**Editorial comments:**

Changes to be made by the Author(s):

*1. Please take this opportunity to thoroughly proofread the manuscript to ensure that there are no spelling or grammar issues. The JoVE editor will not copy-edit your manuscript and any errors in the submitted revision may be present in the published version.*

Response: We thank the journal for the opportunity to elevate the manuscript according to the comments and suggestions. As recommended, we have thoroughly proofread the manuscript to remove spelling or grammar issues.

*2. Please upload Table 5 as an xls/xlsx. Tables 1-4 are just Tables, and not Materials Tables.*

Response: As requested, we have created a new table 5 in xls format.

*3. Please revise the table of the essential supplies, reagents, and equipment. The table should include the name, company, and catalog number of all relevant materials in separate columns in an xls/xlsx file.*

Response: Our manuscript highlights a cloud computing platform for Text Mining. In this manuscript, we do not describe any supplies, reagents, and equipment. Our downloadable code is in the GitHub repository. Users can follow our instructions in the protocol to establish a cloud computing framework for phrase mining and data analysis.

*4. Please remove references from the Abstract.*

Response: As requested by the editors, we have removed all references in the abstract.

*5. JoVE policy states that the video narrative is objective and not biased towards a particular product featured in the video. The goal of this policy is to focus on the science rather than to present a technique as an advertisement for a specific item. To this end, we ask that you please reduce the number of instances of “CaseOLAP” within your text. The term may be introduced but please use it infrequently and when directly relevant. Otherwise, please refer to the term using generic language.*

Response: In the updated version of the manuscript, we have replaced the term “CaseOLAP” with “phrase mining” a general term in Text Mining. Accordingly, the number of instances of “CaseOLAP” has been reduced to a minimum in the revised manuscript.

6. *Please ensure that all text in the protocol section is written in the imperative tense as if telling someone how to do the technique (e.g., “Do this,” “Ensure that,” etc.). The actions should be described in the imperative tense in complete sentences wherever possible. Avoid usage of phrases such as “could be,” “should be,” and “would be” throughout the Protocol. Any text that cannot be written in the imperative tense may be added as a “Note.” However, notes should be concise and used sparingly. Please include all safety procedures and use of hoods, etc.*

Response: We have thoroughly revised our protocol keeping our original framework of procedural steps. As recommended by the editors, the protocol now guides the reader step by step with all commands to run the algorithm and manage intermediate data in creating a cloud computing framework for phrase mining and data analysis.

7. *The Protocol should contain only action items that direct the reader to do something. Please move the discussion about the protocol to the Discussion.*

Response: We have relocated the discussion part of the protocol to the ‘Introduction’ and ‘Discussion’ section. The revised protocol now includes only action items to perform in a sequence.

8. *Please add more details to your protocol steps. Please ensure you answer the “how” question, i.e., how is the step performed? Alternatively, add references to published material specifying how to perform the protocol action.*

Response: We have significantly revised our protocol to guide the user on how to conduct the designated task in a step by step fashion. Samples of results in intermediate steps are provided as figures, tables, and data links to GitHub. Log files are generated at each steps allowing the user to understand how the algorithm is performing the task. If an error occurs, the error messages are printed out in the log files.

*9. Please revise the text to avoid the use of any personal pronouns (e.g., "we", "you", "our" etc.).*

Response: As requested by the editors, we have removed any personal pronouns from the manuscript.

*10. 1.31: How is the checksum retrieved?*

Response: We have moved internal details of the md5-checksum from the protocol to the core of the CaseOLAP package. This allows the user to perform checksum automatically during the download process. Checksum file is retrieved from an FTP address and compared with one calculated from downloaded document to confirm the integrity of the downloaded data.

*11. Please provide all user input commands: File | Save | etc.*

Response: In the revised protocol, we have provided all required steps and commands to create, save, and run programs and data files systematically in allocated directories.

*12. 2.2.2: How is this done?*

Response: Section 2.2.2 in the previous protocol is now section 3.2 in our revised protocol. This parsing step is built on a tree data structure of extracted XML files. For this specific type of data, we have implemented the code to create a key-value pair dictionary for each document. By performing step 3.2, the user automatically gets the parsed data saved as ‘*pubmed.json*’ in the ‘data’ directory.

*13. 2.3: How is this done explicitly?*

Response: Section 2.3 in the previous protocol is now section 4 in the revised protocol. We have explicitly provided the steps to run MeSH to PMID mapping using parsed data. In the first step, the algorithm reads the MeSH Tree as input data, and in the second step performs the MeSH to PMID mapping and finally saves the mapped table as a ‘*mesh2pmid.json*’ file in the ‘data’ directory.

*14. 3.2: Initialize how?*

Response: We have provided all the steps to initiate an ‘indexing’ server using Elasticsearch. Section 3.2 in the previous protocol is now section 5.5 in the revised protocol. With ‘*run\_index\_init.py*’ and the command provided to run this file, the user can initiate indexing of the database with the addressed schema from the ‘index\_init\_config.json’ in the ‘config’ directory.

*15. 3.3: Create how?*

Response: In the revised manuscript, we have provided all the steps to populate indexing server using Elasticsearch. Accordingly, section 3.3 in the previous protocol is now section 5.6 in the revised protocol. With ‘*run\_index\_populate.py*’ and the command to run this file, the user can populate the indexing database with the addressed schema and parsed data.

*16. 4-6: How are these steps explicitly done?*

Response: Sections 4-6 in the previous protocol is now step 6-9 In the revised manuscript. We have provided a step by step breakdown of the process through 4-6 with required code files and commands to run. Based on the user’s entity and category information, the user can perform an entity-category analysis similar to the sample results presented.

*17. Please provide a specific protocol with specific values instead of a generalized one. It helps to have specific search terms for a specific example.*

Response: In the revised manuscript, we have provided all required specific packages with specific versions in ‘*environment.yml*’. Our presented protocol uses the data downloaded from PubMed to produce the output. This protocol is capable of producing quantified phrase based summarizations of phrase-category associations from about 26 million abstracts.

*18. Please highlight 2.75 pages or less of the Protocol (including headings and spacing) that identifies the essential steps of the protocol for the video, i.e., the steps that should be visualized to tell the most cohesive story of the Protocol. Remember that non-highlighted Protocol steps will remain in the manuscript, and therefore will still be available to the reader.*

Response: As requested by the editors, we have limited the highlighted sections of the protocol to 2.75 pages including headings and spacing.

*19. Please ensure that the highlighted steps form a cohesive narrative with a logical flow from one highlighted step to the next. Please highlight complete sentences (not parts of sentences). Please ensure that the highlighted part of the step includes at least one action that is written in imperative tense.*

Response: We have highlighted the portion of the protocol as recommended by the editor.

*20. As we are a methods journal, please revise the Discussion to explicitly cover the following in detail in 3-6 paragraphs with citations:*

*a) Critical steps within the protocol*

Response: The critical steps in selecting user-defined entities and categories in the protocol are discussed in the ‘Discussion’ section **[Line 609-613]**.

b) Any modifications and troubleshooting of the technique

Response: For modification of codes, we have provided a configuration interface in a ‘configuration’ directory, where the user can change the schema or configuration of the process. When this process is successfully completed, a log file provides the detail of internal processing. Or otherwise, if an error occurs, it will show where the error was produced.

c) Any limitations of the technique

Response: The limitation of the current phrase mining technique is discussed in the ‘Discussion’ section **[Line 645-653].**

*d) Any future applications of the technique*

Response: We have discussed the future application of the current phrase mining technique in the ‘Discussion’ section **[Line 655-661]**

*21. Please ensure that the references appear as the following: [Lastname, F.I., LastName, F.I., LastName, F.I. Article Title. Source. Volume (Issue), FirstPage – LastPage (YEAR).] For more than 6 authors, list only the first author then et al.*

Response: We have modified the reference section according to the editor’s recommendations.

**Reviewers' comments:**

Please note that novelty is not a requirement for publication and reviewer comments questioning the novelty of the article can be disregarded.

Please note that the reviewers raised some significant concerns regarding your method and your manuscript. Please thoroughly address each concern by revising the manuscript or addressing the comment in your rebuttal letter.



**Reviewer #1:**

**Major Concerns:**

I am unable to find novelty

Response: JOVE is a journal with the main focus on visualization of experimental protocols. Accordingly, we aimed to highlight the methodology of CaseOLAP which could be beneficial for biomedical communities.



**Reviewer #2:**

**Manuscript Summary:**

This manuscript describes a protocol that has been developed by the authors, the CaseOLAP process, and a data structure called the Text-cube (from prior work). They describe the process of implementing the protocol for an example dataset. A set of figures and tables are included to show the example in detail.

**Major Concerns:**

*1. JOVE is a journal for visualization of experimental protocols. This is most useful when the protocol has many physical steps. This protocol does not. Working with software, especially software that is NOT point and click, is better served by providing script examples and downloadable code for researchers to use directly, which these authors do in their Github repository. If this software has been shown to generally be too hard to setup and use by other researchers, they would be better served using tools like Singularity or Docker and providing images for users to download and use. A screen capture video of this protocol, which seems the most likely visual to be made, is not particularly useful for the most likely users who would be computing researchers.*

Response: We thank our reviewers for the valuable suggestion. We have elevated our protocol to address the above concerns. We have provided a GitHub repository with downloadable code and a sample data directory. Users can download or clone the project directory and set up a python environment on their device and build a cloud-based text mining platform by following the steps in our protocol. The visualization video of this manuscript will demonstrate how to perform the steps in the protocol with our sample results from intermediate steps as well as the final output. We have clearly defined what users need to modify in order to implement this protocol with their set of entities and categories. The video presentation will demonstrate all interfaces in which the user can remodel the whole process based on their entities and categories.

*2. The level of detail in the**protocol is not sufficient. Although the authors do provide links to example python code in Jupyter notebooks, the protocol included in the text itself does not cover many of the minor details that need to be covered to make all of this work. Compare it, for instance, with other software based protocols published by JOVE previously, for instance:* [*www.jove.com/video/51639/high-throughput-image-analysis-tumor-spheroids-user-friendly-software*](https://urldefense.proofpoint.com/v2/url?u=http-3A__www.jove.com_video_51639_high-2Dthroughput-2Dimage-2Danalysis-2Dtumor-2Dspheroids-2Duser-2Dfriendly-2Dsoftware&d=DwQGaQ&c=UXmaowRpu5bLSLEQRunJ2z-YIUZuUoa9Rw_x449Hd_Y&r=KOkeXzUyfou-SwtJpwMQmfdlQ3IecpPh8dAF6FpTFjw&m=R1FMHz9zv8iwVcJmtgUrMrBTzSGv0qh9IoBcPBZq12o&s=pYgHThgSinMbALLmoD588cl_ehWzaRW_sZ96L11IKWc&e=) *which gives a much deeper level of detail of what needs to be done to set things up and use them. Note that if the difference between the protocol here and in the example depends on the nature of this software, then see my comment 1, above. Either there is not enough detail or the* more scriptable nature of this software*argues for a non-visual presentation.*

*Much of the protocol (lines 115 to 302) varies from being what look like draft notes (incomplete) to text describing things that do not seem to be part of the actual protocol implementing the system in hardware and software (280 on integrity is an example), so it is hard to evaluate this as a protocol per se.*

Response: We appreciate the reviewer bringing this issue to our attention, and it does require clarification. We have benefited from the method presented in the link provided by the reviewer. We have elevated our protocol to address the above concerns and have systematically presented the commands to run program files and have provided steps to manage the intermediate data. To allow users to modify the input data and output results, we have provided a configuration directory (‘config’) with data configuration files. We hope, with commands to run programs, data handling instructions, and user based control to data input and output, this protocol sufficiently provides enough scriptability nature to this software.

In the ‘Introduction’ section of the revised manuscript, we have clearly explained the CaseOLAP score calculation with a link to ‘Materials and Methods’ section of the previous publication **(Ref-5)** **[Line 108-122]**. We have also provided a summary of the three sub-scores: Integrity, Popularity, and Distinctiveness in **Table 5**.

The integrity score denotes the quality of the phrases mined from the documents. In our analysis, the phrases (entities) applied to extract information were the mitochondrial protein names (including abbreviations and synonyms), acquired from UniProt (uniprot.org). Thus, the CaseOLAP algorithm was not used to determine the integrity of these phrases, and the integrity score is the same across all proteins. The integrity score for each protein name is 1.0 (the maximum score), because the UniProt protein naming system has been well established and broadly respected and applied. We have clarified about the ‘integrity’ in our ‘Introduction’ section **[Line 110-113].**

*3. The paper's organization makes little sense to me. After "representative results" (lines 304 - 352) there are what appear to be figure captions (lines 353 to 506). If this is meant to be the "representative results" narrative, then it fails to read like a narrative (and, in this case, the figures appear to have no captions!). If these are meant as captions, then the "representative results" are incomplete. But either way this is not a coherently written section.*

Response: We thank the reviewer for pointing out missing “FIGURE AND TABLE LEGENDS”. We have placed this title at the appropriate location in the manuscript.

*4. Neither the introduction nor the discussion at the end cite sufficient references to convince me of the general applicability of this method. While there is some reference in the general literature, my own searching was unable to show that this protocol is something that has wide enough interest for the community at large. This seems to be something of a niche technique. Perhaps if more works can be done by the developers to show general applicability, then this would change.*

Response: We thank the reviewer for providing us an opportunity to elaborate general applicability of this method. As suggested by the reviewer, we have added the comparison of CaseOLAP with other Text Mining techniques in the ‘Discussion’ section with corresponding references **[Line 615-643]**.

CaseOLAP was developed in 2016 **[Ref 1]**. This algorithm is novel in the Text Mining field. The concept of using a Data-Cube**[Ref 8,9,10]** and a Text-Cube**[Ref 2,3,4]** has been evolving since 2005 with new advancements to make data mining more applicable. The concept of Online Analytical Processing (OLAP) **[Ref 11,12,13,14,15]** in data mining and business intelligence goes back to 1993. There are different types of OLAP systems implemented in data mining. For example: (1) Hybrid Transaction/Analytical Processing (HTAP) **[Ref 16, 17]**, (2) Multidimensional OLAP (MOLAP) **[Ref 18, 19]** – Cube based, (3) Relational OLAP (ROLAP) **[Ref 20]**.

Specifically, the CaseOLAP algorithm has been compared with numerous existing algorithms, specifically, with their phrase segmentation enhancements, including TF-IDF+Seg, MCX+Seg, MCX, and SegPhrase. Moreover, RepPhrase (RP, also known as SegPhrase+) has been compared with its own ablation variations, including (1) RP without the Integrity measure incorporated (RP No INT), (2) RP without the Popularity measure incorporated (RP No POP), and (3) RP without the Distinctiveness measure incorporated (RP No DIS). The benchmark results are shown in the study by Fangbo Tao, et. al **[Ref 1]**.

There are still challenges on data mining which can add additional functionality over saving and retrieving the data from the database. Context-aware semantic Analytical Processing (CaseOLAP) systematically implements the Elasticsearch to build an indexing database of millions of documents (Protocol 5). The Text-Cube is a document structure built over the indexed data with user provided categories (Protocol 6). This enhances the functionality to the documents within and across the cell of the Text-Cube and allows us to calculate term frequency of the entities over a document and document frequency over a specific cell (Protocol 8). The final CaseOLAP score utilizes these frequency calculations to output a final score (Protocol 9). In 2018 we implemented this algorithm to study ECM proteins and six heart diseases to analyze protein-disease associations. The details of this study can be found at **[Ref 5]** indicating that CaseOLAP could be widely used in the biomedical community exploring a variety of diseases and mechanisms.

*5. Additionally, the properties claimed for the system (especially lines 509-511) have not been shown statistically, nor has sufficient supporting literature is cited in the text. Therefore, I do not know if there is any reason to expect this method to work when applied to other data sets. While in machine learning and optimization there are the "no free lunch" theorems that suggest that we cannot prove general applicability for any method like this, there are two problems specific to this paper:*

Response: We thank the reviewer for this comment. The CaseOLAP algorithm has been successfully applied to different types of data (e.g., news articles) and the results have been published **[Ref 1].** We applied this algorithm to biomedical documents **[Ref 5]**. The CaseOLAP algorithm can be implemented on any documents that are associated with keywords (e.g., MeSH terms in biomedical publications, keywords in news articles).

Our protocol provides the method to establish a cloud computing platform including the general steps: downloading, parsing, indexing, mapping, counting and score calculation. It is the user’s decision to select proper machine learning approaches to analyze entity-category associations represented by the caseOLAP scores.

*5a. There is no discussion and no apparent citation of another paper that discusses the statistical nature of the results. For example, the authors claim that in figure 6 the difference between 0.14 and 0.17 is a significant difference in their heatmap. Why? What is the basis for this claim? Should I really on the basis of this result claim that Sodium/Potassium-transporting ATPase-s-alpha-3 is significantly more referred to in INFT vs ADLT literature? What are the "meaningful insights" the authors claim (lines 509-511)? How did they show meaningfulness here?*

Response: We thank the reviewer for providing us an opportunity to clarify our results. The results presented in **Figure 5** and **Figure 6** are the top 10 proteins based on their scores. Our result consists of nearly 3000 proteins and their scores in 4 different age groups. We appreciate the reviewer’s suggestion to statistically compare scores between groups. We want to emphasise that the CaseOLAP scores are constructed from three different sub-scores that are normalized within the cell and across the cell. This allows us to compare the score of the entities within the cell and across the cells. In the revised manuscript we have included the significance of the scores and statistical error. Our sample result now reads as *Sodium/Potassium-transporting ATPase-s-alpha-3* is significantly more referred to in *INFT* vs *ADLT* group with ‘0.03’ difference is significant according to a range of mean difference (0.029 to 0.042) with ‘99’% confidence level **[Line 399-406]**.

*5b. Failing to provide some theory supporting the idea that the differences are "meaningful"**in some sense, then the authors need to cite a sufficient number of previous use examples where the system has been used and shown to produce actionable insight/useful summarization or some other positive metric of performance. My reading of the citations does not show that this has been done. If there are other papers I was unable to find them, but ultimately it is the responsibility of the authors to make this point.*

*While I do think the technique may be useful in some sense and is worthy of additional study, it cannot be described at present as an established protocol of general interest to the community.*

Response: We thank the reviewer for this comment. The previous publication **[Ref 1]** by Fangbo Tao, et. al. presents sufficient results demonstrating the performance metric of the algorithm. The CaseOLAP algorithm has been compared with numerous existing algorithms, with their phrase segmentation enhancements, including TF-IDF+Seg, MCX+Seg, MCX, and SegPhrase. Please also see our response to comment 4.

We want to provide the biomedical community with a step by step procedure to establish the cloud computing framework. In our previous publications **[Ref 5]**, we have compared the results with manually searched results from the reactome database and presented the results discussion.

**Minor Concerns:**

1. *While undoubtedly a typesetting issue, the tables are horrendous in appearance and not properly labeled. The bad formatting definitely contributed to the delay in returning the review.*

*Not enough detail is given on the steps described in lines 523 to 527, which are not detailed in the protocol, either, and appear to be critical to the whole example (collecting protein names). This was an important detail of the protocol example not addressed earlier. Perhaps this should be a major issue.*

Response: We thank the reviewer for the comment. Accordingly, we have upgraded our tables regarding label and formatting. We have discussed the user provided entities in the discussion. The user may use a set of entities and categories according to his/her own choice. In our protocol, we have provided steps to set up user-defined entities and categories. The details of protein name selection are highlighted in the ‘introduction’ section **[Line 124-133]**.

*2. Much more detail needs to be given on how the caseOLAP scores are computed and interpreted, or a citation given that does this. This section is not detailed enough to show the general usefulness of these scores.*

Response: We thank the reviewer for this comment. We have revised the ‘Introduction’ section to elaborate the CaseOLAP score calculation. The provided references present in depth descriptions of the score calculation **[Line 108-122]**.

The CaseOLAP score computation is based on user-provided entities and categories. The algorithm compares the entity count data within a cell and across the cells of the Text-Cube. There is not a standard method in the Text Mining literature capable of quantifying entity-category associations based on millions of text documents. We have demonstrated that the CaseOLAP platform fulfills this need.



**Reviewer #3:**

**Manuscript Summary:**

A complete protocol is provided for automatic identification of phrases-category associations using CaseOLAP on the cloud.

**Major Concerns:**

None

**Minor Concerns:**

*1. The JSON file for the parsed sample on the git repo is broken.*

Response: We apologize for this issue. This issue has now been resolved in the revised manuscript.

*2. It will be better to have discussions on the limitation of the system and future development/improvement plans.*

Response: We thank the reviewer for this suggestion. We have included a paragraph addressing the limitations of the platform and future developments in our ‘Discussion’ section.



**Reviewer #4:**

**Manuscript Summary:**

The authors propose platform, namely cloud-based Context-aware Semantic Online Analytical Processing (CaseOLAP), for automatically computing a score that represents how strongly phrase (e.g. portions names, gene names) is associated with a category (e.g. Age group "adult, child") in the biomedical context. The proposed platform includes six steps: Download and extracting of data, Parsing data, Indexing, Text-cube creation, Entity count, CaseOLAP score computation. Authors carry out two case studies using PubMed dataset and CaseOLAP to demonstrate the efficiency of their proposed protocol.

**Major Concerns:**

*1- The authors need to show illustrative example in how**entity including its abbreviation and synonym is counting. Also, they need to clarify how final term frequency is computed in the paper. It is only shown in the code that the final term frequency is computed by summing up the count of all its abbreviation and synonym.*

Response: We appreciate the reviewer’s suggestion to improve the manuscript. Protocol 7.1 provides instructions to specify entities (with abbreviation and synonyms). During the entity count step (Protocol 7.3.), the algorithm reads entity names with all abbreviations and synonyms and creates a key-value dictionary with a key as an entity name and values as a collection of all abbreviations and synonyms.

This entity dictionary is used to search a specific entity in the indexed database and the counting operation computes the sum of all occurrences, including entity the name, synonyms, and abbreviations. The entity count becomes available as a ‘PMID to entity count’ table in the data repository.

In the revised manuscript (‘Introduction’ section), we have provided the references to detail term frequency, document frequency, Text-Cube document structure, and score calculation **[Line 108-122]**.

*2- The authors in step 5.1 "Selection of Entities" mentions that if user doesn't define entities, adopted AutoPhrase approach will be used. Will the adopted method automatically group abbreviations and synonyms of phrase in one entity? The authors need to clarify that.*

Response: We thank the reviewer for this important comment. To avoid the confusion in the protocol, we have moved the discussion about ‘Autophrase’ into the ‘Discussion’ section **[Line 608-610]**. Our protocol is designed to provide the method for user-defined entities and categories. ‘Autophrase’ in the current version is not capable of distinguishing synonyms and abbreviations for a representative entity. ‘Autophrase’ can be implemented to find top phrases in the textual data. The user can manually separate top-phrases as representative entities, its abbreviations, and synonyms. This prepared entity list can be implemented while following our protocol.

*3- CaseOLAP Algorithm relies on statistical information "term frequency" to compute CaseOLAP score. For infrequent but important entity, CaseOLAP score will be small. How will the proposed protocol handle these entities?*

Response: We thank the reviewer for this comment. The final CaseOLAP score is based on three ranking criteria. We have provided a brief introduction and references providing details about these concepts in the ‘Introduction’ section **[Line 108-122]**. The algorithm calculates the term frequency as well as the document frequency to compute the ‘Popularity’ and ‘Distinctiveness’ scores. For infrequent but important concepts, the ‘Popularity’ could be small but the ‘Distiveness’ score could be high based on document frequency, and vice versa.

**Minor Concerns:**

*1- The authors do not provide enough details information about how their approach is more efficient compared to others approaches. They need to show how text-cube technique make their approach is favorable over others approaches. Also, the authors did not name competitive approaches. It would be useful if the authors include some references to those approaches.*

Response: We thank the reviewer for this recommendation. We have included more references to address previous Text Mining methods in comparison to the Text Mining algorithm that is described in this manuscript **[Line 612-640]**. We have added the comparison of this technique with other existing OLAP techniques in the discussion section. Specifically, the CaseOLAP algorithm has been compared with numerous existing algorithms, with their phrase segmentation enhancements, including TF-IDF+Seg, MCX+Seg, MCX, and SegPhrase and RepPhrase. The benchmark results are shown in Figures 5A and 5B in the study by Fangbo Tao et. al. **[Ref 1]**.

