**HealthMegaMall Project Requirement**

The project is related to disease diagnosis based on big data technology and is composed of two parts:

Speech Recognition Application Development, and

Medical Data Analytics.

The purpose of this project is to provide a system, which can be used to guide the user to describe his/her symptoms (using natural language in English language or Chinese) and be guided through the interview until a likely set of diseases are identified as potential causes, which the physician can investigate). The system can then offer a naturopathic plan, foods to take and avoid, plus lifestyle recommendations.

1. **Speech Recognition Application Development:**

The first part of the project is related to speech recognition and application development. This aspect the project aims to achieve a greatly improved human-machine interaction that uses Natural Language Processing (NLP), Speech Recognition and relevant technologies. The system must be able to understand both Chinese and English languages and respond in these languages. Moreover, the speech recognition function must incorporate a medical thesaurus in the back end so that it will accept and understand synonyms. This Speech Recognition function will make it possible for patients that have limited dexterity or limited computer knowledge to naturally interact with the website and a mobile application (to be built) through speech, rather than keyboard operations.

The application to be developed must support both iOS and Android platforms. The application can allow a user login function (if the user so chooses), and provide for database interaction and help the user provide sufficient information (such as follow up or clarifying questions) for a likely accurate diagnosis. The same features must be incorporated in the website.

1. **Medical Data Analytics**

Medical research, clinical trials and peer reviewed publications are rapidly adding to the body of knowledge supporting our database. Sometimes generally accepted medical practices are changed by newer research. It is therefore important for us to read these publications and update our database daily. There are however some 26 million such documents on the National Institute of Health alone. This makes it a big data undertaking.

Our project aims to create the logic and ability for the machine to millions of existing and future medical articles and reports of clinical trials then output those articles that likely contain information beneficial to our database. The machine must understand the data and associate the relationship between symptoms, diseases, foods, naturopathic solutions and lifestyles. We need to search for key terms and calculate their proximity and relevance to list of symptoms and diseases (and their synonyms). The system must decide (based on research and development of relevancy tests and logic) which articles to extract as the most relevant knowledge for our database and highlight the most important sentences and submit them to our Information Coordinator. He will ultimately decide how to include the most important findings in our database. However, the system must index and automatically provide in-text Citation to References. Plus, it must build a List of References and for each item alphabetize it by the first author's last name, or (if no author is listed) the organization or title, the date and year of publication, and the Title of the publication. The correct acknowledgements must be provided for copyright purposes.

These articles must then be added to our internal database in a standardized format and displayed in table form to users.