Peer Review File

Article Information: https://dx.doi.org/10.21037/tlcr-21-531

Reviewer A

Comment 1:

This is an exciting database that the authors are building - with over 60,000 patients it is potentially one of the largest comprehensive databases available.

However, its breadth and ambition is both its strength and its weakness. Because of the number of variables included, the methods are too brief to allow the reader to understand how they are defined and validated. Especially since many of the variables they cite as data that can be extracted reliably are, by themselves, the subjects of papers, clear validation and explanation is key. For example, detection of recurrence is a topic of multiple papers by itself, and in this paper it only gets one sentence.

Reply 1:

Thank you for your comments. We described the method and results on how to define the data and how to validate it in detail. Primarily, we described DFS, PFS, OS, site of metastasis in the Results.

Changes in the text:

Page 8, line 205–208:

The data of brain metastasis from radiology reports of CT or MRI. ROOT conducted parsing pre-defined terminology or sentence, such as “multiple brain metastasis existed in the cerebellum and frontal lobe.”

However, the false-positive, such as a sentence of “no evidence of brain metastasis,” existence of brain metastasis was distinguished using the various operational definition.

Page 9, line 227–231:

PFS and DFS data were collected using data from EMR and radiographic tests. ROOT captured the progression data in EMR, which the clinician recorded as disease progression and switching chemotherapy regimen. The ROOT also captured the data of radiologic progression. In case of changing chemotherapy regimens due to adverse events, ROOT captured why chemotherapy was changed. The data of objective response of each chemotherapy were extracted from EMR by the clinician’s judgment and radiological assessment.

Page 9, line 237–243:

For other examples of ROOT validation, ROOT, the recurrence data included local and distant recurrence after curative intent surgery in patients with stage I-III NSCLC. From the direct chart review, we validated how ROOT extracts data of recurrence from multiple sources, including EMR and radiologic examination such as CT or MRI. The accuracy of data of the recurrence of ROOT was 95% compared to direct chart review by the medical oncologist. In patients diagnosed with early-stage NSCLC and received curative-intent surgery, we compared the data of DFS and site of recurrence in ROOT and direct-chart review by the medical oncologist were compared.

Comment 2:

It is unclear how cancer patients are initially identified and how the site of the primary cancer is determined. Importantly, it is unclear how primary cancer sites are distinguished from metastatic sites. One of the most exciting potential advances they cite is natural language processing to identify metastatic sites, but they don't show any data on to what the accuracy of metastatic site assessment is, or how to distinguish a metastatic site from a new primary.

Reply 2:

Thank you for your comments. ROOT included all new patients diagnosed with histologically confirmed six-type of cancer in daily clinic every 24 h. The ROOT used the diagnostic code of the International Classification of Disease (ICD) and Korean Standard classification of disease (KCD), the national cancer registry, histologic data, and the clinician’s decision recorded in EMR for the final diagnosis of primary cancer.

Changes in the text:

Page 5, Line 106–112:

ROOT included all new patients diagnosed with histologically confirmed six-type of cancer in daily clinic every 24 h. ROOT used the diagnostic code of the International Classification of Disease (ICD) and Korean Standard classification of disease (KCD), the national cancer registry, histologic data, and clinician’s decision recorded in EMR for the final diagnosis of primary cancer. If new patients with NSCLC visited our clinic, they were updated automatically in ROOT when receiving any test or recording in the EMR system.

Comment 3:

It is also unclear how automatic determination of progression-free survival is calculated. Is it based on clinical progression, radiographic progression, or both? Progression can be subjective, even to a treating physician - some clarity in the definition would be helpful especially if it is being automatically updated.

Reply 3:

In the case of PFS calculation, ROOT captured both the clinical data of EMR and radiologic progression. If patients checked CT or MRI at other institutes, we could not catch the progression data. However, the clinician recorded disease progression and switched the regimen of chemotherapy. As an operational definition, we concluded that the disease progression data, if the clinician recorded as disease progression and switched chemotherapy Regimen. In case of changing chemotherapy due to adverse events, ROOT captured the reason for the switch of chemotherapy in EMR.

Changes in the text:

Page 9, line 227–231:

PFS and DFS data were collected using data of EMR and radiographic tests. ROOT captured the progression data in EMR, which the clinician recorded as disease progression and switching chemotherapy regimen. ROOT also captured data of radiologic progression. In case of changing chemotherapy regimens due to adverse events, ROOT captured why chemotherapy was switched. The data of the objective response of each chemotherapy were extracted from EMR by the clinician’s judgment and radiological assessment.

Comment 4:

“When an item appeared in multiple sources, the priority of each source was determined..." how was the priority of each source determined?

The paper would be greatly strengthened by providing more detail about the unstructured variables from the 700 variables extracted and independently validated. In particular, the examples that the authors give - clinical stage and overall survival - are well represented in cancer registry data. It is not surprising that accuracy of those variables is good. However, the variables from unstructured data - such as location of metastasis from radiology reports - is the most subject to inaccuracy and these are the ones that would be most helpful to see validation for.

Reply 4:

Thank you for your comment.

When we need to determine the priority of multiple sources, we validated it several times for each clinical scenario. In the case of patients with early-stage NSCLC who received curative-intent surgery, at least 5–6 data of cTNM in EMR of each department and date of record. We checked and validated which data had the best accuracy of the final cTNM stage.

Natural language processing was conducted based on search terms to extract the data on metastasis, such as the presence and location from radiology report. For example, radiologists described brain metastasis using terminologies, such as “multiple brain metastasis existed in the cerebellum and frontal lobe.” ROOT performed parsing these sentences and defined them as brain metastasis. The false-positive such as a sentence of “no evidence of brain metastasis,” existence of brain metastasis was differentiated using various operational definitions.

Changes in the text:

Page 11, line 288-291

For the extraction of TNM staging, w should determine the priority of date from different times and multiple sources. In the case of early-stage NSCLC who received curative-intent surgery, there were at least five data of TNM. Therefore, we validated it several times for each clinical scenario, which data of TNM stage in EMR represented the final decisions.

Reviewer B

Comment 1:

This article is interesting and has many advantages in advancing the quest for automated data extraction from large EMR data groups.

The author describes the creation of a specialized set of data warehouses for different types of cancer. A series of criteria were developed for each arm and electronic health records (EHR) and other sources were mined for the data. The resulting data warehouse can be used to study patient groups because most of the relevant data for the patient and cancer has already been extracted and ready for analysis.

Putting together fragmented individual medical record data and building big data more effectively has recently attracted the attention of clinical researchers. I think this research is the result of a pioneering study of this trend.

I consider this paper acceptable for publication. However, I would like to ask the author the following minor questions:

Comment 1:

1) The method by which patients were identified is not described. Is it all patients in your institute?

Reply 1:

Thank you for your comment. We included all patients in our institute who were diagnosed with histologically confirmed six-type of cancer.

Changes in the text:

Page 5, line 106-111

ROOT included all new patients diagnosed with histologically confirmed six-type of cancer in daily clinic every 24 h. ROOT used diagnostic code of the International Classification of Disease (ICD) and Korean Standard Classification of Disease (KCD), the national cancer registry, histologic data, and clinician’s decision recorded in EMR for the final diagnosis of primary cancer. If new patients with NSCLC visited our clinic, they were updated automatically in ROOT when receiving any test or recording in the EMR system. In the system of CDW, patients were de-identified to protect their privacy.

Comment 2:

2) Is the language recognized by natural language processing (NLP) Korean? Or is it English and Korean? Are other languages possible?

Reply 2:

Thank you for your comment. The ROOT was recognized only in Korean and English according to the developed algorithm. However, after modification of the algorithm, other languages were also identified.

Reviewer C

Major points:

Comment 1:

- Unsupported claims. Many previous works were in literature in extracting disease specific data into databases. Maybe they are not called a warehouse but it is not clear why authors think their work is the first and not referencing these existing works, nor including other cancer clinical data warehouse works. It is also hard to justify the system being "real time". The only discussion about real-time I can find in the paper was survival data which is understandably not real time but still very valuable.

Reply 1:

Thank you for your comment. When data were updated in the EMR system during the daily clinic, survival data or other results of the test were also updated in real-time in the clinical data warehouse. If patients receive new palliative chemotherapy, the starting date of further palliative chemotherapy is updated. If we make an algorithm for extracting disease-specific data, we don't need to update data manually. In Korea, we have a national cancer registry and national health insurance. We can catch survival data from the national cancer registry, national health insurance, and our institution using the ROOT system. These data were updated every 24 h. If patients die today, we could catch the date of death as today. All data of the ROOT updated in real-time, including the results of the test and EMR records.

As you commented, we clarified how to update clinical data from EMR and merge various data using various formats and references.

Changes in the text:

Page 5, line 106-110:

. ROOT included all new patients diagnosed with histologically confirmed six-type of cancer in daily clinic every 24 h. ROOT used the diagnostic code of the International Classification of Disease (ICD) and Korean Standard Classification of Disease (KCD), the national cancer registry, histologic data, and the clinician's decision recorded in EMR for the final diagnosis of primary cancer. If new patients with NSCLC visited our clinic, they were updated automatically in ROOT when receiving any test or recording in the EMR system.

Comment 2:

- Unclear description of clinical data warehouse. Without going into details that are too technical, there are many aspects of a clinical data warehouse that readers would want to find out in a report - what are the source systems, what is the ETL process, what are the main software components, how often are new patients added and existing data updated, any data standard used, how are patient privacy protected when this warehouse is queried. These should all be clearly given, ideally in the Methods section. Standard terms should be used.

Reply 2:

Thank you for your comment. As you recommended, we described the source systems, the ETL process, the leading software components in the Methods section. In the system of CDW, patients were de-identified to protect their privacy. If new patients with NSCLC visited our clinic, they were updated in CDW system when receiving any test or recording in the EMR system automatically. We protected patients' privacy using de-identifying method.

Changes in the text:

Page 4, lines 101–102:

Our database management system (DBMS) included Darwin-MED (version 11.6), which was an EMR system, and Darwin-C (version1.0.7404), which was a CDW system at SMC.

Page 5, lines 123–126:

ROOT extracted data from CDW-derived Darwin-Med (original data source) according to the specialized algorithm. ROOT conducted daily extraction using programmed SQL and SAP Data Services Designer

14, a developer tool used to create objects through data mapping, transformation, and logic.

Page 5, line 110-111:

In the system of CDW, patients were de-identified to protect their privacy.

Comment 3:

- Algorithm description. Authors should describe the natural language algorithm developed, for example they may belong to “rule-based” methods. From such algorithms readers would be interested to find out how synonyms and negation (e.g.) are handled. The complete QA results should also be presented as in a scientific study.

Reply 2:

Thank you for your comment. We described the example of “role-based” methods.

Page 8, line 205–208:

The data of brain metastasis from radiology reports of CT or MRI. ROOT conducted parsing pre-defined terminology or sentence such as “multiple brain metastasis existed in the cerebellum and frontal lobe.” However, false-positive, such as a sentence of “no evidence of brain metastasis,” the existence of brain metastasis was differentiated using various operational definitions.

Page 9, line 227–231:

PFS and DFS data were collected using data of EMR and radiographic tests. ROOT captured the progression data in EMR, which the clinician recorded as disease progression and switching chemotherapy regimen. ROOT also captured data of radiologic progression. In case of switching chemotherapy regimens due to adverse events, ROOT captured why chemotherapy was switched. The data of objective response of each chemotherapy were extracted from EMR by the clinician’s judgment and radiological assessment.

Page 9, line 237–243:

For other examples of ROOT validation, the recurrence data included local and distant recurrence after curative-intent surgery in patients with stage I-III NSCLC. By direct chart review, we validated how ROOT extracts recurrence data from multiple sources, including EMR and radiologic examination, such as CT or MRI. The accuracy of data of recurrence of ROOT was 95% compared to direct chart review by the medical oncologist. In patients diagnosed with early-stage NSCLC and received curative-intent surgery, we compared the data of DFS and site of recurrence in ROOT and direct-chart review by the medical oncologist.

Minor point

Comment 4:

- Discuss on commonality and differences of the six cancer types in algorithms construction and final data will be a great addition.

Reply 4:

Thank you for your comment. We described on commonality and differences of the six cancer types in algorithms construction and final data.

Page 10-11, line 280–291:

There was a level I key elements that were common in six-type of cancers. The common elements included TNM age, ECOG performance status, TNM staging system, line of palliative chemotherapy, treatment outcome, and genomic profile from NGS. As the specified data for each cohort of six-type of cancer, clinical data were obtained and defined as levels I, II, III. For example, the cohort of NSCLC, the status of major driver mutations, such as EGFR, ALK, BRAF, and ROS-1 mutation, outcomes of matched targeted therapy, a type of resistant mechanism, and method to detect the resistant mechanism were included in level I key elements. In the cohort of head and neck cancer, EBV and HPV status were included. The Masaoka stage and WHO classification were included in the key elements in patients with thymic epithelial tumors. For the extraction of TNM staging, we must determine the priority of date from different times and multiple sources. In the case of early-stage NSCLC who received curative-intent surgery, there were at least five data of TNM. We validated it several times for each clinical scenario in which data of TNM stage in EMR represented the final decisions.