought to understand how certain terms, like "COVID", have influenced research discourse and to explore other significant themes that have emerged alongside or independently of the pandemic.

ii. Theoretical Justification

Latent Dirichlet Allocation (LDA) is a probabilistic model used to uncover latent topics in text data by grouping similar words into topics based on their distribution across documents. It assumes that documents are mixtures of topics, with each word's presence attributed to one topic. Text data was preprocessed in R to generate a document-word matrix, which served as input for the LDA model to infer topics. The beta weights, representing the importance of terms in each topic, were visualized using horizontal bar charts. Annual article counts for each topic were plotted using line charts, which effectively illustrate temporal trends and topic prominence over time.

iii. Interpretation of Plots

Top Terms in Each Topic:

The bar charts represent the most significant terms associated with each identified topic based on their beta weights (importance).

Topic	Main Words	Interpretation
1	"covid," "study" and "post"	focus on COVID-19 and post-pandemic studies
2	"covid," "study" and "patient"	reflecting clinical and cohort-based research during the pandemic
3	"covid," "sar" and "post"	suggesting post-acute sequelae or long-term impacts of COVID-19
4	"covid," "patient" and "cov"	point to studies addressing broader public health concerns.

The recurrence of "covid" across topics confirms its dominance in recent research discourse, while associated terms provide nuances of specific study directions (e.g., clinical focus, health effects, long-term implications).

Annual Trends of Topics Over Time:

The line chart shows the temporal distribution of research articles for each topic from 2019 to 2026. All topics exhibit a sharp rise in publications between 2020 and 2023, correlating with the global focus on COVID-19 during the pandemic. Peaks around 2023 suggest maximum research output for all topics. The steep decline post-2024 likely reflects diminishing COVID-related research as the pandemic's influence wanes.

In summary, the analysis reveals the evolution of research themes centered on COVID-19, emphasizing both immediate and longer-term health impacts, with a strong temporal alignment to the pandemic's progression.

3. Reflection on ethical and governance considerations

Imagine that you were going to request a data set from the Avon Longitudinal Study of Parents and Children (ALSPAC) cohort study (https://www.bristol.ac.uk/alspac/) to explore how participants' health changed over the course of the pandemic. This data set will contain individual-level sensitive medical data from people living in Bristol. What ethical or governance considerations should you take into account?

i. Safe Data

De-identification: The data will be de-identified, meaning direct identifiers (e.g., names, addresses) will be removed, and disclosure control measures will be applied to prevent re-identification of individuals (University of Bristol, n.d.).

Compliance with Legislation: Ensure adherence to data protection regulations, such as GDPR in the UK, which mandates safeguarding personal data.

ii. Safe Projects

Ethical Use: The research proposal must demonstrate that the project aligns with ethical standards, benefits the public, and contributes to scientific knowledge.

Approval Process: The project will require approval from the Children of the 90s Executive, confirming its ethical integrity and appropriateness for the dataset.

iii. Safe People

Researcher Credentials: Only bona fide researchers trained and authorized in handling sensitive data will access the dataset.

Institutional Oversight: Researchers must sign a Data Access Agreement, committing to confidentiality and the secure handling of data.

iv. Safe Settings

Secure Environment: Data access will occur in a secure environment compliant with the Five Safes framework, such as the Data Safe Haven used by Children of the 90s. Data transfer will be encrypted, and researchers are prohibited from sharing data beyond the approved project group.

v. Safe Outputs

Non-Disclosure of Individuals: Outputs will be reviewed to ensure they are non-disclosive, meaning no individual participants can be identified from the published results.

vi. Ethics and Oversight Committees

Ethical Review: The proposal may need to undergo review by the ALSPAC Ethics and Law Committee (ALEC) to assess the ethical implications of the research.

Participant and Public Advisory: Consultation with the ALSPAC Participant and Public Advisory Panel (APPAP) might be required to ensure alignment with participant values and public expectations (University of Bristol, n.d.).

vii. Informed Consent

Purpose-Specific Consent: Confirm whether participants consented to the use of their data for pandemic-related research and secure additional consent if necessary.

4. Reflection on data management

Now imagine that ALSPAC has approved your project to work with the data set you requested. How do you plan to store and manage the data in an appropriate way?

i. Secure Storage

Data Safe Haven: Store the data in a secure environment, which adheres to the Five Safes framework. Use systems approved by your institution's IT or data protection office, meeting national and international data security standards (e.g., ISO 27001) (Office for National Statistics, 2017).

ii. Data Access Control

Authorized Users Only: Restrict access to bona fide researchers specified in the original project proposal. Each user must sign confidentiality agreements.

Access Logs: Maintain detailed logs of who accessed the data, when, and for what purpose.

iii. Data Management Practices

Data Minimization: Only work with the subset of data strictly necessary for your project. Avoid using fields that are not relevant to the research question.

De-identification Integrity: Ensure data remains de-identified throughout its use, with no attempt to re-identify participants.

Regular Checks and Backups: Regularly check data and schedule encrypted backups in secure locations to prevent data loss.

iv. Compliance and Monitoring

Periodic Audits: Allow audits by the Children of the 90s Executive or institutional data protection officers to ensure compliance with agreements and regulations (University of Bristol, n.d.).

Data Breach Protocols: Have a documented response plan for data breaches, including immediate reporting to ALSPAC and relevant authorities (Information Commissioner's Office, n.d.).

v. Data Use Protocols

Safe Settings: Analyze the data only within the secure environment specified in your project proposal, avoiding unauthorized transfers.

Non-Personal Devices: Prohibit downloading data to personal devices or storage systems not covered by the data access agreement.

vi. Outputs Management

Non-Disclosive Results: Submit all outputs to the Children of the 90s Executive for screening before dissemination to ensure no participant is identifiable (University of Bristol, n.d.).

Clear Documentation: Provide metadata and codebooks to describe how the data was processed and analyzed, ensuring transparency without risking confidentiality.

vii. End of Project Protocols

Data Disposal: Securely delete all copies of the dataset at the end of the project, as per the terms of the data access agreement.

References

University of Bristol. (n.d.). *Children of the 90s: Five Safes framework*. Retrieved November 26, 2024, from https://www.bristol.ac.uk/media-library/sites/alspac/documents/participants/CO90s_5_Safes.pdf

Information Commissioner's Office. (n.d.). *A guide to the data protection principles*. Retrieved November 26, 2024, from https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-protection-principles/a-guide-to-the-data-protection-principles/

Office for National Statistics. (2017). *The Five Safes: Data privacy at ONS*. Retrieved from https://blog.ons.gov.uk/2017/01/27/the-five-safes-data-privacy-at-ons/