Peer Review File
Article information: https://dx.doi.org/10.21037/jtd-22-532

Reviewer A

It was a paper on improving the quality of care for pulmonary aspergillosis(PA) through the system in the context of limited medical resources.

1. It is said that the "Diagnostic Coding Reference" was used for PA diagnosis, but I wonder if the medical staff did not enter the "Code", the patient without "Code" is excluded from the PA diagnosis analysis
Example)
1) If the aspergillus Ab IgG or aspergillus Ag test is positive in the EHR, but an appropriate diagnostic code is not entered, I wonder how these patient groups are managed by the system.

Reply 1 1): Thank you very much for your comments. When resident doctors input evidence related to the diagnosis of pulmonary aspergillosis, the system will start and give hints of general examination methods need to be perfected. If it is necessary to modify the inspection method of pulmonary aspergillosis, such as CT, Galactomannan and other general related examinations, relevant examination methods can also be set independently according to the background of examination equipment in different hospitals.

2) If PA was suspected from the imaging findings, but aspergillus Ab IgG or aspergillus Ag tests were not performed, and an appropriate diagnostic code was not entered, I wonder how the system handles it.

Reply 1 2): Thank you very much for your comments. When resident doctors input evidence related to the diagnosis of pulmonary aspergillosis, the system will start and give hints of general examination methods need to be perfected. If it is necessary to modify the inspection method of pulmonary aspergillosis, such as IgG, Galactomannan and other general related examinations, relevant examination methods can also be set independently according to the background of examination equipment in different hospitals.

3) PA is classified in various ways according to the patient's immune status (IA, CPA, or ABPA), and I wonder if it is possible to evaluate the patient's general condition through the system.

Reply 1 3): Thank you very much for your comments. This is a part that we have not designed in the system. Your idea reminds us. In the next process of improving the system, we will add this content, such as further scoring the general condition of each

patient according to the patient's examination and the description of the general condition of the patient in the medical record.

2. PA is one of the disease groups that have recently been receiving increasing attention, and in order to check whether this trend is reflected, I would like you to additionally present the number of patients diagnosed according to each subtype at 1year intervals.

Reply 2: Thank you very much for your comments. We will publish the results on the hospital's official website http://www.gyfyy.com/ each year. For the diagnosis of previous PA patients, we published the results in Emerging Microbes & Infections through papers. Trends of pulmonary fungal infections from 2013 to 2019: an AI-based real-world observational study in Guangzhou, China. As you might expect, pulmonary aspergillosis, the dominant pulmonary fungal disease, continues to increase year by year, with mortality increasing (Figure 1). We provide this document for you in the attachment.

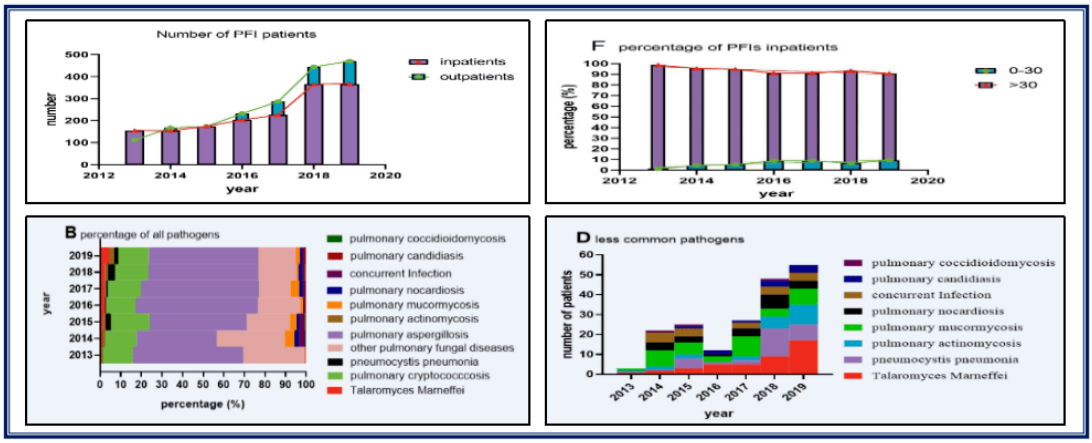

Figure 1: Changes of pulmonary fungal mycosis

3. The personal information of actual patients is listed in Figure 2, so it is necessary to modify the figure to protect personal (or patient) information.

Reply 3: Thank you very much for your comments. For your question, we have blurred the part involving the patient's personal information in Figure 2, making the patient's personal information unidentifiable. We include the ethics review in the appendix.

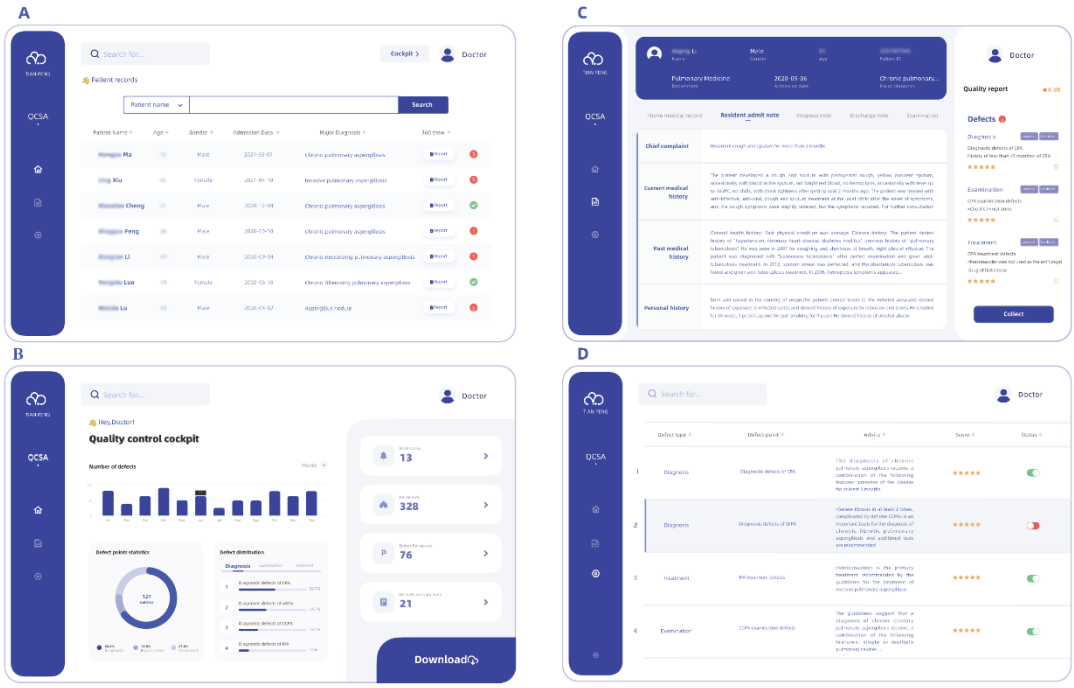

Figure 2 in Manuscript

Changes in the text: We have modified our text as advised (see Page 24, line 531-352)"

4. In Figure 3, 284 unspecified pulmonary aspergillosis was excluded from the analysis by describing it as a patient without a specific subtype. Also, I'm curious about which entity the patients are, and I'd also like to mention if this is a system issue or something else.

Reply 4: Thank you very much for your comments. Dear reviewer, Because the QCSA system did not contain quality control points for diagnosing unclassified PA cases. We excluded 284 cases of unclassified PA cases from 699. "The explanation for this problem is that there is no related quality control item in the system for those patients with unclear diagnosis. The root cause of this problem is that the quality control items for patients with unconfirmed diagnosis were not included in the guide. Since quality control for patients with unconfirmed diagnosis is more complicated and requires more technical requirements, we decide to include the quality control content for patients with unconfirmed diagnosis in the next version of the system.

Reviewer B

The paper by Li et al describes the use of natural language processing to potentially improve outcomes in pulmonary aspergillosis.

This is a potentially hugely beneficial concept in a field where there is clear variability in adherance to guidelines, lack of expert knowledge, rising antimicrobial resistance and poor outcome.
Therefore, potentially using a NLP algorithm to enable and improve antifungal stewardship could have significant improve in outcomes and I congratulate the authors for their work.

To improve the clarity for readers I have made a few suggestions that would improve the manuscript and clarity for readers.

Revisions:
1) Although the overall process is well described, reading the manuscript, I was not particular clear which terms were used to build the NLP algorithm. I can see the diagnostic SNOMED codes in the appendix, but how the NLP algorithm detect defects compared to guidelines? How was that determined using NLP or was manual checking required. In the training, was there any minimal annotation performed, and what were the validation steps performed? It would be helpful I think to have a larger appendix with the relevant language terms used.

Reply 1: Thank you very much for your comments. The identification of defects is divided into several steps. First, the quality control points are extracted from the guidelines, and then the defect trigger logic is set according to the quality control points. For example, if the CPA patient is found to have a disease course of less than 3 months, it is determined to be defective. The refining and setting of this logic are constructed manually by medical experts, and finally the medical records are identified according to the preset logic to determine whether there are defects in the medical records. For different quality control points and their corresponding judgment logic, see“ Appendix I and J” provided in the second point below. In this study, NLP technology is mainly used to extract structured quality control point data from unstructured text, and to standardize data. This technology can not only extract variables such as diseases, clinical manifestations, examinations, and drugs in unstructured texts (such as chief complaint, current medical history, past history, and course records, etc.) The relationship between them, as well as the modification of time adverbs and other information are extracted synchronously. The extracted information is then converted into standard terms through standardized mapping and stored in the database for identification by the algorithm. To standardize terminology, it is necessary to establish

a huge standard terminology base and thesaurus at first. We cooperate with Guangzhou Tianpeng Computer Co., Ltd. They have more than 130,000 terms and more than 6 million synonyms, including diagnosis, drug, examination, test, sign, clinical manifestation, anatomy, unit, occupation, gene, smoking, family Relationships, departments, surgical operations, TCM syndromes, adverse events, limit values, medical consumables and other categories. For an appendix of standard terminology see: "Appendix I". Detailed version：https://cloud.elungcare.com/synonym/index.html

Changes in the text: We have modified our text as advised (see Appendix I and J)

2) Although, the authors give an example of a defect in invasive disease, given the spectrum of aspergillosis analysed which includes chronic aspergillosis and ABPA, it would be important I think to list the defects analysed in each aspergillosis diagnosis. The guidelines are very different for each, and it is very difficult to understand the utility in these conditions without knowing this. Again, perhaps a more detailed appendix would be helpful.

RE: Thank you very much for your comments. 2. See excle"Appendix J", which is the value logic of quality control points.

Changes in the text: We have modified our text as advised (see Appendix J)

3) For this technology to be scaled, as the authors intimate in the discussion there are a number of challenges related to different EHR manufacturers. Could the authors expand this to discuss, if these challenges could be overcome? How would standardization of EHR using SNOMED help this approach. Are there any other solutions that would help. Scalable AI to facilitate antifungal stewardship is a critically important tool for the future and of relevance.

Reply 3: Thank you very much for your comments. We added the following to the discussion: "The most important reason why different hospital medical record systems in China cannot be effectively interconnected is that there is no authoritative and unified terminology standard, which makes clinical descriptions too diverse and non-standardized. The differences between foreign languages and Chinese make some terminology standards such as SNOMED CT not well applied. This greatly affects data interaction and processing, increasing the difficulty of NLP. The good news is that the Chinese Health and Medical Commission is also constantly trying to launch a standardized medical terminology. Some domestic institutions, such as the OMAHA medical terminology system established by the Zhejiang Digital Medical and Health Technology Research Institute, are also committed to solving the medical terminology system and standardization. It is believed that in the future, medical data will become more standardized and semantically interoperable, and NLP algorithms will be more versatile, thereby extending more intelligent application scenarios."

Changes in the text: We have modified our text as advised (see Page 17, line 344-357)"

Reviewer C

This study describes the development of an-EHR based quality control system or pulmonary aspergillosis. The system is able to automatically extract patient's information from EHR. One of the strength of this study is a validation of model's the performance with human experts.
I have several inquiries to the authors as below
1. The QCSA system is based on the clinical practice guidelines from several international societies (IDSA, ECCMID, etc). Obviously, these guideline may change over time based on the most updated evidences. How would this QCSA system follow the update guideline in the future?

Reply 1: Thank you very much for your comments. We are using the add these rules according to the entry into our quality control system, but these items are based on our database as a benchmark, the term includes more than 130000, synonyms number more than 6 million, covers the Chinese in most of the medical standard and nonstandard expressions, it is very important for natural language processing, We in the appendix provided in this article involves the core terms and synonyms for reference, (Appendix I) when the guide to update, we will revised guidelines for entry to import the database for machine learning, the process is a mature, so you can in the shortest possible time to update, and as a result of our system is a system with an Internet connection, So all updates can be pushed to each client in a timely manner.

Changes in the text: We have modified our text as advised (see Appendix I) Detailed version：https://cloud.elungcare.com/synonym/index.html

2. I assume that some of the data in the medical records are in Chinese. How did the authors convert the information to English?

Reply 2: Thank you very much for your comments. Yes, our medical records is based on Chinese, this is our consideration at the early stage of the development, it is well known that Chinese, as the representative of the ideographic system on behalf of the meaning of different words combinations and potential implications is more complex, and the Chinese people in 1.4 billion in the world, as the world's first big language system, and China as a developing country, The imperfection of the medical system requires the assistance of such auxiliary AI diagnosis and treatment system. Based on the above reasons, we set up and expanded the Chinese semantic database and thesaurus, set up quality control rules on this basis, and carry out machine learning.

From the perspective of users, this is also related to the standards of medical records written by Chinese doctors. In the process of writing the article, in order to facilitate reading, the system content has been translated into English. However, the underlying logical structure of the paper is based on the universal computer programming language, so it is not difficult to establish a quality control system based on English or other languages. After all, the Medical term system based on English is perfect, avoiding the programming difficulties caused by a large number of synonyms with the same expression.

3. Page 15; confusion matrix – it may be easier for the reader to see the number in percentages.

Reply 3: Thank you very much for your comments. We have modified the confusion matrix in the text you mentioned into a more clear and convenient percentage number

Table 1. Confusion matrix of verification results between expert team and QCSA

	Defective according to expert	No defect according to the expert
Defective according to QCSA	8.665%	0.342%
No defect according to QCSA	2.622%	88.369%

Note: Of the 877 medical records, 99 pieces were Defective according to expert, and 775 pieces were No defect according to the expert, which were 11.29% and 88.71% of the total medical records, respectively. Of the 99 medical records of Defective according to expert, 76 pieces according to QCSA accounted for 76/99*100%=76.77%, which was 8.67% of the total medical records. There were 23 cases of No defect according to QCSA (23/99*100%=23.23%, 2.62% of the total number of cases). In the 778 No Defect according to the expert, three Defective pieces according to QCSA (3/778*100%=0.39%), which is 0.34% of the total number of pieces. There were 775 No defect according to QCSA, accounting for 775/778*100%=99.61%, 88.37% of the total number of cases. Overall, accuracy =(76+775)/877=0.94; In the sample according to Expert, the probability of being correctly predicted by DCSA was 76/99=0.77, namely, Recall/sensitivity. When Defective according to QCSA, the correct prediction probability is 76/79=0.96 (precision). F1 = 2 * 76 / (76 + 23 + 76 + 3) = 0.85.

Changes in the text: We have modified our text as advised (see Page13, line 263-275)

4. Page 24; conclusion – The system gas good accuracy and expandability….. How would authors come with this conclusion?

Reply 4: Thank you very much for your comments. There is no need to submit this part in the editing requirements, so we delete Summary Points. We have also answered your questions: “The sensitivity and accuracy of QCSA were 0.99 and 0.96, F1 value was 0.85, And the recall rate was 0.77 compared with experts' evaluation. Since our system is developed based on the Internet, we can nest various functions on the server side, such as communication with the administrator and prediction of further diagnostic functions. Therefore, The system has good accuracy and expandability and can be extended to other diseases to help improve The quality of clinical diagnosis and treatment in areas with scarce medical resources.”

5. In supplement – table D2 – QCSA evaluation results of each classification. For the diagnosis of CCPA and CFPA, the diagnosis of these two entities has decreased an overall calculation of recall and F1. Did the authors have explanation on this?

Reply 5: Thank you very much for your comments. This is the result of the system and the reviewer of the same problem cognitive inconsistency. In terms of quality control: The guidelines recommend itraconazole and voriconazole as the first drugs for the treatment of CCPA. According to the system, if voriconazole or itraconazole is selected, there is no defect. The reviewers believe that itraconazole should not be selected when voriconazole can be used in clinical practice. The reason is that voriconazole has a wider antimicrobial spectrum, while itraconazole's resistance rate to fungi increases year by year, so voriconazole is a better choice. So we truthfully record our reviews and count them in our statistics.

CFPA extracted 4 medical records to verify 8 quality control points, and one of them was false negative. If the data amount is small and errors occur, F1 is poor.

6. The development of QCSA demonstrates an excellent effort to standardize the diagnosis and treatment of pulmonary aspergillosis. As a future study, the authors should be able to assess the adoption of clinicians in real world practice. It would certainly increase time that clinicians use to spend on each patient. Adoption of new technology will be challenging in general deployment of QCSA.

Reply 6: Thank you very much for your comments. We added the following to the discussion: “As a new technology, QCSA is a challenge for doctors. It is undeniable that the promotion of a new technology often encounters many problems, such as the learning time of the new technology, the required Internet equipment, and the maintenance and management of the system. Of course, the most important thing is that it also makes doctors see patients for longer. These problems are difficult but not insurmountable. For learning and technical difficulties through on-site and remote teaching can be realized quickly. As for the diagnosis time, the literature shows that the average consultation time of American doctors is more than 20 minutes, ranking second, while the average consultation time of Chinese doctors is less than 5 minutes, ranking third from the bottom among the 67 countries. For fungal lung infections, which are rare and difficult to diagnose, prolonged communication with the patient is required.

How to improve the system in terms of working efficiency and benefit of patients is also our next direction."
Changes in the text: we have modified our text as advised (see Page 16-17, line 358-369)