

13 April, 2019

Dear Prof Yan,

Manuscript Number: 47226

Title: The application of Big Data analysis in gastrointestinal research

Thank you for your letter concerning the revision of our manuscript. The manuscript has been revised according to the reviewers' comments. The changes have been highlighted in the manuscript for your reference. The followings are the point-by-point responses to the reviewers' comments.

We hope that the reviewers and editors will find this revised version acceptable for publication in the World Journal of Gastroenterology.

Yours sincerely,

Wai Kay Seto

Reviewer 1:

This is a review article about Big Data Analysis in GI diseases that presents : 1- Definitions and current context including essence of BDA 2- Areas of application BDA: studying pathogenesis, finding possible drugs and proving therapeutic intervention efficacy. 3- BDA methodology : mostly focused on propensity score 4- Examples of BDA studies in GI

May I recommend the following revision :

a-Table 1 to be substituted with a concept map that contains the same content ; This enables readers to understand the whole idea and details

We thank you for your comments. A concept map is defined as a cluster diagram that illustrates how different ideas are related (i.e. the relationship between different variables). Table 1 describes the advantages and shortcomings of Big Data approach. While we have thought of using a concept map or figures as a presentation, as such, those elements are not related to each other, making a concept map difficult to be constructed. We hence believe the usage of a table best presents the different issues related to Big Data analysis with the necessary clarity for good readership, and deeply appreciate your kind understanding and consideration.

b- Table 3 should include methodological points that the authors mentioned in this article i.e. fulfillment of 3v or using BDA softwares or statistical process for each of included studies.

We thank you for your comments. Important statistical process like propensity score adjustment/matching, and other advanced techniques like machine learning algorithm and natural language processing have been presented. Nearly all studies' electronic database system fulfilled the 3V criteria: volume, velocity and variety. This is because volume refers to the storage space necessary for data recording and storage (but not sample size), velocity refers to the speed of data generation and transformation) and variety (with the data being collected from various sources, e.g. medical records, laboratory results, operation records, endoscopy records, etc). This information was now added in the 5th column of Tables 3-7.

The only exception is Nurses' Health Study (NHSII) and Health Professionals Follow-up Study (HPFS), which are prospective studies without instantaneous update of the clinical information. Some of the clinical data requires regular update of the clinical information by filling in questionnaires by participants, which limits the velocity of data generation and transformation. This has been mentioned on page 15, lines 8-13.

Reviewer 2:

This is a very accurate study about new frontiers in collecting and analysing data in health sciences. Field of applications cited as examples are very useful, as well as

the tables. The references are updated and abundant. English language is correct. It is a valuable paper although not easy to read, worth of publication.

We thank you for your comments.

Reviewer 3:

The article is aimed to review the use of big data research on gastrointestinal and liver diseases with recent published examples. The title is "THE APPLICATION OF BIG DATA ANALYSIS IN GASTROINTESTINAL RESEARCH".

1. This is a topic review.

2. Several factors influence the data analysis. Please discuss these factors.

We thank you for your comments. We have provided a list of factors/shortcomings that could influence the data analysis in **Table 1**, and proposed some solutions to address these problems.

3. What are the new knowledges from this study?

We thank you for your comments. This is a topic review which provides reader with an overview of big data and its application in clinical research in the field of gastroenterology and hepatology. This includes phenomapping of diseases, accurate risk prediction, precision medicine, discovery of new drugs and drug safety surveillance.

We also provide practical tips on the how to use more advanced statistical analysis methods to handle the large volume data, address the common pitfalls of biases and confounding factors from Big Data research. In addition, we provide our viewpoints on future perspective of Big Data research, including some under developed areas like parental-child linkage. These can be found on **page 21, lines 18-30**.

4. Finally, please recommend the readers "How to apply this knowledge for routine clinical practice?".

We thank you for your comments. We have provided a comprehensive list of researches on different aspect of gastrointestinal disease (including gastric cancer, colorectal cancer, liver cancer, inflammatory bowel disease, gastrointestinal bleeding) to illustrate how the knowledge of Big Data approach can be applied for routine clinical practice (**pages 14-21, Tables 3-7**).