Peem Lerdputtipongporn

36757: Advance Data Analysis I

Last updated: 15th May 2022

**Project title:** Detecting Misinformation in Youtube Videos on Diabetes

**Statistics advisor(s):** Nynke Niezink and Larry Wasserman (Department of Statistics and Data Science, CMU).

**External scientists and their affiliations:** Rema Padman (Heinz College of Information Systems and Public Policy, CMU); Anjana Susarla (Broad College of Business, Michigan State University); Xiao Liu (WP Carey Business School, Arizona State University).

**Section 1 – Introduction**

**Introduction**

During the past decade, annual healthcare spending in the US has grown exponentially. The spending reaches $4.1$ trillion — 19.7 percent of the United States’ GDP in 2019 — and is projected to reach $6.2$ trillion by 2028 (CMS). One of the major factors contributing to the rising medical expenditure is the lack of health literacy required to understand medical information, most of which is presented in text. Nowadays, Youtube offers alternative means for physicians to communicate with patients through audios and graphics. Moreover, it serves as a platform where patients can produce content as well as interact with one another. However, the rise of Youtube as alternative means to promote health literacy and disseminate medical information poses a fundamental question: how should one identify if a health video contains misinformation? Answering this question allows healthcare providers to better assist patients by directing them toward clinically appropriate treatment. In this paper, we focus specifically on classifying whether a Youtube video about diabetes contains misinformation.

Diabetes mellitus, also known as diabetes, is a chronic, non-communicable disease whereby defects in insulin discretion cause prolonged high blood glucose (Kharroubi and Darwish 2015). 9.1\% of Americans aged 20 years and above suffer from diabetes, and people with diabetes lose their life expectancy by 0.89 years (Preston et. al., 2018). Some scholars have studied how misinformation about acute diseases such as COVID-19 spread through social media [citation]. However, there is relatively little literature on misinformation about chronic diseases. The scarcity of such literature can be attributed to the fact that assessing the quality of health videos is time-consuming and requires people with medical expertise. To circumvent this problem, Liu et al. (2020) have proposed using semi-supervised learning to classify video in term of its understandability and medical content as well as identify how both qualities affect viewers’ engagement. Using each video’s features and metadata to evaluate its quality, their method scales to large dataset.

However, a video is understandable and contains high medical content may contain misinformation, detracting its audience from seeking appropriate treatment. While physicians may evaluate misinformation in health videos using some guidelines, they cannot manually evaluate the accuracy of every video. Moreover, from a content-creator perspective, it is helpful to know which characteristics of health videos are associated with inaccuracy so that they can improve the quality of their videos. In this project, we seek to use a classifier that uses Youtube video features to identify misinformation. Our work extends the framework by Liu et. al. by using semi-supervised learning to impute labels on misinformation.

**Section 2 — Literature Review**

Many scholars have attempted to study online text misinformation (Muric et. al.’21, Tang et. al. ‘21). Misinformation spreading through text often involves acute diseases, such as seasonal flu, COVID-19, among others. Other scholars attempt to characterize misinformation through user-generated labels on online platform (Knuutila et. al. ’20) or infer misinformation based on the medium network structure (Tang et. al. ’21). Papadamou has written a comprehensive review of current literature involving abhorrent content or misinformation on Youtube (2021). To the best of the author knowledge, we have identified two papers on Youtube misinformation (Jagtap et. al. ’21). Jagtap et. al. claimed video metadata are not indicative whether a video contains misinformation or not, proposing a natural language processing approach and synthetic minority oversampling technique to analyze the video content instead. Hou et. al. uses linguistic, acoustic, and user engagement features for the development of classification models to identify misinformation (2019).

**Section 3 — Dataset Description**

To curate this dataset, we query Youtube under the incognito mode using keywords related to diabetes. The Top 50 search results are stored in our dataset. Our dataset consists of 11,483 videos, of which 573 are labelled according to PEMAT guideline — a guideline to evaluate patient education materials in terms of understandability, actionability, and whether they contain medical information. The labelled videos are randomly sampled, so it might not reflect the distribution of the dataset. For each video, if at least half of the applicable criteria are labelled as positive, it is classified as understandable or actionable. In terms of misinformation labels, two physicians are asked whether a video contains misinformation and provides brief description of the types of misinformation. Examples of comments include “misinformation throughout videos,” “minimal reference to diabetes,” among others.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Understandable (1)** | **Not understandable (0)** | **Total** |
| **Contains medical information (1)** | 348 | 39 | 387 |
| **No medical information (0)** | 84 | 102 | 186 |
| **Total** | 432 | 141 | 573 |

Aside from the labels, our dataset contains 50 features (with some redundant labels). These features can be grouped into the following categories, to name a few:

|  |  |  |
| --- | --- | --- |
| **Video Metadata** | **Video Content** | **Derivatives of video metadata** |
| Caption ID  Number of likes, dislikes, views.  Duration.  Date of publishing.  Channel information. | Category.  Video topic  Subtitle  Title  Speech confidence  Scene count | Cosine similiarity between keywords, titles, and description.  Description readability. |

Please refer to Liu et. al. (2019) for further details on how the data are collected.

**Section 4 — Preliminary Results**

\*\*NOTE: By derivatives, I am referring to features that are not directly provided by Youtube API. For instance, Youtube API provides information on video description, but not how many active verbs appear in it. The former is original variable, whereas the latter is a derivative.

For the 600 labelled videos, their labels are extremely unbalanced. 75% of the videos are understandable, whereas 75% are labelled as misinformation. As of this writing (May 15th), there needs to be further clarification from the labelers whether we should group scientific videos, irrelevant videos, and inaccurate videos under the same category of misinformation. That being said, below are some interesting results:

1. The number of likes, dislikes, views have little correlation to whether a video is understandable, actionable, contains medical information, and is medically accurate (see Figure 3.1). Similar results occur for other groups of variables, such as the number of words (and their derivatives) of the description as well. Therefore, it is unlikely that any of these variables by itself will yield significant results in any predictive models.

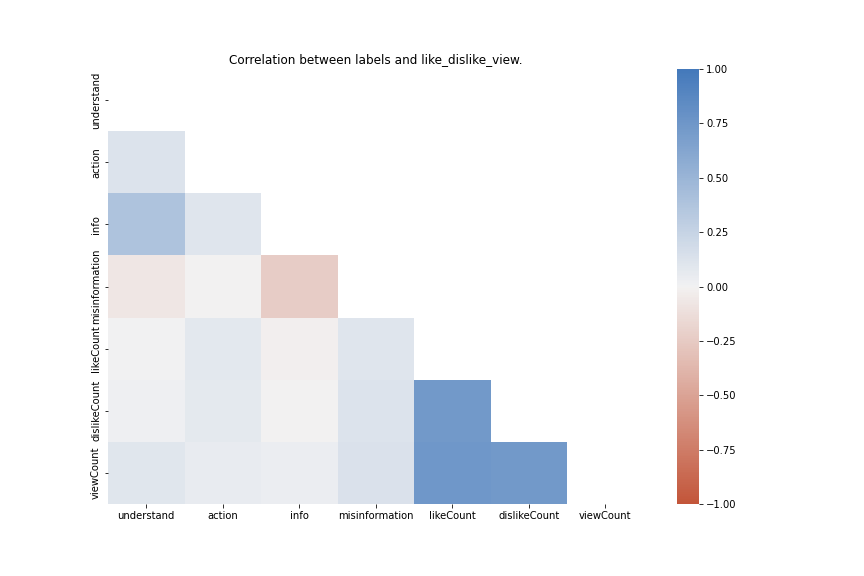


Figure 3.1: A heat map (correlation plot) between likes, dislikes, views suggests

mild correlation.

1. There is little difference in the distribution of feature values between understandable vs not-understandable, actionable vs not-actionable, and informative vs not-informative video. **Please use the code provided in helper\_fn.py to visualize the data.**

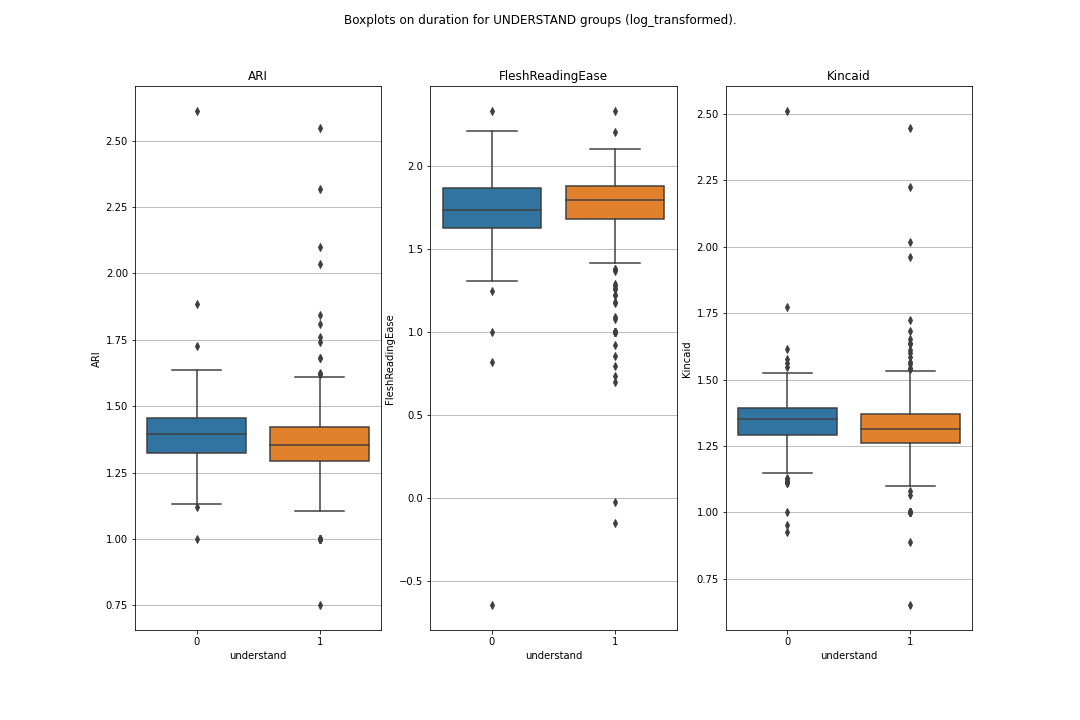


Figure 3.2: A box plot comparing distributions of readability indices (log-transformed) between understandable and not-understandable videos.

1. Our dataset is extremely imbalanced. 75% of the videos are understandable; 68% actionable; 50% informative. Therefore, fitting unweighted off-the-shelf classifier will lead to extremely low detection of videos falling in Class 0. We have attempted to address this problem using weighted logistic regression and synthetic minority sampling. However, this issue is not fully resolved.

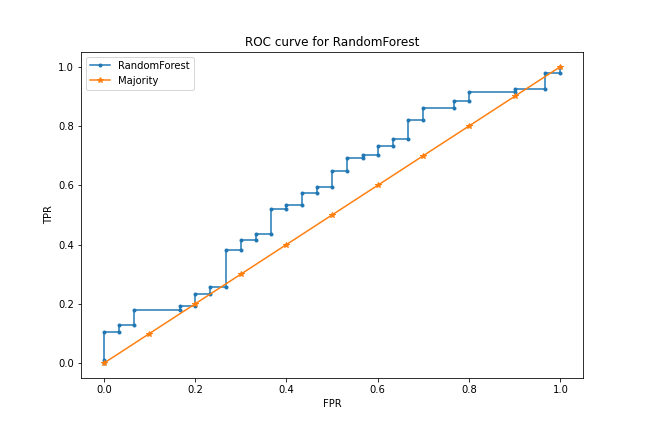


Figure 3.3: An ROC curve of fitting randomForest to classify video understandability. The blue ROC curve suggests that the classifier, trained

on unweighted dataset, has little prediction capability

1. As the categoryID and topicID suggests, our dataset consists of 1,025 videos unrelated to diabetes. These videos appear under queries that are not language-specific, such as nesina (a diabetes drug; also the name of a famous rapper). **Please use the code provided in helper\_fn.py to examine the distribution of topics and categories for each videos.**
2. **(Text-based classification)**

I have attempted to repeat Jagtap et. al.’s proposal of using **texts** to detect misinformation. For each video, we have access to two types of texts: video description and video transcript. In this dataset, video description comes in many forms, be it a promotion link to the channel’s website, an annotated transcript, bunch of hashtags. Therefore, it is impossible to perform topic modelling without substantive pre-processing. Hence, I decided to use derivatives of video description as provided by Prof. Liu to build vanilla classifiers, such as random forests, logistic regression, and SVMs. The model did not achieve great performance — specifically, they almost always predict every video as misinformation due to class imbalance and the lack of strong signal (see Point 1 and Point 2).

In contrast, we have 2,400 videos with transcripts available, 450 of which appear in our labelled dataset on misinformation. I attempted to use Latent Dirichlet Allocation (LDA), a natural-language processing method to topic-modelling on video transcripts. Unfortunately, each topic is different linear combinations of same words — no striking words appear for any topics that point to misinformation. Sentiment analysis also doesn’t work because the free models are usually trained by general texts, not medical texts.

**NOTE:** I was hoping words such as “amazing” “quick cure” to appear, which points to misinformation given that diabetes is a chronic disease. LDA did not yield such results.

Another idea I explored was building features from texts similar to Ghenai et. al’s “Fake Cure, User-Centric Modeling of Health Misinformation in Social Media.” In this paper, they computed from transcripts features such as average syllable per words, number of negative words, etc. Afterwards, they used forward-selection with LASSO penalty to select significant variables. At first, I compared the distribution of word counts between accurate videos and misinformation videos. I used spaCy library to extract word frequency, grammar trees, and something else. Overall, misinformation videos have fewer medical words and more general words (e.g., exercise, I) on average. Using word counts, we can detect scientific videos, but not inaccurate videos. Please refer to PPT for more details.

\*\* Because misinformation as of this writing consists of scientific videos, nonsense videos, medically inaccurate videos, etc., please interpret the comparison between misinformation = 0 and misinformation = 1 with caution.

1. **(Channel-level classification)**

Another approach to detecting misinformation is to weigh highly-viewed videos more than videos with fewer view counts. The rationale is that highly-viewed videos can be a good option if it’s legitimate and should be flagged if it’s misinformation. Below are some observations.

* Top 10 most-viewed videos account for 50% of the total views in this dataset.
* I ranked channels based on how many videos they created in this dataset. Top 20% channels (as ranked by the number of their videos appearing in this dataset) creates 60% of the videos. Examples of which include reputable channels such as Dr. Eric Berg and also pronunciation guide.
* Some videos have outsized likes and dislike counts, such as Skrillex’s song and Jake Logan’s fake videos about a glass that cures color blindness. After adjusting for videos with absurdly high views, the distribution between likes, dislikes, and views hover close to one another. That is, there is no channel that creates too popular or too polarizing videos.

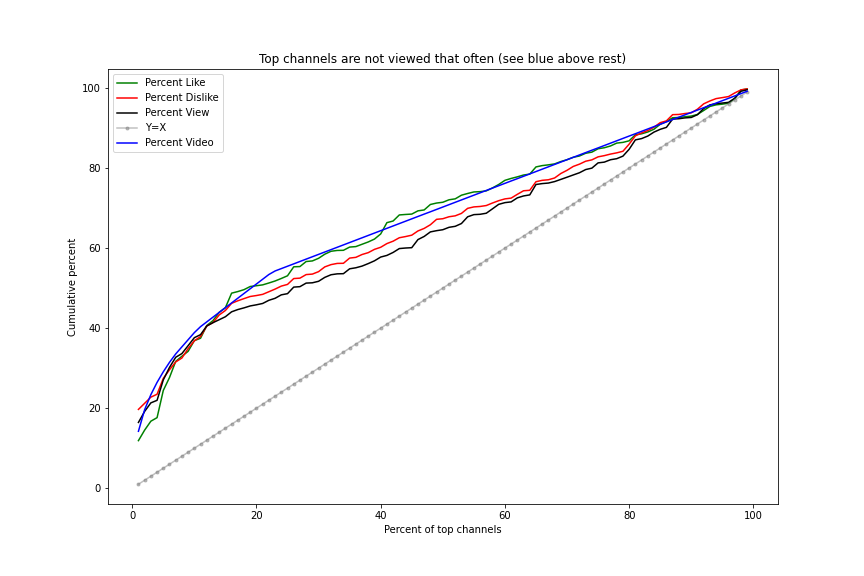


Figure 4.6.1: After removing Top 10 most-viewed videos (all of which are irrelevant to diabetes), the cumulative views/likes/dislikes follow one another closely.

Please see my powerpoint for more details regarding 4.5, 4.6.

**Section 5: Future Directions**

My exploratory data analysis shows that most of video features are not notably different between videos in Class 1 or Class 0 for any of the labels. Moving forward, future collaborators in this project should be cognizant of the following:

* how this dataset is generated.
* how some videos are sampled for labelling for misinformation, actionability, understandability, and whether it contains medical information.

Because Top 50 videos are curated for each diabetes keyword, some irrelevant videos appear in our dataset. The idea is that health-information seekers do not know a priori whether a video is relevant or not, which is plausible. However, because those irrelevant tend to have large views, likes, dislikes, channelSubscribers, and so on, any predictive models I built are skewed by those numbers.

Moving forward, we should define (and operationalize using variables we have) what misinformation means. Is an irrelevant video such as a rap song misinformation? What about scientific video about how to curb hypertension that does not make reference to diabetes in any way? The categoryID and topicID are great indicators for video relevance, so an easy-fix is to filter out videos using both IDs. Having pinpointed definition of misinformation, one might explore:

* Is misinformation strictly an artifact of **video content/metadata**? Should likes, dislikes, views be considered in building predictive models?
* To what extent is misinformation a **channel-level construct?** 
  + Idea 1: Measuring authority on diabetes using the total number of videos populating this diabetes dataset (or ratio of videos in this dataset/total videos created) doesn’t work.
    - **Logic:** Suppose the dataset is representative of all diabetes videos. Then, channels with fewer videos about diabetes will have fewer of their videos appearing in this dataset. Using number of videos as proxies for expertise in diabetes, we should give more credence to channels whose videos appear in this dataset many times.
    - **What works:** This idea would weigh channels specializing in diabetes such as “All About Diabetes and Related” more heavily. It also handles irrelevant channels such as “British Got Talent” very well.
    - **What doesn’t work:** Trustworthy sources such as “Khan Academy” and “PBS News Hours” have created many science videos, of which a few happen to be about diabetes.
  + Using engagement levels for each channel also doesn’t work. Too many videos have few views, and too few videos have many views.
* Is there a **network or some homophily** between misinformation videos and accurate videos? Because we have no info on which channels subscribe to which nor which videos tend to be recommended after having watched certain videos, there exist only few ways to create a network.
  + Idea 1: Connect through knowledge graph using categoryID, topicID.
    - **Logic:** Filter out irrelevant videos.
    - **What doesn’t work:** Because some categories such as health dominate the graph, we have only a few clusters, which doesn’t work for predicting misinformation.
  + Idea 2: Connect videos from the same channels together.
    - **What doesn’t work:** A priori, it’s a bunch of complete graphs. Moreover, the majority of channels create have one video populating this dataset.
* **Text-features**, sentiment-analysis, other NLP approach.
  + This approach is promising once we have pinpointed the definition of misinformation.
  + Challenge: Extensive pre-processing. We don’t have the transcripts for 6,300 videos. Some videos are also not in English.
* How to handle irrelevant videos?

I hope this summary document is useful for future work on this project ☺ .

**Appendix 1: Description of all variables in the dataset.**

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Group** | **Description** |
| captid (string) | Metadata | ID of Youtube caption. |
| captsLastUpdated (text, number): | Metadata | The last time a video caption is updated; is stored in the YYYY-MM-DD: Time format. |
| categoryId (encoding of text) | Content | The category each Youtube video belongs to, such as person, blog, science. See Youtube API for more info. <https://gist.github.com/dgp/1b24bf2961521bd75d6c> |
| contentDefinition (string) | Content | Consists of two types: hd (high definition) or sd (standard definition). 68% are hd. |
| 'contentLicensed (Boolean) | Content | Whether the content on that channel is licensed. If a video is licensed, it cannot be used for commercial purposes without the permission of the video creator. |
| description (string) | Metadata | Description of videos seen below the rectangle box (e.g., importance of lipid metabolism). |
| keyword (text) | Search | Keyword used to query that video. Some videos appear under more than one keyword. |
| 'publishedAt' (text, number): | Metadata | Date and time a video is published. |
| rank (number) | Search | The rank at which the video appears if searched using keywords in incognito mode. |
| relevantTopicIds (encoding of text): | Content | TopicIDs created by Youtube based on knowledge graph. |
| 'subtitle' (text): | Content | The transcript of what is said in the video (e.g., “In this video, we will talk about”). In this dataset, only 700 videos have subtitles. |
| 'title' (text): | Metadata | Title of the video (e.g., “Importance of insulin”). |
| 'video\_duration (text, number) | Metadata | Duration of the video |
| 'topicIds' (encoding of text): | Content | All except 3 are NaN |
| 'Kincaid' (number)\*\*: | Derivative | Flesch-Kincaid readability index. Measures the minimum education level required to comprehend the text.  \*\* The higher Kincaid is, the more difficult a text is to comprehend. See formula using link above. |
| 'FleshReadingEase (number)\*\*: | Derivative | Flesch-reading ease index. Measures how easy it is to read a text. The higher the reading score, the easier a piece of text is to read) |
| 'ARI' (number)\*\*: | Derivative | Automated readability index. The higher ARI is, the more difficult it is to read the text. |
| 'word\_count' (number) \*\* | Derivative | How many words appear in the description; calculated using Python library.  \*\* NOTE: Cannot handle special symbols, related to readability. |
| 'word\_unique (number)\*\* | Derivative | Number of unique words in the description of the video. |
| 'transition\_words (number)\*\*: | Derivative | Number of transition words (think: and, so, but, however, etc.) in the description of the video. |
| 'summary\_words (number)\*\* | Derivative | Number of transition words (think: and, so, but, however, etc.) in the description of the video. |
| 'active\_verb' (number)\*\* | Derivative | How many active verbs appear in the description. |
| 'sentence\_count' (number) \*\* | Derivative | How many sentences appear in the description. This is calculated using Python library. |
| channelCommentCount (number) | Metadata | How many comments a video receives.  Some channels have extremely high commentCounts, such as British GOT. |
| channelDescription (text): | Metadata | Description of each channel (created by channel owner in About). |
| channelId (text) | Metadata | ID of the video’s channel. |
| channelPublishedat (text): | Metadata | Time at which channel was established. |
| channelSubscriberCount (number): | Metadata | How many subscribers a channel has. |
| channelTitle (text) | Metadata | Name of the channel (e.g., All about Diabetes and Related). |
| channelVideoCount (number): | Metadata | How many videos that video’s channel has posted. |
| channelViewCount (number): | Metadata | How many views a channel posting that particular video has received. |
| 'keyword\_title\_cosine' (Number) | Derivative | Cosine similarity between title and keyword. Doesn’t suffer from out of range problem. |
| 'keyword\_decription\_cosine' (Number): | Derivative | Cosine simiarlity between keyword and description. |
| comment\_title\_cosine' | Derivative | Cosine simiarlity between comment and title. |
| Comment\_description\_  cosine' (Number) | Derivative | Cosine simiarlity between comment and description. |
| viewCount (number) | Metadata | How many views the video receives. |
| likeCount (number): | Metadata | How many likes a video receives. |
| dislikeCount (number) | Metadata | How many dislikes a video receives. |
| commentCount (text) | Metadata | How many comments have been made on each video. |
| postive\_comment\_count' (Number) | Derivative | Of the most recent 100 comments, how many are positive comments? |
| Negative\_comment\_  count' (Number) | Derivative | Of the most recent 100 comments, how many are negative comments? |
| neutral\_comment\_count' (Number): | Derivative | Of the most recent 100 comments, how many are neutral comments? |
| comment\_unique\_words(Number) | Derivative | The number of unique words appearing in the most recent 100 comments. |
| comment\_total\_words' (Number): | Derivative | The total number of words appearing in the most recent 100 comments. |
| Scene\_count (Number) | Derivative/Content | The number of scenes appearing throughout the video. |
| Object\_count (Number) | Derivative/Content | The number of medical objects appearing throughout the video. |
| Text\_confidence (Number) | Derivative/Content | The average confidence level in Youtube auto-generated transcription. |
| Speech\_confidence (Number) | Derivative/Content | The quality of audio level in Youtube subtitle. |

**Citations:**

<https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NHE-Fact-Sheet>

Holman: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7077778/>