Data Science and Engineering (DSE) Record, Volume 4, issue 1.

Development of Data Processing and Visualization for Bacterial and Antibiotic Susceptibility Profile

Chanchanok Aramrat [1] and Pruet Boonma [2]

[1] Master's Degree Program in Data Science, Chiang Mai University, Chiang Mai, Thailand
[2] Department of Computer Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai, Thailand
chanchanok_aram@cmu.ac.th

Abstract. Bacterial data are under-utilized in Maharaj Nakorn Chiang Mai hospital. Bacterial data contains information regarding the bacteria that are isolated from various biological samples collected in routine clinical cares. The data can be used to create bacterial profiles and antibiotics susceptibility profiles which help doctor decide on the most appropriate antibiotics agent to be given to patients with infection. The aims of this study were to develop an application which create bacterial profiles and antibiotics susceptibility profiles by utilizing the hospital bacterial data. To do this, the study was sub-divided into 4 parts 1. Development of ETL process to prepare data for utilization, 2. Data quality assessment, 3. Development of pilot application utilizing prepared data to create bacterial profiles and antibiotics susceptibility profiles, and 4. Feasibility assessment of the pilot application.

All data was extracted from Maharaj Nakorn Chiang Mai hospital database from 2017 to 2018 with an assistance from hospital information technology (IT) personnel. All extracted data was explored and compile into one table to be utilized by the pilot application. The pilot application was written in Google Collaboratory. Overall, the data quality was good. There was some missing data but should barely affect reliability and performance of the application. For feasibility assessment, the pilot application was given to 6 doctors conveniently selected from all doctors working in the hospital for test uses. Later, the doctors were interviewed and asked to provide feedbacks on the pilot application. The application received positive review overall. Improvement points were addressed focusing on data cleaning and preprocessing, minimizing any potential bias.

This study provides insight into the development processes of the pilot application that provide bacterial profiles and antibiotics susceptibility profiles to doctors. Modifications are required before such an application can be used in clinical practice.

Keywords: Bacterial profile, Antibiotics susceptibility profile, Antibiogram, Extract-Transform-Load, Data quality assessment.

1 Introduction

Health data is defined as any data related to health conditions, reproductive outcomes, causes of death, and quality of life [1] [2]. Health data ranges from individual's demographics to medical records documents during health care visits to various individual's biological samples analysis results. Health data is essential in clinical practices, health care system management, and health-related policy decision making [3] [4] [5]. Health data is also utilized in medical research [6] [7].

Microorganism isolation, identification, and profiling from biological samples are laboratory investigation routinely performed in hospitals. They help identified potential microorganisms that cause infections so that appropriate antimicrobial agent can be administered to patients. Microorganism profile and antimicrobial susceptibility profile are essential for physicians when providing care to patients with infection. Microorganism profile shows frequencies of each type of microorganism isolated from biological specimens. Antimicrobial susceptibility profile shows the percentage of a particular type of microorganism is susceptible to each type of antibiotic agent. Both microorganism profile and antimicrobial susceptibility profile can be wildly diverse in patients with different demographic, different health conditions, and different geological location to various mechanisms [8] [9].

Maharaj Nakorn Chiang Mai hospital is a medical school and medical center in Northern Thailand. Many of the health data generated here are recorded as electronic data. Microorganism isolated data is among the various health data stored electronically. However, there is no automatic process for utilizing these microorganism isolated data to create microorganism profile and antimicrobial susceptibility profile to assist physicians in clinical practice in the hospital.

The following are the main objectives of this study

1. To develop ETL for utilizing bacterial culture data from Maharaj Nakorn Chiang Mai hospital electronic medical record to analyze for bacterial profiles and antibiotic susceptibility profiles
2. To evaluate data quality of the bacterial culture data
3. To create a application that summarized data of bacterial profiles and antibiotic susceptibility profiles of Maharaj Nakorn Chiang Mai hospital that are filterable (can specified sub-group characteristics)
4. Assess feasibility of the pilot application

2 Literature Review

2.1 Bacterial profile, antibiotic susceptibility profile, and how they are utilized

Antibiotics contain substances that kill bacterial or inhibit bacterial proliferation. Therefore, antibiotic agents are usually given to patients that are infected with bacteria. The importance of selecting an appropriate antibiotic agent comes from the fact that antibiotics administration for bacterial infections is one of the driving forces that

increase the prevalence of multi-drug resistance organisms (MDROs) [10] [11] [12]. MDROs are microorganisms, primarily bacteria, that develop resistance to one or more classes of antibiotic agents. MDROs are less affected by antibiotic agents that they are resistant to. This makes diseases that are caused by MDROs more resilient to antibiotic administration, making treatment more complicated. Patients with MDROs infection have longer hospital length of stay, require more health resources, and have higher mortality rates [10] [13].

The rationale for antibiotics usage nowadays is to use as much specific bacterial-killing activity as possible [10]. Selecting an antibiotic agent that targets specific bacteria rather than a broad-spectrum agent that attacks a wide range of bacteria is one component for appropriate antibiotic agents. Bacterial profile and antibiotic susceptibility profile provide crucial information that helps guide physicians to decide best optimal appropriate antibiotic agent for a given patient.

2.2 Extract-Transform-Load (ETL)

ETL is a data pipeline processes used for extract data from multiple sources, combine the extracted data together, then load the combined data into a data warehouse or other target system [14]. This data pipeline processes are important to make the most uses out of available data.

Development of ETL pipeline is a challenging process. Many variables need to be considered during the development process. It starts from understand data structure of data sources, plan how data will be cleaned and transformed, all the way up to decide target data warehouse operating system. The number of variables need to be considered can escalate quickly throughout the development process. Below are 10 steps of ETL development framework presented by [15] to help guide with the development of ETL pipeline.

Table 1: Steps of ETL development framework

	Step	Brief description
1	Draw the high-level plan	Design on overall processes, from data sources to target table. No detail in each process needs to be decided/presented yet.
2	Choose an ETL tool	Choose an available ETL tool to use for the ETL pipeline. Using an ETL tool is preferred over pure coding. Although, there are some learning to be done when start using new tool, but after some period of time, the tool would become useful when things need to be added or edited to the ETL pipeline later.
3	Develop default strategies	Design a detailed processes of how data be extracted from data sources and how should the data be transferred to the ETL pipeline.

4	Drill down by target table	Design a detailed processes of how the data, after transferred to the ETL pipeline, be processed/manipulated to create target table.
5	Populate dimension tables with historic data	Further build up ETL pipeline by execute the designed processes for dimension tables using historical data
6	Perform the fact table historic load	Execute the designed processes for fact table using historical data
7	Dimension table incremental processing	Connect the ETL pipeline to the data sources. Further build up ETL pipeline.
8	Fact table incremental processing	Update fact table and set up process to periodically update fact table.
9	Aggregate table and OLAP loads	Create aggregate table and OLAP loads.
10	ETL system operation and automation	Implement ETL pipeline. Make the pipeline to be automatic as much as possible, e.g., automatic update of data, algorithm to handle unexpected errors.

2.3 Data quality assessment

Health-related data are used by health professionals and health policy makers to make clinical decisions and plan health-related policy. It is important to have reliable health data – data that accurately reflect what is happening to a patient or health care system.

There are many frameworks for DQA available which are all similar. According to the reference **"DATA QUALITY ASSESSMENT HANDBOOK"** [16], DQA is divided into seven dimensions as follow:

Table 2: Dimensions, definitions, and brief analysis method in DQA

	Dimension	Definition	Method
1	Accuracy	Degree in which data correctly describes real-world event/object.	Compare with actual real-world data would be ideal (Primary research). If not able to obtain real-world data, a reliable surrogate data would be the next reference choice.
2	Reliability	Degree to which the values (measurement, calculation, or any specification) within the data are stable, consistent, and repeatable over time.	Established through primary research. Identify how a clinical value is measured. Estimate reliability of the measurement method.

3	Consistency	Information of the same feature is represented stored in the same format.	Explore how values in a feature are stored. Compare its format to one other within the same feature.
4	Completeness	Degree in which data are complete	Identify all blank data that should not left blank
5	Relevance	Data fitness to serve its purpose in a given context	Established through primary research. Explore how much information is useful to the users.
6	Accessibility	Data are easily accessed.	Describe how user or IT personnel are able to access to the data.
7	Timeliness	The data are up to date as much as the intended use need it to be.	Explore how fast new input data are updated within the data storage system.

3 Methodology

Methodology is divided into 4 components following 4 main objectives that together pieces into a bigger picture of process required to utilizing bacterial culture data from a hospital

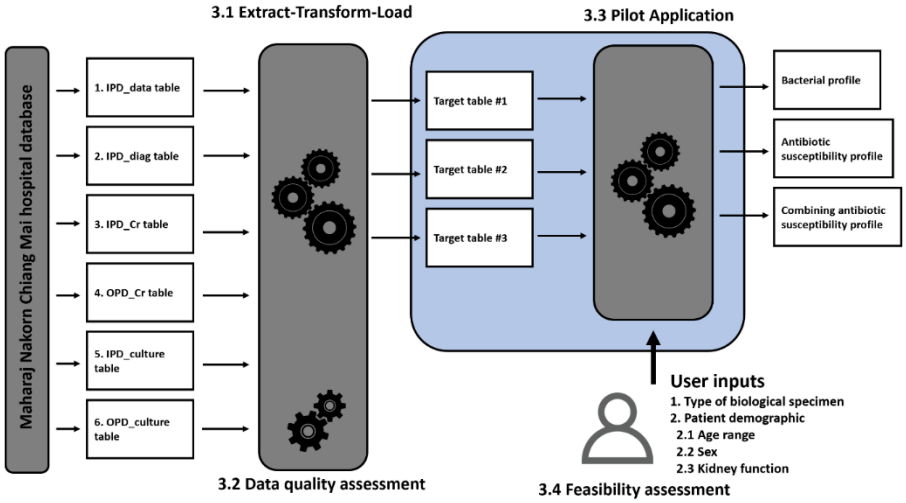


Figure 1: High-level overview methodology of the project

3.1 Data extraction, transformation, and loading

Data was extracted from Maharaj Nakorn Chiang Mai hospital database using an automated software developed by Information Technology (IT) team of Maharaj Nakorn Chiang Mai hospital. All identifiable information was encrypted.

Data structure of each table will be explored. Numbers of rows and columns were counted. All features were statistically described. Numerical data, means and standard errors were calculated. Distributions of the numerical data were presented in histograms. For categorical data, frequencies of the categories were reported. For data regarding date and time, date was aggregated into months then presented as histogram, time was aggregated into hours and presented as histogram. For text data, the pattern of the data was described. Missing data was counted.

[15] was used as a guidance for the development of ETL process to transform all the extracted tables into necessary target tables.

3.2 Data quality assessment

Framework presented in "DATA QUALITY ASSESSMENT HANDBOOK" was used for performing data quality assessment [16]. The 6 dimensions will be assessed including accuracy, reliability, consistency, completeness, relevance, and accessibility. Timeliness dimension was not assessable since we did not have access to the real database.

3.2.1. Accuracy

Accuracy assessment was done by comparing data from the extracted tables and actual data in the hospital information system (HIS). The data was large that was impractical to compare them all manually. Subset of data was sampled for manual comparison. five hospital numbers were randomly chosen and re-identification for manual comparison.

3.2.2. Reliability

Reliability assessment was done by unstructured interview with people responsible for entry data into the hospital database. This would include people who involved in the patient registration processes and lab technicians.

3.2.3. Consistency

All features will be explored for any inconsistency format/pattern. The finding will be described qualitatively.

3.2.4. Completeness

All features will be counted for missing data or incompleteness. The finding will be described with simple descriptive statistics.

3.2.5. Relevance

Relevance assessment will be assessed. Availability of necessary data to develop ETL process and pilot application from the extracted data will be explored. This also includes doctor interviews whether there should be any additional information provided to the end users. The interviewing process is included in feasibility assessment (3.4) of this project.

3.2.6. Accessibility

Accessibility assessment will not be systematically assessed, instead it will be described qualitatively by the principal investigator of this project.

3.3 Pilot application

Pilot application was done in Google Collaboratory [17]. All target tables were uploaded to google drive that connect to the same Google Collaboratory. All target tables were used to create 3 summary tables called bacterial profile, antibiotic susceptibility profile, and combining antibiotic susceptibility profile.

3.3.1 Bacterial profile shows the prevalent distribution of isolated bacteria

3.3.2 Antibiotic susceptibility profile shows how a bacterium is susceptible to each antibiotic agent

3.3.3 Combining antibiotic susceptibility profile shows overall coverage of each antibiotic agent

All summary tables were filterable by the users – users could give specific characteristics of subpopulation they interested in, and all the summary tables would recalculate based on the characteristics given. Characteristics that could be specified included age, sex, and kidney function at admission. Type of specimen that the bacteria was isolated could also be given to the application.

3.4 Feasibility assessment of the pilot application

Total of 6 physicians were conveniently sampled from all the physicians working in Maharaj Nakorn Chiang Mai hospital. All participating physicians were asked for their consent prior to the assessment, and they are free to leave the assessment process at any time. Consented physicians were given an URL link to Google Collaboratory with instructions **(Appendix 1: Instruction on how to use the pilot application).** They were free to use the application whenever they like for 2-3 weeks, then they were appointed for semi-structured interview and provide any feedback. The guide for interview questions in shown in **Appendix 2: Interview guide.** All information gathered will be combined together and described qualitatively.

3.5 Ethical considerations

This study received ethical approval from Chiang Mai University ethical committee, Chiang Mai, Thailand (FAM-2565-08889).

4 Results

4.1 Data extraction, transformation, and loading.

Total of 6 relational tables were extracted from Maharaj Nakorn Chiang Mai hospital database as follows:

Table 3: Summary of numbers of rows and columns of all extracted tables

	Name of the table	Description	Number of rows	Number of columns
1	IPD_data table	Contains information regarding patients who were admitted in the hospital during 2017 to 2018	92,436	25
2	IPD_diag table	Contains information regarding diagnosis of patients who were admitted in the hospital during 2017 to 2018	307,454	5
3	IPD_Cr table	Contains information regarding the creatinine level of all patients admitted during 2017 and 2018. Creatinine is a biomarker that reflects performance of kidney function	199,425	8
4	OPD_Cr table	Contains information regarding the blood creatinine level of all patients during outpatient department (OPD) visit	348,169	7
5	IPD_culture table	Contains information regarding all the bacterial culture results of all patients admitted to Maharaj Nakorn Chiang Mai hospital between 2017 and 2018	557,798	16
6	OPD_culture table	Contains information regarding all the bacterial culture results of all patients admitted to Maharaj Nakorn Chiang Mai hospital between 2017 and 2018	215,257	15

All descriptive information on all the extracted tables is shown in the full report of this study **(Supplement material 1).**

All extracted tables were transformed to one target table then uploaded to google drive that linked with the pilot application Google Collaboratory.

4.2 Data quality assessment

4.2.1. Accuracy

Table 4: Manual comparisons of the 5 sampled patients

Table's name	Number of data points identified in the extracted data	Number of data points identified in the hospital electronic medical record	Number of matched result	Accuracy rate (%)
IPD_data	132	120	120	100.00
IPD_diag	66	66	66	100.00
IPD_Cr	90	90	90	100.00
OPD_Cr	12	12	12	100.00
IPD_culture	1,204	1,204	1,204	100.00
OPD_culture	14	14	14	100.00

Overall accuracy of the extracted data based-on sampling of 5 patients for manual comparison is 100.00%.

4.2.2. Reliability

All patient's demographic information in Maharaj Nakorn Chiang Mai medical records is from governmental documents/information.

Bacterial culture results and Creatinine level results – According to a Maharaj Nakorn Chiang Mai hospital laboratory technician, both laboratory investigations had passed the International Organization for Standardization (ISO) 15189: Medical laboratories since 2006. ISO 15189 specifies requirements for quality and competence in medical laboratories [18].

All date and time records were electronically documented into the hospital electronic system.

4.2.3. Consistency

Information stored in each feature was consistent. Date and time information was store in the same format in all the extracted tables. Creatinine level was stored with milligram/deciliter (mg/dl) as value unit. All information store in **"IPD_Cr table"** were in the same pattern as **"OPD_Cr table".** All information store in **"IPD_culture table"** were in the same pattern as **"OPD_culture table".**

Detail information regarding the consistency is in full report of this study **(Supplement material 1).**

4.2.4. Completeness

In **"IPD_data table",** 18 out of 25 features had no missing value. Five features had less than 1% missing rate. Two features contain more than 30% missing rate. Features with missing data were not utilized by the pilot application

In **"IPD_diag table"**, there was no missing data.

In **"IPD_Cr table"**, there was no missing data.

In **"OPD_Cr table"**, there was no missing data.

In **"IPD_culture table"**, 7 out of 16 features had no missing value. The rest feature had more than 25% missing rate. There was one feature contain more than 99% missing rate.

In **"OPD_culture table"**, 5 out of 15 features had no missing value. There was one feature with less than 1% missing rate. The rest feature had more than 25% missing rate. There was one feature contain 100% missing rate.

Detail information regarding the completeness is in full report of this study **(Supplement material 1).**

4.2.5. Relevance

All necessary information required to develop the pilot application are included in the extracted data. Information regarding admitting medical ward was included in the **"IPD_data table"** but was not utilized, however, infectious doctors commented that this information is very important and should be included in the pilot application as bacterial profiles and antibiotics susceptibility profiles may be different from a medical ward to the others.

4.2.6. Accessibility

Data in Maharaj Nakorn Chiang Mai hospital's database was not open to public access. To access the data, one of the following conditions must be met:

1. The data was requested by health personnel working in the hospital with the goal to utilize the data to improve health-related service(s) of Maharaj Nakorn Chiang Mai hospital.

2. Receive ethic approval from the Maharaj Nakorn Chiang Mai hospital ethical committee to extract the requested data from the hospital database.

All data extraction must be done through Maharaj Nakorn Chiang Mai hospital's IT personnel.

4.3 Pilot application

The application was created in Google Collaboratory within the same Google drive as the target table. Sample images of the application are shown in **Appendix 3: Sample images of the pilot application.**

4.4 Feasibility assessment of the pilot application

The application was stored in the project's Google drive. Instruction on how to use the application is shown in **Appendix 1: Instruction on how to use the pilot application.**

Total of 7 Doctors were invited to participate in feasibility assessment, only 6 agree to participate. Of all the 6 doctors, there are 2 internists with 1.5 years of clinical experience; 2 family doctors with 4 and 5 years of clinical experience; and 2 infectious doctors with 6 and 14 years of clinical experience.

Table 5: Summary of doctor's role and years of clinical experience

Doctor No	Role	Years of clinical experience
1	Internist	1.5
2	Internist	1.5
3	Family doctor	4
4	Family doctor	5
5	Infectious doctor	6
6	Infectious doctor	14

The summary of interview results is shown below

Table 6: Summary of comments, feedbacks, and opinion from internists and family doctors

	Doctor 1 (Internist)	**Doctor 2 (Internist)**	**Doctor 3 (Family doctor)**	**Doctor 4 (Family doctor)**
1. Clinical experience (year)	1.5	1.5	4	5
2. Do the results shown in the application similar to what you experience in clinical practice? **If there is something different, could you specify?**	-	Similar to clinical experience	Similar to clinical experience	Similar to clinical experience
3. Do you think the application would be any useful in your clinical practice? **Please provide reason(s)**	The application would be useful. It would help doctor deciding on which empirical antibiotic agent should be given to a patient	The application would be useful. It provides an organized framework to present data	The application would be useful. It helps processing data from the laboratory department	The application would be useful. It provides information on which empirical antibiotic agent to use. However, there is a suggestion to do internal validation with other source of data or do external validation with other similar hospital.

4. Do you think the application at the current state is enough to be implemented into clinical practice? **Please provide reason(s)**	No. The source of data should include outpatient department (OPD) as well. The application should be easier to use – minimize number of steps required to run the application	Not yet.	May be. However, the application should be more user friendly	May be. Information on the internal and/or external validation may require
5. How should the application be improved?	The application should utilize data from OPD. The application should be accessible via mobile devices and may not require Gmail log-in	Modify input panel to make it easier to fill-in	Minimize the steps required to access the application	The interface should be improved to make it more intuitive to use without the need for user manual.
6. Other comments	-	-	-	-

Table 7: Summary of comments, feedbacks, and opinion from infectious doctors

	Doctor 5 **(Infectious doctor)**	**Doctor 6** **(Infectious doctor)**
1. Clinical experience (year)	6	14
2. Do the results shown in the application similar to what you experience in clinical practice? **If there is something different, could you specify?**	Similar to routine practice	-
3. Do you think the application would be any useful in your clinical practice? **Please provide reason(s)**	It is useful in assisting with the decision-making process for empirical antibiotic agents	It is useful in assisting with the decision-making process for empirical antibiotic agents while waiting for the bacterial culture result
4. Do you think the application at the current state is enough to be implemented into clinical practice? **Please provide reason(s)**	May need to make the application more user friendly	Must be careful when interpreting the information shown by the application as some antibiotic agents are rarely used in clinical practice but are relatively frequently presented.
5. How should the application be improved?	1. Medical ward in which the patient admitted should be add in the input features 2. The result calculations should only include patient that actually have infection. In this case, the application include all data without know whether those data is actually from patient with infection. 3. If possible, the application may suggest antibiotic agent and it dosage to the user	1. Medical ward should be added into the input features 2. Label prescription indication for each antibiotics agent so that these information can be used to filter out irrelevant suggestions 3. Include Gram stain in the input feature might also be useful
6. Other comments	-	Suggest consulting with laboratory personnel before real-world implementation

5 Discussion

Data extraction was done through Maharaj Nakorn Chiang Mai hospital IT personnel consultation. The data structure of the extracted tables was derived from several discussion sessions with the IT personnel. The main reason for the data to be separated into multiple relational tables was that the hospital health information system (HIS) was primary created to support hospital health care services. As each aspect of the health services required different storing data format, thus, resulted in multiple relational tables.

Overall, the quality of the extracted data was relatively good for providing summary profiles of bacteria isolated within the hospital. However, from the assessment, there were some points of concern.

Some extracted tables contained relatively high proportion of missing data. Upon manual exploration from the accuracy domain, missing value in the extracted data was matched with empty value from the hospital's database. True missing value in the context of bacterial culture data was difficult to determine, as missing values may result from not having bacterial growth, or there was bacterial growth but laboratory technician decide not to perform antibiotic susceptibility test due to various reasons. This might indicate an improvement point in database design.

The way the HIS recorded type of specimen was text-based in the majority of the bacterial culture data, making classification of type of specimen problematic. Method for classification the type of specimen presented in this project was developed from exhaustive trial and error process, the process could only classify approximately 85% of the entire dataset. Improving data encodings for type of specimen would benefit uses of this aggregated data. Nevertheless, in the perspective of providing health care services, the quality of the data was excellent for the tasks.

All functions of the application on Google Collaboratory seemed to work as intended. The platform was stable; allow easy access and sharing. The platform was suitable for the pilot project. However, for actual implementation of the application. The platform should be carefully considered.

The "**Combining antibiotic susceptibility profile**" was discouraged by infectious doctor as the summary data shown in the table deviated clinically significantly from expected result.

Tables shown from the application were built using the previous table data. The **"antibiotic susceptibility profile"** (2nd and 3rd table) were built from **"Bacterial profile"** (1st table) and the **"Combining antibiotic susceptibility profile"** was the built from **"antibiotic susceptibility profile"**. At each step of data selection and aggregation hided biases within. The **"Combining antibiotic susceptibility profile"** would suffer from biases the most, to the point where infectious doctor discourage using it clinically.

Generally, the application received positive reviews. Non-infectious doctors commented that the summarized data provided by the application match with their knowledge and experience, and their feedbacks seemed to be focused on optimizing user experience. While infectious doctors really discuss down to the mechanism of

data collection and generation, and pointed out many potential for biases and provided suggestions to counter these biases, which is discussed further in **Limitation**.

The application had filtering function that allowed users to specified populational characteristics of interest. All interviewed doctors agreed that the function is useful, however, which health characteristics should be used for the filter may need further discussions.

Both infectious doctors suggested adding medical ward in which patients were admitted. This feature, according to both doctors, is very important for determining which initial antibiotics to use. As each medical ward admitted patients with primarily different conditions (e.g. orthopedics ward would admitted patients with primarily musculoskeletal problem, while neurosurgery ward would admitted patients with primarily structural brain problem.) resulting in different infectious disease distribution, thus, different bacterial profiles. Medical ward in which patients were admitted can be found in table "IPD_data table". They contain information regarding the place where patients were admitted.

There have been other organizations as well that publish bacterial profiles and antibiotics susceptibility profiles. National Antimicrobial Resistant Surveillance Center, Thailand (NARST) have been collecting samples and regularly publish these profiles in the national level for microbial surveillance and monitoring purposes, while World Health Organization (WHO) have been doing the same in global level. The organizations also published their methodology for samples collection and susceptibility testing which may be a more appropriate approach compared to the method used in this project. This issue had also been raised during the interview with the infectious doctors. However, they did not totally agree with the methods used by NARST or WHO as they thought that the methods were designed for surveillance and monitoring purposes but not for supporting frontline clinical decision making. Further discussions are needed.

Limitation

The application treated each biological sample record as if they were independent from one other, which is not true in real-life situation. A dozen of biological samples could be originated from the same person over the course of an admission. This caused the results to deviate toward group of people that biological samples had been collected multiple times and not representing the true information. This is a problem of repeated measures and, according to both infectious doctors, is difficult to dealt with. To minimize this problem in future work, algorithms on biological sample record selection must be created, with inputs from infectious doctor and laboratory technician.

Results in the addition table (the 4th table) was also subject to biases. One main problem was the record selection as mentioned in previously. Another problem was that for a type of biological specimen, there could be multiple different combinations of antibiotic agents that were tested against for the susceptibility test. This was evident from the result table "Antibiotic susceptibility profile, count (3rd table)" that for a bacterial name, the number of counts in the denominators were different across different antibiotic agents. This means that for a same type of biological specimen, in some specimen, there were some information hidden (not tested). This leads to an unreliable estimate of coverage percentages. One way to fix the problem is to test the same

combination of antibiotic agents in all biological specimens. Another way is to understand why laboratory technicians test different combinations of antibiotic agents in the first place, so that counter measures could be included in the application. Or, at least, inform users of this biases so that they would be caution when interpreting the result table.

This project only explored bacteria-related data store in Maharaj Nakorn Chiang Mai hospital. Any result shown here are specific to the local context. Generalizing any of the findings need to be cautioned.

Considerations for future work

For a successful project implementation, a multidisciplinary team needs to be assembled. The team should at least include one health provider and one engineer. The health provider should have experience in clinical practice and has some knowledge regarding database management. The engineer should have experience in data warehouse set up. Multiple sessions of expert consultations should be done, e.g., hospital's stakeholders consultation for resources support; hospital's IT personnel consultation for ETL pipeline set up, preferably setting up in the hospital database server so that IT personnel can help with system maintenance; infectious disease and laboratory technician consultation for ideas on how to best utilize bacterial data. All targeted users of the application should be involved in the testing phase of the implementation to optimized user experience.

Conclusion

This pilot project demonstrated that ETL pipeline and bacterial profile visualization application can be done. All key components and problems with potential solutions were addressed. Data quality of Maharaj Nakorn Chiang Mai hospital was in good quality. Feasibility assessment of the application from interviewing doctors showed potential for implementation in clinical practice. Filtering function of the application was said to be beneficial. Hospital's stakeholder consultation, and expert consultation are needed to ensure a successful implementation and maintenance of the pipeline and application in clinical practice.

References

1. "What is Health Data." IGI Global. https://www.igi-global.com/dictionary/health-data/42215 (accessed 23 June, 2022).
2. ""health data"," in *McGraw-Hill Concise Dictionary of Modern Medicine*, ed: The McGraw-Hill Companies, Inc.
3. K. Bookman *et al*., "Embedded Clinical Decision Support in Electronic Health Record Decreases Use of High-cost Imaging in the Emergency Department: EmbED study," (in eng), ***Acad Emerg Med***, vol. 24, no. 7, pp. 839-845, Jul 2017, doi: 10.1111/acem.13195.
4. M. S. Patel et al., "Using Active Choice Within the Electronic Health Record to Increase Influenza Vaccination Rates," (in eng), *J Gen Intern Med*, vol. 32, no. 7, pp. 790-795, Jul 2017, doi: 10.1007/s11606-017-4046-6.
5. J. R. Lakin, E. Isaacs, E. Sullivan, H. A. Harris, R. D. McMahan, and R. L. Sudore, "Emergency Physicians' Experience with Advance Care Planning Documentation in the Electronic

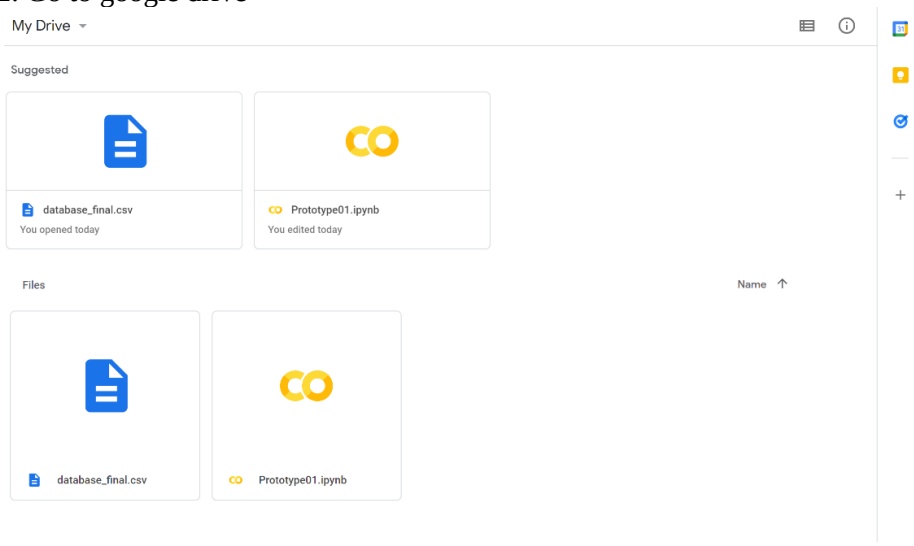
Medical Record: Useful, Needed, and Elusive," (in eng), *J Palliat Med*, vol. 19, no. 6, pp. 632-8, Jun 2016, doi: 10.1089/jpm.2015.0486.
6. J. P. DeShazo and M. A. Hoffman, "A comparison of a multistate inpatient EHR database to the HCUP Nationwide Inpatient Sample," (in eng), *BMC Health Serv Res*, vol. 15, p. 384, Sep 15 2015, doi: 10.1186/s12913-015-1025-7.
7. E. Aref-Eshghi et al., "Identification of Dyslipidemic Patients Attending Primary Care Clinics Using Electronic Medical Record (EMR) Data from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) Database," (in eng), *J Med Syst*, vol. 41, no. 3, p. 45, Mar 2017, doi: 10.1007/s10916-017-0694-7.
8. World Health Organization, "Antimicrobial Resistance: Global Report on Surveillance," World Health Organization, France, 2014. [Online]. Available: https://apps.who.int/iris/handle/10665/112642
9. R. Molla, M. Tiruneh, W. Abebe, and F. Moges, "Bacterial profile and antimicrobial susceptibility patterns in chronic suppurative otitis media at the University of Gondar Comprehensive Specialized Hospital, Northwest Ethiopia," (in eng), *BMC Res Notes*, vol. 12, no. 1, p. 414, Jul 15 2019, doi: 10.1186/s13104-019-4452-4.
10. World Health Organization, "Global action plan on antimicrobial resistance," Geneva, 2015. [Online]. Available: https://apps.who.int/iris/handle/10665/193736
11. E. Castro-Sánchez, L. S. Moore, F. Husson, and A. H. Holmes, "What are the factors driving antimicrobial resistance? Perspectives from a public event in London, England," (in eng), *BMC Infect Dis*, vol. 16, no. 1, p. 465, Sep 2 2016, doi: 10.1186/s12879-016-1810-x.
12. H. Nikaido, "Multidrug resistance in bacteria," (in eng), *Annu Rev Biochem*, vol. 78, pp. 119-46, 2009, doi: 10.1146/annurev.biochem.78.082907.145923.
13. M. Serra-Burriel *et al.*, "Impact of multi-drug resistant bacteria on economic and clinical outcomes of healthcare-associated infections in adults: Systematic review and meta-analysis," (in eng), *PLoS One*, vol. 15, no. 1, p. e0227139, 2020, doi: 10.1371/journal.pone.0227139.
14. IBM Cloud Education. "ETL (Extract, Transform, Load)." IBM. https://www.ibm.com/cloud/learn/etl (accessed 6 March, 2022).
15. R. Kimball and M. Ross, *The data warehouse toolkit : the definitive guide to dimensional modeling*, Third edition ed. Indianapolis, IN: John Wiley & Sons, Inc., 2013, pp. xxxiv, 564 pages.
16. "DATA QUALITY ASSESSMENT HANDBOOK." [Online]. Available: https://www.ungm.org/UNUser/Documents/DownloadPublicDocument?docId=945103
17. "Google Colaboratory: Frequently Asked Questions." Google. https://research.google.com/colaboratory/faq.html (accessed 6 March, 2022).
18. "ISO 15189:2012(en) Medical laboratories — Requirements for quality and competence." The International Organization for Standardization. https://www.iso.org/obp/ui/#iso:std:iso:15189:ed-3:v2:en (accessed 13 July, 2022).

Appendix

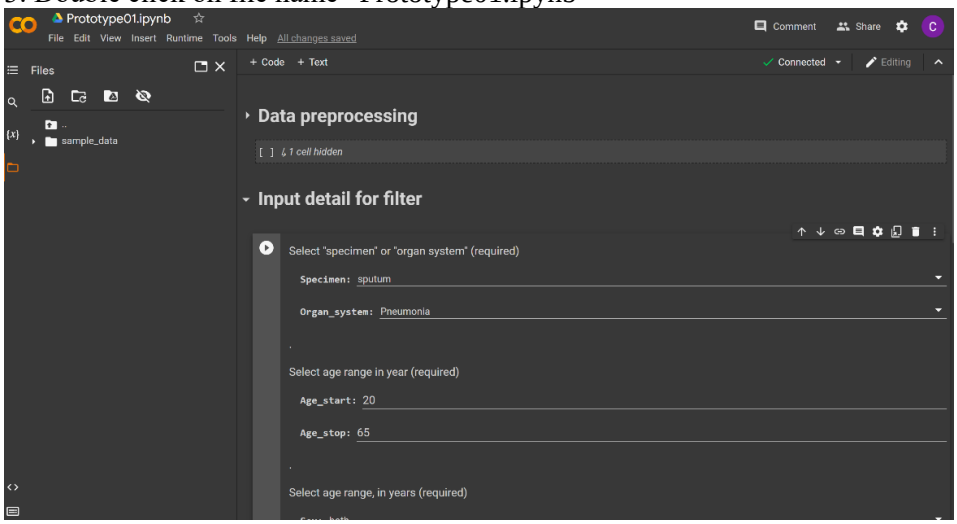
Appendix 1: Instruction on how to use the pilot application

1. Log in to the following g-mail

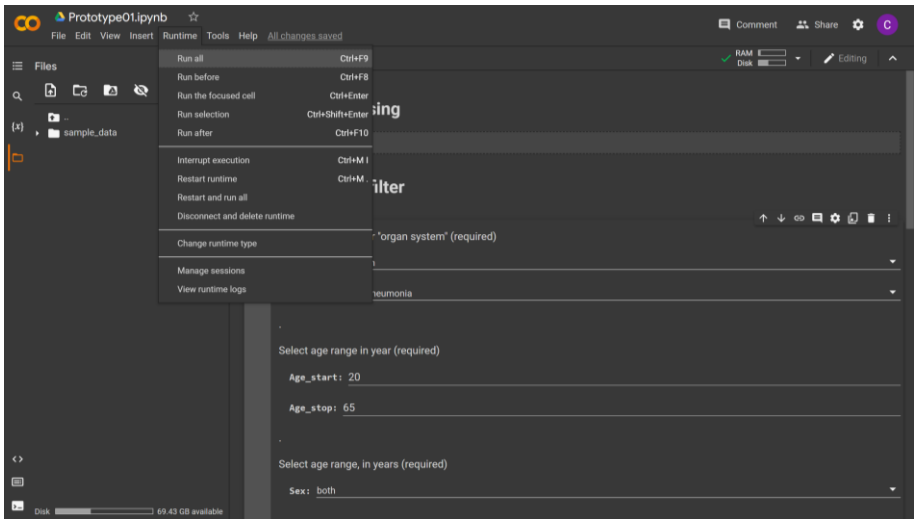
2. Go to google drive



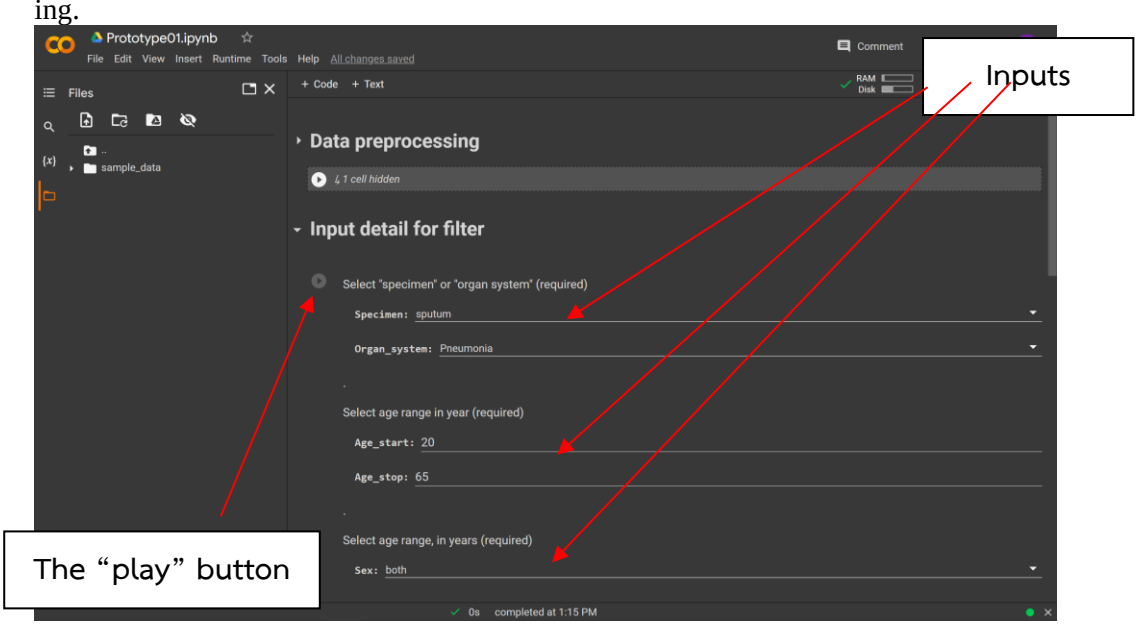
3. Double click on file name “Prototype01.ipynb”



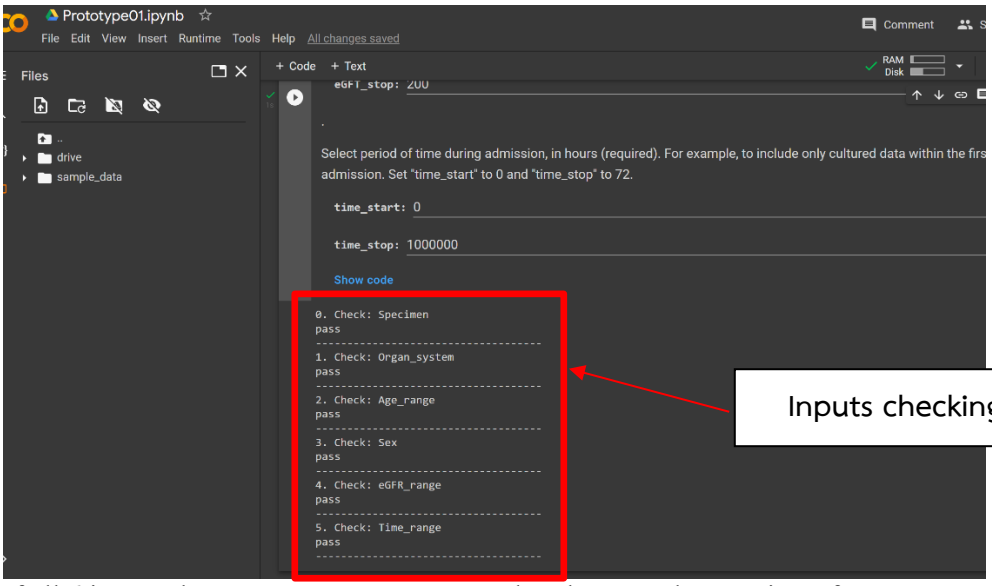
4. Go to Runtime -> Run all. Then wait all processes have been run. This should take approximately no more than 2 minutes. If there is any error appear, please contact me.



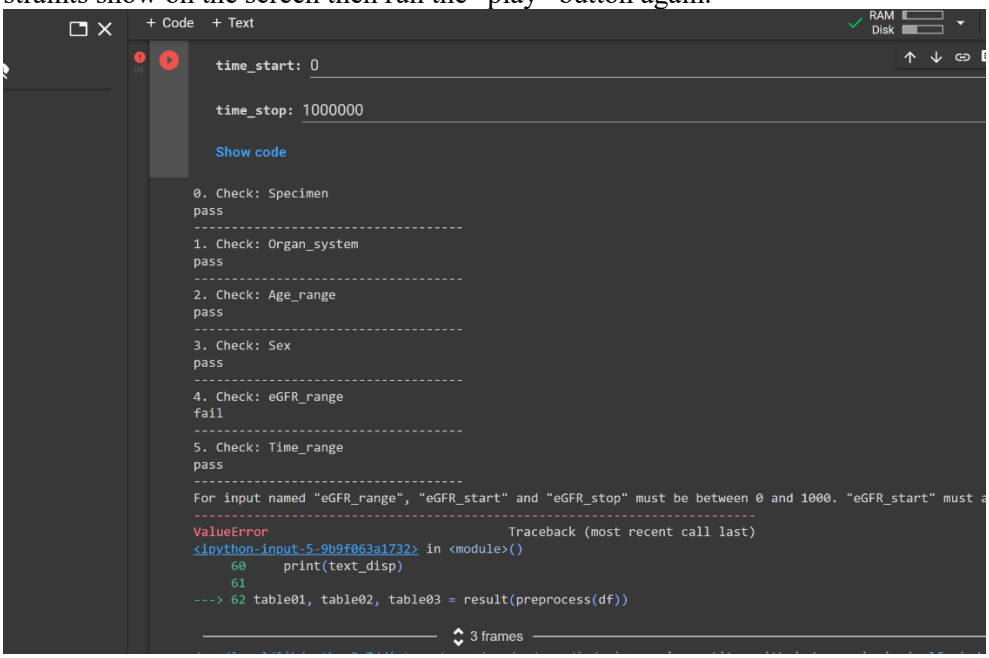
The file is divided into 3 parts – “Data preprocessing”, “Input detail for filter”, and “Results”.
In “Input detail for filter” part, you can try fill in different values. The data will be filtered accordingly to your inputs. After you satisfy with your input values, click on the “play” button just below the “Input detail for filter” heading.



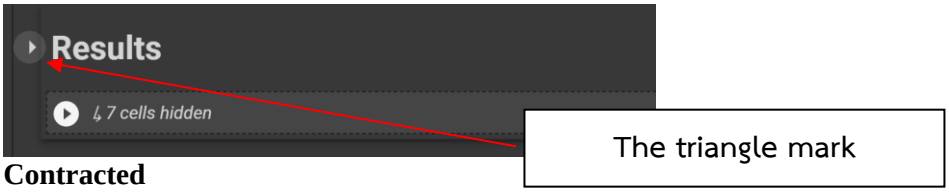
After it finish running, scroll down to check whether the inputs pass all the requirements.



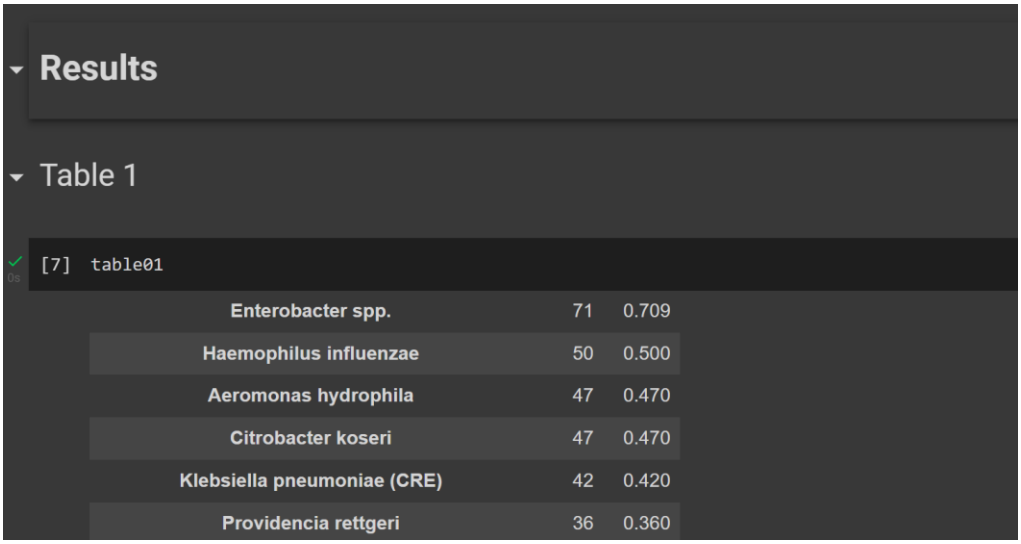
If all 6 inputs show “pass”, you can proceed to the “Results” section. If at least one of the inputs shows “fail”, please change your inputs to match constraints show on the screen then run the “play” button again.



In the “Results” section, you can see the triangle mark in front of the heading, this helps contract and expand items within the “Results” section. Within the “Results” section, there are 4 tables which report bacterial profile of the inputs data (table 1), antibiogram (table 2), and the coverage of each antimicrobial agents (table 3). You must run the “play” button to update tables. You can run them individually when the section is expanded, or run them simultaneously when the section is contracted.



Contracted



Expanded

Interpret results as you please.
You can try change any input values and see how the results change.

Appendix 2: Interview guide
แบบสอบถาม

1. ท่านเป็นแพทย์สาขาอะไร?

☐ General doctor ☐ Internal medicine doctor ☐ Infectious disease doctor

☐ Other specialty โปรดระบุ _____

2. ท่านมีประสบการณ์ดูแลผู้ป่วยในฐานะแพทย์มาแล้วกี่ปี?

3. ผลที่ได้ในส่วน Results เหมือนหรือแตกต่างจากประสบการณ์ของท่าน? หากมีกรณีที่แตกต่าง ขอท่านโปรดยกตัวอย่างกรณีดังกล่าว?

4. ท่านคิดว่าโปรแกรมนี้จะมีประโยชน์ในเวชปฏิบัติของท่านหรือไม่? โปรดให้เหตุผลประกอบ

5. ท่านคิดว่าโปรแกรมนี้ในขณะนี้น่านำไปใช้ในเวชปฏิบัติของท่านหรือไม่? โปรดให้เหตุผลประกอบ

6. ท่านคิดว่าโปรแกรมนี้ควรพัฒนาให้ดีขึ้นกว่านี้อย่างไร?

7. ข้อเสนอแนะอื่นๆ

Appendix 3: Sample images of the pilot application