# 암 연구를 위한 빅데이터 플랫폼의 최근 동향 및 문제점

벨무루간 아레시 발라지, 염성웅, 김경백
전남대학교 전자컴퓨터공학부

# Recent Trends and Issues in Bigdata Platforms for Cancer Research

Velmurugan Arresh Balaji, Sungwung Yeom, Kyungbaek Kim
Dept. Electronics and Computer Engineering, Chonnam National University
E-mail:arreshvnass@gmail.com, yeomsw0421@gmail.com, kyungbaekkim@jnu.ac.kr

## Abstract

The Cancer Patients are feeling better than they have felt in the past. Because of the tremendous changes in the field of oncology, the survival rate of the patients have been increased and the survival period is now longer for many cancers. The toxicity level is getting less for the drugs and treatment and it has been effective now a days. And there is improvement in the supportive care too. Hence, the patients are living a better quality of life with high survival period. Through effective preventive measures and advancing technologies, there is an reduction in some chronic cancers from its incidence (for example, smoking related lung cancer and hepatitis related hepatocellular carcinoma). Indeed, cancer is no longer a death sentence for most, and some declare that it is the new chronic disease. This paper discuss about the recent trends in the big data platform related to cancer like improvement in managing the personal information of the cancer patients, chat-bot or web interfaces, and issues like patients privacy, configuring clinical cancer registries and clinical datasets merging, validating and processing.

## I. INTRODUCTION

In the clinical Oncology, there is an vast improvement happened by the arrival of big data platforms. With the help of these advanced techniques, we can integrate and access various form of data sources from multiple institutions across the world. In the vision of precision medicine, the generic data warehouses holds insufficient data and they are lagging in handling the unstructured data. For researches on precision medicines, a essential platform should be developed, that collect data such as electronic medical records, genomic sequences, tumour biopsy specimens, etc., from the public domain which were independently organized and linked to cancer Registries[1]. a personalized mobile Health based Pulmonary Rehabilitation platform will improve the exercise capacity, dyspne symptoms and quality of life in patients with Non-Small Cell Lung cancer. A platform like this will enhance only when its rehabilitive efficacy is corroborated by evidence based clinical results [2]. An Integrated Pharmacogenomic platform of human cancer cell lines and tissues will help the biologists to investigate the connectivity of small molecules and genomic features related to the cancer cell lines and real cancer tissues [3]. An online Medical Suggestions System for the improvement of patients health by means of web interfaces or Chatbots which are trained based on the deep learning concepts will support in overcoming the limitation of cold interaction between users and the software simulating human behaviour [5]. The Genomic Common Data model will minimizes the privacy issues when conducting multicentre studies by integrating statistical results of the same analysis code rather than sharing the clinical sequencing data directly [6].

## II. RECENT TRENDS IN CANCER TREATMENT

The Korea Cancer Big Data Platform (K-CBP),a National Cancer Control Initiative(NCCI) of the Republic of Korea is an integrated system of four distinct National Cancer Control Activities conducted by the NCCI. It holds the Clinical data from 515,780 cancer patients, NGS data from 280 patients, metadata of blood samples from 32,760 subjects, and metadata of tissue samples from 17,813 subjects were merged on the basis of the alternative key. On the web platform of the K-CBP, a user can query for the clinical registry data, analyze the features of different datasets. The K-CBP attempts to establish and implement basic datasets for cancer-related clinical studies and precision medicine. For Data merging, based on primary key of the dataset, each datasets are matched. The data is trasferred to integrated Data warehouse by means of alternative keys. To avoid affecting of patient's privacy De-identification (decryption of the personal information) is done before importing the data from K-cbp, the direct identifiers of the source data will be replaced by alternative patient keys. The quality of data is also improved through the structuring of unstructured data, creation of datasets, and use of a data validation system for the ETL process.

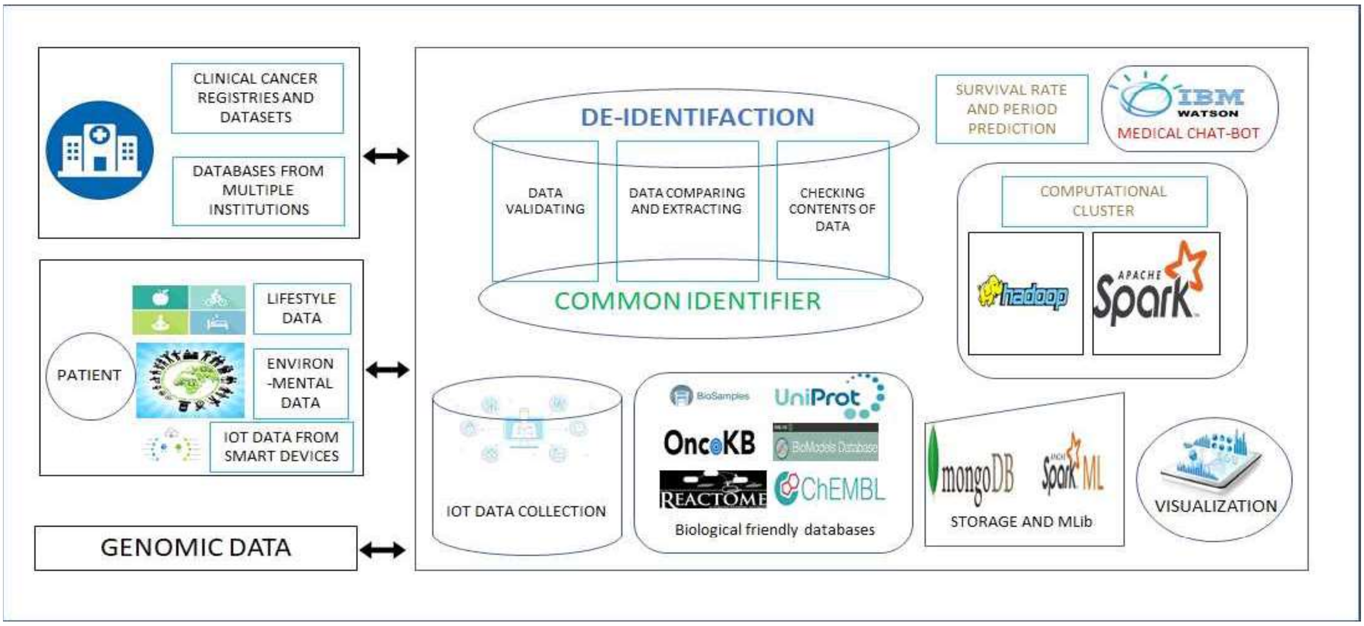
The IPCT (Integrated Pharmacogenomic Platform of Human Cancer Cell Lines and Tissues) is a biological database platform that integrates data about cancer cell lines, small molecules, human pathways, experimental results, and cancer somatic mutations. Th IPCT webportal gives access to research the connection between the data points presented in the CTRP, CCLE, Expression Atlas, REACTOME, and cBio Portal databases in an integrated fashion. The database of IPCT contains nearly 860 cell lines, 481 small molecules, 2500 differential expression studies, 2000 human pathways, and 151 cancer studies. Moreover, the IPCT contains of 8,214,573 unique connections between the different data points. The overall database size is 20 GB. With the help of graph, the connectivity of the given small molecules, cell lines, and genes can be viewed conveniently by the users. The platform enables the researchers to investigate the connectivity of small molecules and genomics features in relationship with cancer cell lines and real cancer tissues. It also highlights the genomic features sensitive to a specific drug and the percentage of cancer patients affected by that drug. Notably, IPCT can also identify cancer cell lines that are truly representative of real cancer tissues.

Today, chat-bots are used in dialogue-based systems for several practical purposes, including customer services or data collection. The most performing bots use sophisticated natural language processing (NLP) techniques and are trained with deep neural networks to better emulate human behaviors. The HOLMes, an Health Online Medical Suggestion Platform which supports eHealth applications using a trained machine learning algorithm, deployed on a cluster-computing framework, which provides medical suggestion via a chat-bot module. A total of 16,733 patients' prevention records have been collected from the clinical teams of each prevention pathway, led by 13 expert physicians, The Spark.ML library provides all the data-preparation functionalities and the machine learning algorithms to train the random forest models using the collected training data. T It is worth noting that the second-level evaluation leads to better results (improving the AUC by about 12%) because it exploits more specific features provided by physicians or obtained via further examination. We want to note that using machine learning in a clinical scenario requires taking into account the explainability of the choices made by the model.

## III. RECENT ISSUES AND DISCUSSION

The Korea Cancer Big Data Platform (K-CBP) is an National Cancer Control Initiative of the Republic of Korea is an integrated system of four distinct National Cancer Control Activities conducted by the NCCI : The Korea Central Cancer Registry, National Cancer Screening Program, Financial Aid Program for Cancer Patients, and Cancer Hospice and Palliative Care Program. The problems associated with this platform are as follows. First, the genomic data in the k-cbp is not sufficient to provide the confidence in the results. Hence, There is a limitation for rare and intracable cancers. Second, There is no Common Identifiers to integrate the source of data from multiple institutions. It is necessary to select a common identifier that covers multiple institutions. Most common identifiers that satisfy the conditions are not free from the laws and regulations imposed on the medical data. Third, For Next Generation Sequencing, the patient's data on lifestyle and environment are essential and very useful for obtaining precision medicine. The access to Environmental and Lifestyle data are usually limited interms of incorporation into the cancer big data platform because of the different clinical data Legal regulations in Korea that pertain to permission to access data, including personal information, should be eased. However, the legislation that concerns the authority to access these data should be considered and specific legal boundaries against the risk of using personal information should be proposed.

The two major limitations of IPCT are to make effective use of all information are the Extensively Heterogeneity of the data and the Lack of data Integration as like the issues of the previous platform. For instance, the European Bio informatics Institute's RDF platform is a state-of-art-example that has enabled the integration of six different biological databases, including UniProt, the Expression Atlas, BioSamples. One of the key questions for any biologist is whether the genomic features of cancer cell lines that are sensitive to drugs are relevant in real cancer tissues. To address this, biologists previously had the search through multiple heterogeneous databases, which is a challenge job even for researchers with advanced computer skills. In Figure(1), the idea has been discussed to solve the above mentioned problems.



FIGURE(1): PROPOSED ARCHITECTURE FOR THE CANCER BIGDATA PLATFORM

## IV. CONCLUSION

The growth of Data in the 21st Century has made the idea of big data platform a reality in the clinical oncology. Many Organizations around the world have already started extracting benefits from there huge pool of data. Despite the benefits, there are still some drawbacks in achieving the precision diagnosis of Non-small Cell Lung Cancer due to insufficient data caused by lagging of integrating multi-layer data sources all across the world due to privacy issues. Moreover, the Clinical registries devoloped by the public domains are not convincing in case of rare and intraceable cancers. If the integration of relational datasets are computed precisely, it will enablethe oncology experts to utilize high-quality base materials necessary for precision medicine.

## Acknowledgements

This research was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF)& funded by the Korean government (MSIT) (NRF-2019M3E5D1A02067961). This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2019-2016-0-00314) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation). This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2017R1A2B4012559).

## References

[1] Hyo Soung Cha, et al, "The Korea Cancer Big Data Platform (K-CBP) for Cancer Research." International Journal of Environmental Research and Public Health 2019 doi:10.3390

[2] Wonjun, et al. "Mobile Health Management Platform – Based.Pulmonory Rehabilitation for Patients With Non-Small Cell Lung Cancer: Prospective Clinical Trial" JMIR Mhealth Uhealth 2019. doi: 10.2196/12645.

[3] Muhammad Shoaib, et al. "IPCT: Integrated Pharmacogenomic Platform of Human Cancer Cell Lines and Tissues." MDPI genes 2019. doi:10.3390/genes10020171

[4]Ying Liu, et al. "A Novel Cloud-Based Framework for the Elderly Healthcare Services Using Digital Twin" 2019 IEEE Access doi:10.1109/ACCESS.2019.2909828

[5] Flora Amato, et al. "HOLMeS: eHealth in the Big Data and Deep Learning Era" MDPI Information,2019.doi:10.33390/info 10020034.

[6] Seo Jeong Shin, et al "Genomic Common Data Model for Seamless Interoperation of Biomedical Data in Clinical Practice: Retrospective Study" Journal of Medical Interest Research 2019 vol. 21, issue 3, el3249,p.1