Cancer Big Data Initiative: A pilot study
- Construction of Korean Cancer Control Statistics Information System -

Hyunsoon Cho, Byung-Woo Kim, Hyun-Joo Kong, Chang-Mo Oh, Kyu-Won Jung, Young-Joo Won*

Cancer Registration and Statistic Branch, Division of Cancer Registration and Surveillance, National Cancer Center, Goyang, Republic of Korea

중앙암등록본부 Korea Central Cancer Registry

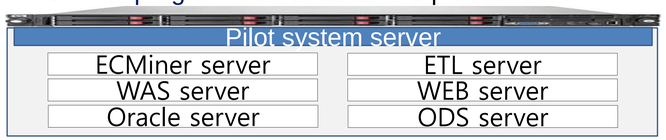
BACKGROUND

- A cancer control program is based on reliable and accurate statistics and evidences.
- According to the environmental change, data release and usages through integration of cancer related data were requested.
  - Increased public demand for cancer related information
  - Need for realizing 'Government 3.0' of which the ultimate goal is to provide customized services for individual citizens.
  - As a part of cancer big data management and utilization effort, we conducted a pilot project.
- Needs for the meaningful new index for cancer control program
  - impossible at its current level but possible to be produced regularly when integrated cancer DB is satisfied

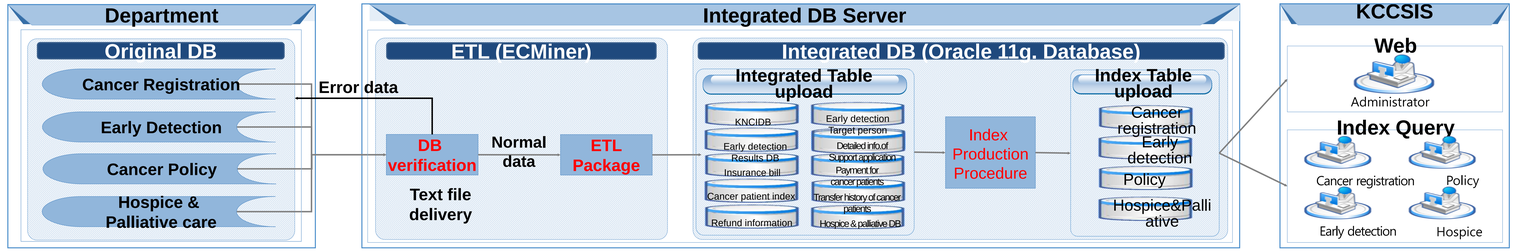
Aim: creating an integrated cancer control database (DB) by combining the cancer management project DBs from four business departments within the National Cancer Control Institute (NCCI), and to establish new index service using the combined DB.

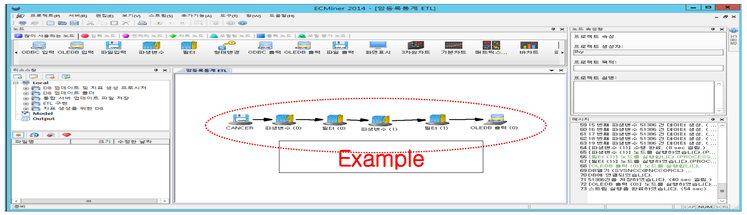
METHODS

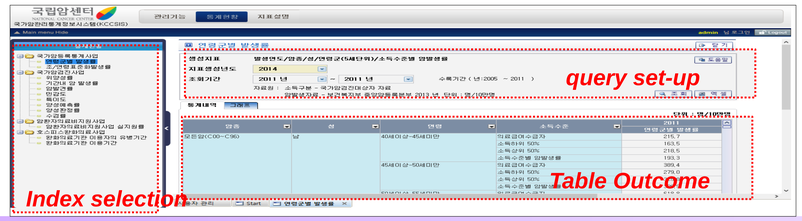
- Fifteen indices were selected for a new service through the integrated DB.
- Standardization of the raw data was essential for data integration, data structures and variable definitions were examined.
- Comprehensive data cleaning was conducted.
- The actual integrated DB was created by an IT company.
- ECminer program was used for the processes.

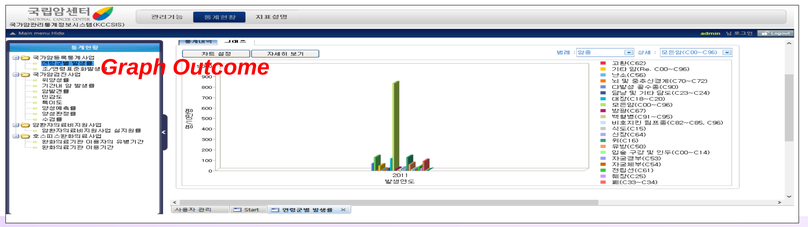


- ETL : Extract Transform Load
- WAS : Web Application Server
- WEB : Web Server
- ODS : Operational Data Store, Temporary store









RESULTS

- Through analysis of the current computing environment, a new server was introduced that considered the DB capacity of integrating data from four sources.
- Method that does not violate the Personal Data Protection Act was used.
- For DB standardization, a standard term dictionary, domain dictionary, code dictionary, and conversion mapping specification were prepared.
- Data was downloaded as txt file format from the four business departments DBs and uploaded through the ETL (Extraction, Transformation, Loading**)** process to the standardized integrated DB.

- Create new indices
  - Age-standardized cancer incidence rates based on income level.
  - False-positive rates, cancer detection rates, positive predictive values, screening rate.
  - The percentage of people eligible for cancer policy support who are actually receiving support.
  - The length of illness from cancer diagnosis to death for individuals who used palliative hospice medical facilities.

DISCUSSION & CONCLUSIONS

- The National Cancer Incidence DB(NCIDB) is useful in planning health services but has limitations to provide more detailed information about cancer patients.
- The KCCSIS will contribute to the advanced national cancer control program through integration and utilization of scattered cancer management information.
- It is the first attempt to integrate cancer control database and develop web-based services in Korea.