GEMINI 2.0 : A Visualizing tool for Data quality between Common Data Model Databases

Chungsoo Kim, PharmD[1], Jun Hyeong Kim[2], Doyeop Kim, BE[1], Seng Chan You, MD, MS[2], Seongwon Lee, PhD[1], Rae Woong Park, MD, PhD[1,2,3]

[1] Dept. of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, South Korea;
[2] Dept. of Biomedical informatics, Ajou University School of Medicine, Suwon, South Korea;
[3] FEEDER-NET(Federated E-health Big Data for Evidence Renovation Network)

Introduction

- Analyzing the characteristics of databases without leakage of institutional information is an important issue in distributed research networks.
- As the OMOP CDM is introduced to various data partners, the heterogeneity of data conversion policies raises concerns. So the need for a program to compare data characteristics and assess data quality through it has emerged.
- We have released General ExaMINation and visualizing application for paired Institutions (GEMINI) 1.0 through the OHDSI Poster in 2017. GEMINI has shown the possibility of being used for data quality management in addition to comparing characteristics without data leakage. We released GEMINI 2.0 by adding the report generation function and applying the improved User Interface.

GEMINI 2.0

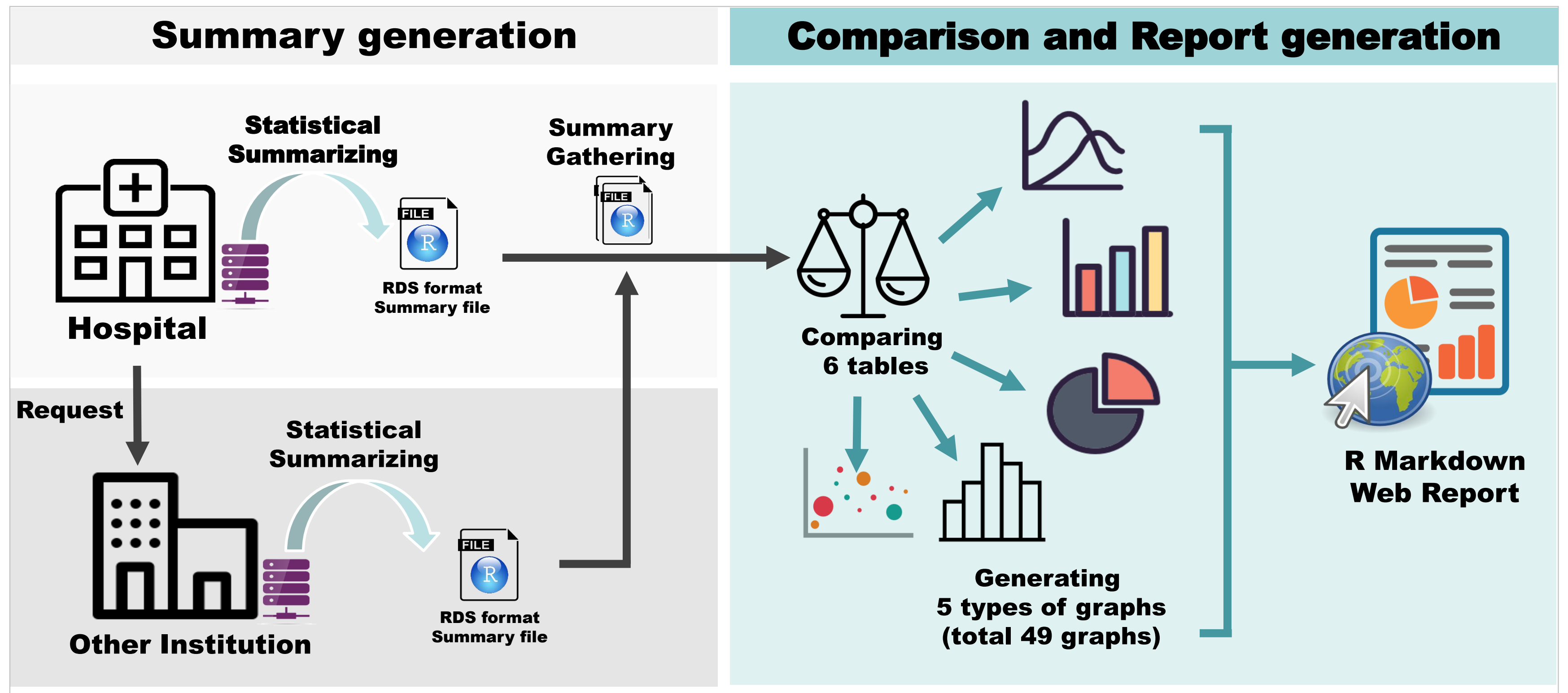

Summary generation | **Comparison and Report generation**

Figure 1. Overall process of GEMINI 2.0

- Developed based on R
- Generates a statistical summary file in R data format
- After gathering the summary file, GEMINI 2.0 generates a web report including 49 comparison graphs (ratio, percentage, frequency) from 6 tables (Person, Death, Visit_Occurrence, Condition occurrence, Drug_exposure and Drug_era).
- All codes are available at : https://github.com/ABMI/GEMINI

Method

- **As a tool for Data Comparison and Visualization**
  - **Databases**
    National Health Insurance Service(NHIS) - National Sample Cohort
    Ajou University School of Medicine (AUSOM)
  - **Study design**
    For data characteristic comparison, we compared NHIS and AUSOM databases.
- **As a tool for Data Quality Management**
  - **Databases**
    National Health Insurance Services(NHIS) – National Sample Cohort
  - **Study design**
    For data quality assessment, we extracted, transformed and loaded(ETL) NHIS to OMOP CDM version 5.3. NHIS ETL version 1(NHIS_Ver.1) was conducted in 2017 and NHIS ETL version 2(NHIS_Ver.2) was conducted in 2018 with corrected standardized vocabulary and ETL rule. we compared NHIS_Ver.1 and NHIS_Ver.2

Result

- **Result for Data comparison and visualization**
  A comparison of NHIS_Ver.1 and AUSOM shows usability of GEMINI 2.0 as a data visualization and comparison tool.
  - Graph (A) : Total records comparison between institutions. AUSOM was higher than NHIS_Ver.1 in Drug exposure (59.6%, 43.6%) and was lower than NHIS_Ver.1 in Condition (14.7%, 26.7%).
  - Graph (B) : Visit concept comparison between institutions. In NHIS_Ver.1, 98.4% were outpatients, while in AUSOM, 90.5% were outpatient. and we found that the proportion of inpatient and emergency room patients was higher in AUSOM. (1.3%,0.3% vs 4.1%,5.4%)

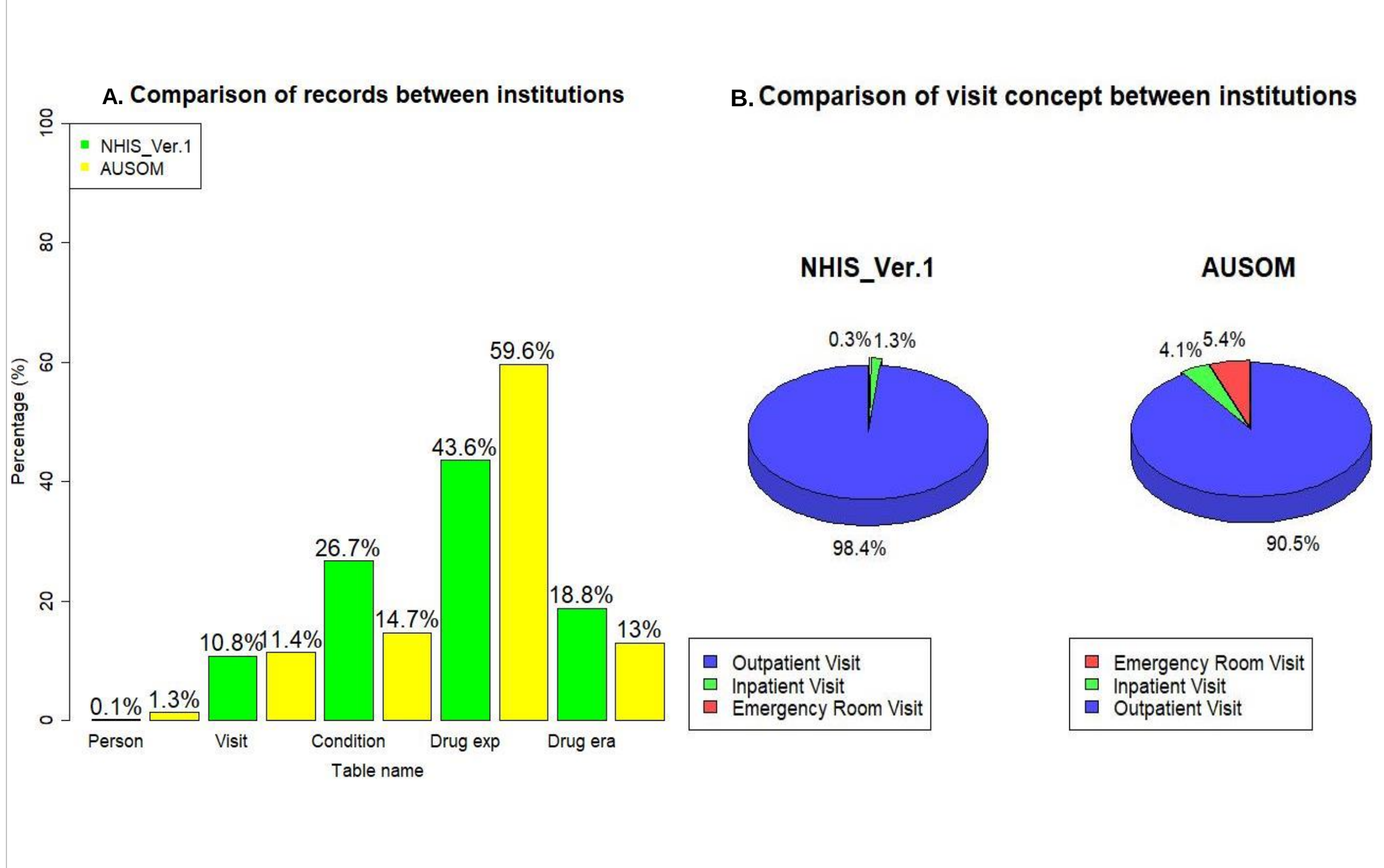


Figure 2. Examples of GEMINI results (NHIS_Ver.1 vs AUSOM)

Graph(A) - NHIS_Ver.1 recorded rates of 0.1% for Person, 10.8% for Visit, 26.7% for Condition, 43.6% for Drug_exposure and 18.8% for Drug_era. In the case of AUSOM, 1.3% for Person, 11.4% for Visit, Condition 14.7%, Drug_exposure 59.6% and Drug_era 13%. Characteristically, for AUSOM, the Drug_exposure rate was found to be higher than that of NHIS_Ver.1. (AUSOM – 59.6%, NHIS_Ver.1 – 43.6%)

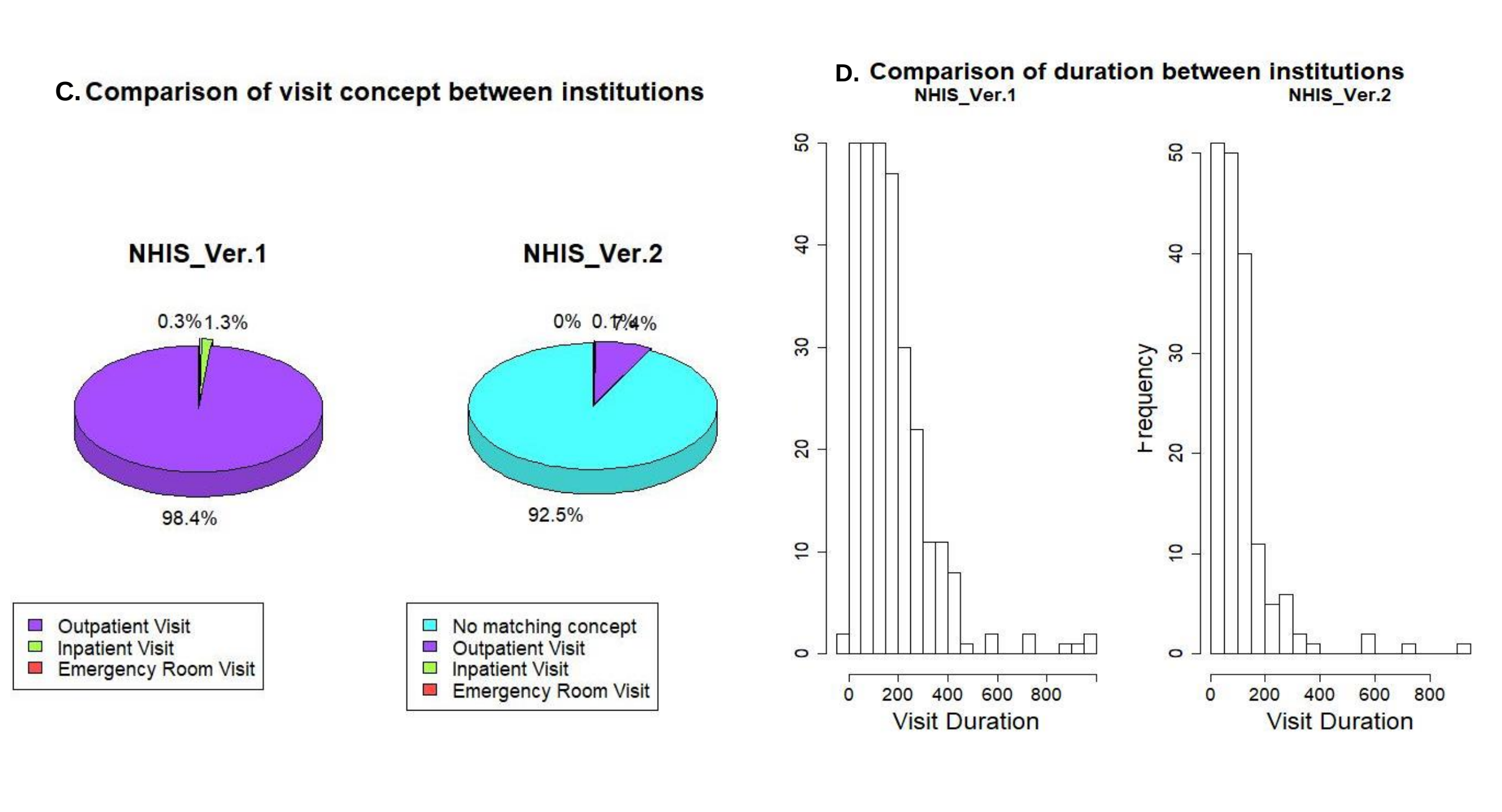
Graph(B) - When comparing the visit concept of the two institutions, there were 3 types(Outpatient, Inpatient, Emergency Room). NHIS_Ver.1 showed rates of 98.4% of Outpatient, 1.3% of Inpatient, 0.3% of Emergency room. The AUSOM data showed a higher ratio in Emergency room(5.4%) and Inpatient(4.1%).

- **Result for Data quality management**
  A comparison of NHIS_Ver.1 and NHIS_Ver.2 shows usability of GEMINI 2.0 as a data quality assessment tool.
  - Graph (C) : Comparison of Visit concept between institutions. In NHIS_Ver.2, 92.5% were no matching concept, 7.4% outpatients, 0.1% inpatients and 0% in emergency rooms. This means that conversion of the visit concept during the data conversion process has not been performed properly.
  - Graph (D) : Comparison of duration between institutions. A negative value was found in the visit duration graph of NHIS_Ver.1. This problem was solved in NHIS_Ver.2.

Figure 3. Examples of GEMINI results (NHIS_Ver.1 vs NHIS_Ver.2)

Graph (C) – In the case of NHIS_Ver.1, 98.4% outpatient, 1.3% inpatient, and 0.3% in emergency room, but 92.5% in NHIS_Ver.2 were no matching concept, 7.4% outpatient, 0.1% inpatient and 0% in emergency room. This means loss of data and also means that the data conversion process should be reviewed.

Graph (D) – The shapes of the two graphs are similar, but negative values were found in NHIS_Ver.1. However, in NHIS_Ver.2, it was confirmed that the negative value disappeared.



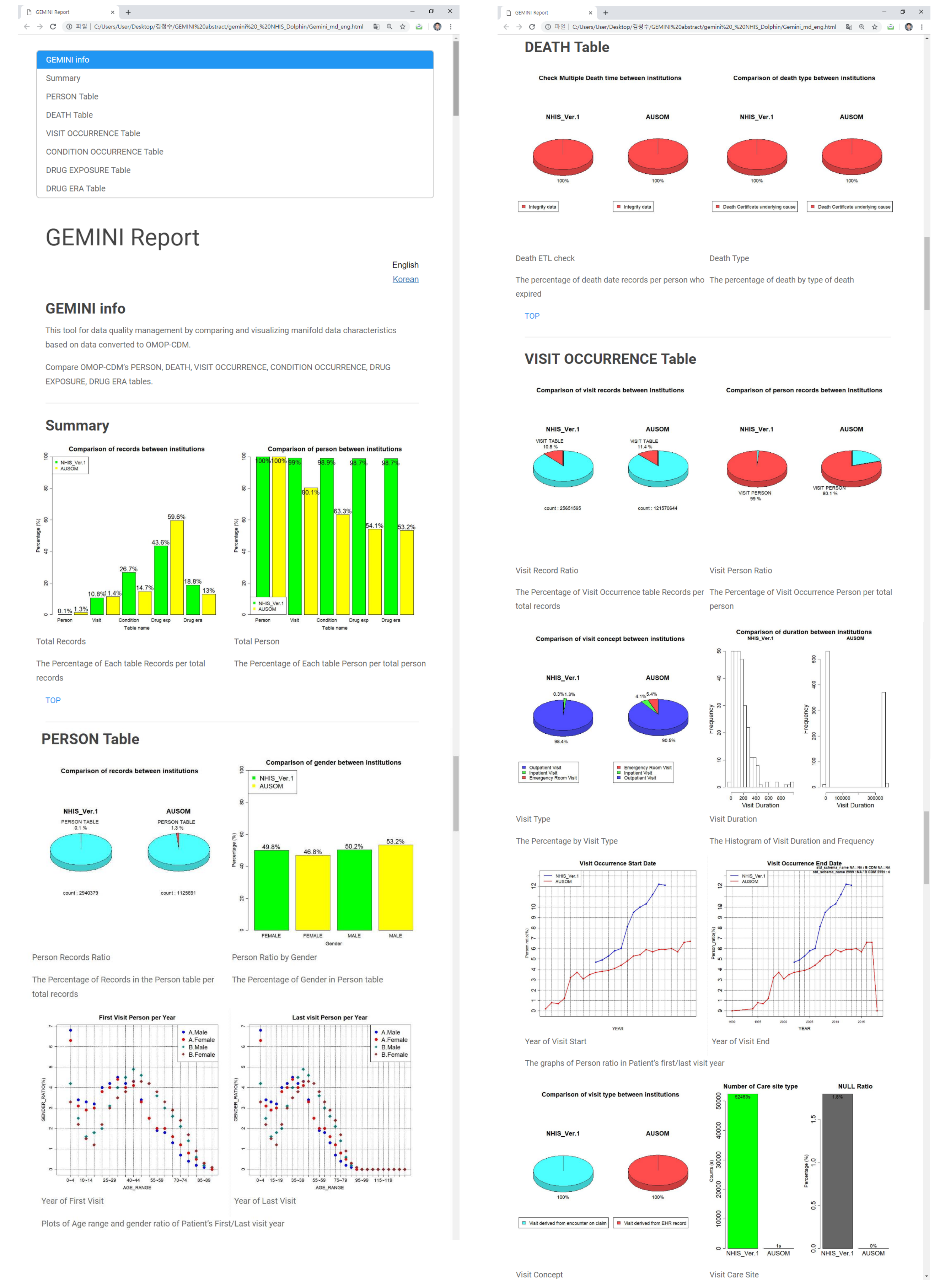
Conclusion

- We developed the OMOP-CDM comparison tool, GEMINI 2.0, which can help researchers intuitively perform qualitative evaluation of their data and compare their data with other institution's data.
- GEMINI 2.0 allows data comparison to identify problems that can occur due to data conversion and to use them for quality management.
- We expect that GEMINI 2.0 helps to maintain consistency of data conversion policy. In the future, we will develop GEMINI 2.0 to enable simultaneous comparisons of multicenter databases.

[Supplement] - GEMINI 2.0 Web Report

GEMINI 2.0 uses R markdown to create webpage reports. The report generates 49 comparative graphs for each of the 6 tables, with summaries comparing the entire data. Statistical comparison values are represented by 5 types of graphs : line graph, pie graph, bar graph, scatter plot and histogram.

Figure 4. GEMINI 2.0 Web Report


GEMINI info
Summary
PERSON Table
DEATH Table
VISIT OCCURRENCE Table
CONDITION OCCURRENCE Table
DRUG EXPOSURE Table
DRUG ERA Table

GEMINI Report

GEMINI info

Summary

PERSON Table

DEATH Table

VISIT OCCURRENCE Table

Acknowledgements : This work was supported by the Bio Industrial Strategic Technology Development Program (20001234) funded By the Ministry of Trade, Industry & Energy (MOTIE, Korea) and a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea [grant number: HI16C0992]. / **Conflict of interest :** none / **Contact** : veritas@ajou.ac.kr