



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

Ph.D. Dissertation of Hyo Jung Kim

# Clinical Genome Data Model towards Precision Medicine

정밀의학을 위한 임상유전체데이터모델

August 2020

Graduate School of Medicine  
Seoul National University  
Interdisciplinary Program of Medical Informatics

Hyo Jung Kim

# Clinical Genome Data Model towards Precision Medicine

Hyo Jung Kim

Submitting a Ph.D. Dissertation of Medical  
Informatics


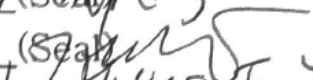

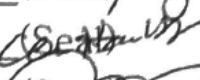

May 2020

Graduate School of Medicine  
Seoul National University  
Interdisciplinary Program of Medical Informatics

Hyo Jung Kim

Confirming the Ph.D. Dissertation written by  
Hyo Jung Kim

June 2020

Chair	<u>Jinwook Choi</u> (Seal) 
Vice Chair	<u>Ju Han Kim</u> (Seal) 
Examiner	<u>Hyung-Sun Yoon</u> (Seal) 
Examiner	<u>Hae-Young Lee</u> (Seal) 
Examiner	<u>Hyunwook Nam</u> (Seal) 

# Abstract

## Clinical Genome Data Model towards Precision Medicine

Hyo Jung Kim

Interdisciplinary Program of Medical Informatics

Graduate School of Medicine

Seoul National University

**Background** The transition to precision medicine and personalized medicine is accelerating owing to progress in genomic technology and the consequent accumulation of genomic information. However, the clinical application of genomic information remains limited, and its spread rate has been slower than expected. This lag has been attributed to complex causes, including 1) a knowledge gap between medical experts and bioinformaticians, 2) separation of the bioinformatics workflow from clinics, and 3) unique characteristics of genomic data. Nevertheless, current informational approaches to link genomic data to clinical fields mostly address the data structure problem.

**Objective** We aimed to develop a genomic data model allowing for more interactive support in clinical decision-making. Informational modeling was



used as a knowledge communication scheme from the highly intellectual product of bioinformatics to a representative data component of a clinical decision.

**Methods** Reliability-related attributes were derived through failure mode and effect analysis (FMEA). This study involved a multidisciplinary working group that conducted clinico-genomic workflow analyses and attributes extraction. Based on these data, an entity-attribute model was then developed through abstraction and normalization.

**Results** The outputs of FMEA were a dataflow snapshot obtained from next-generation sequencing, the information process map extended to the clinico-genomic context, and the set of attributes. Next, an entity-attribute model consisting of eight entities and 49 attributes was identified to develop the final genome data model, including: a linkage identifier to clinical information, experiment-related information, bioinformatics protocol-related information, physical location information, expression, annotation, actor information, and timeline information.

**Conclusion** The proposed genome data model could serve as a data-layer infrastructure supporting the intellectual interplay between medical experts and informative decision-making. Given the importance of recognizing a genome information system as a component of the clinical information system to realize precision medicine, the model could help enhance integration of genomic data in clinical settings.

**Keyword:** Precision medicine, Knowledge translation, Knowledge engineering, Hospital information system integration, Genome data model, Interactive pharmacogenomic clinical decision support

**Student Number:** 2015-30615

# Table of Contents

<b>Abstract .....</b>	<b>i</b>
<b>List of Tables and Figures .....</b>	<b>vii</b>
<b>General Introduction.....</b>	<b>1</b>
<b>Chapter 1. Clinical Genome Data Model: Data Level Integration of Patient Specific Genomic and Clinical Data for Multifaceted Utilization</b>	
<b>1.1. Introduction .....</b>	<b>4</b>
<b>1.2. Purpose of Research.....</b>	<b>9</b>
<b>1.3. Materials and Method .....</b>	<b>1 1</b>
1.3.1. The Production Process of Bringing Genomic Information to Bedside Care .....	1 2
1.3.2. FMEA: An Attribute-Clarified Framework .....	1 3
1.3.3. Logical Data Modeling.....	1 6
1.3.4. Demo Datasets for the real-world data validation.....	1 7
<b>1.4. Results .....</b>	<b>1 8</b>
1.4.1. Dataflow diagram based on an NGS workflow .....	1 9
1.4.2. Extending the NGS process under a clinicogenomic context	2 2
1.4.3. The cGDM.....	2 7
1.4.4. Validation of the cGDM.....	3 4

**Chapter 2. Pharmacogenomic Clinical Decision Support: Modular Implementation of CPIC Guideline**

**2.1. Introduction ..... 4 1**

**2.2. Purpose of Research..... 4 4**

**2.3. Material and Methods ..... 4 5**

    2.3.1 Material: CPIC guideline as knowledge resource ..... 4 5

    2.3.2. Data Collection..... 4 6

    2.3.3. Clinical decision support service architecture..... 4 7

**2.4. Results ..... 4 9**

    2.4.1. Collected CPIC guideline and exploratory analysis..... 4 9

    2.4.2. Data integration and modeling ..... 5 3

    2.4.3. CDS Rule Extraction..... 5 9

    2.4.4. Structured database construction..... 6 0

    2.4.5. PGx CDS service module..... 6 2

**Chapter 3. Clinical Application of Clinical Genome Data Model: Integrating Star Allele and HLA Data Models**

**3.1. Introduction ..... 6 5**

**3.2. Purpose of Research..... 6 8**

**3.3. Material and Methods ..... 6 9**

<b>3.4. Results</b> .....	7 0
3.4.1. Summary of collected dataset .....	7 0
3.4.2. HLA data model .....	7 2
<b>General Discussion</b> .....	7 3
The GDM as an Infrastructure for a GIS.....	7 4
Current Approach to Genomic Data Management.....	7 6
The cGDM: A Step Beyond the Capabilities of Existing Systems	8 0
Unrecognized Ambiguity in the Interdisciplinary Knowledge Interplay	
.....	8 2
Adoption of FMEA to Information Processing.....	8 5
Limitations .....	8 7
<b>Supplementary Information</b> .....	8 9
<b>Bibliography</b> .....	9 9
<b>Abstract in Korean</b> .....	1 0 6

# List of Tables and Figures

## Chapter 1

Figure 1.1 Data-level linkage structure between conventional HIS and GIS .....	8
Figure 1.2 Data flowchart based on a next-generation sequencing workflow .....	21
Figure 1.3 Failure mode identification: mapped next-generation sequencing process extended to a clinico-genomic context .....	24
Figure 1.4 How implementation of the cGDM provides interactive clinical decision support in clinical information system .....	26
Figure 1.5 The Clinical Genome Data Model: Structured data modeling with entities and attributes .....	32
Figure 1.6 Semantic search implementation based on the CGDM .....	33
Figure 1.7 Entity-relationship diagram of the CGDM implemented in RDBMS .....	36
Figure 1.8 The conceptual map of genomic decision support system based on the cGDM .....	40
Table 1.1 Extracted classes and related attribute sets from each step of clinic-genomic context for the Entity-Attribute model .....	29
Table 1.2 Summary of imported genomic data from various data sources in the cGDM databases .....	37

## Chapter 2

Figure 2.1. The configuration of the study environment .....	46
Figure 2.2. Modular implementation of PGx CDS overview .....	47
Figure 2.3. Gene allele definition table example .....	54
Figure 2.4. Diplotype-Phenotype table example and its meta-data structure .....	57
Figure 2.5. Snapshot of CPIC guidelines content structure converted to be computable .....	58
Figure 2.6. Collection of ‘Flow chart’ over available 15 guidelines ....	59
Figure 2.7. Entity-relationship diagram of reconstructed database based on CPIC contents .....	61
Figure 2.8. PGx CDS module architecture .....	63
Figure 2.9. PGx CDS module integration scenario with dataflow .....	64
Table 2.1. The collected CPIC guideline overview .....	51
Table 2.2. Dataset list and its availability over guidelines .....	52
Table 2.3. Reference Sequence Information for Locus assignment.....	55
Table 2.4. Gene allele definition table data profiles .....	56

## **Chapter 3**

Figure 3.1 HLA Database design merge in the cGDM schema ..... 72

Table 2.1 Extracted field list gathered from the EHR records ..... 71

## **General Discussion**

Table 4.1 Comparison table of characteristics of related resources .... 78

## **Supplementary Information**

Supplementary Figure S1. PGx CDS mock-up application based on the cGDM architecture ..... 89

Supplementary Table S1. Table Specification of the cGDM ..... 90

Supplementary Table S2. IUPAC nucleotide code table for processing double/triple based code ..... 98

Supplementary Table S3. Number of HLA alleles ..... 99



# General Introduction

One of the significant tasks of medical informatics for the implementation of precision medicine is supporting clinicians by integrating personal genomic information with other clinical evidence so that constantly-evolving knowledge and inherently complex genomic data can be handled on-demand at the point of care. The transition to precision medicine and personalized medicine was expected to be accomplished within a few years due to the outstanding high-throughput sequencing capabilities of next-generation sequencing and the accumulation of knowledge about its interpretation. The prior studies present that this delay can be attributed to complicated factors, such as knowledge gaps between medical experts and bioinformatics, the separated workflow between clinical practice and bioinformatics analysis, the unique quantitative and qualitative data structure of genomic data, which can make interpretation more complicated. In an attempt to solve this problem, there is an increasing demand for the integration of personal genomic information in the electronic medical records. However, it has not been proposed as a sustainable, scalable, and interoperable method for storage, management, and processing the genomic data concerning clinical utilization.

In this study, the current barriers were explored through literature review, and related concepts and methods were investigated about these phenomena. Moreover, we addressed the immediate task of storing,

processing, and delivering data based on next-generation sequencing analysis methods to prepare for multifaceted clinical utilization. Data modeling is the first and most crucial step in the multi-tiered design of information systems. The point is that the final product reliability, such as specific clinical decision support algorithms or integrated information systems, is hardly improved over the designed reliability on the lower level of architecture.

Chapter 1 proposed a clinical genomic data model based on Deoxyribonucleic Acid (DNA) level data extracted from next-generation sequencing (NGS) technology. The multidisciplinary discussion reveals a set of genetic knowledge expressions that can be preserved and delivered the meaning for clinical decision making. In Chapter 2, the CPIC guideline, which is a knowledge of how to use available genomic test results to optimize drug therapy for individuals, is structured. Furthermore, we propose a modular drug genome clinical decision support system by linking the patient's genomic information and data-level information flow constructed in Chapter 1. Chapter 3 deals with the design and implementation of structured information about the HLA gene as one of the extensions to accommodate the diversity of naming systems as the discoveries that reveal their clinical significance in bioinformatics continue. The sustainability and scalability of the clinical genomics data model were verified by design and expand knowledge expression for HLA nomenclature.

In this study, we explored multidisciplinary space where medical informatics can contribute to precision medicine, and an approach that encompasses aspects of knowledge expression, functional realization, and usability of information systems was attempted.

# **Chapter 1. Clinical Genome Data Model: Data Level Integration of Patient Specific Genomic and Clinical Data for Multifaceted Utilization\***

## **1.1. Introduction**

As the field of medicine transitions from experience-based medicine to data-driven medicine, an apparent paradigm shift to precision medicine is underway, driven by the development of technologies in fields including medical information technology and computer engineering<sup>1,2</sup>. Genomic information is one of the most critical component of precision medicine, given its power to explain individual variability<sup>3</sup>. However, the practical clinical use of genomic information remains limited because its circulation is suboptimal, with each data processing step tending to be independently performed and thus isolated. To narrow this gap, many organizations have attempted to identify and develop methods to more effectively link genomic data to clinical information and thereby facilitate its use<sup>4-6</sup>. However, several challenges must be surmounted before realizing this goal.

First, a mismatch exists between the structure of genomic and clinical data. Genomic data based on next-generation sequencing (NGS) technology is stored as a number of file types at various stages of the bioinformatics

---

\* The main body of the dissertation chapter 1 published as following paper: Kim, H. J., Kim, H. J., Park, Y., Lee, W. S., Lim, Y., & Kim, J. H. (2020). clinical Genome Data Model (cGDM) provides interactive clinical Decision Support for precision Medicine. Scientific reports, 10(1), 1-13.

analysis, with flexible file specifications to accommodate the broad range of research interests in bioinformatics<sup>7</sup>. Raw genomic data can contain up to several tens of gigabytes of sequence information, each stored as a long string of data, and therefore cannot be used directly in this form in clinical practice without further processing. Since data processing to determine clinical relevance is both computationally intensive and time-consuming, genomic information is not readily accessible relative to other types of clinical data. Thus, for precision medicine and personalized medicine, pre-processed genomic data needs to be linked with other clinical information and provided at the appropriate time. In order to resolve this issue, a structured database is needed to store and appropriately manage genomic information for easy accessibility.

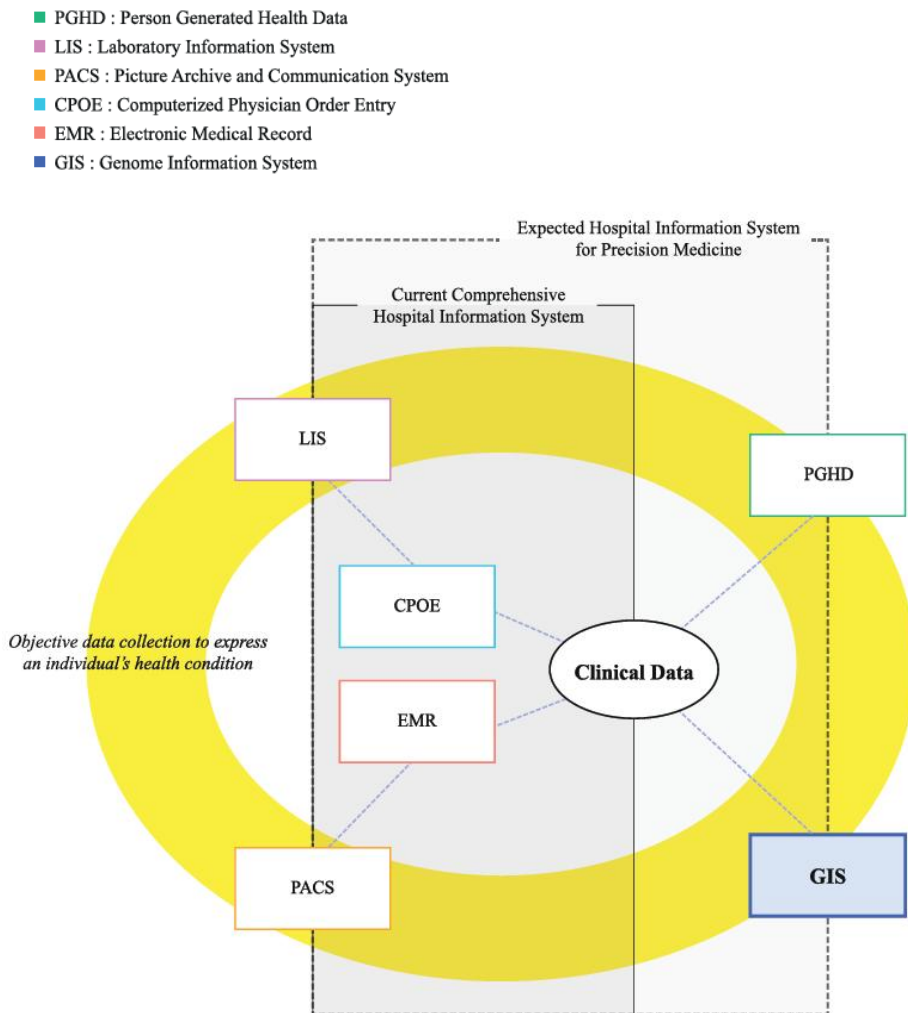
Second, genomic data has different properties than conventional observational data used in clinical settings. Therefore, genomic data must be clarified by considering procedural dimensions. Since genomic workflows contain a large number of pipelines for information processing, significant differences between the interpretation of processed data and data obtained from different information systems relative to the clinical workflow are inevitable<sup>8</sup>. Accordingly, a robust data model is required to serve as an information system to systematically manage genomic data, encompassing the detailed processes of data processing, analysis, and filtering. Additionally, information on the reliability and accuracy of these analyses

results, along with the detailed analytical process and equipment used, must also be systematically stored and managed, as it is an essential criterion for clinical decision-making<sup>9</sup>. Moreover, because genomic data is less variable than observational data, information integration will allow for maximization of the utility of the collected genomic information for clinical use.

The third challenge, majorly hindering the integration of genomic data with clinical information, is difficulty in mapping the two types of data for medical interpretation. The presence of biomarkers for specific diseases or drug reactions is a critical factor in clinical decision-making<sup>10</sup>. In the case of targeted sequencing, the data processor is informed about biomarkers related to the panel prior to analysis. In clinical practice, reannotation of patient genetic information according to updated biomarker discoveries from the biomedical research community is continuously required at the population level. Thus, a structured data model with consistent data representation would enable the rapid adoption of both evolving biomedical knowledge and individual medical records, which can be delivered to the point of care through agile data processing. Furthermore, patient genomic data expressing specific biomarkers should be readily accessible from the information system along with clinician-confirmed interpretations<sup>10,11</sup>.

Personal-health status can be converted to a composition of multi-layered, multi-dimensional digitalized information for utilization in an information system that facilitates handling big data (Fig. 1). Indeed, vast

amounts of data and associated metadata from multiple medical measuring technologies, such as laboratory tests or imaging studies, have already been successfully merged in clinical information systems. Overall, although genomic information represents the most sound and intensive health-related signals provided by the human body throughout life, the weak links to medical practice highlighted above contribute to its underutilization in clinical decision-making. Therefore, it is necessary to effectively link and integrate clinical information with personal genomic information, helping to accelerate the shift to personalized medicine.



**Figure 1.1 Data-level linkage structure between conventional HIS and GIS**  
 From a software engineering perspective, a comprehensive hospital information system comprises components that represent separated data collection routes and distinguishing characters of the data. We suggest the concept of GIS to illustrate the implementation of the cGDM. This architecture supports both information and functional integration, even with existing clinical information systems.



## **1.2. Purpose of Research**

The proposed GDM is based on an entity-attribute model to effectively manage and maximize the use of genomic data in clinical practice. Through the development of this method, we focused on equal weighting to the clinical perspective and bioinformatics process analysis as business continuity, starting from the initial clinical intention to bioinformatics information processing associated with a knowledge-related protocol, finally offering a deliverable and interpretable form to the point-of-care clinician. The GDM was designed based on DNA level data from next-generation sequencing (NGS) technology to deliver processed genomic data of patients from different pipelines by applying an appropriate information scale and granularity at the clinical level.

Toward this end, we began by redefining the obstacles to the spread of genomic information into routine care, including reliability problems of proposed measurement data that could cause hesitation in clinical decision-making, and data structure problems that have hindered the integration of genomic data into existing information systems. From a clinical perspective, we focused on the reliability of information as well as the problem of a heterogeneous data structure. In this context, we define a bioinformatics process not as a “measurement,” but rather as a “production” to transition a physical form of existence to an interpretable human representation.

Overall, we aimed to develop a model with appropriate information granularity and scale, which would minimize the possibility of misinterpretation at the point of care by formal and procedural variation related to the production process.

### **1.3. Materials and Method**

The study material was genomic information with clinical relevance based on NGS technology. A failure mode and effect analysis (FMEA) approach was adopted as the analysis process and attributes-extracting method, which was accomplished by assembling a multidisciplinary working group. From November 2017 to July 2018, process mapping, failure identification, and related attribute extraction were performed by the FMEA method at over 18 team meetings. An entity-attribute model was then developed by reconstruction of the attribute set derived from the FMEA.

### **1.3.1. The Production Process of Bringing Genomic Information to Bedside Care**

Here, we define a genomic test as a series of team-based information production processes, in which the meaning of the information is expanded, represented, and reproduced by reference to an external knowledge base, rather than through direct extraction of inherent information. Despite the invariant nature of a personal genome, genomic information presented to a clinician may vary according to specific processing protocols adopted<sup>7,12-14</sup>. This variability raises reliability issues for the use of genomic test results as clinical evidence<sup>15</sup>.

As artifacts from production, genome information processed for clinical use may pose a likelihood of misinterpretation due to information distortion, omissions, and fragmented senses. Furthermore, information reliability is a critical factor determining the ability of clinicians to utilize the genomic information<sup>16</sup>. Thus, our approach in developing this cGDM for focussed on the multi-dimensional scope of information, including procedural factors, derived from NGS technology.

### **1.3.2. FMEA: An Attribute-Clarified Framework**

FMEA is a systematic prospective risk factor analysis approach that predicts and prevents possible errors, improving quality across team-based processes<sup>17</sup>. When used for advanced investigation, the method has advantages enabling exploration of uncertain, unforeseen complex workflows at an early stage<sup>18,19</sup>. Since its introduction in 1963, broad subtype applications of FMEA have been performed in broad domains including reliability engineering<sup>20,21</sup>, behaviour modeling<sup>22</sup>, software engineering<sup>23</sup>, conceptual design<sup>24</sup>, and knowledge management and representation<sup>25,26</sup>. In particular, FMEA has been applied as a method of knowledge representation to extract process reliability-related attributes and to structure and map entities and attributes<sup>22,26-28</sup>. In this study, the FMEA approach was adopted for workflow analysis and the attribute-extracting method.

### **1.3.2.1 The working group**

A multidisciplinary expert team was formed from the areas of bioinformatics, medical informatics, and medicine. The participants included three bioinformaticians, two medical informaticians with clinical informatics and application expertise, and one medical doctor. The medical doctor has experience in both clinical practice and conducting translational research from the perspective of both biomedical science and clinical practice.

### **1.3.2.2 Workflow analysis**

Over a period of nine months, process mapping, failure identification, and related attribute extraction were conducted using FMEA at over 18 team meetings. Structured data modeling for enhancement of data accessibility was then conducted using a logical data model, with the attribute set derived from the FMEA workflow diagram.

We chose the conventional FMEA workflow analysis<sup>21,28</sup> and adapted it for cGDM development. Conventional FMEA consists of two main steps. First, the failure mode is identified through 1) assembling a multi-disciplinary team with at least one expert from each domain over the target production process, 2) combining components and process function in order to derive a workflow diagram, and 3) listing the modes that may lead to failure at each step. The second part involves modifying the process itself with consideration of priority, including 1) evaluating the severity and occurrence ranking of each failure mode and 2) proposing a modified workflow or audition guideline.

In this study, risk estimation and priority-scoring steps were not designed, since our purpose was to review the fragment of metadata composition that may cause unintended information distortion of misinterpretation.

### **1.3.3. Logical Data Modeling**

Data models are the basis of computation ability for intelligent information systems<sup>29</sup>. The database design process can generally be divided into logical and physical database design<sup>30</sup>. The physical data model requires a clear and specific description over logical design, which depends on the existing development environment. Thus, we designed this cGDM as a logical data model based on the FMEA results to support data-level integration with any existing clinical information systems.

Logical data modeling methods are comprised of abstraction and normalization. Database abstraction refers to aggregation and generalization that occur at the points of intersection<sup>31</sup>. We first abstracted the attributes derived from FMEA and expressed the factors corresponding to each step in the workflow. Then, normalization was performed to prevent duplication and inconsistency of data elements considering their names, scale, and relations.



### **1.3.4. Demo Datasets for the real-world data validation**

Two of representative public accessible dataset are selected for the development of the demo databases: The 1000 Genomes Project of the International Genome Sample Resource (IGSR) with population code "CEU" (Utah Residents with Northern and Western European Ancestry)<sup>32</sup>, the pancreatic cancer data from The Cancer Genome Atlas (TCGA\_PAAD)<sup>33</sup>.

Collected datasets were VCF and MAF file format, and the Extract-Transformation-Load (ETL) process of the genomic data was performed by two bioinformaticians with Python 2.7.16. ANNOVAR 2016Oct24 version was used as a clinical annotation tool for the 1000 Genome Project CEU dataset. The resulting dataset imported in a table within the MySQL server database by two medical informaticians. We ran the SQL scripts in MySQL 5.6.46 on a Server with 8GB of RAM and an NVIDIA tesla c1060 / Quad-core CPU running run on CentOS Linux release 7.7.1908. The final outputs took the form of SQL tables and functions.

## **1.4. Results**

This section primarily consists of Failure Mode and Effects Analysis (FMEA) results and entity-attribute modeling. FMEA output is presented in two diagrams: a dataflow diagram that focusses on the derivation of the contents of the genetic test based on NGS sequencing technology, and an information process map that extends the viewpoint to the level of clinico-genomic context. At this step, the protocol entity of the former dataflow diagram was subclassified to reveal the procedural dimension in information processing. Moreover, the set of attributes involved in each step of information transfer was identified. Finally, the cGDM are suggested as a result of structured data modeling based on the attribute set.

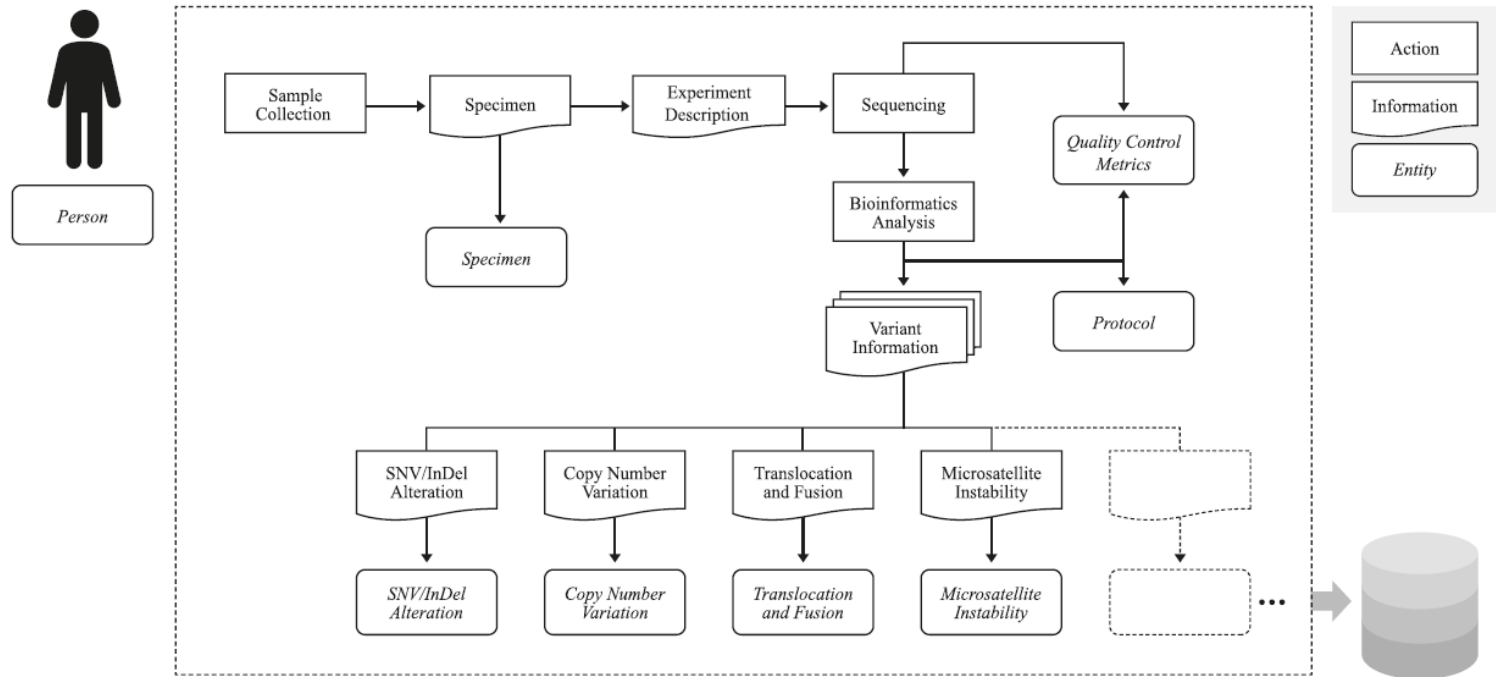
### **1.4.1. Dataflow diagram based on an NGS workflow**

A workflow diagram was derived in order to illustrate the data flow in which the genomic information inherent in the human body is converted to a genomic test result. (Fig 2.) At this stage, the clinical view is minimized, with both the flow of information and the process of analyzing the specimen after the sample collection across experimental laboratory and computational analysis drawn on a large scale.

The subtypes of processed variant information in the parallel structure, used to cope with the growing body of knowledge in bioinformatics, are listed at the bottom of Fig. 1. Variant information can be called in multiple types depending on the perspective and purpose of the analysis. For example, there are four types of genetic variation: single nucleotide variation (SNV), small insertion/deletion (InDel), copy number variation (CNV), and translocation/fusion. There are predictive biomarkers as well such as microsatellite instability (MSI) and tumor mutation burden (TMB).

As the amount of NGS technology-based knowledge increases, more subclasses representing novel perspectives can be added. Scalable data modeling to support the differentiation of knowledge over time is essential not only for expressiveness but also for reducing the burden of information systems maintenance.

In summary, we linked the separate offline workflows at this step that occurred in different places until genomic data could be provided as processed data. The workflow diagram provided the basis for detailed analysis and discussion.



**Figure 1.2 Data flowchart based on a next-generation sequencing workflow**

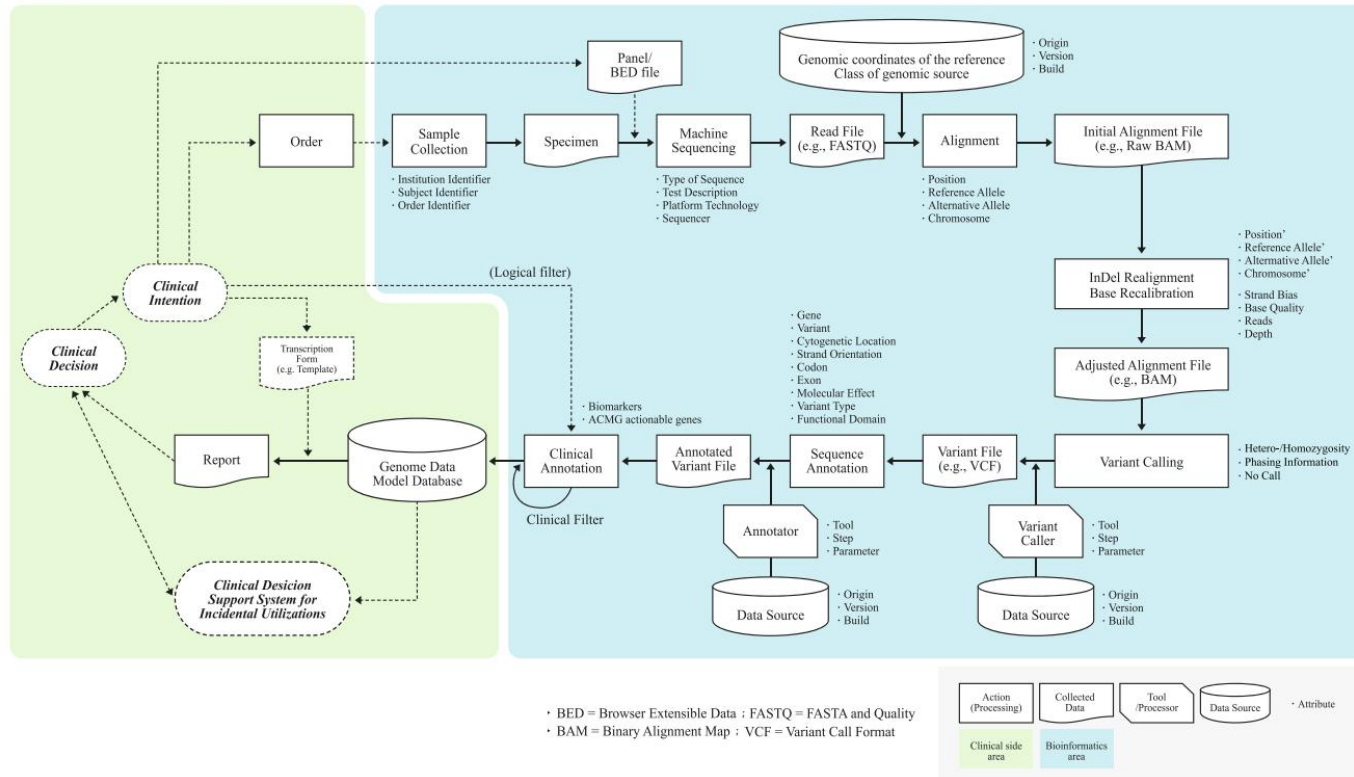
The objects shown in this diagram are classified into three class types- 'Action', 'Information', and 'Entity'. 'Action' was first posted with respect to what occurred in each expert domain and the resulting 'information' was displayed as a result of each action. Finally, 'Entity' was defined as the captured information class at each stage of the workflow. Subtypes of 'Variant Information' were drawn scalable to accommodate the potential extension of subclasses.

## **1.4.2. Extending the NGS process under a clinico-genomic context**

After establishing a consensus on a larger scale, we extended the information flow to the clinical context in detail. At this stage, the standpoint of the workflow analysis was clinical decision making. Hence, the workflow diagram started with a clinical decision. We extended the flow between several actions in the clinico-genomic context involving multiple entities identified, and detailed analysis was performed. In this process, the output data file format and detailed processes for handling output files, along with the tools required for linking to external knowledge databases, are also described.

The working group discussed mechanisms for extraction of the entity-attribute set which would avoid probable information distortion and omission. We considered that the genomic data model for clinical use should be the knowledge communication scheme, thus preserving its reliability-related factors. At a minimum, the genomic data model must provide sufficient information to decide whether the confidence level of the genomic test result justifies its consideration as clinical evidence. For this function, failure was defined as that which causes misinterpretation or non-use of the genomic data for clinical decision. The process of producing clinical evidence from genomic data at the bioinformatics area (Fig. 3) shows a pattern that is a series of repeated representations of information converted

by reference knowledge bases and data processing rules. Thus, failure modes can be classified as incomplete specifications in three meta-categories: origin, reference, or symbol. Due to the nature of the semantic interpretation, any fragmentation of symbol causes not only loss of information but also assignment information to direct the origin<sup>12,13</sup>.

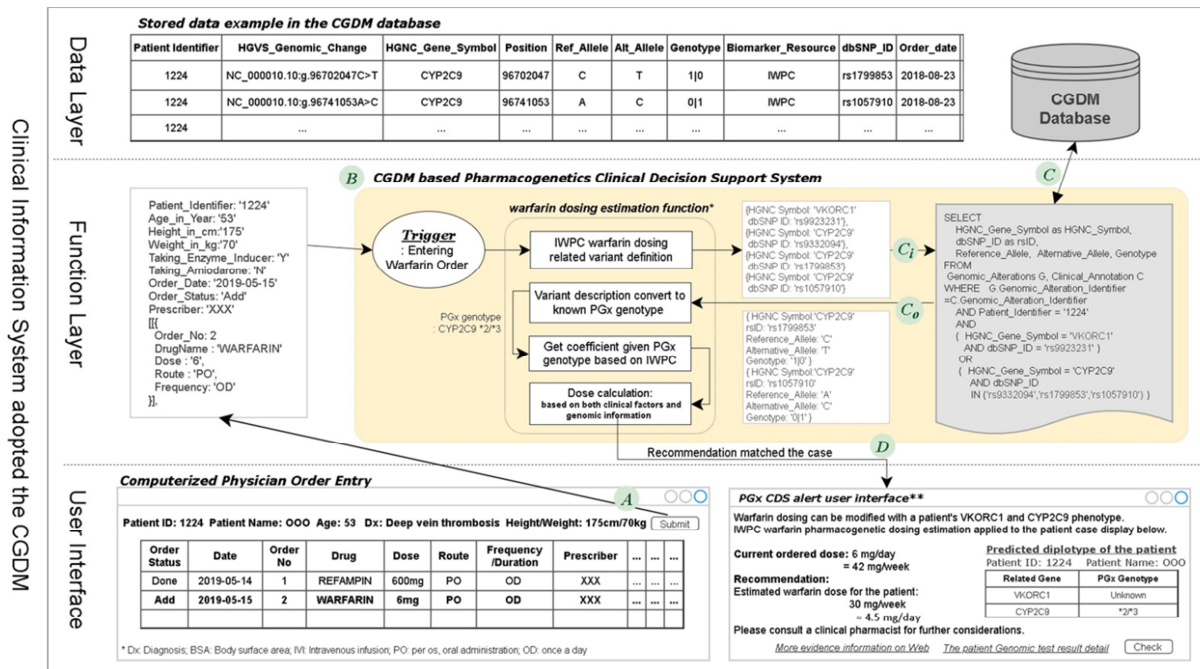


**Figure 1.3 Failure mode identification: mapped next-generation sequencing process extended to a clinico-genomic context**

In the bioinformatics area (cyan background), information may be distorted by the insufficient representation of origin, processing rule, and external reference. To prevent this failure, identification and semantics, related attributes are listed under the boxes. In the clinical area (yellow background), the data model functions as a communication scheme for the collaborative process implemented in the hospital information system. Data-level integration facilitates just-in-time queries and reuse of data.



We conducted workflow analysis to extrapolate general descriptors of the related attributes with the goal of preserving information during production and delivery processes from clinical intention to clinical utilization. Figure 2 provides a more detailed data-level view, including how genomic information is realized as clinical evidence in a case based on a structured data model. The structured genome data model can support a report via presentation on a variety of transcription forms (report forms), which are optimized for initial intent. Furthermore, additional utilization paths are accessible in the clinical-information system. As shown in Fig. 2, data-level integration helps the amplification of the incidental utilization. (Fig. 4) To illustrate, consider a patient who orders whole-genome sequencing to screen for cancer biomarkers at their first visit. When the patient receives a prescription for antibiotics a year later at a visit for other symptoms, that same genomic test result can be re-used from a pharmacogenomics perspective for safer and more efficient drug prescription. The clinical decision support system plays a vital role by just-in-time display of the matching information with pre-defined rule and knowledge-based processing<sup>6,34,35</sup>. A computational genome data model is a prerequisite for this implementation<sup>35-37</sup>. Finally, we introduce a logical data model in the next step of the study.



\* Knowledge reference : International Warfarin Pharmacogenetics Consortium. (2009). Estimation of the warfarin dose with clinical and pharmacogenetic data. New England Journal of Medicine, 360(8), 753-764.

**Figure 1.4 How the implementation of the cGDM provides interactive clinical decision support in clinical information system**

A: When a doctor enters a prescription, a dataset for the prescription is generated and transmitted for storage. B: The dataset is passed along to the CDS module to search for the relevant knowledge base in accordance with a predefined set of rules. In this case, we internalized the systematic reference to the IWPC algorithm\* integrated with the CGDM database. C: The PGx CDS module based on the cGDM selects the patient-specific warfarin dosing related variant information which matches the IWPC algorithm in real-time. The cGDM produces an effect as a knowledge representation backbone as well as a genomic data storage scheme in the process. (e.g., Expression converted from input variables (Ci) to output variable (Co) for further processing.) D: The recommendation, which personalized dosing results from the IWPC warfarin PGx estimation based on both clinical and genomic factors, are delivered to the prescriber. Trackable links for each origin of the used genomic data and evidence in the algorithm are also provided.

### 1.4.3. The cGDM

Finally, the cGDM was designed as an entity-attribute model consisting of 8 entities and 46 attributes (Fig. 5). For a structured data model of the identified clinico-genomic attributes, logical modeling was conducted to ensure data-level linkage with conventional primary clinical databases. In order to define the entity-attribute model based on the action and collected data, tool/processor classes and the attributes of each class from Fig. 2, we define three types of classes as protocol and related attributes (Table 1). Since the cGDM is designed to support data-level integration with the existing system, only the minimum subject identifier is defined as ‘linkage identifier to clinical information.’ To represent the procedural dimension, which is stressed in the study, we combined two workflow analyses on different scales. For example, the entity ‘Protocol’ as a part of the procedural dimension is explicitly represented in Fig. 2, then expressed again as a list of lower steps in Fig. 3. Since clinical observation is typically considered as the collection of events<sup>38</sup>, the logical composition of the date/time and actor identifier related to the clinico-genomic context were declared.

The derived classes and entities in Table 1 were used to declare final entities and attributes in the cGDM (Fig. 5). The mapped Actions and Action-related classes (Collected Data and Tool/Processor) are categorized into subdomains and related attributes for each step in Table 1. In Table 1,

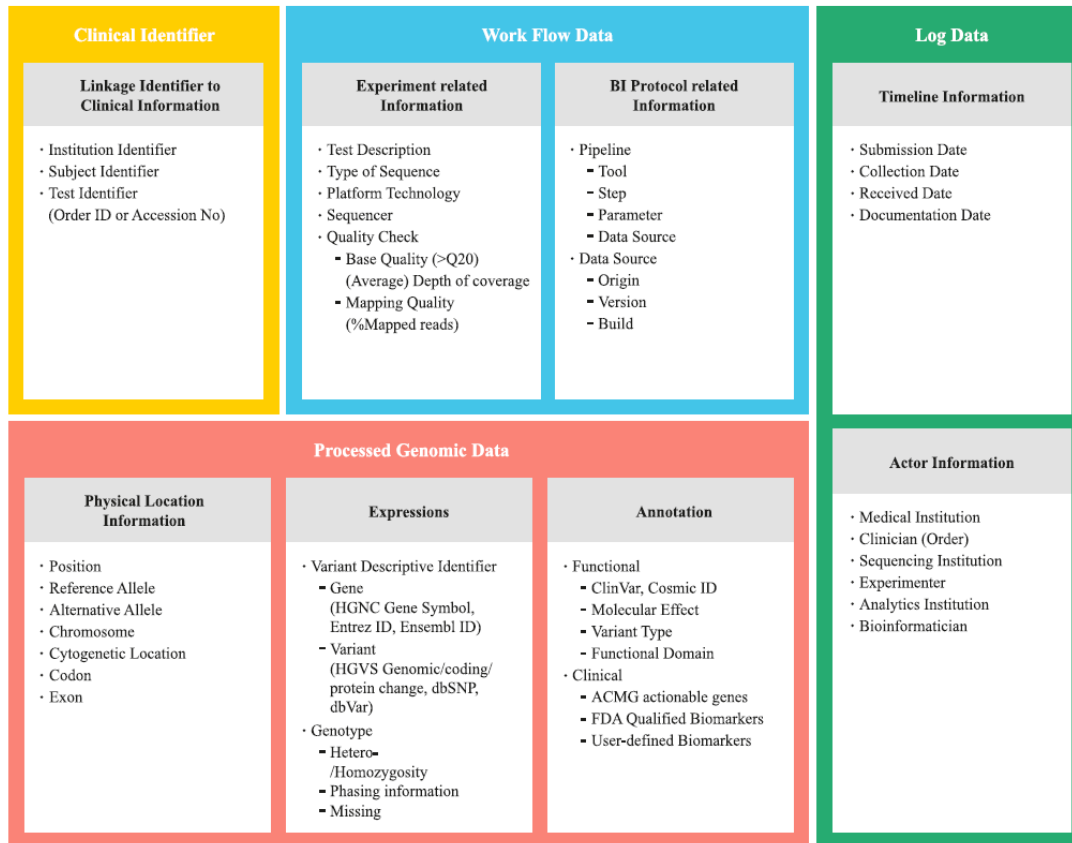
action and its result are grouped into one step, and the related attributes are represented by the attributes classified in the corresponding step. For normalization, related attributes are categorized to create one or more new groups called 'entities' for each step, and they are the basis for defining 'Entities' in the Entity-Attribute model (Fig. 5). For example, 'Physical information according to the coordinate system' is one of the three subdomains of the action 'Sequence Annotation.' It can include an attribute set (Cytogenic location, Codon, Exon) representing physical location information for each variant. However, this "Physical information according to coordinate system" can be a subdomain in other steps besides "Sequence Annotation". And even though it is the same subdomain, the related-attribute set may be different depending on which step or action. In summary, each step identified in the entire clinico-genomic process can include multiple entities, and one entity can be related to multiple steps. Even in the same entity, the configuration of the related attribute as a factor affecting each step may vary from step to step.

**Table 1.1 Extracted classes and related attribute sets from each step of clinic-genomic context for the Entity-Attribute model.** The processes in the clinico-genomic workflow shown in Figure 2 are listed in order and associated with the classes, related attribute sets for each process. This table is an intermediate result between the result of FMEA and the final logical model. Derived related attributes are abstracted within each class and grouped into entities.

Action	Class		Related Attribute	Entity
	Collected Data	Tool/ Processor		
Sample Collection			Institution Identifier Subject Identifier Test Identifier (Order ID or Accession No)	Linkage Identifier to Clinical Information
			Submission Date	Timeline Information
			Medical Institution Clinician	Actor Information
Specimen				
Machine Sequencing			Test Description Type of Sequence Platform technology Sequencer	Experiment Related Information
			Collection Date	Timeline Information
			Sequencing Institution Experimenter	Actor Information
Read File				
Alignment			Position Reference allele Alternative allele Chromosome	Physical(Location) information according to coordinate system
			Analytics Institution Bioinformatician	Actor Information
Initial Alignment File				
InDel Realignment / Base Recalibration			Position <sup>c</sup> Reference allele <sup>c</sup> Alternative allele <sup>c</sup> Chromosome <sup>c</sup>	Physical(Location) information according to coordinate system
			Base quality(>Q20) (Average) Depth of coverage Mapping Quality (%Mapped reads)	Quality Check information

		Received Date	Timeline Information
		Analytics Institution Bioinformatician	Actor Information
	Adjusted Alignment File		
Variant Calling		Hetero- /Homozygosity Phasing information Missing	Genotype Expressions
		Analytics Institution Bioinformatician	Actor Information
	Variant Caller	Tool Step Parameter	Pipeline information
		Origin Version Build Parameter	Data source
	Variant File		
Sequence Annotation		Gene (HGNC Gene Symbol, Entrez ID, Ensembl ID) Variant (HGVS(genomic, coding, protein change + version), dbSNP, dbVar)	Variant Descriptive Expressions
		Cytogenetic location Codon Exon	Physical(Location) information according to coordinate system
		ClinVar, COSMIC ID Molecular Effect Variant Type Functional Domain	Functional Annotation
		Analytics Institution Bioinformatician	Actor Information
	Annotator	Tool Step Parameter	Pipeline information
		Origin Version Build	Data source
	Annotated Variant File		

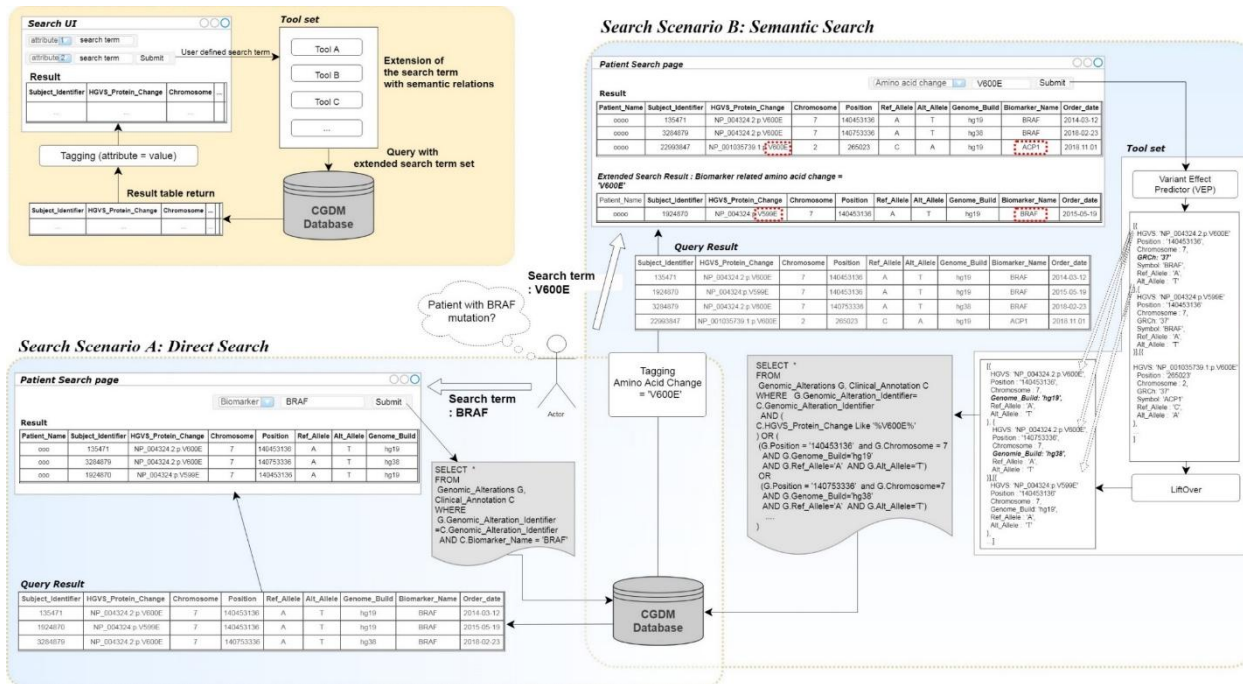
Clinical Annotation	ACMG actionable genes	Clinical Annotation
	FDA qualified biomarkers	
	User-defined biomarkers	
	Analytics Institution Bioinformatician	Actor Information
	Documentation Date	Timeline Information



**Figure 1.5 The Clinical Genome Data Model: Structured data modelling with entities and attributes**

The cGDM is designed as a logical data model of 8 entities and 46 attributes. The objects and related attributes derived through FMEA are integrated into a logical data model through abstraction and normalization.





**Figure 1.6 Semantic search implementation based on the CGDM**

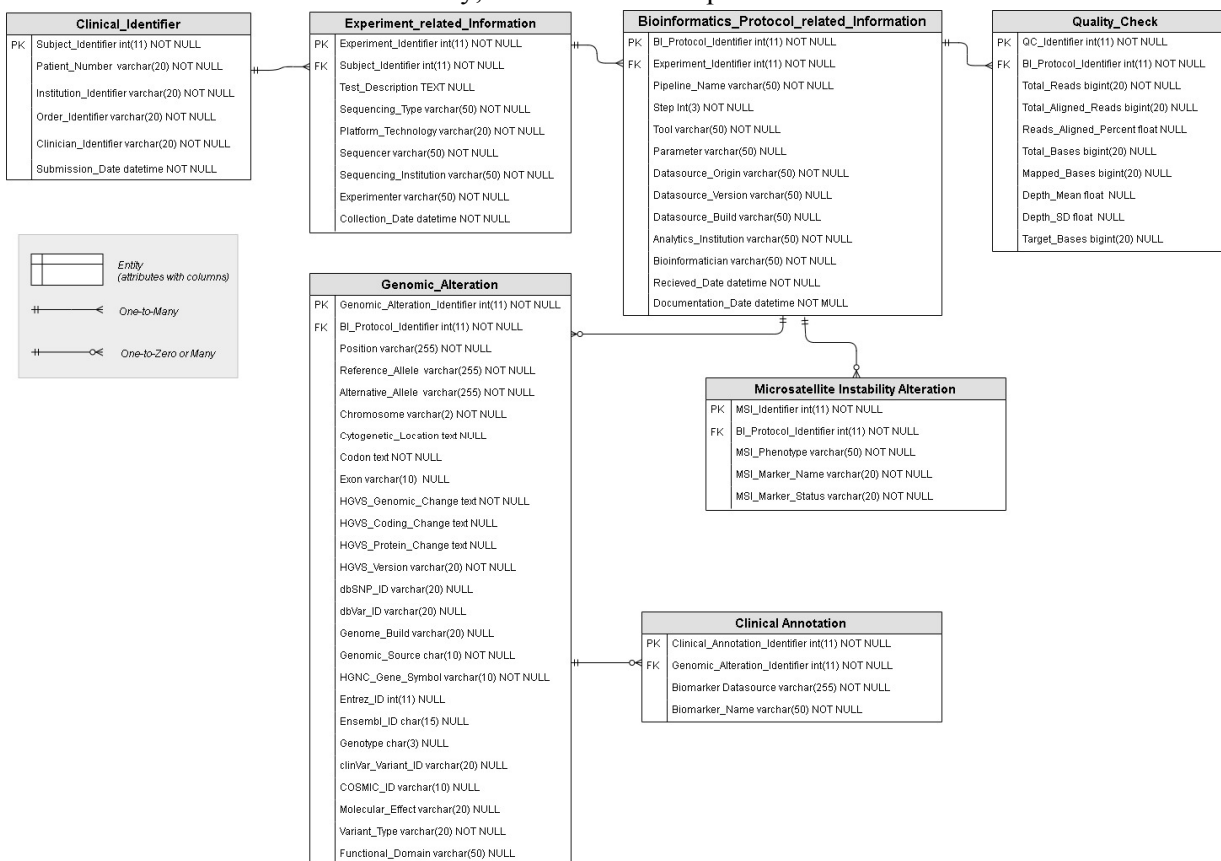
Even if the user does not know all the nomenclature or metadata relevant to the genomic data to be searched, search function based on the CGDM can use information entered in the search fields in order to derive an extended search result. Through the generated SQL syntax, the user can determine which genomic metadata (such as chromosome and position, genome build version, HGVS ID) can be associated and extended to the outcome of the patient's data. In addition to the attributes "Biomarker" and "HGVS ID" presented in the example, multiple data queries can be made with a single attribute or combination of attributes presented in the CGDM. Therefore, by using these user interfaces with the data model, it is possible to trace and verify whether the queried genomic data of the patient represent more reliable information.

#### **1.4.4. Validation of the cGDM**

Here, the cGDM was finalized in the form of a logical model, which allows adaptation to the diverse development environments of existing heterogeneous clinical information systems. Logical model can play an essential role to generalize the complex phenomenon by abstraction and enhance understanding core ideas the model deliver between different stakeholders of in the complex system<sup>39</sup>. Whereas, the drawback of this approach is that physical modeling layer is needed in order to the data model implementation and validation. Thus, we design a physical data model implemented in relational database to evaluate the model validity for real-world data and to proof of concept how implementation of the cGDM enables interactive clinical decision support in clinical information system shown as Fig3 (Left side; Clinical decision support system for incidental utilization).

**Figure 1.7. Entity-relationship diagram of the CGDM implemented in RDBMS**

The entity-relation for the physical model as a diagram (ERD) was presented based on the table shown in Supplementary Table 1. The diagram shows the entities and the attributes that describes the entity, and the relationship between the entities is also defined.



### **1.4.4.1. Implementation of the real world data**

This physical data model of the cGDM is provided in forms of entity-relationship diagram and table (Supplementary Information Table 1; Fig 1.7). Also, one-click executable data definition language script is also freely accessible on a web page (<https://github.com/SNUBI-HyojungKim/cGDM-Clinical-Genome-Data-Model>).

For the data model validation with real-world data, we built pilot databases based on the cGDM and uploaded genomic data of over 2,000 patients for multiple diseases, including acute lymphoblastic leukaemia, solid cancers, and depression cases (Table 2, internal databases). However, the pilot dataset related researches remains undergoing, we have built two representative demo datasets for open source (Table 2, demo databases) 1) 1000 genome CEU (Utah Residents with Northern and Western European Ancestry) population dataset for whole genome sequencing (n=99, 47.67 GB), 2) TCGA PAAD (Pancreatic Adenocarcinoma) dataset for somatic mutation (n=155, 9.41 MB). We believe those well-known public dataset has advantages on data validation issue. Every demo dataset and source codes are freely available from at the Github page as mentioned above.

**Table 1.2 Summary of imported genomic data from various data sources in cGDM databases.**

The databases are categorised into internal and demo database. The specifications of the database tables are informed in Table 1. This table presents row counts of each database table and data volumes of each database. The internal databases includes 3 private datasets (cancer panel, leukemia and depression) and 2 public datasets (TCGA COAD and TCGA LUAD). The demo databases includes 2 public datasets (1000 Genome Phase3 CEU and TCGA PAAD).

Table name		Database							Summary
		Internal database					Demo database (public license)		
		Cancer Panel	Leukemia	Depression	TCGA COAD	TCGA LUAD	1KGP P3 CEU	TCGA PAAD	
	Type of sequencing	cancer panel	WES	WES	somatic mut.	somatic mut.	WGS	Somatic mut.	7 data sets WGS/WES/ targeted panel
	CLINICAL_IDENTIFIER	10	503	1,000	459	522	99	155	2,748
	EXPERIMENT_RELATED_INFORMATION	10	517	1,000	459	522	99	155	2,762
	BIOINFORMATICS_PROTOCOL_RELATED_INFORMATION	10	517	1,000	459	522	99	155	2,762
	GENOMIC_ALTERATION	2733	29,279,631	842,199,347	361,933	318,947	229,525,363	56,159	1,101,744,113
	MICROSATELLITE_INSTABILITY	0	0	0	0	0	0	775	775
	CLINICAL_ANNOTATION	40	267	108	123	97	1	12	648
	QUALITY_CHECK	10	517	1,000	0	0	0	0	1,527
Data volume	database total	2 MB	8.2 GB	144.7 GB	48.4 MB	42.6 MB	47.7 GB	9.4 MB	201.5 GB
	per test	0.2 MB	8.12 MB	144.7 MB	0.1 MB	0.1 MB	481 MB	0.6 MB	91.8 MB

Real-world data validation is designed to cover all three types of NGS tests (targeted panel, WES, WGS) and both cases of somatic mutations and germline variants. The storage capacity of data was reduced when converted into relational database with cGDM schema by 30% compared to the prepared data file in VCF format. Interestingly, as the data size of the genomic alteration table per test increased, the gap in data size by converting narrowed or overturned. The circumstance is due to the addition of multiple indexes for in-time query performance. Table indexing was generally required when an average of more than 30,000 rows per test occurs.

### **1.4.4.2. How the implementation of the cGDM enables interactive clinical decision support**

One of the major challenges of healthcare informatics is supporting clinicians who need to handle constantly evolving knowledge and inherently complex genomic data. Patient genomic data in static document format or in structured model but in which has vague designation of the variant limits functionality of clinico-genomic information system<sup>40</sup>. The cGDM could address the issue by working as a data-level infrastructure for interactive clinical decision support along with external knowledge bases (Fig.6). For the cGDM's programmability test, we developed a pharmacogenomic clinical decision support function running on the cGDM database which reflects the knowledge of the IWPC warfarin dosing algorithm. The source code is freely available at <https://github.com/SNUBI-HyojungKim/cGDM-Clinical-Genome-Data-Model>. Figure 7 illustrates both of logical information flow in back-end system and its appearance on the user interface. A query performance test is conducted with the algorithm procedure over 99 individuals in 1KGP P3 CEU database. The SQL stored procedure has executed in MySQL on a server with 8GB of RAM and quad-core CPU running Linux CentOS 6. The average query out duration was  $0.013 \pm 0.008$  second range from 0.00001 to 0.0460.

## Integrated in clinical workflow

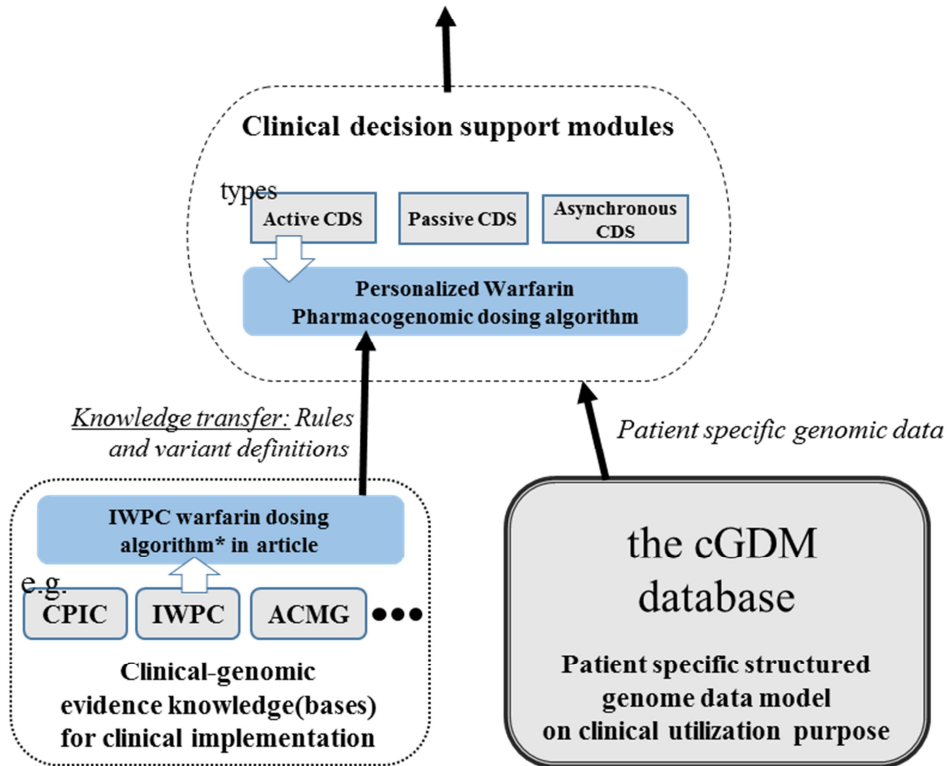


Figure 1.7 The conceptual map of genomic decision support system based on the cGDM

While the accumulation of confirmatory knowledge could seem relatively slow compared to the speed of the vast discovery of the bioinformatics field, the benefits and impacts the two will have on patients when they are seamlessly connected are evident. The cGDM brings this process into computational space.



# **Chapter 2. Pharmacogenomic Clinical Decision Support: Modular Implementation of CPIC Guideline**

## **2.1. Introduction**

As the development of sequencing technology and the results of research on pharmacogenomics (PGx) accumulate, efforts are being made to apply personalized drug prescriptions and dose adjustments in the clinical field. The same drug may cause adverse reactions due to congenital or acquired causes, and drug adverse reactions are a major obstacle to the safe and effective use of drugs. “The social costs and health disadvantages of these adverse drug reactions are well known. PGx use cases are of particular interest because over half of all primary care patients are exposed to PGx relevant drugs. Studies have found that 7% of U.S. Food and Drug Administration (FDA)-approved medications and 18% of the 4 billion prescriptions written in the United States per year are affected by actionable PGx variants that nearly all individuals (98%) have at least one known, actionable variant by current Clinical Pharmacogenetics Implementation Consortium (CPIC) guidelines and that when pharmacogenes with at least one known, actionable, inherited variant are considered, over 97% of the U.S. population has at least one high-risk diplotype with an estimated impact on nearly 75 million prescriptions”<sup>41</sup>. Therefore, it is expected that

applying knowledge about the drug genome to avoid predictable adverse reactions to patients and maximizing the effects of drug treatment prior to drug prescriptions would improve patient safety and quality of treatment.

Various efforts are being made to establish a knowledge resource of pharmacogenomic knowledge that can be applied in clinical practice and to connect it to clinical information systems. Representatives are as follows; Clinical Pharmacogenetics Implementation Consortium (CPIC) of the Pharmacogenomics Research Network (PharmGKB)<sup>42</sup> and the Dutch Pharmacogenetics Working Group (DPWG)<sup>43</sup>, International warfarin pharmacogenetics consortium (IWPC)<sup>44</sup>, Canadian Pharmacogenomics Network for Drug Safety (CPNDS)<sup>45</sup>. Efforts have been made to implement informed decision making using pharmacogenomic information in clinical settings based on these refined knowledge resources. In particular, recent attempts at systematic clinical implementation have been reported by the European Consortium<sup>46</sup>, the IGNITE Network Pharmacogenetics Working Group<sup>47,48</sup>, and the United Kingdom<sup>49</sup>. In order for PGx to become routine in practice, attention has been paid to establishing a PGx decision support system integrated with EHR.

However, it has not been proposed as a sustainable, scalable, and interoperable design among different sites. When considering the complexity of dealing with the volatility of PGx knowledge and the considerable amount of information in patient-specific genomic data as an

extension of the clinical context, PGx clinical decision support pipeline focused on knowledge representation is needed. Moreover, data processing methods is needed to provide PGx test result on demands. Clinical decision support (CDS) holds great promise for genomics but has had limited utility because executing CDS has required manual entry of genetic conditions into the problem list for decision support<sup>50</sup>.

In the study, we aim to develop a PGx CDS pipeline linking between clinical actionable drug-gene interaction knowledge and personal genomic data. First of all, we transform CPIC guideline knowledge resources into a machine-readable structured database. Finally, we suggest a PGx CDS service design based on the data model layer, both on CPIC guideline knowledge resources and personal genomic data.

## 2.2. Purpose of Research

We propose PGx CDS that enables modular implementation between heterogeneous existing clinical information systems. Modeling of medical knowledge and representation of and reasoning about medical knowledge are the significant steps of the construction of CDS tool<sup>70</sup>. Although CPIC guidelines supporting the clinical application of pharmacogenomics knowledge provide reliable content, considerable modeling activities are required to transform knowledge from human-interpretable form to a machine-readable form for consistent application.

Thus, we firstly collected, integrated CPIC guideline contents. Data integration gives a unified landscape by combining data from disconnected resources<sup>51</sup> In this process, modeling the relationship between the sources and the global schema is, therefore, a crucial aspect. Then, we transform CPIC guideline knowledge resource to the machine-readable structured database along with content analysis. Exploratory analysis of the collected dataset reveals the rules or properties that the content implicitly implied. Finally, we propose a modular PGx CDS service by capturing the explicit and implicit knowledge flow of the CPIC knowledge resource through the modeling process and seamlessly unites actionable drug-gene interaction knowledge with patient genomic information on computational space.

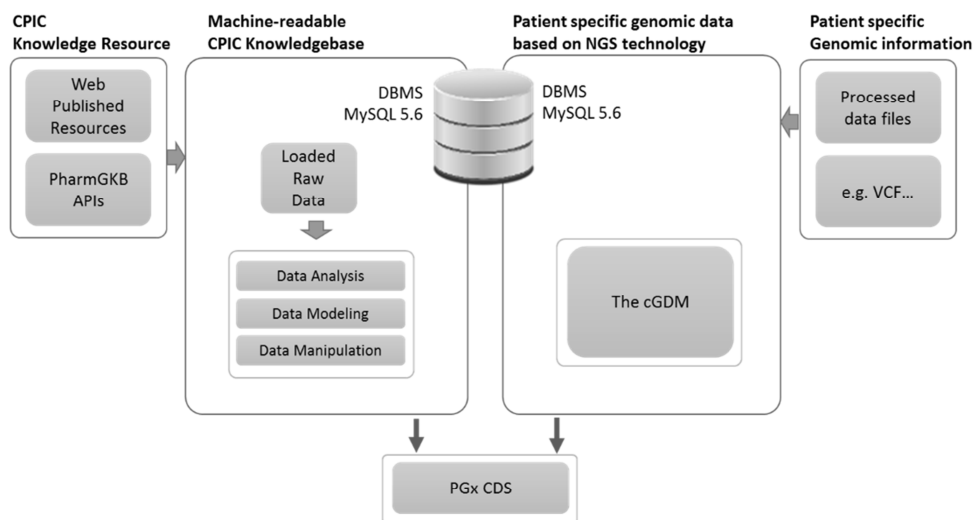
## **2.3. Material and Methods**

### **2.3.1 Material: CPIC guideline as knowledge resource**

The CPIC was formed in 2009 as a shared project between PharmGKB (<https://www.pharmgkb.org>) and the Pharmacogenomics Research Network (PGRN) (<http://www.pgrn.org>). One of the goals of CPIC is to provide peer-reviewed, updated, evidence-based, freely accessible guidelines for gene-drug pairs<sup>6</sup>. All CPIC guidelines adhere to a standard format, and the terms used in CPIC guidelines to describe allele function and phenotype are standardized<sup>7,52</sup>. An ultimate goal for CPIC guidelines is to provide actionable guidelines for clinicians to make more precision decisions for specific drugs when genetic results are available. As a result of the admirable contribution of the consortium, it provides the most world-widely adoptable clinical pharmacogenomic implementation knowledge base. Efforts are underway to make CPIC guidelines more machine-readable, including making the guidelines available in various file formats<sup>53</sup>.

## 2.3.2. Data Collection

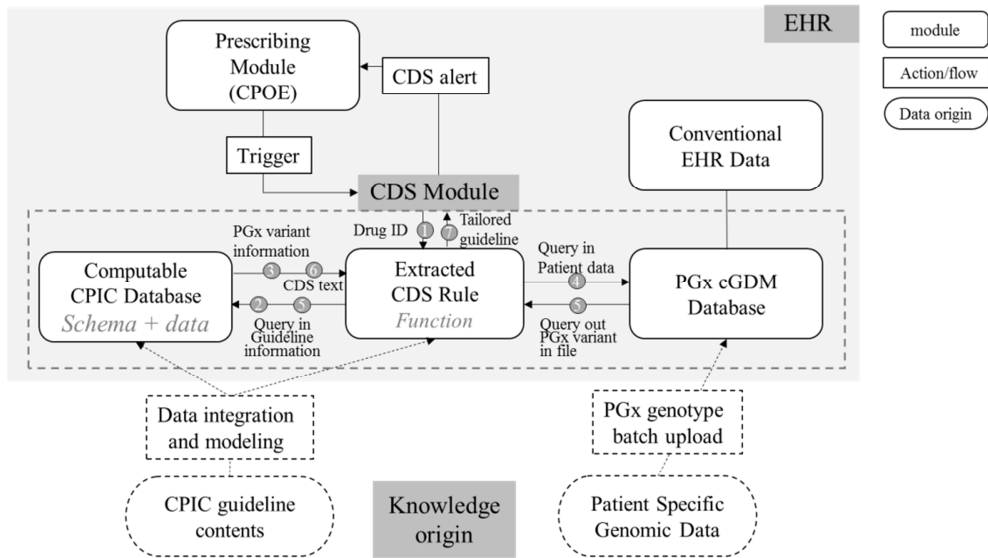
CPIC guideline datasets are first collected between July 10th and August 30th, and updated between 2019 March 15th and March 30th in 2020, via open assessed CPIC webpages and PharmGKB APIs. Collected data items are as follows; guideline list (drug-gene pair information included), drug resource mapping, gene resource mapping, gene allele definition, gene diplotype phenotype, clinical decision support guidelines. Except for the guideline list, other data formats are downloaded in comma-separated values form. Collected datasets are imported to a relational database management system (MySQL 5.6) for exploratory analysis and data-driven restructuring.



\*CPIC: The Clinical Pharmacogenetics Implementation Consortium; DBMS: database management system; PGx CDS: pharmacogenomics clinical decision support system; VCF: variant call format

**Figure 2.1. The configuration of the study environment**

### 2.3.3. Clinical decision support service architecture



**Figure 2.2. Modular implementation of PGx CDS overview**

As discussed in Chapter 1, we perceive patient-specific genomic information as a sub-dimension of representation that reflects the patient's health status. Therefore, we consider the data level integration so that the service architecture ensures agile combined and computation with other sub-dimensional information.

Among collected 6 CPIC content categories, guideline title, drug resource mapping, gene resource mapping, and gene allele definition are used to construct a computable CPIC database (Figure 2.2, middle-left). Others, gene diplotype-phenotype and clinical decision support guideline categories, are applied to CDS rule function that matches PGx variant definition and patient genomic information and selects a personalized PGx CDS to alert given drug prescribing condition. The cGDM is adopted as a

patient-specific genome data model, developed in Chapter 1, to serve as a data layer infrastructure supporting the intellectual interplay between medical experts and informed decision-making.



## **2.4. Results**

### **2.4.1. Collected CPIC guideline and exploratory analysis**

The CPIC guidelines reviewed for machine-readable data conversion are a total of 24 guideline entries (Table 2.1) published to date on the official website<sup>42,54</sup>. Each guideline contains specific information related to certain gene-drug pairs; unique 20 genes and 62 drugs. Each guideline gives well-curated knowledge in forms of procedural subcategories such as drug resource mapping, gene resource mapping, gene allele definition, gene diplotype-phenotype, allele frequency, clinical decision support guidelines. However, mainly due to differences in how each gene affects the drug efficacy or biological characters, the composition of the provided items are varied.

Table 2.2 shows representative CPIC content items and their dataset availability according to each guideline. In the case of drug and gene resource mapping, every dataset is available. HLA-A and HLA-B gene allele definitions are not defined in CPIC standard format due to its unique biological character and high complexity. Gene diplotype-phenotype tables are not provided when the former form of information is not describable, or the only haplotype is existed (G6PD), or the structural variants have a meaningful proportion in the PGx gene. When the items reflect the PGx

drug-gene interpretation process, ensuring the entire item shows the feasibility of building a seamless digitalized pipeline. To explicit clinical decision support workflow and recommendation text files, guidelines that have complete data items are 10; 1) CYP2D6, CYP2C19 and Tricyclic Antidepressants (for 2 of 7 drugs), 2) CYP2D6 and Atomoxetine, 3) TPMT, NUDT15 and Thiopurines, 4) DPYD and Fluoropyrimidines, 5) CYP2D6, CYP2C19 and Selective Serotonin Reuptake Inhibitors, 6) RYR1, CACNA1S and Volatile anesthetic agents and Succinylcholine, 7) CYP2B6 and efavirenz, 8) CYP2D6 and Ondansetron and Tropisetron, 9) CYP2D6 and Tamoxifen, CYP2C19 and Voriconazole, 10) CYP2C9 and NSAIDs (for 7 of 15 drugs).

**Table 2.1. The collected CPIC guideline overview**

CPIC Guideline Title	Drug or Ingredient (unique n = 62)	Gene (n = 20)
HLA-B and Abacavir	abacavir	HLA-B
HLA-B and Allopurinol	allopurinol	HLA-B
CYP2D6, CYP2C19 and Tricyclic Antidepressants	amitriptyline, clomipramine, desipramine, doxepin, imipramine, nortriptyline, trimipramine	CYP2C19, CYP2D6
UGT1A1 and Atazanavir	atazanavir	UGT1A1
CYP2D6 and Atomoxetine	atomoxetine	CYP2D6
TPMT, NUDT15 and Thiopurines	azathioprine, mercaptopurine, thioguanine	TPMT, NUDT15
DPYD and Fluoropyrimidines	capecitabine, fluorouracil, tegafur	DPYD
HLA-A, HLA-B and Carbamazepine and Oxcarbazepine	carbamazepine, oxcarbazepine	HLA-A, HLA-B
CYP2D6, CYP2C19 and Selective Serotonin Reuptake Inhibitors	citalopram, escitalopram, fluvoxamine, paroxetine, sertraline	CYP2D6, CYP2C19
CYP2C19 and Clopidogrel	clopidogrel	CYP2C19
CYP2D6 and Codeine	codeine	CYP2D6
RYR1, CACNA1S and Volatile anesthetic agents and Succinylcholine	desflurane, enflurane, halothane, methoxyflurane, isoflurane, sevoflurane, succinylcholine	RYR1, CACNA1S
CYP2B6 and efavirenz	efavirenz	CYP2B6
CFTR and Ivacaftor	ivacaftor	CFTR
CYP2D6 and Ondansetron and Tropisetron	ondansetron, tropisetron	CYP2D6
IFNL3 and Peginterferon-alpha-based Regimens	peginterferon alfa-2a, peginterferon alfa-2b, ribavirin	IFNL3
CYP2C9, HLA-B and Phenytoin	phenytoin	CYP2C9, HLA-B
G6PD and Rasburicase	rasburicase	G6PD
SLCO1B1 and Simvastatin	simvastatin	SLCO1B1
CYP3A5 and Tacrolimus	tacrolimus	CYP3A5
CYP2D6 and Tamoxifen	tamoxifen	CYP2D6
CYP2C19 and Voriconazole	voriconazole	CYP2C19
CYP2C9, VKORC1, CYP4F2 and Warfarin	warfarin	CYP2C9, VKORC1, CYP4F2
CYP2C0 and NSAIDs	aspirin, diclofenac, celecoxib, flurbiprofen, aceclofenac, ibuprofen, indomethacin, lornoxicam, lumiracoxib, meloxicam, metamizole, nabumetone, naproxen, piroxicam, tenoxicam	CYP2C8, CYP2C9

**Table 2.2. Dataset list and its availability over guidelines**

CPIC Guideline Title	Original Publication Date	Most Recent Update Date	Drug Resource Mapping	Gene Resource Mapping	Gene Allele Definition	Gene Diplotype-phenotype	Clinical Decision Support
HLA-B and Abacavir	April 2012	May 2014			Not available	Not available	Not available
HLA-B and Allopurinol	February 2013	June 2015			Not available	Not available	Not available
CYP2D6, CYP2C19 and Tricyclic Antidepressants	May 2013	October 2019					(2/7)
UGT1A1 and Atazanavir	September 2015	November 2017					Not available
CYP2D6 and Atomoxetine	February 2019	October 2019					
TPMT, NUDT15 and Thiopurines	March 2011	February 2019					
DPYD and Fluoropyrimidines	December 2013	January 2020					
HLA-A, HLA-B and Carbamazepine and Oxcarbazepine	September 2013	December 2017			Not available	Not available	
CYP2D6, CYP2C19 and Selective Serotonin Reuptake Inhibitors	August 2015	October 2019					
CYP2C19 and Clopidogrel	August 2011	March 2017					Not available
CYP2D6 and Codeine	February 2012	October 2019					Not available
RYR1, CACNA1S and Volatile anesthetic agents and Succinylcholine	November 2018	September 2019				Not applicable*	
CYP2B6 and efavirenz	April 2019	No updates					
CFTR and Ivacaftor	March 2014	May 2019				Not available	Not available
CYP2D6 and Ondansetron and Tropisetron	December 2016	October 2019					
IFNL3 and Peginterferon-alpha-based Regimens	February 2014	No updates				Not available	Not available
CYP2C9, HLA-B and Phenytoin	November 2014	No updates			Not available	Not available	Not available
G6PD and Rasburicase	August 2014	September 2018				Not available	Not available
SLCO1B1 and Simvastatin	October 2014	No updates					Not available
CYP3A5 and Tacrolimus	July 2015	No updates					Not available
CYP2D6 and Tamoxifen	January 2018	October 2019					
CYP2C19 and Voriconazole	December 2016	No updates					
CYP2C9, VKORC1, CYP4F2 and Warfarin	December 2016	No updates				Not applicable*	Not available
CYP2C9 and NSAIDs	March 2020	No updates	(7/15)	(1/2)	(1/2)	(1/2)	(7/15)
Number of available files grouped by guidelines			23	23	20	15	11

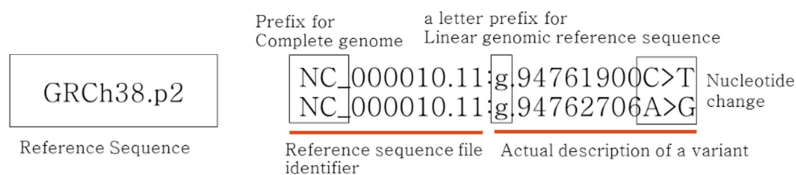
## 2.4.2. Data integration and modeling

In this section, we briefly examine each CPIC content item in terms of its attribute and value set. On top of that, *CPIC guideline title* contains drug-gene pair information at ingredient or drug class level. *Drug resource mapping* file provides for each drug of ingredient, respectively, which has four attributes; ‘Drug or Ingredient,’ ‘Source,’ ‘Code Type,’ ‘Code.’ Source attribute has a member of RxNorm, DrugBank, ATC, PharmGKB. In summary, this item provides definitions of drugs that can be identified in four representative external drug knowledge bases. *Gene resource mapping* file is also expressed in the same attribute set, and provides unique indexes of 4 different external genome knowledge bases for each gene; PharmGKB, Ensembl, NCBI, HGNC.

*The Gene allele definition table* can be divided into four districts when clustered with similar value properties as below (Figure 2.3). This table is a collection of PGx variant information in a gene. For example, we can start \*4 in the C district. At the same line in D district, we can find the alternative allele Y and G. In the first line of those, reference allele C and A are shown. We could make the exact HGVS nomenclature when combine assigned A+B district. In this case, CYP2C19 \*4 consists of two variants; NC\_000010.11:g.94761900C>T and NC\_000010.11:g.94762706A>G. This expression is interoperable with any line of A+B, for example, rs12248560 and rs28399504 in terms of rsID from NCBI dbSNP. The machine cannot

interpret the table, evidently. We naturally extracted codified token from part A. As a consequence, we abstracted each value pattern and named its properties. As a consequence of data modeling and reconstruction, district A of gene allele definition table over 17 gene files results in Table 2.3.

(a) HGVS nomenclature for CYP2C19 \*4 variant



(b) CYP2C19 allele definition table from CPIC (and PharmGKB)

	A	B	C	D	E	F	G	H	I	J	K	L
1	GENE: CYP2C19											
2	Nucleotide change to gene from <a href="http://www.pharmvar.org">http://www.pharmvar.org</a>											
3	-806C>T	1A>G	7C>T	10T>C	50T>C	55A>C	83A>T	151A>G	12401C>T	12416C>T	12455G>C	
4	Effect on protein (NP_000760.1)	5' region	M1V	P3S	F4L	L17P	I19L	K28I	S51G	R73C	H78Y	G91R
5	Position at NC_000010.11 (Homo sapiens chromosome 10, chr10:38,929)											
6	g.94761900C>T	g.94762706A>G	g.94762712C>T	g.94762715T>C	g.94762755T>C	g.94762760A>C	g.94762788A>T	g.94762856A>C	g.94775106C>T	g.94775121C>T	g.94775160G>C	
7	Position at NG_008384.3 (CYP2C19 RefSeqGene, forward relative to chromosome)											
8	g.4220C>T	g.5026A>G	g.5032C>T	g.5035T>C	g.5075T>C	g.5080A>C	g.5108A>T	g.5176A>G	g.17426C>T	g.17441C>T	g.17480G>C	g.17480G>C
9	rs12248560	rs28399504	rs367543002	rs367543003	rs55752064	rs17882687			rs145328984		rs118203756	
10	CYP2C19 Allele											
11	*1	C	A	C	T	T	A	A	A	C	C	G
12	*2											
13	*3											
14	*4	Y	G									
15	*5											
16	*6											
17	*7											
18	*8											
19	*9											
20	*10											
21	*11											
22	*12											
23	*13											
24	*14					C						
25	*15						C					
26	*16											
27	*17	T										
28	**											

Figure 2.3. Gene allele definition table example

- (a) Variant expression in HGVS nomenclature and its meaning.
- (b) Gene allele definition table collected from CPIC guideline contents. File has for distinctive areas; A) Reference Sequence level related values; B) Detail location and variant information given A; C) Star allele nomenclature; D) actual variant information at locus A+B

**Table 2.3. Reference Sequence Information for Locus assignment**

HGNC_Gene_Symbol	Chromosome	Reference_Sequence_Source	Reference_Assembly	Complete Genomic Molecule ID	Genomic Region ID	Protein ID
CACNA1S	1	NCBI RefSeq	GRCh38.p7	NC_000001.11	NG_009816.1	NP_000060.2
CFTR	7	NCBI RefSeq	GRCh38.p2	NC_000007.14	NG_016465.3	NP_000483.3
CYP2B6	19	NCBI RefSeq	GRCh38.p2	NC_000019.10	NG_007929.1	NP_000758.1
CYP2C19	10	NCBI RefSeq	GRCh38.p2	NC_000010.11	NG_008384.3	NP_000760.1
CYP2C9	10	NCBI RefSeq	GRCh38.p2	NC_000010.11	NG_008385.1	NP_000762.2
CYP2D6	22	NCBI RefSeq	GRCh38.p2	NC_000022.11	NG_008376.3	NP_000097.3
CYP3A5	7	NCBI RefSeq	GRCh38.p2	NC_000007.14	NG_007938.1	NP_000768.1
CYP4F2	19	NCBI RefSeq	GRCh38.p2	NC_000019.10	NG_007971.2	NP_001073.3
DPYD <sup>+</sup>	1	NCBI RefSeq	GRCh38.p2	NC_000001.11	NG_008807.2	NP_000101.2
G6PD	X	NCBI RefSeq	GRCh38.p2	NC_000023.11	NG_009015.2	
IFNL3 <sup>+</sup>	19	NCBI RefSeq	GRCh38.p2	NC_000019.10	NG_042193.1	
NUDT15	13	NCBI RefSeq	GRCh38.p7	NC_000013.11	NG_047021.1	NP_060753.1
RYR1	19	NCBI RefSeq	GRCh38.p2	NC_000019.10	NG_008866.1	NP_000531.2
SLCO1B1	12	NCBI RefSeq	GRCh38.p2	NC_000012.12	NG_011745.1	NP_006437.3
TPMT	6	NCBI RefSeq	GRCh38.p2	NC_000006.12	NG_012137.2	NP_000358.1
UGT1A1	2	NCBI RefSeq	GRCh38.p2	NC_000002.12	NG_002601.2	NP_000454.1
VKORC1	16	NCBI RefSeq	GRCh38.p2	NC_000016.10	NG_011564.1	

\* HLA-A, HLA-B, CYP2C8 Allele Definition Tables are not available

<sup>+</sup> source - <https://www.pharmgkb.org/page/pgxGeneRef>

Table 2.4 shows information density and terminology variation in the value field of the gene allele definition table. Among 17 available PGx gene variant information, 11 genes adopted star allele nomenclature<sup>55</sup>, and G6PD has its own nomenclature, and WHO class to designate distinctive functions on drug reaction mechanism<sup>56</sup>, two genes have a single PGx variant. Almost of PGx variant over 17 genes are single nucleotide variant (SNV) or insertion/deletion (InDel), but CYP2B6 and CYP2D6 include 14 and 4 copy number variants respectively. The number of different loci that appear in CPIC guideline contents is 702.

**Table 2.4. Gene allele definition table data profiles**

HGNC Gene Symbol (n=20)	No of Loci	No of assigned designation	Matrix size	Example values	
CACNA1S	2	2	4	Reference	c.520C>T
CFTR	40	42	1,640	2789+5G->A	S977F
CYP2B6 <sup>+</sup>	38	38	1,444	*1	*38
CYP2C19	34	34	1,156	*1	*37
CYP2C8				not available	
CYP2C9	58	61	3,538	*1	*61
CYP2D6 <sup>+</sup>	128	146	18,560	*1	*9xN, *139
CYP3A5	8	8	64	*1	*9
CYP4F2	2	2	4	*1	*3
DPYD	15	93	1395	Reference	c.1003G>T (*11)
G6PD	173	187	32,351	202G>A_376A>G_1264C>G	Yunan <sup>++</sup>
HLA-A				not available	
HLA-B				not available	
IFNL3	single variant(g.39248147C>T)			rs12979860 reference (C)	rs12979860 variant (T)
NUDT15	17	19	323	*1	*19
RYR1	43	48	2,064	Reference	c.1021G>A
SLCO1B1	29	37	1,073	*10	*9
TPMT	39	43	1,677	*1	*9
UGT1A1	5	10	50	*1	*80+*37
VKORC1	single variant(g.3109638C>T)			rs9923231 reference (C)	rs9923231 variant (T)

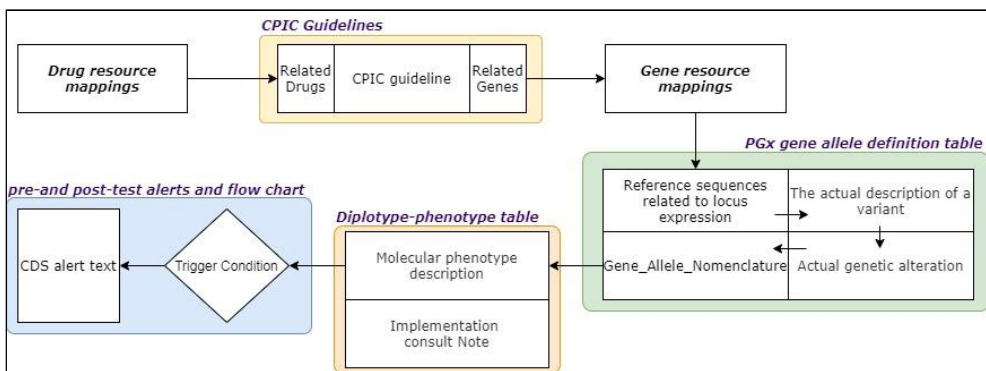
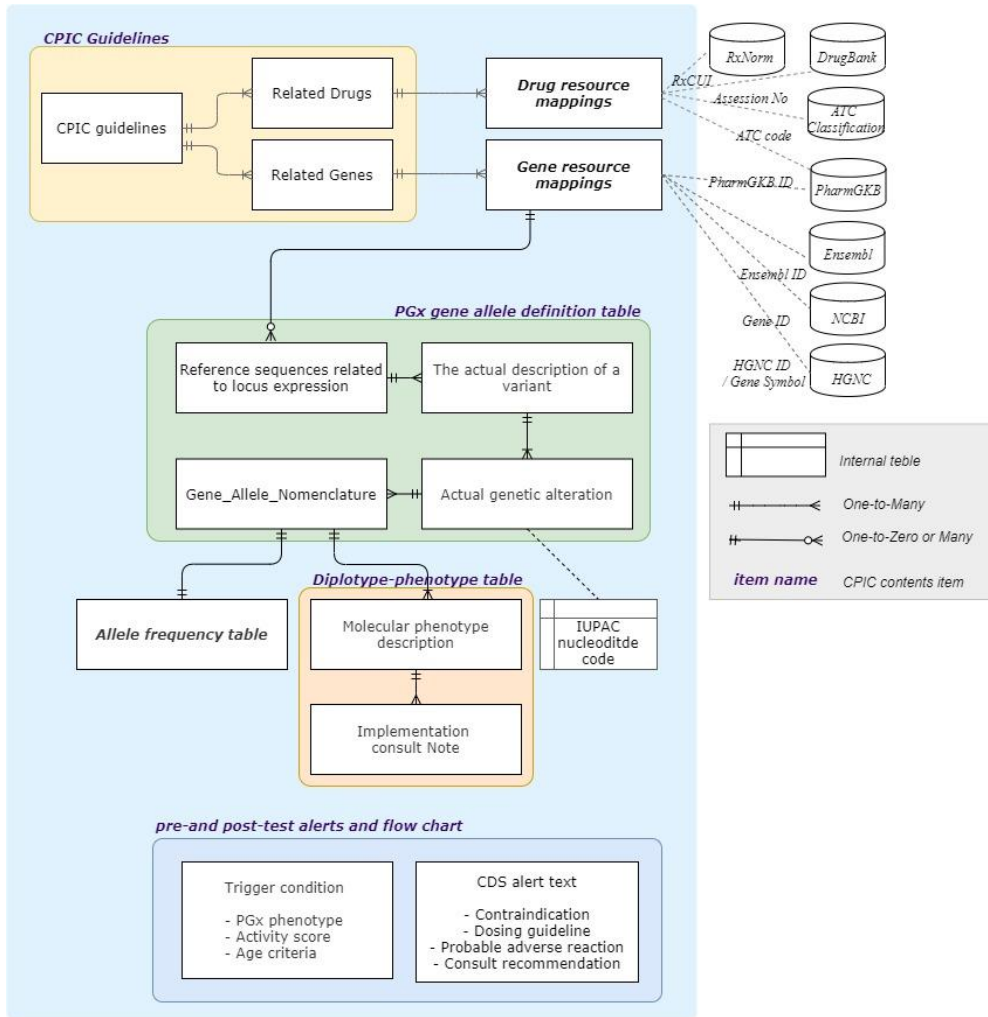
\* Star allele available gene count: N=11 (CYP2B6; CYP2C19; CYP2C9; CYP2D6; CYP3A5; CYP4F2; DPYD; NUDT15; SLCO1B1; TPMT; UGT1A1)

<sup>+</sup> CYP2D6 and CYP2B6 include 14 and 4 copy number variants respectively

<sup>++</sup> G6PD Genetic Variant Nomenclature and WHO Class







**Figure 2.5. Snapshot of CPIC guidelines content structure converted to be computable**

### 2.4.3. CDS Rule Extraction

The pre-and post-test alert file consists of two sheets; 'Pre- and post-test alerts,' 'Flow Chart.' Flow chart helps end-user's understanding also easily convert to a conditional phrase in computer language. However, the trigger condition, a particular exact subset, is offered by the 'Pre- and post-test alerts' sheet. In other words, conditional trigger information for CDS function is distributed in two sheets. Firstly, 'Flow Chart' has one common condition whether the patient's genomic information is available or not. There are two exceptions over three guidelines; one is filtering weight over 40 kg criteria in case of 'CYP2B6 and efavirenz', the other has branched alert message between for pediatrics and adults in case of 'CYP2D6 and Atomoxetine' and 'CYP2C19 and Voriconazole'. The latter type of exception does not appear in 'Flow chart' but implied to provide two alert text message columns in 'Pre- and post-test alerts.' Through this separation and regrouping process, we constructed trigger condition, alert message, and trigger condition-alert message relation.

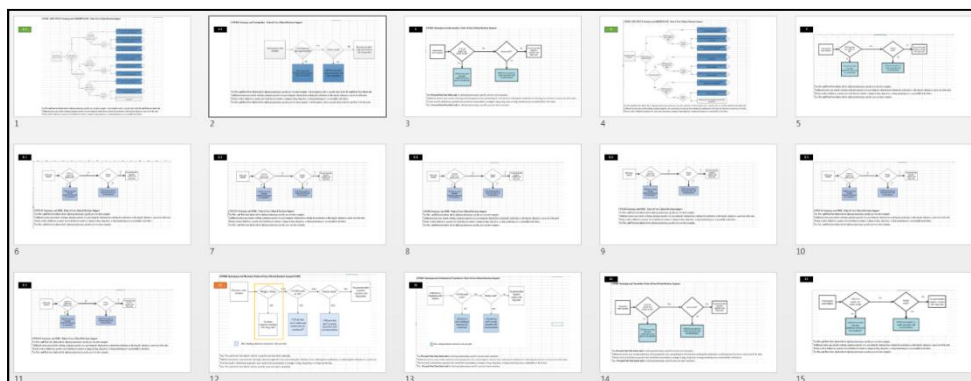
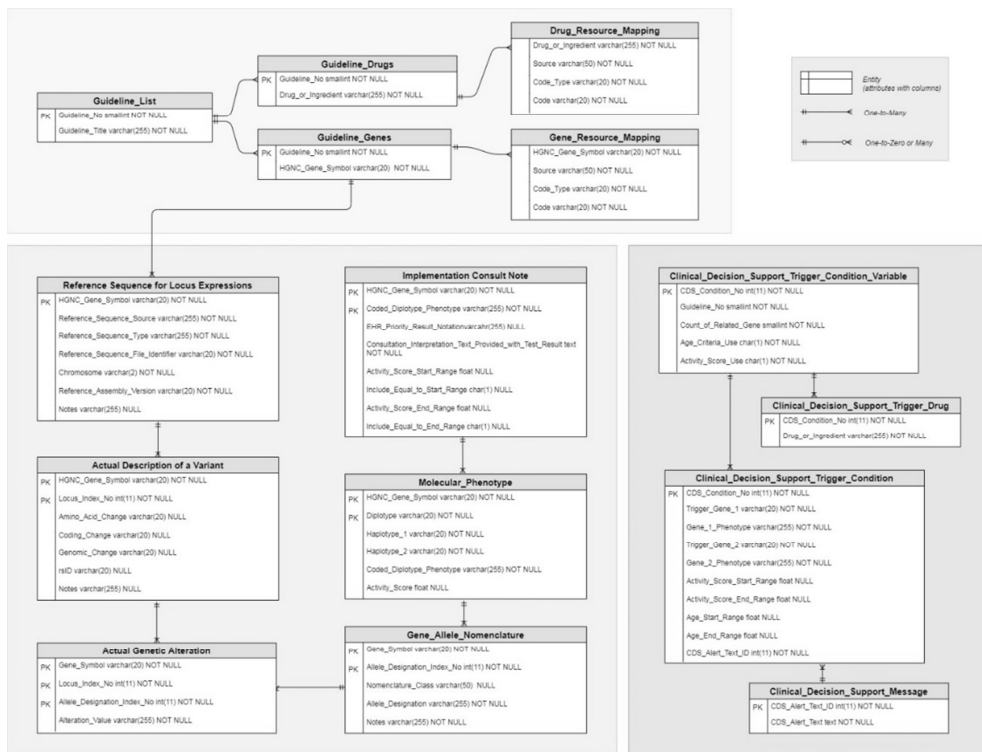


Figure 2.6. Collection of 'Flow chart' over available 15 guidelines

#### **2.4.4. Structured database construction**

Finally, we have constructed a machine-readable CPIC guideline database in the form of a relational database. The database includes 15 tables and 46 unique attributes (Figure 2.7). Interestingly, the left and right parts of the ERD are separated.

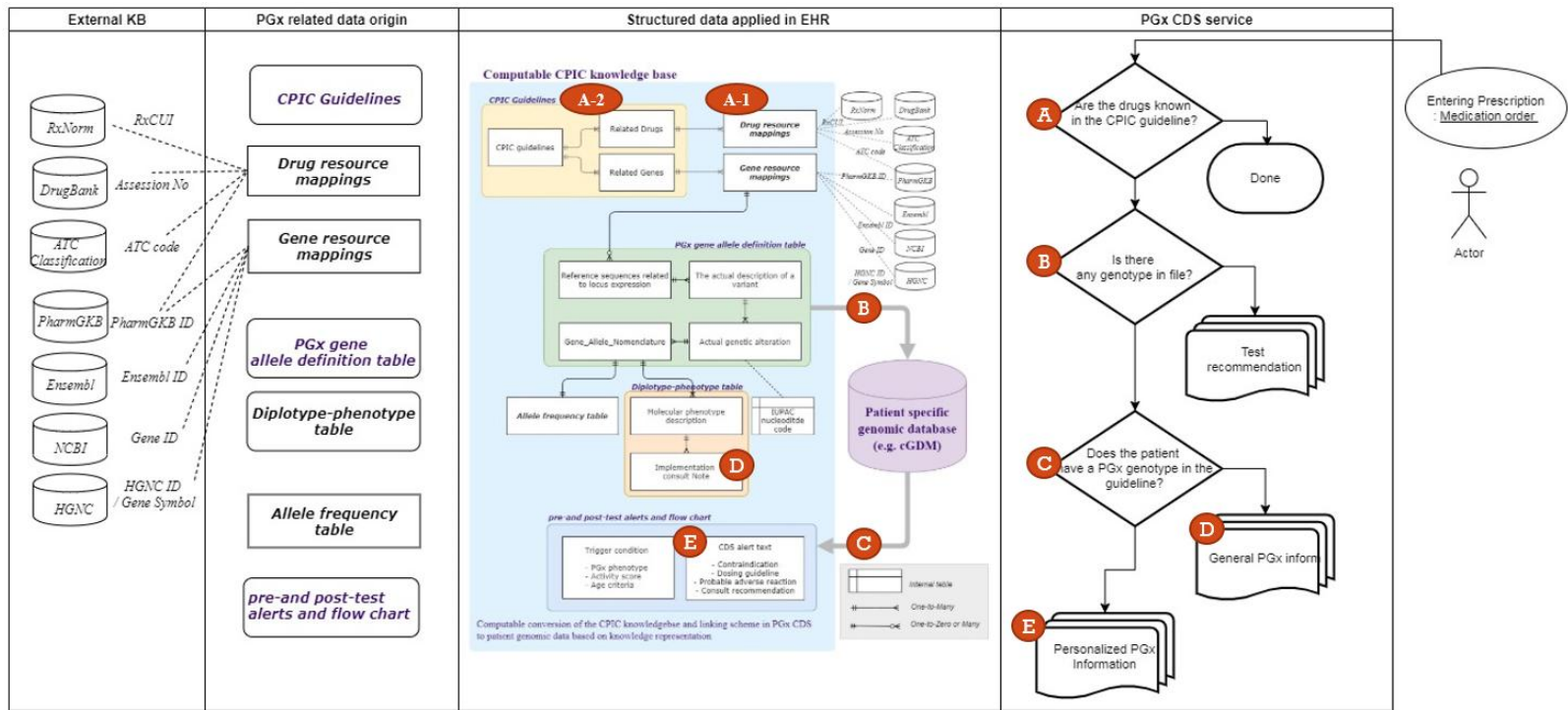
The left side represents the knowledge that declares PGx related variant definition and converts those findings into interpretable codified phenotypes for each drug-gene pair for which the guideline is targeted. The right part is a guide that provides a tailored CDS message when an individual's codified phenotype and prescribing drug ingredient is known. The CDS message contents could break down a set of properties comprised of contraindication, dose adjustment guidelines, probable adverse reactions, and consult recommendations to the clinical pharmacist for further consideration. However, in this study, the CDS alert text was not structured because the distribution of the corresponding attributes when segmented by sentence was irregular.



**Figure 2.7. Entity-relationship diagram of reconstructed database based on CPIC contents**

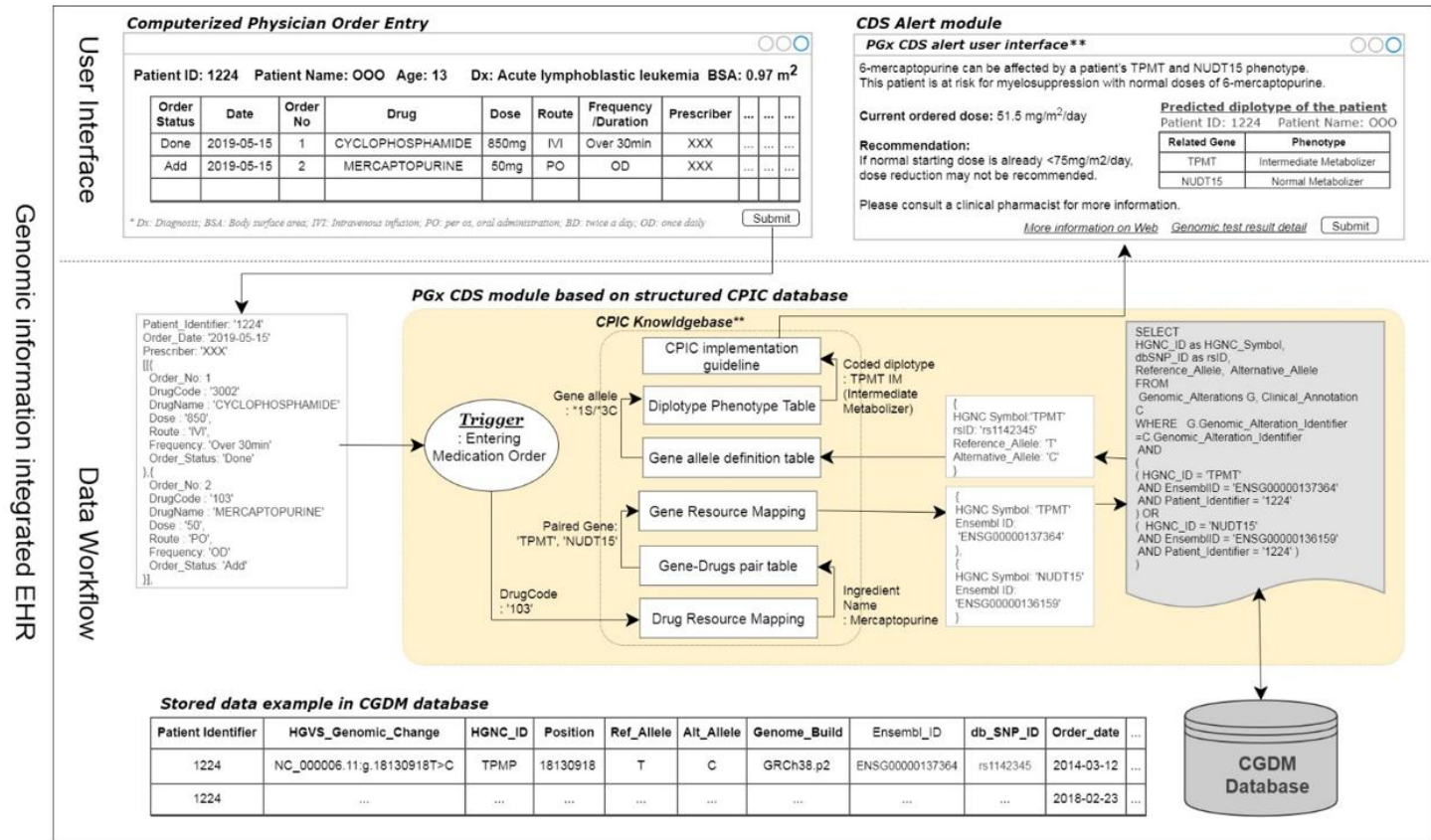
### **2.4.5. PGx CDS service module**

Figure 2.8 shows the developed PGx CDS service module. When the system is evoked, the CDS module looks at patient genome data stored in the EHR server and returns potential phenotype according to the CPIC PGx variant definition. Also, the module queries out individualized recommendations for the prescriber. The novelty of this modular solution is the machine-readable conversion of the CPIC guideline and seamless function execution in a single EHR system. Data modeling reveals four components of the CPIC knowledge resource. The first is targeted phenomena identifier, CPIC guideline title, and drug-gene pair information. The effort to provide curated and filtered PGx variant definition lists with expert knowledge with clinical relevance. Then, they try to capture related annotation systems for interpretation, such as the star allele system. This information is presented in the nomenclature field in the Gene-allele definition table and codified data field in the Diplo-type-phenotype table. Final CDS alert texts are given with the assumption that a person who looks at guidelines knows the specific genotype information. A data flow crack is found in here, but we could bridge this gap with the patient-specific genome database proposed in Chapter 1. Finally, seamless PGx CDS are enabled shown in Fig. 2.9. Through the data collection and reconstruction process, we could briefly explore the colossal landscape of their accomplishment. For enhancing usability, CPIC does process standardization along with the development of new guidelines.



**Figure 2.8. PGx CDS module architecture**

(A) In this step, service refers to the data in (A-1) and (A-2) to check if the prescribed drug has relevance to the pharmacogenomic guideline. (B) Execute a query into a patient-specific genome database by referring to pharmacogenomic variant information declared in the CPIC knowledge base. (C) The search result includes the possession of genomic information of the patient is returns in the form of a phenotype. (D) Provide general guidance on the drug-dielectric guidelines. (E) Provide individualized PGx CDS alert message



**Figure 2.9. PGx CDS module integration scenario with dataflow**



## **Chapter 3. Clinical Application of Clinical Genome Data Model: Integrating Star Allele and HLA Data Models**

“An ideal nomenclature would be one that is entirely unambiguous. One might hope that a geneticist of the year 2493 could pick up a 1993 copy of *The American Journal of Human Genetics* and quickly understand, from the designation of a mutation and without extensive study of other sources, the location of a nucleotide change. However, the complexity of the genome and its functions is such that a perfect nomenclature is unachievable.<sup>57</sup>” (Ernest Beutler, 1993)

### **3.1. Introduction**

As Beutler envisioned, the perceived complexity of the genomics is expanding, and a perfect nomenclature is not achieved yet. However, there is some accomplishment, such as the HGVS nomenclature and star allele system that helps effective communication between scientists. The HGVS<sup>58</sup> nomenclature has advantages in figure out a specific locus from the nomenclature. Nevertheless, it does not specify a specific reference sequence. Thus the same variant could be described using different reference sequences, which might cause confusion. Furthermore, the expression is not scalable enough to express functional combinations. Thus star allele nomenclature was introduced in 2006<sup>55</sup>. The star allele nomenclature could contain multiple-locus in one name (so-called star), and one locus could be placed in redundant stars. The star-allele nomenclature is the result of efforts to standardize genetic polymorphism annotation for the cytochrome P450

genes. As clinical pharmacogenetic testing becomes widespread, this system has played a vital role in effectively delivering the patient's genotype and predicted clinical phenotype. As genomics research expands, the system remains a valuable tool for the broader community of genetic researchers to exploit our ever-improving ability to catalog variability in the human genome<sup>55</sup>. However, as scientific discoveries accumulate, the number of assigned stars is increased, and the complexity of the naming system itself is also expanded. For example, \*1 is mostly accepted as a reference sequence functionality, but a few exceptions occur as known population distribution of the variants are changed. In addition, there are highly curated representative registries according to research interest so we could use those naming system as an auxiliary identifier. We prove the concept in Chapter 2 using PGx variant definition construction and interoperable interpretation in the data of the patient-specific genomic information in cGDM.

Furthermore, there are independent nomenclatures such as the human leukocyte antigen (HLA) system. The HLA system<sup>59</sup> is known to be the most polymorphic in human. The HLA polymorphism is not evenly spread throughout the molecule but is clustered in the antigen-binding groove<sup>60</sup>. HLA is a protein that plays a vital role in our body's immune function with a wide variety of allele types.<sup>61</sup> HLA diversity is particularly important in organ transplantation because transplant recipients and donors with different serological HLA proteins will exhibit organ transplant

rejection<sup>62</sup>. Therefore, transplant recipients must perform HLA screening before transplantation. Recently, HLA diversity has been reported to cause severe drug hypersensitivity as well as organ transplantation<sup>63</sup>. However, the HLA results of transplant patients and donors have not been used to predict future adverse drug reactions. This is because the HLA test is performed in various ways, from a simple serological test to an NGS test. Besides, while the nomenclature that represents the HLA test results is continuously updated, the test results simply have been stored in free text in the electronic medical record (EMR)<sup>64</sup>.

## **3.2. Purpose of Research**

Firstly, the HLA database is designed to be used in clinical practice with data-driven approach. Construction of HLA DB linked in hospital information system could bring clinical pharmacogenomics information to physicians. Secondly, the HLA database is covering multiple test methods enable to protect from the harm due to the non-use of health-related data<sup>65</sup>. Ultimately, we try to validate the model consistency to cope with the evolving annotation systems by construction of HLA database.

### **3.3. Material and Methods**

We used the dataset extracted the results of the HLA test performed and demographics of patients using SUPREME® between February 2002 and June 2018, a clinical data warehouse of Seoul National University Hospital<sup>66</sup>. With a data-driven approach, we could extract clinical context enriched entities and attributes. Also, HLA nomenclature has been adopted as the primary material for designing and elaborating the HLA entity.

We designed the cGDM HLA as a physical data model in a relational database on MySQL 5.6 in an agile manner. Data-driven modeling is comprised of data mining and clarification of implicit properties and relations<sup>67</sup>.

## **3.4. Results**

### **3.4.1. Summary of collected dataset**

Collected dataset from SUPREME<sup>®</sup> has 11,287 records for 11,144 patients; 4,039 male and 7,105 female patients, including 2,642 high-resolution tests, 5,835 low-resolution tests, and 2,810 tests. Gathered data fields are shown in Table 3.1 below. We filtered these fields with data existence, and remove its redundancy. Then, the reclassification of each field was conducted compared to the cGDM schema. Unlike the expectation that it will be a true subset of the existing cGDM schema, except for the HLA nomenclature, unique properties remain that called 'related patient.' This is caused by a unique clinical context when the HLA test ordered, organ transplantation. In this case, donor-recipient tag information or family relationship information has significant meaning for test result application. For internal integrity, we decide to capture this information with the appended entity for further use.

**Table 3.1. Extracted field list gathered from the EHR records**

<b>Document item name</b>	<b>full name or example data</b>
MRN	patient identification no
PatientDOB	Birthdate
PatientName	patient name
PatientSex	patient sex
TestCode	test code
TestDate	test date
TestName	test name
Name	name (data not found)
PatientType	donor/recipient
diagnosis	dx (data not found)
RelatedPatientsNo	relatives(data not found)
A1_gene	A11
A1_allele	*11
A2_gene	A24
A2_allele	*24
B1_gene	B7
B1_allele	*07
B2_gene	B62
B2_allele	*15
C1_gene	Not tested
C1_allele	Not tested
C2_gene	Not tested
C2_allele	Not tested
DR1_gene	DR1
DR1_allele	*01:01g
DR2_gene	DR4
DR2_allele	*04:03g
DQ1_gene	Not tested
DQ1_allele	Not tested
DQ2_gene	Not tested
DQ2_allele	Not tested
RelatedPatientName	NA

### 3.4.2. HLA data model

HLA entity is added in forms of tokenized HLA nomenclature. HLA gene classes and its subtypes are represented in Supplementary information 2. Because this nomenclature is logically well developed, one of the major challenges was in its version control. Opportunely, the HLA community provides a version conversion tool and table as a text file. We parsed the HLA test results from the dataset with nomenclature logic and normalized its values with mass conversion when we uploaded the dataset to the DBMS table.

Common entities with the cGDM

Extension entity represent HLA nomenclature

Subject			
1	Subject Identification Number	Subject_ID	GDM generate
2	Subject name	Subject_Name	Name
3	Patient Identification Number	Patient_ID	MRN
4	Birth date	Birth_Date	PatientDOB
5	Gender	Gender	PatientSex
6	Race	Race	
7	Ethnicity	Ethnicity	
8	Institution code	Institution_Code	GDM generate
9	Register Identification Number	Register_ID	GDM generate
10	Submission date	Submission_Date	GDM generate

Specimen			
1	Specimen Identification Number	Specimen_ID	GDM generate
2	Subject Identification Number	Subject_ID	GDM generate
3	Specimen origin type	Origin_Type	
4	Body site	Body_Site	
5	Body site code	Body_Site_Code	
6	Physical type	Physical_Type	
7	Physical type code	Physical_Type_Code	
8	Specimen type	Specimen_Type	
9	Specimen block Identification Number	Block_ID	SampleNo1
10	Specimen accession Identification Number	Accession_ID	
11	Collection date	Collection_Date	DateOfTest1
12	Received date	Received_Date	
13	Differentiation state	Differ_State	

Protocol			
1	Protocol Identification Number	Protocol_ID	GDM generate
2	Specimen Identification Number	Specimen_ID	GDM generate
3	Test name	Test_Name	TestName
4	Type of sequencing	Sequencing_Type	GDM generate
5	Ordered date	Order_Date	
6	Order Identification Number	Order_ID	
7	Lab name	Lab_Name	
8	Reagent	Reagent	
9	Received Date	Received_Date	
10	Bioinformatician	Bioinformatician	
11	Analytics institution	Analytics_institution	
12	Sequencer Identification Number	Sequencer_ID	
13	Panel Identification Number	Panel_ID	
14	Pipeline Identification Number	Pipeline_ID	
15	SNV/indel pipeline detail iden	SNV_inDel_PD_ID	
16	Copy Number Variation pipeline	CNV_PD_ID	
17	Translocation pipeline detail iden	Translocation_PD_ID	
18	Microsatellite Instability Alterat	MSI_PD_ID	
19	Tumor mutation burden pipeline	TMB_PD_ID	
20	Document creation date	Docu_Creation_Date	
21	Document version	Docu_Version	

HLA Star Allele			
1	Star Allele Identification Number	Star_Allele_ID	GDM generate
2	Protocol Identification Number	Protocol_ID	
3	A1_gene	A1_gene	A11
4	A1_allele	A1_allele	*11
5	A2_gene	A2_gene	A24
6	A2_allele	A2_allele	*24
7	B1_gene	B1_gene	B7
8	B1_allele	B1_allele	*07
9	B2_gene	B2_gene	B62
10	B2_allele	B2_allele	*15
11	C1_gene	C1_gene	Not tested
12	C1_allele	C1_allele	Not tested
13	C2_gene	C2_gene	Not tested
14	C2_allele	C2_allele	Not tested
15	DR1_gene	DR1_gene	DR1
16	DR1_allele	DR1_allele	*01:01g
17	DR2_gene	DR2_gene	DR4
18	DR2_allele	DR2_allele	*04:03g
19	DQ1_gene	DQ1_gene	Not tested
20	DQ1_allele	DQ1_allele	Not tested
21	DQ2_gene	DQ2_gene	Not tested
22	DQ2_allele	DQ2_allele	Not tested

Figure 3.1. HLA Database design merged in the cGDM schema



## General Discussion<sup>†</sup>

The rapid accumulation of genome information has led to a paradigm shift in medicine. Nevertheless, significant barriers remain to overcome inflection points. Through multi-disciplinary analysis and consideration of this phenomenon, we determined two main causes: 1) reliability-related result variance among numerous pipelines and processes, and 2) the unique data structure of genome information. Since these two causes have mutual influences, an integrative solution may be more effective than a point solution. Moreover, we foresee that GIS will become an essential component of an integrated clinical information system in the precision medicine era. In this context, this cGDM could serve as a genomic information representation scheme enabling the intellectual interaction between medical experts and informed decision making, ultimately contributing to the enhancement of personal genomic data utilization at the point of care.

---

<sup>†</sup> The part of the dissertation general discussion published in following paper: Kim, H. J., Kim, H. J., Park, Y., Lee, W. S., Lim, Y., & Kim, J. H. (2020). clinical Genome Data Model (cGDM) provides interactive clinical Decision Support for precision Medicine. *Scientific reports*, 10(1), 1-13.

## **The GDM as an Infrastructure for a GIS**

We recommend the GDM as a genomic information representation scheme for clinical purposes. To ensure the convenient and appropriate clinical use of genomic data, medical informatics technology is needed as part of the infrastructure supporting the integration of clinic and genomic layers of information<sup>68,69</sup>. Given the multi-level and multi-dimensional nature of health, clinicians must perform decision-making for a given case based on a collection of segmented data representing a person's health, including laboratory data, imaging, and observation data assessed by experts. Currently, a clinical information system is typically used as a core tool for supporting this knowledge in a management process. To broaden perspectives in the era of precision medicine, we propose a concept of genome information system (GIS) as an integral component of an expected clinical information system for precision medicine (Fig. 1.1).

The cGDM can serve as a data-level infrastructure for implementation of the GIS. When decision makers face unfamiliar health-status measurements, determining clinical significance and meaning is challenging<sup>69,70</sup>. The cGDM was designed to preserve genomic information at an appropriate information scale and granularity covering the procedural dimension, which is related to the confidence level as a clinical measurement for clinical application. The design of the cGDM allows processed genomic data for a general purpose to be stored and merged with

existing clinical data, providing outputs in an interoperable data format. Likewise, sequencing analysis, data processing, and presentation of processed information can be managed in a form that can be explicitly confirmed. Once data are uploaded to the cGDM-based database, they serve as a supportive backbone for any downstream functional applications such as report generation or a clinical decision support system. (e.g., Fig 8; Fig 3)

To develop a system for the systematic management of genomic data, it is necessary to unify its data structure with that of other existing components of clinical information systems, ensuring sufficient reliability for identifying the original data generation process<sup>71</sup>

## **Current Approach to Genomic Data Management**

The Health Level 7 (HL7) clinical genomics working group provided a model for health information exchange and Fast Health Interoperability Resources (FHIR) genomics, a model that integrates genetic and clinical information via the HL7 interfacing standard<sup>70,72</sup>. FHIR provides standards for medical and genomic information exchange and offers open-source and open application programming interfaces (APIs) that can easily be applied in clinical fields among heterogeneous data sources. FHIR and FHIR genomics have made substantial contributions toward the implementation of medical information exchange and are drawing electronic health records vendors' attention in this respect.

The Global Alliance for Genomics & Health (GA4GH) was established in 2013 to develop public tools that enable the responsible, voluntary, and secure sharing of clinical and genomic data<sup>73</sup>. The federated approach of GA4GH does not involve the storage and management of data in centralized data repositories. Instead, it provides an API that enables users to request and share data while holding data for institutions<sup>74</sup>.

The FHIR and GA4GH consortium of HL7 were developed with the intention to facilitate the exchange of genomic and clinical data among multiple sites. Both resources have a common character as a form of information exchange at the communication level. These systems use the latest web technologies such as the representational state transfer (REST)

API to make it easier for developers to implement clinical applications or information systems in the healthcare industry.

The International Organization for Standardization (ISO) Technical Committee 215 (Medical Information) has proposed genomic information standards. ISO 27720:2009 (GSVML; General Sequence Variation Markup Language) is a standard that defines how genetic sequencing variation information is exchanged based on XML. The scope of this standard is in the data exchange format and does not include the database schema. Although all genetic sequencing is within the standard's scope, the SNP is the main target of this standard. Another standard for more specific clinical utilization of genomic information is ISO/TS 20428 Health information - Data elements and their metadata for describing structure information in electronic health records established in 2017. Additionally, ISO/CD TS 23357 Genomic informatics – clinical genomics data sharing specification for next generation sequencing is under development state.

**Table 4.1 Comparison table of characteristics of related resources**

Resource	Publication (year)	Data management scope			Computability			Purpose	Organization
		Storage	Exchange	Clinical data linkage	Patient identification	for CDS rule	for report generation		
cGDM	2020	O	X	O	O	O	O	Data level EHR integration	SNUBI
OMOP G-CDM	2019	O	X	O	X	X	X	Federated Research Network	OHDSI
FHIR Genomics	2020 (2015~)	X	O	O	O	O *	O	Information Exchange	HL7
GA4GH Genomics API	in progress (2015~)	X	O	X	X	X	X	Data interchange for bioinformatics research	GA4GH
ISO/TS 20428:2017	2017	X	O	O	O	X	O	Structuring sequencing report	ISO/TC215 (Health Informatics)
ISO/TS 25720:2009	2009	X	O	X	X	X	O	SNP data exchange	ISO/TC215 (Health Informatics)
GDC	2017	X	O	X	X	X	X	Cancer related genomic data sharing	NIH NCI

\*via SMART on FHIR, CDS Hooks, HL7 Inforbutton

cGDM: clinical Genome Data Model; OMOP G-CDM: Observational Medical Outcomes Partnership Genome Common Data Model; FHIR: Fast Healthcare Interoperability Resources; GA4GH: Global Alliance for Genomics and Health; ISO/TS 20428:2017: Health informatics - Data elements and their metadata for describing structured clinical genomic sequence information in electronic health records; ISO/TS 25720:2009: Health informatics - Genomic Sequence Variation Markup Language(GSVML); API: Application Programming Interface; GDC: Genomic Data Common; SNP: Single Nucleotide Polymorphism; SNUBI: Seoul National University Biomedical Informatics; OHDSI: Observational Health Data Science and Informatics; HL7: Health Level Seven; NIH: National Institutes of Health; NCI: National Cancer Institute

Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM) aims to conduct distributed research across observational databases in multiple institutions using a common data model approach. Genomic Common Data Model (G-CDM) proposed as an extension part of OMOP-CDM represents genomic information<sup>75</sup>. Focused on research purposes, the granularity and scale of knowledge representation have limited for multifaceted clinical application.

The almost resources discussed earlier focus on data exchange formats for utilization rather than on EHR integration of genomic information. Therefore, the system is being developed by designing functions first rather than expressing knowledge of the genomic information itself, and by further defining the element whenever the function is added. This development methodology has strength for easy and fast software function development. On the one hand, however, all of reviewed resources are on a separate layer from the ground level schema in data management.

## **The cGDM: A Step beyond the Capabilities of the Existing Systems**

To develop a system for the systematic management of genomic data, it is necessary to unify the data structure with that of other components of clinical information systems, and to ensure sufficient reliability for identifying the data generation process<sup>18</sup>. Conventional systems have focussed on data structure unification issues first, to harmonise heterogeneous systems among separate institutions<sup>76</sup>. By contrast, our model was designed to achieve both clinico-genomic knowledge representation accompanied by traceability of the genomic data, to enable determination the clinical significance of a genomic test result provided to a clinician.

Through the developed cGDM, standardization and integration of the structure of genomic data can be realized, along with tracing of the information in a step-by-step manner until the data related to the target are extracted according to clinical or research requirements. To secure the clarity of genomic information, we defined the basis for each attribute and focused on designing an entity set that can accurately represent the genomic data to be delivered to the target user, without information distortion, through composition of the basis.

To allow better assessment of the meaningfulness of genomic information, we defined the basis for each attribute and focused on



designing an entity set that accurately represents the genomic data that are delivered to the target user, without information distortion. Furthermore, the cGDM is adaptable as a data-level extension to any existing information system, regardless of database system or application platform. Effectiveness and feasibility of genomic data management in the computational environment in terms of the data-level EHR integration approach by the cGDM were also broadly evaluated in Chapter 2 and 3.

## **Unrecognized Ambiguity in the Interdisciplinary Knowledge Interplay**

Accumulation of basic, translational, and regulatory science is a prerequisite to implementing personalized medicine in routine care<sup>22</sup>. As a basic science, bioinformatics has witnessed explosive and rapid progress since the completion of the Human Genome Project. In the context of regulatory science, there are currently several ongoing efforts within the bioinformatics and molecular biology domains,<sup>10,11,77</sup> with great maturation in the body of knowledge during the last decade, including principles and recommendations related to NGS technology. These efforts have focussed primarily on the standardization of bioinformatics protocols and the file structures for intra- or interlaboratory communication.

Translational science represents the next challenge for the realization of actual health promotion with personalized medicine<sup>78</sup>. In the context of clinico-genomics, translational approaches ultimately target the syntactic and semantic interoperability between genomics and clinical practice, to ensure business continuity in terms of knowledge management<sup>23,24,79</sup>. Previous approaches have stressed a need for structural transformation to overcome the currently low adaptation of genomic information for clinical decision-making. However, the other major cause, the knowledge gap, has yet to be seriously considered because the solution appears obvious: the education of medical experts in bioinformatics

principles.

Nevertheless, this raises the question of the specific level of bioinformatics knowledge required in clinical practice. Our working group agreed that clinicians do not need to be bioinformatics experts to implement precision medicine. Preferably, the key is education on how to understand genomic data and confidence levels, and then be provided with sufficient information to make clinical decisions. Based on this perspective, we identified a previously unrecognised ambiguity related to the knowledge interplay between bioinformatics and medical practices (Fig. 3). Although the genome is the most concrete type of observational data representing an individual's inheritance, the genomic information delivered to clinicians is rarely transformed to a human-readable form and is also rarely a direct representation of the genomic sequence. Instead, this information is more of an intellectual product, processed in a purpose-weighted result file structure. Thus, the question of reliability of the genomic information must be addressed before it is adopted by the physician, similar to other types of conventional observational data.

Considering the knowledge gap in this clinico-genomic context, unrecognised ambiguities may occur on each side. For example, when linking the outputs of bioinformatics to clinical fields, the indicator of information quality moves from internal consistency within the same protocol to external consistency between different protocols. Thus, to

accomplish the final goal of precision medicine, more discussion is needed about how data will cross this intermediate space, then about how to best represent and deliver crossover information.

## **Adoption of FMEA to Information Processing**

To best of our knowledge, the methodology proposed herein has not yet been applied in the field of genetic information processing. FMEA is the most commonly used methodology for determining reliability of manufacturing and design processes<sup>17,20,21,80,81</sup>. We perceive the result of genetic testing not as an output of static measurement, but rather as an output of an intellectual production process. When conducting bioinformatics analyses, there is no requirement for unification among the processes, since the internal consistency within each process guarantees scientific rigour. Moreover, the flexible data specifications used in the bioinformatics field have the advantage of supporting various research applications<sup>7</sup>, but that advantage becomes an obstacle to data integration for comprehensive clinical decision making. In addition, relevant external knowledge, tools, platforms, and analytical techniques cannot be unified because they are still under development. Considering this large interdisciplinary hyperspace, our approach aims to improve the quality of information delivery while responding to an enormous, growing body of knowledge that has yet to be integrated within its own basic-science field. Therefore, the FMEA was adopted to derive and clarify a set of metadata designed to prevent information from being distorted.

To facilitate the use of genomic test results in clinical practice, it is

essential to integrate genomic data into clinical decision support systems regarding data volume and knowledge management<sup>6,34,37,82</sup>. Data modeling is the first and most crucial step in the multi-tiered design of information systems. The final product reliability, for example specific clinical decision support algorithms or integrated information systems, is hardly improved over the designed reliability on the lower level of architecture (data-level)<sup>20</sup>. This viewpoint was projected to the study design. An important consideration is that the analytic scheme presented here can help to enhance clinico-genomic understanding for experts on both the medical and bioinformatics sides of the workflow. (see Methods Section) Throughout the development of this method, we focussed on equally weighting the clinical perspective and bioinformatics process analysis in the context of business continuity, starting from our initial clinical intention through bioinformatics information processing by a knowledge-based protocol, finally offering a deliverable and interpretable form to the point-of-care clinician.

## **Limitations**

Multi-omics data have a fundamental limitation of unification, which is derived from the difference of knowledge expression forms related to the processing methodology, final processed data depending on the target layer, and its biological characteristics. In addition, prior to NGS, there were already several structured models according to differences in data scale and technical maturity. The entity and attribute set defined in the GDM is derived from analysis of the workflow of NGS. Therefore, we do not consider the elements of other technology-based workflows in multi-omics layers.

The methods, equipment, data processing and analytical techniques for extracting data from targets in nature will continue to evolve and accumulate. The cGDM was designed to be flexible and able to readily adapt to technological changes. However, an eventual failure in responding to these changes cannot be excluded and represents a potential limitation of this study.

Several standard models have been generated, based on differences in data scale and technical maturity, prior to the development of NGS technology. Thus, we have not considered multi-omics data. Focussing on NGS technology-based workflow helped us to determine an optimized information scale and granularity for the clinical level, and to design a model to generalise and process genomic data based on individual patients.

The cGDM could be extended to be a part of technology-wide data model integration for multi-omics data management.

The data model proposed in this study aims to clarify blind points within the interdisciplinary genomic-clinical interface, connecting separated expertise within a single platform to provide a broad perspective that covers the information reliability required for clinical evidence. In particular, we have made a novel attempt to adopt the FMEA method for a systematic meta-level data design process. Future work will focus on the development of functional systems to conduct real-world validation, including a data-upload pipeline from processed genome data files, as well as a clinical decision support tools based on the cGDM.



# Supplementary Information

## Supplementary Figure S1. PGx CDS mock-up application based on the cGDM architecture

**Stored data example in the cGDM database**

Patient Identifier	HGV5_Genomic_Change	HGNC_Gene_Symbol	Position	Ref_Allele	Alt_Allele	Genotype	Biomarker_Resource	dbSNP_ID	Order_date
1224	NC_000010.10:g.96702047C>T	CYP2C9	96702047	C	T	1 0	IWPC	rs1799853	2018-08-23
1224	NC_000010.10:g.96741053A>C	CYP2C9	96741053	A	C	0 1	IWPC	rs1057910	2018-08-23
1224	...	...	...	...	...	...	...	...	...

1000 genome phase 3 CEU

**Pt Name (Pt No)**  
Gender-Age-Race-Height(cm)-Weight(kg)

- Mario Speedwagon (PA06984)  
M-21-Unknown-187-79
- Petey Cruiser (PA06985)  
F-22-Unknown-189-61
- Anna Sthesia (PA06986)  
M-23-Unknown-161-97
- Paul Molve (PA06989)  
F-24-Unknown-156-45
- Anna Mull (PA06994)  
M-25-Unknown-194-65
- Gail Forcewind (PA07000)  
F-26-Unknown-168-81
- Paige Turner (PA07037)  
F-27-Unknown-185-63
- Bob Frapples (PA07048)  
M-28-Unknown-172-95

Patient No: PA06989    Name: Paul Molve    F / 24  
Dx: Deep vein thrombosis    156cm / 45kg   

Date	Order No	Drug	Dose	Dose Unit	Route	Frequency/Duration	Prescriber
2020-07-10	1	warfarin	6	mg	PO	OD	Dr. Kim

**PGx CDS message**

Warfarin dosing can be modified with a patient's VKORC1 and CYP2C9 phenotype.  
IWPC warfarin pharmacogenetics dosing estimation applied to the patient's case display below.

Current ordered dose: 6 mg/day (= 42.0 mg/week)  
Recommendation: 3.7 mg/day = 26.0 mg/week

Related Gene	PGx Genotype
VKORC1	A/G
CYP2C9	*2/*2

Please consult a clinical pharmacist for further considerations.

[More evidence information on Web](#)

leathavatar:snubi.org:8080/admin/map/menu/iso#

### Supplementary Table S1. Table Specification of the cGDM

The logical entities and attributes expressed in Figure 1.5 were converted into physical entities and attributes. Here, we provided our physical data model as the following table. The required data type, description, and example value for each attribute defined are described. All of the logical entities and attributes in Figure 1.5 have been transformed and defined in the physical model presented here. So, by applying this sort of conversion to physical model, each researchers can construct a genomic database according to the environment of the existing information system.

#### CLINICAL IDENTIFIER Table specification

#	Logical Name	Physical Name	PK	Required	Data Type	Description	Example
1	Subject Identifier	Subject_Identifier	PK	Yes	int(11)	Arbitrary person identifier defined in the CGDM database	1
2	Patient Number	Patient_Number		Yes	varchar(20)	Patient number of existing HIS database used to link with the CGDM database	12345678
3	Medical Institution Identifier	Institution_Identifier		Yes	varchar(20)	An abbreviation of the hospital name where the patient data linked with the CGDM database	SNUH
4	Order Identifier	Order_Identifier		Yes	varchar(20)	Unique key value represents an order of existing HIS database used to link with the CGDM database	602489471
5	Clinician Identifier	Clinicain_Identifier		Yes	varchar(20)	Unique key value represents a physician of existing HIS database used to link with the CGDM database	A2068494
6	Submission Date	Submission_Date		Yes	datetime	Date of the beginning of the data production period (e.g. ordered date)	2018-08-17 13:44

**EXPERIMENT RELATED INFORMATION Table specification**

#	Logical Name	Physical Name	PK	Required	Data Type	Description	Example
1	Experiment Identifier	Experiment_Identifier	PK	Yes	int(11)	Arbitrary identifier of the experiment defined in the CGDM database	11
2	Subject Identifier	Subject_Identifier	FK	Yes	int(11)	Arbitrary person identifier defined in the CGDM database	1
3	Test Description	Test_Description		No	text	Detailed description for ordered test	
4	Type of sequencing	Sequencing_Type		Yes	varchar(50)	Library strategy for genome sequencing	{WGS, WES, Targeted sequencing, etc.} <sup>72</sup>
5	Platform technology	Platform_Technology		Yes	varchar(20)	The technology platform used to identify the variant	NGS
6	Sequencer	Sequencer		Yes	varchar(50)	Sequencing equipment	Illumina Hiseq 2500
7	Sequencing Institution	Sequencing_Institution		Yes	varchar(50)	Name of sequencing institution	SNUBI
8	Experimenter	Experimenter		Yes	varchar(50)	Name of the primary experimenter	BJ Min
9	Collection Date	Collection_Date		Yes	datetime	Date of the sample collection	2018-09-03 11:00

**BIOINFORMATICS PROTOCOL RELATED INFORMATION Table specification**

#	Logical Name	Physical Name	PK	Required	Data Type	Description	Example
1	Bioinformatics Protocol Identifier	BI_Protocol_Identifier	PK	Yes	int(11)	Arbitrary identifier of the bioinformatics protocol defined in the CGDM database	121

2	Experiment Identifier	Experiment_Identifier	FK	Yes	int(11)	Arbitrary identifier of the experiment defined in the CGDM database	11
3	Pipeline Name	Pipeline_Name		Yes	varchar(50)	Name of the pipeline	SNUBI WXS data pipeline
4	Step (of the pipeline)	Step		Yes	int(3)	The order in which the steps are executed	1
5	Tool (of the pipeline)	Tool		Yes	varchar(50)	Procedure description	(alignment, sort, deduplication, variant calling, etc.)
6	Parameter (of the pipeline)	Parameter		Yes	varchar(50)	The name of tools	GATK
7	Datasource origin (used in the pipeline)	Datasource_Origin		Yes	varchar(50)	The version of tools	v2.5-2
8	Datasource version (used in the pipeline)	Datasource_Version		No	varchar(50)	Preset parameters used for the step	stand_call_conf=30,stand_emit_conf=10
9	Datasource Build (used in the pipeline)	Datasource_Build		No	varchar(50)	The source of databases	1kG, Mills, dbSNP137
10	Analytics Institution	Analytics_Institution		Yes	varchar(50)	Name of the bioinformatics analytics institution	SNUBI
11	Bioinformatician	Bioinformatician		Yes	varchar(50)	Name of the primary bioinformatician	YM Park
12	Received Date	Received_Date		Yes	datetime	Date of the raw data file (eg. BAM file) received	2018-09-15 17:35
13	Documentation Date	Documentation_Date		Yes	datetime	Date of the processed data stored in the CGDM database	2018-09-22 11:22

**QUALITY CHECK Table specification**

#	Logical Name	Physical Name	PK	Required	Data Type	Description	Example
1	Quality Check Identifier	QC_Identifier	PK	Yes	int(11)	Arbitrary identifier of the quality check matrix in the CGDM database	123
2	Bioinformatics Protocol Identifier	BI_Protocol_Identifier	FK	Yes	int(11)	Arbitrary identifier of the bioinformatics protocol in the CGDM database	121
3	Total Reads	Total_Reads		Yes	bigint	Total number of reads	100720000
4	Total Aligned Reads	Total_Aligned_Reads		No	bigint	Total number of aligned reads	99168912
5	% Reads Aligned	Reads_Aligned_Percent		No	float	Percentage of reads aligned	98.46 ( = 4/3)
6	Total Bases	Total_Bases		No	bigint	Total number of bases	7260000
7	Total Mapped Bases	Mapped_Bases		No	bigint	Total number of mapped bases	7050000
8	Average on target depth	Depth_Mean		No	float	Mean on target depth	71.94
9	Standard deviation on target depth	Depth_SD		No	float	Standard deviation of on target depth	16.54
10	On Target Bases	Target_Bases		No	bigint	On target bases	2640000

**GENOMIC ALTERATION Table specification**

#	Logical Name	Physical Name	PK	Required	Data Type	Description	Example
1	Genomic Alteration Identifier	Genomic_Alteration_Identifier	PK	Yes	int(11)	Arbitrary identifier of the genomic alteration defined in the CGDM database	14009

2	Bioinformatics Protocol Identifier	BI_Protocol_Identifier	FK	Yes	int(11)	Arbitrary identifier of the bioinformatics protocol defined in the CGDM database	121
3	Position	Position		Yes	varchar(255)	The genomic position where the alteration occurs	180888597
4	Reference allele	Reference_Allele		Yes	varchar(255)	The base found in the reference genome	A
5	Alternative allele	Alternative_Allele		Yes	varchar(255)	Any base other than the reference	T
6	Chromosome	Chromosome		Yes	varchar(2)	The chromosome where the alteration occurs	7
7	Cytogenetic location	Cytogenetic_Location		No	text	Cytogenetic band that the location of the alteration maps to	17q12
8	Codon	Codon		No	text	The codon where the alteration is identified	12
9	Exon	Exon		No	varchar(10)	The exonic location where the alteration is identified	19
10	HGVS genomic change	HGVS_Genomic_Change		Yes	text	Description of the nucleotide change for a genomic sequence (supplied by HGVS)	NG_007873.3:g.176429T>A
11	HGVS coding change	HGVS_Coding_Change		No	text	Description of the nucleotide change for a coding DNA sequence (supplied by HGVS)	NM_004333.4:c.1799T>A

12	HGVS protein change	HGVS_Protein_Change	No	text	Description of the nucleotide change for a protein sequence (supplied by HGVS)	NP_004324.2:p.Val600Glu
13	HGVS version	HGVS_Version	Yes	varchar(20)	The version number of HGVS	HGVS version 15.11
14	dbSNP Identification Number	dbSNP_ID	No	varchar(20)	The identification tag (supplied by NCBI dbSNP)	rs56046546
15	dbVar Identification Number	dbVar_ID	No	varchar(20)	The identification tag (supplied by NCBI dbVar)	nsv1123397
16	Genome build	Genome_Build	No	varchar(20)	Genomic coordinates of the reference	GRCh37/hg19
17	Genomic source	Genomic_Source	Yes	varchar(10)	Class of genomic source	{Somatic, Germline, Unknown, etc.}
18	HGNC gene symbol	HGNC_Gene_Symbol	No	varchar(20)	The official gene symbol approved by the HGNC	ALK, JMJD7-PAL2G4B
19	Entrez gene ID	Entrez_ID	No	integer	Entrez Gene ID (supplied by NCBI)	238
20	Ensembl gene ID	Ensembl_ID	No	char(15)	Ensembl Gene ID (supplied by Ensembl)	ENSG00000171094
21	Genotype	Genotype	No	char(3)	Allelic state of the given variant	0 1, 0 0, . ., etc
22	clinVar Variation Identification Number	clinVar_Variant_ID	No	varchar(20)	The identification tag (supplied by clinVar)	188275
23	COSMIC Identification Number	COSMIC_ID	No	varchar(10)	The identification tag (supplied by COSMIC)	COSM476
24	Molecular Effects	Molecular_Effect	No	varchar(50)	Effects of mutations on protein function	{Missense, Nonsense, Frameshift, Promoter, etc}

25	Variant type	Variant_Type		Yes	varchar(20)	The type of variant in a sequence of DNA	{Substitution, Deletion, Duplication, Insertion, InDel, Inversion, Conversion, etc.}
26	Functional Domain	Functional_Domain		No	varchar(50)	The functional domain where the alteration occurs	ATP-binding domain

### CLINICAL ANNOTATION Table specification

#	Logical Name	Physical Name	PK	Required	Data Type	Description	Example
1	Clinical Annotation Identifier	Clinical_Annotation_Identifier	PK	Yes	int(11)	Arbitrary identifier of the clinical annotation defined in the CGDM database	22
2	Genomic Alteration Identifier	Genomic_Alteration_Identifier	FK	Yes	int(11)	Arbitrary identifier of the genomic alteration defined in the CGDM database	14009
3	Biomarker Datasource	Biomarker_Datasource		Yes	varchar(255)	Name of datasource for biomarkers of genomic data	ACMG actionable genes
4	Biomarker Name	Biomarker_Name		Yes	varchar(50)	Name of predictive indicator from biomarker datasource	EGFR Exon 19 Deletion

### MICROSATELLITE INSTABILITY Table specification

#	Logical Name	Physical Name	PK	Required	Data Type	Description	Example
1	Microsatellite Instability Identifier	MSI_Identifier	PK	Yes	int(11)	Arbitrary identifier of microsatellite instability defined in the CGDM database	14



2	Bioinformatics Protocol Identifier	BI_Protocol_Identifier	FK	Yes	int(11)	Arbitrary identifier of the bioinformatics protocol defined in the CGDM database	121
3	MSI phenotype	MSI_Phenotype		Yes	varchar(50)	Distinct phenotype of the microsatellite instability	{Microsatellite Stable (MSS), MSI-Low (MSI-L), MSI-High (MSI-H), Indeterminate MSI}
4	MSI marker name	MSI_Marker_Name		Yes	varchar(20)	Name of the MSI marker	BAT26
5	MSI marker status	MSI_Marker_Status		Yes	varchar(20)	Determined MSI status	Positive

**Supplementary Table S2. IUPAC nucleotide code table for processing double/triple based code**

Symbol	Meaning
a	a; adenine
c	c; cytosine
g	g; guanine
t	t; thymine in DNA; uracil in RNA
m	a or c
r	a or g
w	a or t
s	c or g
y	c or t
k	g or t
v	a or c or g; not t
h	a or c or t; not g
d	a or g or t; not c
b	c or g or t; not a
n	a or c or g or t

\*reference: Cornish-Bowden, A. Nucl Acid Res 13, 3021-3030 (1985)  
[https://www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/MolBioReview/iupac\\_nt\\_abbreviations.html](https://www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/MolBioReview/iupac_nt_abbreviations.html) 에서 재인용

**Supplementary Table S3. Number of HLA alleles**

<i>Category</i>	<i>Locus</i>	<i>Allele number</i>	<i>Protein number</i>	<i>Null allele number</i>
<i>Class I</i>	HLA-A	673	527	46
	HLA-B	1077	911	38
	HLA-C	360	283	8
	HLA-E	9	3	0
	HLA-F	21	4	0
	HLA-G	36	14	1
	Pseudogenes	39		
	Total	2215	1742	93
<i>Class II</i>	HLA-DRA	3	2	0
	HLA-DRB	669	546	8
	HLA-DQA1	34	25	1
	HLA-DQB1	93	68	1
	HLA-DPA1	27	16	0
	HLA-DPB1	128	114	2
	HLA-DMA	4	4	0
	HLA-DMB	7	7	0
	HLA-DOA	12	3	1
	HLA-DOB	9	4	0
	Total	986	789	13
<i>MHC-like</i>	MICA	64	54	0
	MICB	30	19	2
	Total	94	73	2

\* reference: Shiina, T., Hosomichi, K., Inoko, H., & Kulski, J. K. (2009). The HLA genomic loci map: expression, interaction, diversity and disease. *Journal of human genetics*, 54(1), 15-39. Table 4. Number of HLA alleles

# Bibliography

- 1 Ginsburg, G. S. & Willard, H. F. Genomic and personalized medicine: foundations and applications. *Translational research* **154**, 277-287 (2009).
- 2 Downing, G. J., Boyle, S. N., Brinner, K. M. & Osheroﬀ, J. A. Information management to enable personalized medicine: stakeholder roles in building clinical decision support. *BMC medical informatics and decision making* **9**, 44 (2009).
- 3 Collins, F. S. & Varmus, H. A new initiative on precision medicine. *New England Journal of Medicine* **372**, 793-795 (2015).
- 4 Dewey, F. E. *et al.* Clinical interpretation and implications of whole-genome sequencing. *Jama* **311**, 1035-1045 (2014).
- 5 McCarty, C. A. *et al.* The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC medical genomics* **4**, 13 (2011).
- 6 Masys, D. R. *et al.* Technical desiderata for the integration of genomic data into Electronic Health Records. *J Biomed Inform* **45**, 419-422, doi:10.1016/j.jbi.2011.12.005 (2012).
- 7 Lubin, I. M. *et al.* Principles and Recommendations for Standardizing the Use of the Next-Generation Sequencing Variant File in Clinical Settings. *J Mol Diagn* **19**, 417-426, doi:10.1016/j.jmoldx.2016.12.001 (2017).
- 8 Kho, A. N. *et al.* Practical challenges in integrating genomic data into the electronic health record. *Genet Med* **15**, 772-778, doi:10.1038/gim.2013.131 (2013).
- 9 Kassakian, S. Z., Yackel, T. R., Gorman, P. N. & Dorr, D. A. Clinical decisions support malfunctions in a commercial electronic health record. *Applied clinical informatics* **8**, 910-923 (2017).
- 10 Yang, W. *et al.* Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic acids research* **41**, D955-D961 (2012).
- 11 Roukos, D. H. Next-generation, genome sequencing-based biomarkers: concerns and challenges for medical practice. *Biomarkers in medicine* **4**, 583-586 (2010).
- 12 Roy, S. *et al.* Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists. *J Mol Diagn* **20**, 4-27, doi:10.1016/j.jmoldx.2017.11.003 (2018).

- 13 Oliver, G. R., Hart, S. N. & Klee, E. W. Bioinformatics for clinical next generation sequencing. *Clin Chem* **61**, 124-135, doi:10.1373/clinchem.2014.224360 (2015).
- 14 Gargis, A. S. *et al.* Good laboratory practice for clinical next-generation sequencing informatics pipelines. *Nat Biotechnol* **33**, 689-693, doi:10.1038/nbt.3237 (2015).
- 15 Han, P. K. J. *et al.* A taxonomy of medical uncertainties in clinical genome sequencing. *Genet Med* **19**, 918-925, doi:10.1038/gim.2016.212 (2017).
- 16 Simianu, V. V. *et al.* Understanding clinical and non-clinical decisions under uncertainty: a scenario-based survey. *BMC medical informatics and decision making* **16**, 153 (2016).
- 17 Shebl, N. A., Franklin, B. D. & Barber, N. Is failure mode and effect analysis reliable? *Journal of patient safety* **5**, 86-94 (2009).
- 18 Singh, V., Pungotra, H., Singh, S. & Gill, S. S. Prioritization of Failure Modes in Process FMEA using Fuzzy Logic. *International Journal Of Enhanced Research In Science Technology & Engineering* **2** (2013).
- 19 Certa, A., Hopps, F., Inghilleri, R. & La Fata, C. M. A Dempster-Shafer Theory-based approach to the Failure Mode, Effects and Criticality Analysis (FMECA) under epistemic uncertainty: application to the propulsion system of a fishing vessel. *Reliability Engineering & System Safety* **159**, 69-79 (2017).
- 20 Teng, S.-H. & Ho, S.-Y. Failure mode and effects analysis: an integrated approach for product design and process control. *International journal of quality & reliability management* **13**, 8-26 (1996).
- 21 Gilchrist, W. Modelling Failure Modes and Effects Analysis. *International Journal of Quality & Reliability Management* **10**, doi:10.1108/02656719310040105 (1993).
- 22 Eubanks, C. F., Kmenta, S. & Ishii, K. in *ASME Design Engineering Technical Conferences*. 14-17.
- 23 Reifer, D. J. Software failure modes and effects analysis. *IEEE Transactions on reliability* **28**, 247-249 (1979).
- 24 Vajna, S. in *ASME 2003 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. 375-382 (American Society of Mechanical Engineers).
- 25 Sayyadi Tooranloo, H., Ayatollah, A. S. & Alboghobish, S. Evaluating knowledge management failure factors using intuitionistic fuzzy FMEA approach. *Knowledge and Information Systems* **57**, 183-205, doi:10.1007/s10115-018-1172-3 (2018).
- 26 Cabanes, B., Hubac, S., Le Masson, P. & Weil, B. in *14th International*

*Design Conference (DESIGN 2016).*

- 27 Chandrasegaran, S. K. *et al.* The evolution, challenges, and future of knowledge representation in product design systems. *Computer-aided design* **45**, 204-228 (2013).
- 28 Blount, G., Kneebone, S. & Kingston, M. Selection of knowledge-based engineering design applications. *Journal of Engineering Design* **6**, 31-38 (1995).
- 29 Tamisier, T. & Feltz, F. A Data Model for Knowledge Representation in Collaborative Systems. *Data Science Journal* **6**, S225-S233 (2007).
- 30 Navathe, S. B. & Schkolnick, M. in *Proceedings of the 1978 ACM SIGMOD international conference on management of data.* 144-156 (ACM).
- 31 Smith, J. M. & Smith, D. C. Database abstractions: aggregation and generalization. *ACM Transactions on Database Systems (TODS)* **2**, 105-133 (1977).
- 32 Consortium, G. P. A global reference for human genetic variation. *Nature* **526**, 68 (2015).
- 33 National Cancer Institute GDC Data Portal TCGA PAAD dataset. at <https://portal.gdc.cancer.gov/projects/TCGA-PAAD> (2017)
- 34 Sen, A., Al Kawam, A. & Datta, A. Emergence of DSS efforts in genomics: Past contributions and challenges. *Decision Support Systems* **116**, 77-90 (2019).
- 35 Overby, C. L., Tarczy-Hornoch, P., Hoath, J. I., Kalet, I. J. & Veenstra, D. L. in *BMC bioinformatics.* S10 (BioMed Central).
- 36 Hoffman, M. A. & Williams, M. S. Electronic medical records and personalized medicine. *Human genetics* **130**, 33-39 (2011).
- 37 Castaneda, C. *et al.* Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. *Journal of clinical bioinformatics* **5**, 4 (2015).
- 38 Dinu, V. & Nadkarni, P. Guidelines for the effective use of entity-attribute-value modeling for biomedical databases. *Int J Med Inform* **76**, 769-779, doi:10.1016/j.ijmedinf.2006.09.023 (2007).
- 39 Peleg, M. The Role of Modeling in Clinical Information System Development Life Cycle. *Methods of information in medicine* **50**, 7-10 (2011).
- 40 Williams, M. S. *et al.* Genomic Information for Clinicians in the Electronic Health Record: Lessons Learned from ClinGen and eMERGE. (2019).
- 41 Dolin, R. H., Boxwala, A. & Shalaby, J. A Pharmacogenomics Clinical

- Decision Support Service Based on FHIR and CDS Hooks. *Methods Inf Med* **57**, e115-e123, doi:10.1055/s-0038-1676466 (2018).
- 42 Relling, M. V. & Klein, T. E. CPIC: Clinical Pharmacogenetics Implementation Consortium of the Pharmacogenomics Research Network. *Clin Pharmacol Ther* **89**, 464-467, doi:10.1038/clpt.2010.279 (2011).
- 43 Swen, J. *et al.* Pharmacogenetics: from bench to byte. *Clinical Pharmacology & Therapeutics* **83**, 781-787 (2008).
- 44 Consortium, I. W. P. Estimation of the warfarin dose with clinical and pharmacogenetic data. *New England Journal of Medicine* **360**, 753-764 (2009).
- 45 Ross, C. J. *et al.* The Canadian Pharmacogenomics Network for Drug Safety: a model for safety pharmacology. *Thyroid* **20**, 681-687 (2010).
- 46 Blagec, K. *et al.* Implementing pharmacogenomics decision support across seven European countries: The Ubiquitous Pharmacogenomics (U-PGx) project. *J Am Med Inform Assoc* **25**, 893-898, doi:10.1093/jamia/ocy005 (2018).
- 47 Cavallari, L. H. *et al.* Multi-site investigation of strategies for the clinical implementation of CYP2D6 genotyping to guide drug prescribing. *Genet Med*, doi:10.1038/s41436-019-0484-3 (2019).
- 48 Cicali, E. J. *et al.* Challenges and lessons learned from clinical pharmacogenetic implementation of multiple gene-drug pairs across ambulatory care settings. *Genet Med*, doi:10.1038/s41436-019-0500-7 (2019).
- 49 Pearce, C. *et al.* Delivering genomic medicine in the United Kingdom National Health Service: a systematic review and narrative synthesis. *Genet Med*, doi:10.1038/s41436-019-0579-x (2019).
- 50 Walton, N. A., Johnson, D. K., Person, T. N. & Chamala, S. Genomic Data in the Electronic Health Record. *Advances in Molecular Pathology* **2**, 21-33, doi:10.1016/j.yamp.2019.07.001 (2019).
- 51 Lenzerini, M. in *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems.* 233-246.
- 52 Kho, A. N. *et al.* Practical challenges in integrating genomic data into the electronic health record. *Genetics in Medicine* **15**, 772 (2013).
- 53 Caudle, K. E. *et al.* Standardizing terms for clinical pharmacogenetic test results: consensus terms from the Clinical Pharmacogenetics Implementation Consortium (CPIC). *Genet Med* **19**, 215-223, doi:10.1038/gim.2016.87 (2017).
- 54 The Clinical Pharmacogenetics Implementation Consortium (CPIC®), CPIC guidelines. <https://cpicpgx.org/guidelines/> [accessed 2020-05-10]

- 55 Robarge, J., Li, L., Desta, Z., Nguyen, A. & Flockhart, D. The star-allele nomenclature: retooling for translational genomics. *Clinical Pharmacology & Therapeutics* **82**, 244-248 (2007).
- 56 Minucci, A. *et al.* Glucose-6-phosphate dehydrogenase (G6PD) mutations database: review of the "old" and update of the new mutations. *Blood Cells Mol Dis* **48**, 154-165, doi:10.1016/j.bcmd.2012.01.001 (2012).
- 57 Beutler, E. The designation of mutations. *American journal of human genetics* **53**, 783 (1993).
- 58 den Dunnen, J. T. & Antonarakis, S. E. Nomenclature for the description of human sequence variations. *Hum Genet* **109**, 121-124, doi:10.1007/s004390100505 (2001).
- 59 Bodmer, J. G. *et al.* Nomenclature for factors of the HLA system, 1989. *Immunobiology* **180**, 278-292 (1990).
- 60 Klein, J. & Sato, A. The HLA system. *New England Journal of Medicine* **343**, 702-709 (2000).
- 61 Mosaad, Y. Clinical role of human leukocyte antigen in health and disease. *Scandinavian journal of immunology* **82**, 283-306 (2015).
- 62 Mahdi, B. M. A glow of HLA typing in organ transplantation. *Clinical and translational medicine* **2**, 6 (2013).
- 63 Fan, W.-L. *et al.* HLA association with drug-induced adverse reactions. *Journal of immunology research* **2017** (2017).
- 64 Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in medicine* **17**, 405-423 (2015).
- 65 Jones, K. H. *et al.* The other side of the coin: Harm due to the non-use of health-related data. *Int J Med Inform* **97**, 43-51, doi:10.1016/j.ijmedinf.2016.09.010 (2017).
- 66 Lee, K. H., Kim, H. J., Kim, Y.-J., Kim, J. H. & Song, E. Y. Extracting Structured Genotype Information from Free-Text HLA Reports Using a Rule-Based Approach. *Journal of Korean Medical Science* **35** (2020).
- 67 Solomatine, D., See, L. M. & Abrahart, R. in *Practical hydroinformatics* 17-30 (Springer, 2009).
- 68 Warner, J. L., Jain, S. K. & Levy, M. A. Integrating cancer genomic data into electronic health records. *Genome medicine* **8**, 113 (2016).
- 69 Pennington, J. W. *et al.* Genomic decision support needs in pediatric primary care. *Journal of the American Medical Informatics Association* **24**, 851-856 (2017).



- 70 Heale, B. S. *et al.* Integrating genomic resources with electronic health records using the HL7 Infobutton standard. *Applied clinical informatics* **7**, 817-831 (2016).
- 71 Hamburg, M. A. & Collins, F. S. The path to personalized medicine. *New England Journal of Medicine* **363**, 301-304 (2010).
- 72 Alterovitz, G. *et al.* FHIR Genomics: enabling standardization for precision medicine use cases. *NPJ genomic medicine* **5**, 1-4 (2020).
- 73 Page, A. *et al.* Genomics. A federated ecosystem for sharing genomic, clinical data. Global Alliance for Genomics and Health. *Science* **352**, 1278-1280 (2016).
- 74 Lawler, M. *et al.* All the world's a stage: facilitating discovery science and improved cancer care through the Global Alliance for Genomics and Health. *Cancer discovery* **5**, 1133-1136 (2015).
- 75 Shin, S. J. *et al.* Genomic Common Data Model for Seamless Interoperation of Biomedical Data in Clinical Practice: Retrospective Study. *J Med Internet Res* **21**, e13249, doi:10.2196/13249 (2019).
- 76 Haarbrandt, B. *et al.* HiGHmed—an open platform approach to enhance care and research across institutional boundaries. *Methods of information in medicine* **57**, e66-e81 (2018).
- 77 Rector, A. L. Thesauri and formal classifications: terminologies for people and machines. *Methods of information in medicine* **37**, 501-509 (1998).
- 78 Barile, S., Polese, F., Saviano, M. & Carrubbo, L. in *Innovating in Practice* 417-438 (Springer, 2017).
- 79 Tooranloo, H. S., Ayatollah, A. S. & Alboghobish, S. Evaluating knowledge management failure factors using intuitionistic fuzzy FMEA approach. *Knowledge and Information Systems*, 1-23 (2018).
- 80 DeRosier, J., Stalhandske, E., Bagian, J. P. & Nudell, T. Using health care failure mode and effect analysis™: the VA National Center for Patient Safety's prospective risk analysis system. *The Joint Commission journal on quality improvement* **28**, 248-267 (2002).
- 81 Deandrea, S. *et al.* Implementation of Failure Mode and Effects Analysis to the specimens flow in a population-based colorectal cancer screening programme using immunochemical faecal occult blood tests: a quality improvement project in the Milan colorectal cancer screening programme. *BMJ Open Qual* **7**, e000299 (2018).
- 82 Overby, C. L. *et al.* Developing a prototype system for integrating pharmacogenomics findings into clinical practice. *Journal of personalized medicine* **2**, 241-256 (2012).
- \* Kim, H. J., Kim, H. J., Park, Y., Lee, W. S., Lim, Y., & Kim, J. H. (2020).

clinical Genome Data Model (cGDM) provides interactive clinical Decision Support for precision Medicine. Scientific reports, 10(1), 1-13.

### [Website]

The Clinical Pharmacogenetics Implementation Consortium (CPIC®), CPIC guidelines. <https://cpicpgx.org/guidelines/> [accessed 2020-05-10]

Health Level Seven, FHIR Genomics, <https://www.hl7.org/fhir/genomics.html> [accessed 2020-07-03]

The Global Alliance for Genomics and Health alliance, GA4GH Genomics API, <https://ga4gh-schemas.readthedocs.io/en/latest/> [accessed 2020-07-03]

International Organization for Standardization, IOS 25720:2009 Genomic Sequence Variation Markup Language(GSVML) <https://www.iso.org/standard/43182.html> [accessed 2020-07-03]

International Organization for Standardization, ISO/TS 20428:2017 Data elements and their metadata for describing structured clinical genomic sequence information in electronic health records <https://www.iso.org/standard/67981.html> [accessed 2020-07-03]

National Institution NHI NCI GDC <https://gdc.cancer.gov/developers/gdc-data-model> [accessed 2020-07-03]

### [Acknowledgement]

The main body of the dissertation chapter 1 and part of the general discussion has been published as the following paper: Kim, H. J., Kim, H. J., Park, Y., Lee, W. S., Lim, Y., & Kim, J. H. (2020). clinical Genome Data Model (cGDM) provides interactive clinical Decision Support for precision Medicine. Scientific reports, 10(1), 1-13.

# 국문 초록

## 정밀의학을 위한 임상유전체데이터모델

김 효 정

서울대학교 의과대학

의료정보학 협동과정

진료 현장에서 의사결정을 내려야 하는 임상에게 개인 유전체 정보를 다른 임상 근거들과 통합하여 보다 쉽게 다룰 수 있도록 구조화하여 지원하는 것은 정밀의학 구현을 위한 의료정보학의 주요 과제 중 하나이다. 차세대 염기서열 분석법과 같은 대량신속처리 유전체 기술의 등장과 그에 따른 해석정보의 축적으로 정밀 의학 및 개인 맞춤형 의학으로의 전환이 가시화 되는 듯 보였으나, 차세대염기서열 분석 기술 기반의 개인유전체 정보의 임상 활용은 여전히 제한적이다. 선행연구에서는 임상현장에서 유전체정보의 활용이 더딘 이유로 의료 전문가와 생물정보학자들 사이의 지식 격차, 진료 현장과 생물정보학 작업절차 간의 분리, 유전체 데이터만의 독특한 양적, 질적 자료구조의 특성과 같은 복합적인 원인을 제시하고 있다. 이러한 문제를 해결하고자 하는 시도로서 개인유전체정보를 병원정보시스템에 통합해야 한다는 요구가 높아지고 있으나 임상현장에서 활용하는 것을 목적으로 하는 지속가능하고 상호운용가능한 저장, 관리, 처리 방식에 대한 구체적인 논의는 부족한 실정이다.

본 연구에서는 임상정보시스템에 개인 유전체 정보가 통합되어 임상에 적용되기까지 현재의 장벽들을 문헌고찰을 통해 재탐색하고 관련된 개념과 방법들을 고찰하였다. 그리고 차세대 염기서열 분석방법을 기반으로 한 데이터를 어떻게 임상에서 활용하기 쉽도록 저장하고 처리하고 전달할 것인가 하는 당면한 과제에 단계적으로 접근하였다. 정보시스템 설계에 있어 데이터 모델의 설계는 최종시스템의 기능이 데이터 모델에 표현된 정보량 안에서 제한된다는 점에서 가장 일차적이며 중요한 단계이다. 따라서 1장에서는 다학제적 논의를 통해 임상 의사결정에 활용할 수 있는 유전체 지식표현을 논리적 데이터모델의 형태로 도출하여 차세대염기서열분석기술 기반의 임상유전체데이터모델(cGDM; clinical Genome Data Model)을 제안하였다. 2장에서는 약물치료를 개인별로 최적화하기 위해 이용 가능한 유전체검사결과를 사용하는 방법에 대한 지식체인 CPIC guideline을 구조화하여 1장에서 구축한 환자의 유전체 정보와 데이터 레벨의 정보흐름을 구현함으로써 모듈 방식의 약물유전체 임상 의사결정지원시스템을 제시한다. 3장에서는 생명정보학에서 임상적 의미를 드러내는 발견들이 지속됨에 따른 명명체계의 다양함을 수용하는 확장 체계의 하나로서 HLA gene에 대한 구조화된 정보 설계와 구현을 다루었다. 즉, HLA nomenclature를 대상으로 지식표현을 설계, 확장하여 임상유전체데이터모델의 지속가능성과 확장성을 검증하였다.

본 연구에서는 중개과학으로서 의료정보학이 정밀의료에 기여할 수 있는 다학제적공간을 탐색하고 정보시스템의 지식표현, 기능구현, 사용성 측면을 포괄하는 접근을 시도하였다. 본 연구의 결과로 제시된 임상유전체데이터모델은 논리적인 데이터모델 수준에서 설계되어 기존 병원정보시스템에 사용된 개발 언어에 제약을 받지 않고 데이터 수준의 확장체계로 활용할 수 있다. 즉, 정형화된 데이터를 기반으로

임상정보를 처리하는 기존의 다양한 정보시스템 아키텍처의 설계에 통합되어 각 기관 혹은 사용자의 필요에 맞게 CDSS나 서식에 연결하는 등 다양한 기능의 구현을 지원할 수 있다. 또한 연구용 데이터의 수집과 분석에 사용될 수도 있어 개인유전체분석결과를 실질적인 데이터 순환 사이클에 연결하는 데 기여할 수 있다. 궁극적으로, 의료전문가와 정보를 활용한 임상 의사결정간의 지적상호작용을 지원하는 데이터 계층 인프라를 제공한다.

**Keyword:** 정밀의료, 지식중개, 지식공학, 통합병원정보시스템, 유전체데이터모델, 약물유전체정보를 활용한 임상 의사결정지원시스템

**Student Number:** 2015-30615

Ph.D. Dissertation of Hyo Jung Kim

# Clinical Genome Data Model towards Precision Medicine

정밀의학을 위한 임상유전체데이터모델

August 2020

Graduate School of Medicine  
Seoul National University  
Interdisciplinary Program of Medical Informatics

Hyo Jung Kim

# Clinical Genome Data Model towards Precision Medicine

Hyo Jung Kim

Submitting a Ph.D. Dissertation of Medical  
Informatics


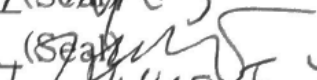
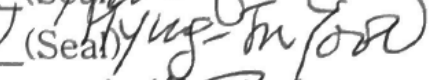
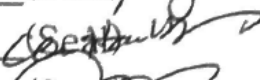

May 2020

Graduate School of Medicine  
Seoul National University  
Interdisciplinary Program of Medical Informatics

Hyo Jung Kim

Confirming the Ph.D. Dissertation written by  
Hyo Jung Kim

June 2020

Chair	<u>Jinwook Choi</u> (Seal) 
Vice Chair	<u>Ju Han Kim</u> (Seal) 
Examiner	<u>Hyung-Sun Yoon</u> (Seal) 
Examiner	<u>Hae-Young Lee</u> (Seal) 
Examiner	<u>Hyunwook Nam</u> (Seal) 

# Abstract

## Clinical Genome Data Model towards Precision Medicine

Hyo Jung Kim

Interdisciplinary Program of Medical Informatics

Graduate School of Medicine

Seoul National University

**Background** The transition to precision medicine and personalized medicine is accelerating owing to progress in genomic technology and the consequent accumulation of genomic information. However, the clinical application of genomic information remains limited, and its spread rate has been slower than expected. This lag has been attributed to complex causes, including 1) a knowledge gap between medical experts and bioinformaticians, 2) separation of the bioinformatics workflow from clinics, and 3) unique characteristics of genomic data. Nevertheless, current informational approaches to link genomic data to clinical fields mostly address the data structure problem.

**Objective** We aimed to develop a genomic data model allowing for more interactive support in clinical decision-making. Informational modeling was



used as a knowledge communication scheme from the highly intellectual product of bioinformatics to a representative data component of a clinical decision.

**Methods** Reliability-related attributes were derived through failure mode and effect analysis (FMEA). This study involved a multidisciplinary working group that conducted clinico-genomic workflow analyses and attributes extraction. Based on these data, an entity-attribute model was then developed through abstraction and normalization.

**Results** The outputs of FMEA were a dataflow snapshot obtained from next-generation sequencing, the information process map extended to the clinico-genomic context, and the set of attributes. Next, an entity-attribute model consisting of eight entities and 49 attributes was identified to develop the final genome data model, including: a linkage identifier to clinical information, experiment-related information, bioinformatics protocol-related information, physical location information, expression, annotation, actor information, and timeline information.

**Conclusion** The proposed genome data model could serve as a data-layer infrastructure supporting the intellectual interplay between medical experts and informative decision-making. Given the importance of recognizing a genome information system as a component of the clinical information system to realize precision medicine, the model could help enhance integration of genomic data in clinical settings.

**Keyword:** Precision medicine, Knowledge translation, Knowledge engineering, Hospital information system integration, Genome data model, Interactive pharmacogenomic clinical decision support

**Student Number:** 2015-30615

# Table of Contents

<b>Abstract .....</b>	<b>i</b>
<b>List of Tables and Figures .....</b>	<b>vii</b>
<b>General Introduction.....</b>	<b>1</b>
<b>Chapter 1. Clinical Genome Data Model: Data Level Integration of Patient Specific Genomic and Clinical Data for Multifaceted Utilization</b>	
<b>1.1. Introduction .....</b>	<b>4</b>
<b>1.2. Purpose of Research.....</b>	<b>9</b>
<b>1.3. Materials and Method .....</b>	<b>1 1</b>
1.3.1. The Production Process of Bringing Genomic Information to Bedside Care .....	1 2
1.3.2. FMEA: An Attribute-Clarified Framework .....	1 3
1.3.3. Logical Data Modeling.....	1 6
1.3.4. Demo Datasets for the real-world data validation.....	1 7
<b>1.4. Results .....</b>	<b>1 8</b>
1.4.1. Dataflow diagram based on an NGS workflow .....	1 9
1.4.2. Extending the NGS process under a clinicogenomic context	2 2
1.4.3. The cGDM.....	2 7
1.4.4. Validation of the cGDM.....	3 4

## **Chapter 2. Pharmacogenomic Clinical Decision Support: Modular Implementation of CPIC Guideline**

<b>2.1. Introduction .....</b>	<b>4 1</b>
<b>2.2. Purpose of Research.....</b>	<b>4 4</b>
<b>2.3. Material and Methods .....</b>	<b>4 5</b>
2.3.1 Material: CPIC guideline as knowledge resource .....	4 5
2.3.2. Data Collection.....	4 6
2.3.3. Clinical decision support service architecture.....	4 7
<b>2.4. Results .....</b>	<b>4 9</b>
2.4.1. Collected CPIC guideline and exploratory analysis.....	4 9
2.4.2. Data integration and modeling .....	5 3
2.4.3. CDS Rule Extraction.....	5 9
2.4.4. Structured database construction.....	6 0
2.4.5. PGx CDS service module.....	6 2

## **Chapter 3. Clinical Application of Clinical Genome Data Model: Integrating Star Allele and HLA Data Models**

<b>3.1. Introduction .....</b>	<b>6 5</b>
<b>3.2. Purpose of Research.....</b>	<b>6 8</b>
<b>3.3. Material and Methods .....</b>	<b>6 9</b>

<b>3.4. Results .....</b>	<b>7 0</b>
3.4.1. Summary of collected dataset .....	7 0
3.4.2. HLA data model .....	7 2
<b>General Discussion .....</b>	<b>7 3</b>
The GDM as an Infrastructure for a GIS.....	7 4
Current Approach to Genomic Data Management.....	7 6
The cGDM: A Step Beyond the Capabilities of Existing Systems	8 0
Unrecognized Ambiguity in the Interdisciplinary Knowledge Interplay .....	8 2
Adoption of FMEA to Information Processing.....	8 5
Limitations .....	8 7
<b>Supplementary Information .....</b>	<b>8 9</b>
<b>Bibliography.....</b>	<b>9 9</b>
<b>Abstract in Korean .....</b>	<b>1 0 6</b>

# List of Tables and Figures

## Chapter 1

Figure 1.1 Data-level linkage structure between conventional HIS and GIS .....	8
Figure 1.2 Data flowchart based on a next-generation sequencing workflow .....	21
Figure 1.3 Failure mode identification: mapped next-generation sequencing process extended to a clinico-genomic context .....	24
Figure 1.4 How implementation of the cGDM provides interactive clinical decision support in clinical information system .....	26
Figure 1.5 The Clinical Genome Data Model: Structured data modeling with entities and attributes .....	32
Figure 1.6 Semantic search implementation based on the CGDM .....	33
Figure 1.7 Entity-relationship diagram of the CGDM implemented in RDBMS .....	36
Figure 1.8 The conceptual map of genomic decision support system based on the cGDM .....	40
Table 1.1 Extracted classes and related attribute sets from each step of clinic-genomic context for the Entity-Attribute model .....	29
Table 1.2 Summary of imported genomic data from various data sources in the cGDM databases .....	37

## Chapter 2

Figure 2.1. The configuration of the study environment .....	46
Figure 2.2. Modular implementation of PGx CDS overview .....	47
Figure 2.3. Gene allele definition table example .....	54
Figure 2.4. Diplotype-Phenotype table example and its meta-data structure .....	57
Figure 2.5. Snapshot of CPIC guidelines content structure converted to be computable .....	58
Figure 2.6. Collection of ‘Flow chart’ over available 15 guidelines ....	59
Figure 2.7. Entity-relationship diagram of reconstructed database based on CPIC contents .....	61
Figure 2.8. PGx CDS module architecture .....	63
Figure 2.9. PGx CDS module integration scenario with dataflow .....	64
Table 2.1. The collected CPIC guideline overview .....	51
Table 2.2. Dataset list and its availability over guidelines .....	52
Table 2.3. Reference Sequence Information for Locus assignment.....	55
Table 2.4. Gene allele definition table data profiles .....	56

## **Chapter 3**

Figure 3.1 HLA Database design merge in the cGDM schema ..... 72

Table 2.1 Extracted field list gathered from the EHR records ..... 71

## **General Discussion**

Table 4.1 Comparison table of characteristics of related resources .... 78

## **Supplementary Information**

Supplementary Figure S1. PGx CDS mock-up application based on the cGDM architecture ..... 89

Supplementary Table S1. Table Specification of the cGDM ..... 90

Supplementary Table S2. IUPAC nucleotide code table for processing double/triple based code ..... 98

Supplementary Table S3. Number of HLA alleles ..... 99



# General Introduction

One of the significant tasks of medical informatics for the implementation of precision medicine is supporting clinicians by integrating personal genomic information with other clinical evidence so that constantly-evolving knowledge and inherently complex genomic data can be handled on-demand at the point of care. The transition to precision medicine and personalized medicine was expected to be accomplished within a few years due to the outstanding high-throughput sequencing capabilities of next-generation sequencing and the accumulation of knowledge about its interpretation. The prior studies present that this delay can be attributed to complicated factors, such as knowledge gaps between medical experts and bioinformatics, the separated workflow between clinical practice and bioinformatics analysis, the unique quantitative and qualitative data structure of genomic data, which can make interpretation more complicated. In an attempt to solve this problem, there is an increasing demand for the integration of personal genomic information in the electronic medical records. However, it has not been proposed as a sustainable, scalable, and interoperable method for storage, management, and processing the genomic data concerning clinical utilization.

In this study, the current barriers were explored through literature review, and related concepts and methods were investigated about these phenomena. Moreover, we addressed the immediate task of storing,

processing, and delivering data based on next-generation sequencing analysis methods to prepare for multifaceted clinical utilization. Data modeling is the first and most crucial step in the multi-tiered design of information systems. The point is that the final product reliability, such as specific clinical decision support algorithms or integrated information systems, is hardly improved over the designed reliability on the lower level of architecture.

Chapter 1 proposed a clinical genomic data model based on Deoxyribonucleic Acid (DNA) level data extracted from next-generation sequencing (NGS) technology. The multidisciplinary discussion reveals a set of genetic knowledge expressions that can be preserved and delivered the meaning for clinical decision making. In Chapter 2, the CPIC guideline, which is a knowledge of how to use available genomic test results to optimize drug therapy for individuals, is structured. Furthermore, we propose a modular drug genome clinical decision support system by linking the patient's genomic information and data-level information flow constructed in Chapter 1. Chapter 3 deals with the design and implementation of structured information about the HLA gene as one of the extensions to accommodate the diversity of naming systems as the discoveries that reveal their clinical significance in bioinformatics continue. The sustainability and scalability of the clinical genomics data model were verified by design and expand knowledge expression for HLA nomenclature.

In this study, we explored multidisciplinary space where medical informatics can contribute to precision medicine, and an approach that encompasses aspects of knowledge expression, functional realization, and usability of information systems was attempted.

# **Chapter 1. Clinical Genome Data Model: Data Level Integration of Patient Specific Genomic and Clinical Data for Multifaceted Utilization\***

## **1.1. Introduction**

As the field of medicine transitions from experience-based medicine to data-driven medicine, an apparent paradigm shift to precision medicine is underway, driven by the development of technologies in fields including medical information technology and computer engineering<sup>1,2</sup>. Genomic information is one of the most critical component of precision medicine, given its power to explain individual variability<sup>3</sup>. However, the practical clinical use of genomic information remains limited because its circulation is suboptimal, with each data processing step tending to be independently performed and thus isolated. To narrow this gap, many organizations have attempted to identify and develop methods to more effectively link genomic data to clinical information and thereby facilitate its use<sup>4-6</sup>. However, several challenges must be surmounted before realizing this goal.

First, a mismatch exists between the structure of genomic and clinical data. Genomic data based on next-generation sequencing (NGS) technology is stored as a number of file types at various stages of the bioinformatics

---

\* The main body of the dissertation chapter 1 published as following paper: Kim, H. J., Kim, H. J., Park, Y., Lee, W. S., Lim, Y., & Kim, J. H. (2020). clinical Genome Data Model (cGDM) provides interactive clinical Decision Support for precision Medicine. Scientific reports, 10(1), 1-13.

analysis, with flexible file specifications to accommodate the broad range of research interests in bioinformatics<sup>7</sup>. Raw genomic data can contain up to several tens of gigabytes of sequence information, each stored as a long string of data, and therefore cannot be used directly in this form in clinical practice without further processing. Since data processing to determine clinical relevance is both computationally intensive and time-consuming, genomic information is not readily accessible relative to other types of clinical data. Thus, for precision medicine and personalized medicine, pre-processed genomic data needs to be linked with other clinical information and provided at the appropriate time. In order to resolve this issue, a structured database is needed to store and appropriately manage genomic information for easy accessibility.

Second, genomic data has different properties than conventional observational data used in clinical settings. Therefore, genomic data must be clarified by considering procedural dimensions. Since genomic workflows contain a large number of pipelines for information processing, significant differences between the interpretation of processed data and data obtained from different information systems relative to the clinical workflow are inevitable<sup>8</sup>. Accordingly, a robust data model is required to serve as an information system to systematically manage genomic data, encompassing the detailed processes of data processing, analysis, and filtering. Additionally, information on the reliability and accuracy of these analyses

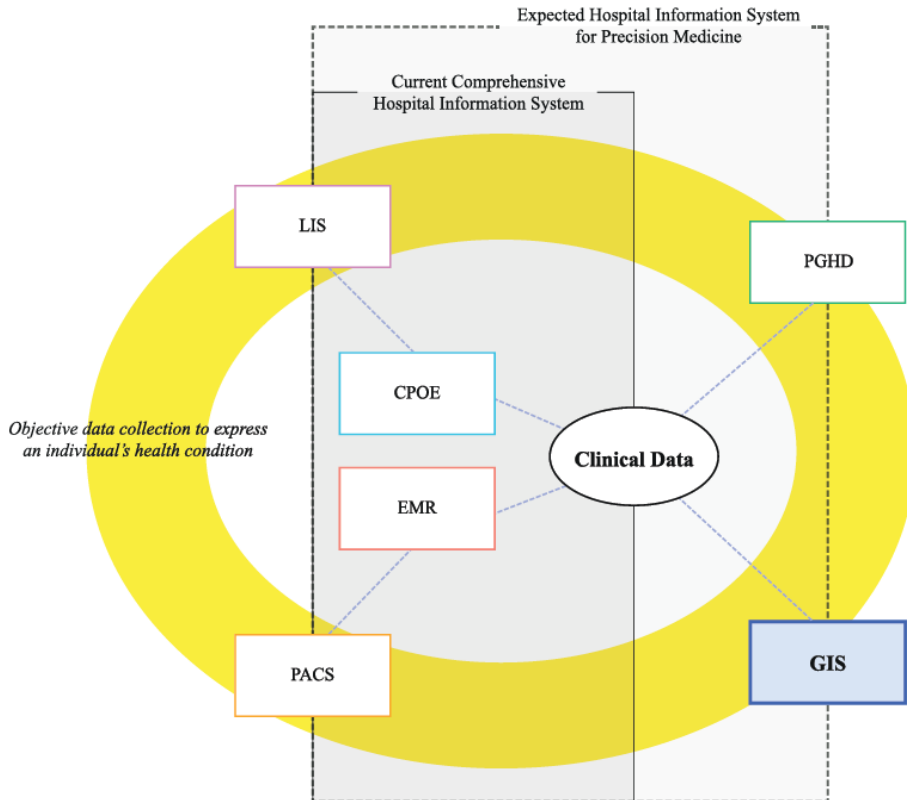
results, along with the detailed analytical process and equipment used, must also be systematically stored and managed, as it is an essential criterion for clinical decision-making<sup>9</sup>. Moreover, because genomic data is less variable than observational data, information integration will allow for maximization of the utility of the collected genomic information for clinical use.

The third challenge, majorly hindering the integration of genomic data with clinical information, is difficulty in mapping the two types of data for medical interpretation. The presence of biomarkers for specific diseases or drug reactions is a critical factor in clinical decision-making<sup>10</sup>. In the case of targeted sequencing, the data processor is informed about biomarkers related to the panel prior to analysis. In clinical practice, reannotation of patient genetic information according to updated biomarker discoveries from the biomedical research community is continuously required at the population level. Thus, a structured data model with consistent data representation would enable the rapid adoption of both evolving biomedical knowledge and individual medical records, which can be delivered to the point of care through agile data processing. Furthermore, patient genomic data expressing specific biomarkers should be readily accessible from the information system along with clinician-confirmed interpretations<sup>10,11</sup>.

Personal-health status can be converted to a composition of multi-layered, multi-dimensional digitalized information for utilization in an information system that facilitates handling big data (Fig. 1). Indeed, vast

amounts of data and associated metadata from multiple medical measuring technologies, such as laboratory tests or imaging studies, have already been successfully merged in clinical information systems. Overall, although genomic information represents the most sound and intensive health-related signals provided by the human body throughout life, the weak links to medical practice highlighted above contribute to its underutilization in clinical decision-making. Therefore, it is necessary to effectively link and integrate clinical information with personal genomic information, helping to accelerate the shift to personalized medicine.

- PGHD : Person Generated Health Data
- LIS : Laboratory Information System
- PACS : Picture Archive and Communication System
- CPOE : Computerized Physician Order Entry
- EMR : Electronic Medical Record
- GIS : Genome Information System



**Figure 1.1 Data-level linkage structure between conventional HIS and GIS**

From a software engineering perspective, a comprehensive hospital information system comprises components that represent separated data collection routes and distinguishing characters of the data. We suggest the concept of GIS to illustrate the implementation of the cGDM. This architecture supports both information and functional integration, even with existing clinical information systems.



## **1.2. Purpose of Research**

The proposed GDM is based on an entity-attribute model to effectively manage and maximize the use of genomic data in clinical practice. Through the development of this method, we focused on equal weighting to the clinical perspective and bioinformatics process analysis as business continuity, starting from the initial clinical intention to bioinformatics information processing associated with a knowledge-related protocol, finally offering a deliverable and interpretable form to the point-of-care clinician. The GDM was designed based on DNA level data from next-generation sequencing (NGS) technology to deliver processed genomic data of patients from different pipelines by applying an appropriate information scale and granularity at the clinical level.

Toward this end, we began by redefining the obstacles to the spread of genomic information into routine care, including reliability problems of proposed measurement data that could cause hesitation in clinical decision-making, and data structure problems that have hindered the integration of genomic data into existing information systems. From a clinical perspective, we focused on the reliability of information as well as the problem of a heterogeneous data structure. In this context, we define a bioinformatics process not as a “measurement,” but rather as a “production” to transition a physical form of existence to an interpretable human representation.

Overall, we aimed to develop a model with appropriate information granularity and scale, which would minimize the possibility of misinterpretation at the point of care by formal and procedural variation related to the production process.

### **1.3. Materials and Method**

The study material was genomic information with clinical relevance based on NGS technology. A failure mode and effect analysis (FMEA) approach was adopted as the analysis process and attributes-extracting method, which was accomplished by assembling a multidisciplinary working group. From November 2017 to July 2018, process mapping, failure identification, and related attribute extraction were performed by the FMEA method at over 18 team meetings. An entity-attribute model was then developed by reconstruction of the attribute set derived from the FMEA.

### **1.3.1. The Production Process of Bringing Genomic Information to Bedside Care**

Here, we define a genomic test as a series of team-based information production processes, in which the meaning of the information is expanded, represented, and reproduced by reference to an external knowledge base, rather than through direct extraction of inherent information. Despite the invariant nature of a personal genome, genomic information presented to a clinician may vary according to specific processing protocols adopted<sup>7,12-14</sup>. This variability raises reliability issues for the use of genomic test results as clinical evidence<sup>15</sup>.

As artifacts from production, genome information processed for clinical use may pose a likelihood of misinterpretation due to information distortion, omissions, and fragmented senses. Furthermore, information reliability is a critical factor determining the ability of clinicians to utilize the genomic information<sup>16</sup>. Thus, our approach in developing this cGDM for focussed on the multi-dimensional scope of information, including procedural factors, derived from NGS technology.

### **1.3.2. FMEA: An Attribute-Clarified Framework**

FMEA is a systematic prospective risk factor analysis approach that predicts and prevents possible errors, improving quality across team-based processes<sup>17</sup>. When used for advanced investigation, the method has advantages enabling exploration of uncertain, unforeseen complex workflows at an early stage<sup>18,19</sup>. Since its introduction in 1963, broad subtype applications of FMEA have been performed in broad domains including reliability engineering<sup>20,21</sup>, behaviour modeling<sup>22</sup>, software engineering<sup>23</sup>, conceptual design<sup>24</sup>, and knowledge management and representation<sup>25,26</sup>. In particular, FMEA has been applied as a method of knowledge representation to extract process reliability-related attributes and to structure and map entities and attributes<sup>22,26-28</sup>. In this study, the FMEA approach was adopted for workflow analysis and the attribute-extracting method.

### **1.3.2.1 The working group**

A multidisciplinary expert team was formed from the areas of bioinformatics, medical informatics, and medicine. The participants included three bioinformaticians, two medical informaticians with clinical informatics and application expertise, and one medical doctor. The medical doctor has experience in both clinical practice and conducting translational research from the perspective of both biomedical science and clinical practice.

### **1.3.2.2 Workflow analysis**

Over a period of nine months, process mapping, failure identification, and related attribute extraction were conducted using FMEA at over 18 team meetings. Structured data modeling for enhancement of data accessibility was then conducted using a logical data model, with the attribute set derived from the FMEA workflow diagram.

We chose the conventional FMEA workflow analysis<sup>21,28</sup> and adapted it for cGDM development. Conventional FMEA consists of two main steps. First, the failure mode is identified through 1) assembling a multi-disciplinary team with at least one expert from each domain over the target production process, 2) combining components and process function in order to derive a workflow diagram, and 3) listing the modes that may lead to failure at each step. The second part involves modifying the process itself with consideration of priority, including 1) evaluating the severity and occurrence ranking of each failure mode and 2) proposing a modified workflow or audition guideline.

In this study, risk estimation and priority-scoring steps were not designed, since our purpose was to review the fragment of metadata composition that may cause unintended information distortion of misinterpretation.

### **1.3.3. Logical Data Modeling**

Data models are the basis of computation ability for intelligent information systems<sup>29</sup>. The database design process can generally be divided into logical and physical database design<sup>30</sup>. The physical data model requires a clear and specific description over logical design, which depends on the existing development environment. Thus, we designed this cGDM as a logical data model based on the FMEA results to support data-level integration with any existing clinical information systems.

Logical data modeling methods are comprised of abstraction and normalization. Database abstraction refers to aggregation and generalization that occur at the points of intersection<sup>31</sup>. We first abstracted the attributes derived from FMEA and expressed the factors corresponding to each step in the workflow. Then, normalization was performed to prevent duplication and inconsistency of data elements considering their names, scale, and relations.



### **1.3.4. Demo Datasets for the real-world data validation**

Two of representative public accessible dataset are selected for the development of the demo databases: The 1000 Genomes Project of the International Genome Sample Resource (IGSR) with population code "CEU" (Utah Residents with Northern and Western European Ancestry)<sup>32</sup>, the pancreatic cancer data from The Cancer Genome Atlas (TCGA\_PAAD)<sup>33</sup>.

Collected datasets were VCF and MAF file format, and the Extract-Transformation-Load (ETL) process of the genomic data was performed by two bioinformaticians with Python 2.7.16. ANNOVAR 2016Oct24 version was used as a clinical annotation tool for the 1000 Genome Project CEU dataset. The resulting dataset imported in a table within the MySQL server database by two medical informaticians. We ran the SQL scripts in MySQL 5.6.46 on a Server with 8GB of RAM and an NVIDIA tesla c1060 / Quad-core CPU running run on CentOS Linux release 7.7.1908. The final outputs took the form of SQL tables and functions.

## **1.4. Results**

This section primarily consists of Failure Mode and Effects Analysis (FMEA) results and entity-attribute modeling. FMEA output is presented in two diagrams: a dataflow diagram that focusses on the derivation of the contents of the genetic test based on NGS sequencing technology, and an information process map that extends the viewpoint to the level of clinico-genomic context. At this step, the protocol entity of the former dataflow diagram was subclassified to reveal the procedural dimension in information processing. Moreover, the set of attributes involved in each step of information transfer was identified. Finally, the cGDM are suggested as a result of structured data modeling based on the attribute set.

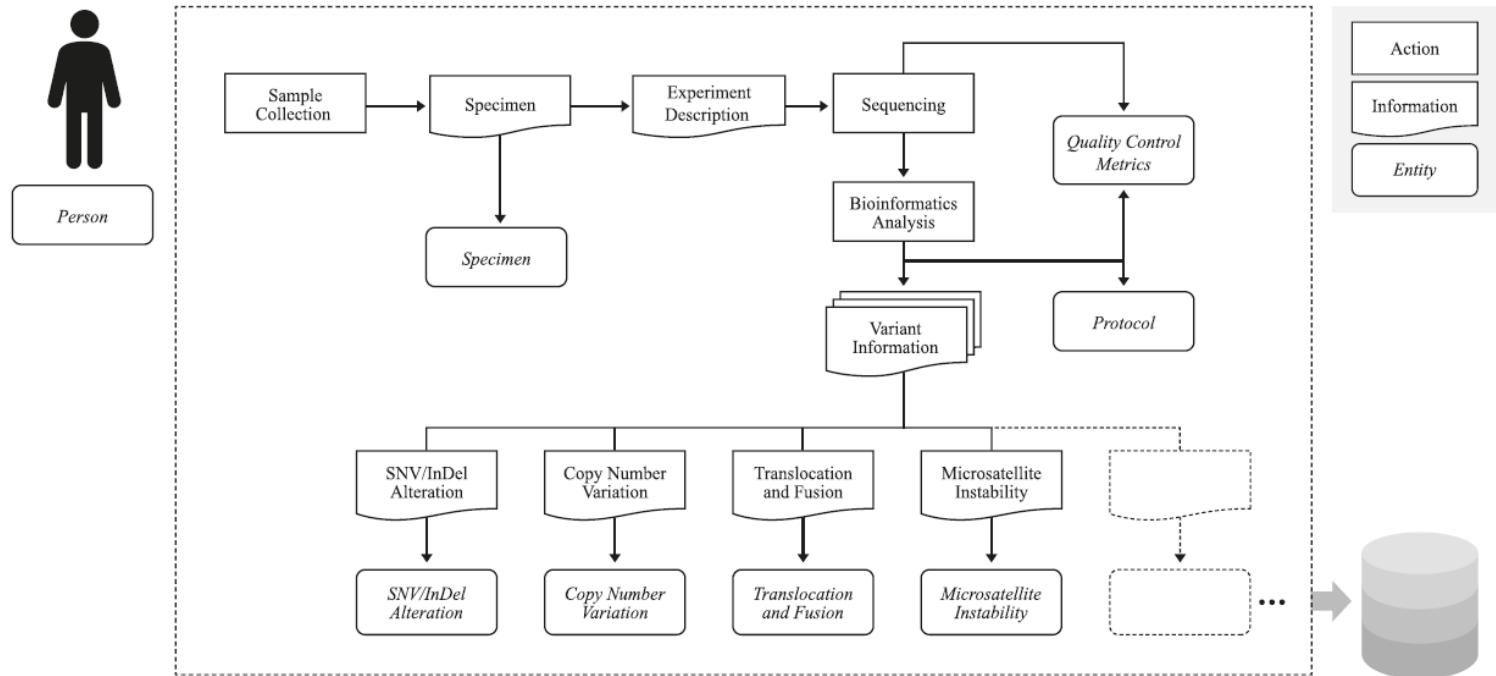
### **1.4.1. Dataflow diagram based on an NGS workflow**

A workflow diagram was derived in order to illustrate the data flow in which the genomic information inherent in the human body is converted to a genomic test result. (Fig 2.) At this stage, the clinical view is minimized, with both the flow of information and the process of analyzing the specimen after the sample collection across experimental laboratory and computational analysis drawn on a large scale.

The subtypes of processed variant information in the parallel structure, used to cope with the growing body of knowledge in bioinformatics, are listed at the bottom of Fig. 1. Variant information can be called in multiple types depending on the perspective and purpose of the analysis. For example, there are four types of genetic variation: single nucleotide variation (SNV), small insertion/deletion (InDel), copy number variation (CNV), and translocation/fusion. There are predictive biomarkers as well such as microsatellite instability (MSI) and tumor mutation burden (TMB).

As the amount of NGS technology-based knowledge increases, more subclasses representing novel perspectives can be added. Scalable data modeling to support the differentiation of knowledge over time is essential not only for expressiveness but also for reducing the burden of information systems maintenance.

In summary, we linked the separate offline workflows at this step that occurred in different places until genomic data could be provided as processed data. The workflow diagram provided the basis for detailed analysis and discussion.



**Figure 1.2 Data flowchart based on a next-generation sequencing workflow**

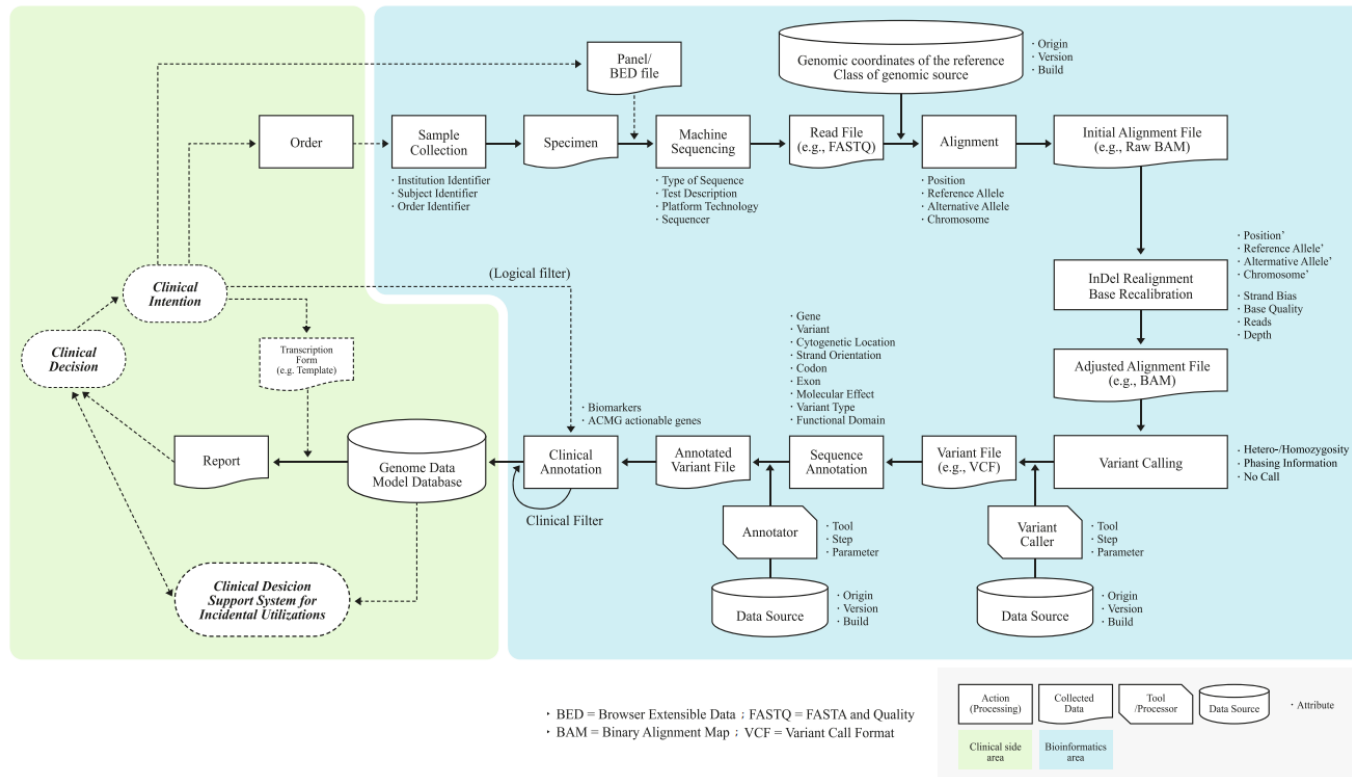
The objects shown in this diagram are classified into three class types- 'Action', 'Information', and 'Entity'. 'Action' was first posted with respect to what occurred in each expert domain and the resulting 'information' was displayed as a result of each action. Finally, 'Entity' was defined as the captured information class at each stage of the workflow. Subtypes of 'Variant Information' were drawn scalable to accommodate the potential extension of subclasses.

## **1.4.2. Extending the NGS process under a clinico-genomic context**

After establishing a consensus on a larger scale, we extended the information flow to the clinical context in detail. At this stage, the standpoint of the workflow analysis was clinical decision making. Hence, the workflow diagram started with a clinical decision. We extended the flow between several actions in the clinico-genomic context involving multiple entities identified, and detailed analysis was performed. In this process, the output data file format and detailed processes for handling output files, along with the tools required for linking to external knowledge databases, are also described.

The working group discussed mechanisms for extraction of the entity-attribute set which would avoid probable information distortion and omission. We considered that the genomic data model for clinical use should be the knowledge communication scheme, thus preserving its reliability-related factors. At a minimum, the genomic data model must provide sufficient information to decide whether the confidence level of the genomic test result justifies its consideration as clinical evidence. For this function, failure was defined as that which causes misinterpretation or non-use of the genomic data for clinical decision. The process of producing clinical evidence from genomic data at the bioinformatics area (Fig. 3) shows a pattern that is a series of repeated representations of information converted

by reference knowledge bases and data processing rules. Thus, failure modes can be classified as incomplete specifications in three meta-categories: origin, reference, or symbol. Due to the nature of the semantic interpretation, any fragmentation of symbol causes not only loss of information but also assignment information to direct the origin<sup>12,13</sup>.

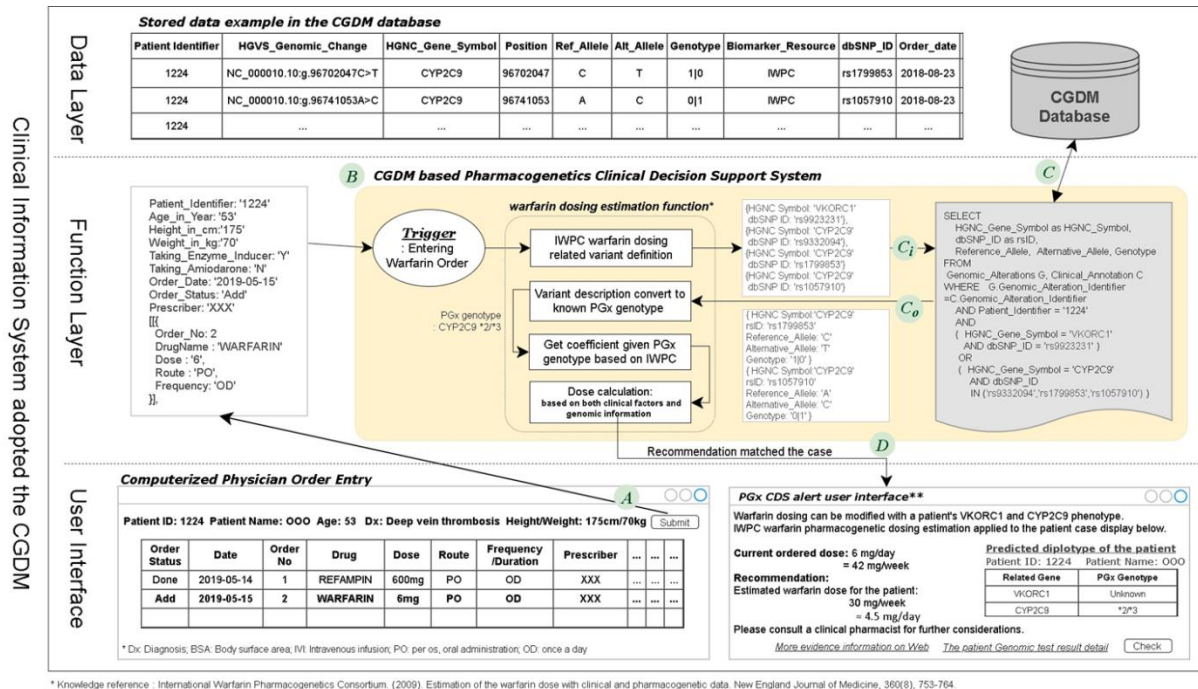


**Figure 1.3 Failure mode identification: mapped next-generation sequencing process extended to a clinico-genomic context**

In the bioinformatics area (cyan background), information may be distorted by the insufficient representation of origin, processing rule, and external reference. To prevent this failure, identification and semantics, related attributes are listed under the boxes. In the clinical area (yellow background), the data model functions as a communication scheme for the collaborative process implemented in the hospital information system. Data-level integration facilitates just-in-time queries and reuse of data.



We conducted workflow analysis to extrapolate general descriptors of the related attributes with the goal of preserving information during production and delivery processes from clinical intention to clinical utilization. Figure 2 provides a more detailed data-level view, including how genomic information is realized as clinical evidence in a case based on a structured data model. The structured genome data model can support a report via presentation on a variety of transcription forms (report forms), which are optimized for initial intent. Furthermore, additional utilization paths are accessible in the clinical-information system. As shown in Fig. 2, data-level integration helps the amplification of the incidental utilization. (Fig. 4) To illustrate, consider a patient who orders whole-genome sequencing to screen for cancer biomarkers at their first visit. When the patient receives a prescription for antibiotics a year later at a visit for other symptoms, that same genomic test result can be re-used from a pharmacogenomics perspective for safer and more efficient drug prescription. The clinical decision support system plays a vital role by just-in-time display of the matching information with pre-defined rule and knowledge-based processing<sup>6,34,35</sup>. A computational genome data model is a prerequisite for this implementation<sup>35-37</sup>. Finally, we introduce a logical data model in the next step of the study.



### **1.4.3. The cGDM**

Finally, the cGDM was designed as an entity-attribute model consisting of 8 entities and 46 attributes (Fig. 5). For a structured data model of the identified clinico-genomic attributes, logical modeling was conducted to ensure data-level linkage with conventional primary clinical databases. In order to define the entity-attribute model based on the action and collected data, tool/processor classes and the attributes of each class from Fig. 2, we define three types of classes as protocol and related attributes (Table 1). Since the cGDM is designed to support data-level integration with the existing system, only the minimum subject identifier is defined as ‘linkage identifier to clinical information.’ To represent the procedural dimension, which is stressed in the study, we combined two workflow analyses on different scales. For example, the entity ‘Protocol’ as a part of the procedural dimension is explicitly represented in Fig. 2, then expressed again as a list of lower steps in Fig. 3. Since clinical observation is typically considered as the collection of events<sup>38</sup>, the logical composition of the date/time and actor identifier related to the clinico-genomic context were declared.

The derived classes and entities in Table 1 were used to declare final entities and attributes in the cGDM (Fig. 5). The mapped Actions and Action-related classes (Collected Data and Tool/Processor) are categorized into subdomains and related attributes for each step in Table 1. In Table 1,

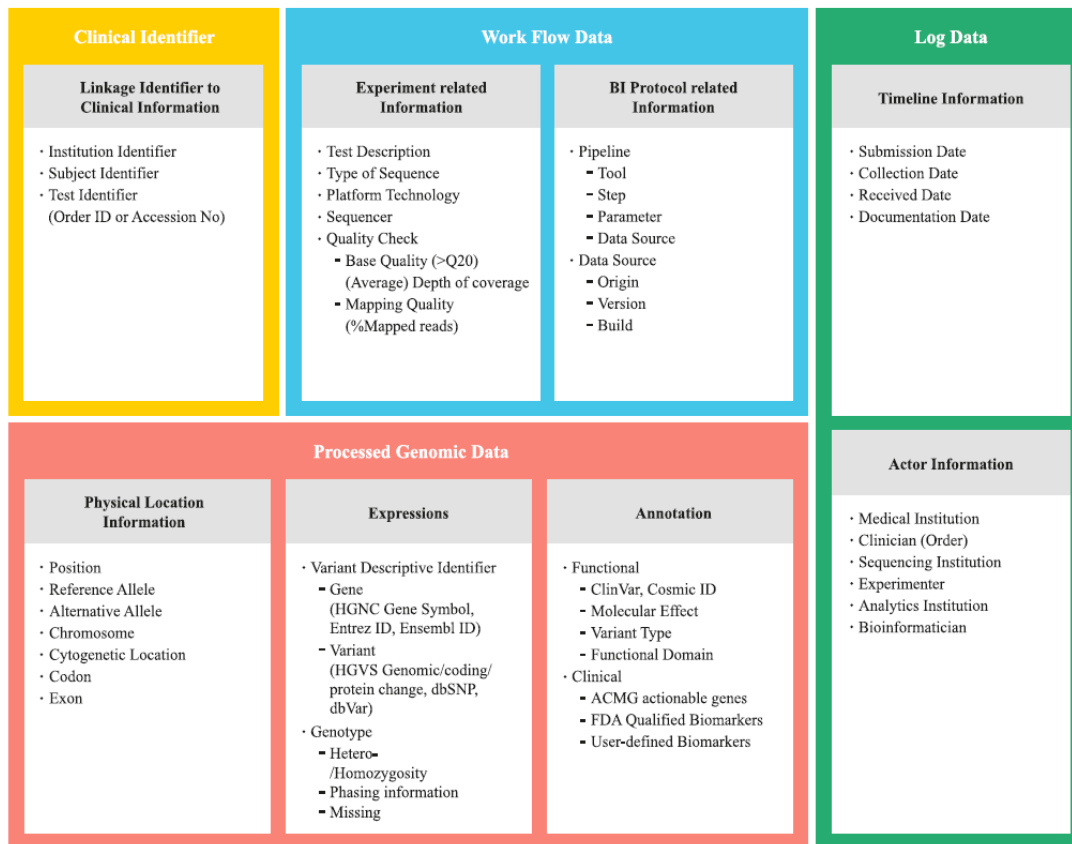
action and its result are grouped into one step, and the related attributes are represented by the attributes classified in the corresponding step. For normalization, related attributes are categorized to create one or more new groups called 'entities' for each step, and they are the basis for defining 'Entities' in the Entity-Attribute model (Fig. 5). For example, 'Physical information according to the coordinate system' is one of the three subdomains of the action 'Sequence Annotation.' It can include an attribute set (Cytogenic location, Codon, Exon) representing physical location information for each variant. However, this "Physical information according to coordinate system" can be a subdomain in other steps besides "Sequence Annotation". And even though it is the same subdomain, the related-attribute set may be different depending on which step or action. In summary, each step identified in the entire clinico-genomic process can include multiple entities, and one entity can be related to multiple steps. Even in the same entity, the configuration of the related attribute as a factor affecting each step may vary from step to step.

**Table 1.1 Extracted classes and related attribute sets from each step of clinic-genomic context for the Entity-Attribute model.** The processes in the clinico-genomic workflow shown in Figure 2 are listed in order and associated with the classes, related attribute sets for each process. This table is an intermediate result between the result of FMEA and the final logical model. Derived related attributes are abstracted within each class and grouped into entities.

Action	Class		Related Attribute	Entity
	Collected Data	Tool/ Processor		
Sample Collection			Institution Identifier Subject Identifier Test Identifier (Order ID or Accession No)	Linkage Identifier to Clinical Information
			Submission Date	Timeline Information
			Medical Institution Clinician	Actor Information
Specimen				
Machine Sequencing			Test Description Type of Sequence Platform technology Sequencer Collection Date	Experiment Related Information Timeline Information
			Sequencing Institution Experimenter	Actor Information
Read File				
Alignment			Position Reference allele Alternative allele Chromosome	Physical(Location) information according to coordinate system
			Analytics Institution Bioinformatician	Actor Information
Initial Alignment File				
InDel Realignment / Base Recalibration			Position <sup>c</sup> Reference allele <sup>c</sup> Alternative allele <sup>c</sup> Chromosome <sup>c</sup>	Physical(Location) information according to coordinate system
			Base quality(>Q20) (Average) Depth of coverage Mapping Quality (%Mapped reads)	Quality Check information

		Received Date	Timeline Information
		Analytics Institution	Actor Information
	Adjusted Alignment File		
Variant Calling		Hetero- /Homozygosity	Genotype Expressions
		Phasing information	
		Missing	
		Analytics Institution	Actor Information
		Bioinformatician	
	Variant Caller	Tool Step	Pipeline information
		Parameter	
		Origin	
		Version	Data source
		Build	
		Parameter	
	Variant File		
Sequence Annotation		Gene (HGNC Gene Symbol, Entrez ID, Ensembl ID)	
		Variant (HGVS(genomic, coding, protein change + version), dbSNP, dbVar)	Variant Descriptive Expressions
		Cytogenetic location	Physical(Location) information according to coordinate system
		Codon	
		Exon	
		ClinVar, COSMIC ID	
		Molecular Effect	Functional Annotation
	Variant Type		
	Functional Domain		
		Analytics Institution	Actor Information
		Bioinformatician	
	Annotator	Tool Step	Pipeline information
		Parameter	
		Origin	
		Version	Data source
		Build	
	Annotated Variant File		

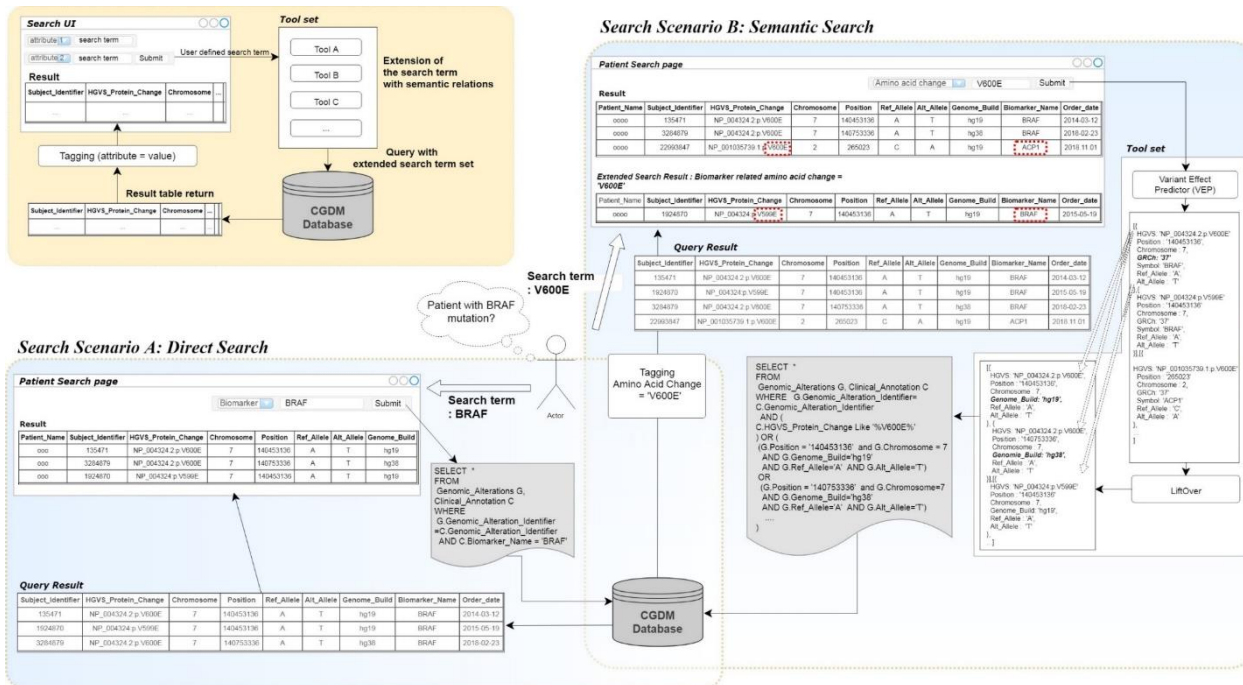
Clinical Annotation	ACMG actionable genes	
	FDA qualified biomarkers	Clinical Annotation
	User-defined biomarkers	
	Analytics Institution Bioinformatician	Actor Information
	Documentation Date	Timeline Information



**Figure 1.5 The Clinical Genome Data Model: Structured data modelling with entities and attributes**

The cGDM is designed as a logical data model of 8 entities and 46 attributes. The objects and related attributes derived through FMEA are integrated into a logical data model through abstraction and normalization.





**Figure 1.6 Semantic search implementation based on the CGDM**

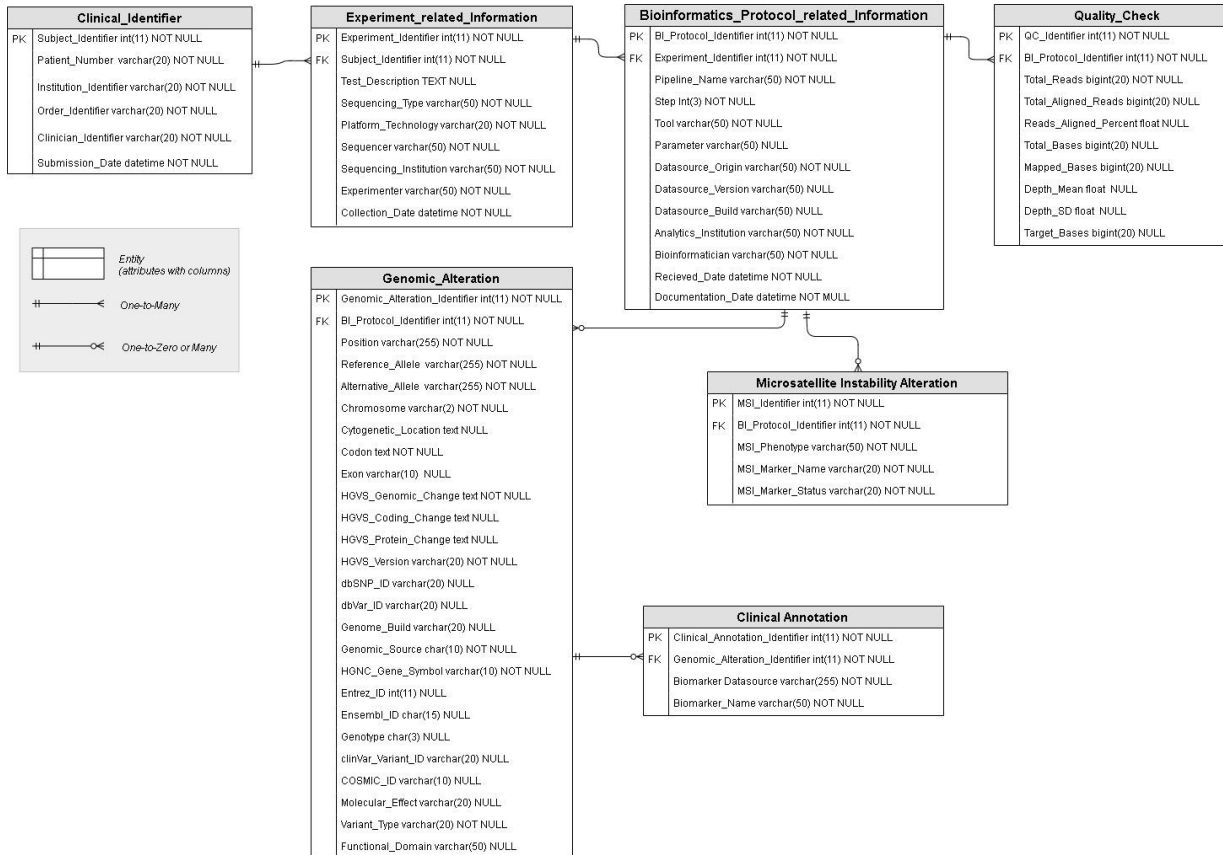
Even if the user does not know all the nomenclature or metadata relevant to the genomic data to be searched, search function based on the CGDM can use information entered in the search fields in order to derive an extended search result. Through the generated SQL syntax, the user can determine which genomic metadata (such as chromosome and position, genome build version, HGVS ID) can be associated and extended to the outcome of the patient's data. In addition to the attributes "Biomarker" and "HGVS ID" presented in the example, multiple data queries can be made with a single attribute or combination of attributes changed in the CGDM. Therefore, by using these user interfaces with the data model, it is possible to trace and verify whether the queried genomic data of the patient represent more reliable information.

#### **1.4.4. Validation of the cGDM**

Here, the cGDM was finalized in the form of a logical model, which allows adaptation to the diverse development environments of existing heterogeneous clinical information systems. Logical model can play an essential role to generalize the complex phenomenon by abstraction and enhance understanding core ideas the model deliver between different stakeholders of in the complex system<sup>39</sup>. Whereas, the drawback of this approach is that physical modeling layer is needed in order to the data model implementation and validation. Thus, we design a physical data model implemented in relational database to evaluate the model validity for real-world data and to proof of concept how implementation of the cGDM enables interactive clinical decision support in clinical information system shown as Fig3 (Left side; Clinical decision support system for incidental utilization).

**Figure 1.7. Entity-relationship diagram of the CGDM implemented in RDBMS**

The entity-relation for the physical model as a diagram (ERD) was presented based on the table shown in Supplementary Table 1. The diagram shows the entities and the attributes that describes the entity, and the relationship between the entities is also defined.



### **1.4.4.1. Implementation of the real world data**

This physical data model of the cGDM is provided in forms of entity-relationship diagram and table (Supplementary Information Table 1; Fig 1.7). Also, one-click executable data definition language script is also freely accessible on a web page (<https://github.com/SNUBI-HyojungKim/cGDM-Clinical-Genome-Data-Model>).

For the data model validation with real-world data, we built pilot databases based on the cGDM and uploaded genomic data of over 2,000 patients for multiple diseases, including acute lymphoblastic leukaemia, solid cancers, and depression cases (Table 2, internal databases). However, the pilot dataset related researches remains undergoing, we have built two representative demo datasets for open source (Table 2, demo databases) 1) 1000 genome CEU (Utah Residents with Northern and Western European Ancestry) population dataset for whole genome sequencing (n=99, 47.67 GB), 2) TCGA PAAD (Pancreatic Adenocarcinoma) dataset for somatic mutation (n=155, 9.41 MB). We believe those well-known public dataset has advantages on data validation issue. Every demo dataset and source codes are freely available from at the Github page as mentioned above.

**Table 1.2 Summary of imported genomic data from various data sources in cGDM databases.**

The databases are categorised into internal and demo database. The specifications of the database tables are informed in Table 1. This table presents row counts of each database table and data volumes of each database. The internal databases includes 3 private datasets (cancer panel, leukemia and depression) and 2 public datasets (TCGA COAD and TCGA LUAD). The demo databases includes 2 public datasets (1000 Genome Phase3 CEU and TCGA PAAD).

		Database							Summary
		Internal database					Demo database (public license)		
		Cancer Panel	Leukemia	Depression	TCGA COAD	TCGA LUAD	1KGP P3 CEU	TCGA PAAD	
Table name	Type of sequencing	cancer panel	WES	WES	somatic mut.	somatic mut.	WGS	Somatic mut.	7 data sets
									WGS/WES/ targeted panel
CLINICAL_IDENTIFIER		10	503	1,000	459	522	99	155	2,748
EXPERIMENT_RELATED_INFORMATION		10	517	1,000	459	522	99	155	2,762
BIOINFORMATICS_PROTOCOL_RELATED_INFORMATION		10	517	1,000	459	522	99	155	2,762
GENOMIC_ALTERATION		2733	29,279,631	842,199,347	361,933	318,947	229,525,363	56,159	1,101,744,113
MICROSATELLITE_INSTABILITY		0	0	0	0	0	0	775	775
CLINICAL_ANNOTATION		40	267	108	123	97	1	12	648
QUALITY_CHECK		10	517	1,000	0	0	0	0	1,527
Data volume	database total	2 MB	8.2 GB	144.7 GB	48.4 MB	42.6 MB	47.7 GB	9.4 MB	201.5 GB
	per test	0.2 MB	8.12 MB	144.7 MB	0.1 MB	0.1 MB	481 MB	0.6 MB	91.8 MB

Real-world data validation is designed to cover all three types of NGS tests (targeted panel, WES, WGS) and both cases of somatic mutations and germline variants. The storage capacity of data was reduced when converted into relational database with cGDM schema by 30% compared to the prepared data file in VCF format. Interestingly, as the data size of the genomic alteration table per test increased, the gap in data size by converting narrowed or overturned. The circumstance is due to the addition of multiple indexes for in-time query performance. Table indexing was generally required when an average of more than 30,000 rows per test occurs.

### **1.4.4.2. How the implementation of the cGDM enables interactive clinical decision support**

One of the major challenges of healthcare informatics is supporting clinicians who need to handle constantly evolving knowledge and inherently complex genomic data. Patient genomic data in static document format or in structured model but in which has vague designation of the variant limits functionality of clinico-genomic information system<sup>40</sup>. The cGDM could address the issue by working as a data-level infrastructure for interactive clinical decision support along with external knowledge bases (Fig.6). For the cGDM's programmability test, we developed a pharmacogenomic clinical decision support function running on the cGDM database which reflects the knowledge of the IWPC warfarin dosing algorithm. The source code is freely available at <https://github.com/SNUBI-HyojungKim/cGDM-Clinical-Genome-Data-Model>. Figure 7 illustrates both of logical information flow in back-end system and its appearance on the user interface. A query performance test is conducted with the algorithm procedure over 99 individuals in 1KGP P3 CEU database. The SQL stored procedure has executed in MySQL on a server with 8GB of RAM and quad-core CPU running Linux CentOS 6. The average query out duration was  $0.013 \pm 0.008$  second range from 0.00001 to 0.0460.

## Integrated in clinical workflow

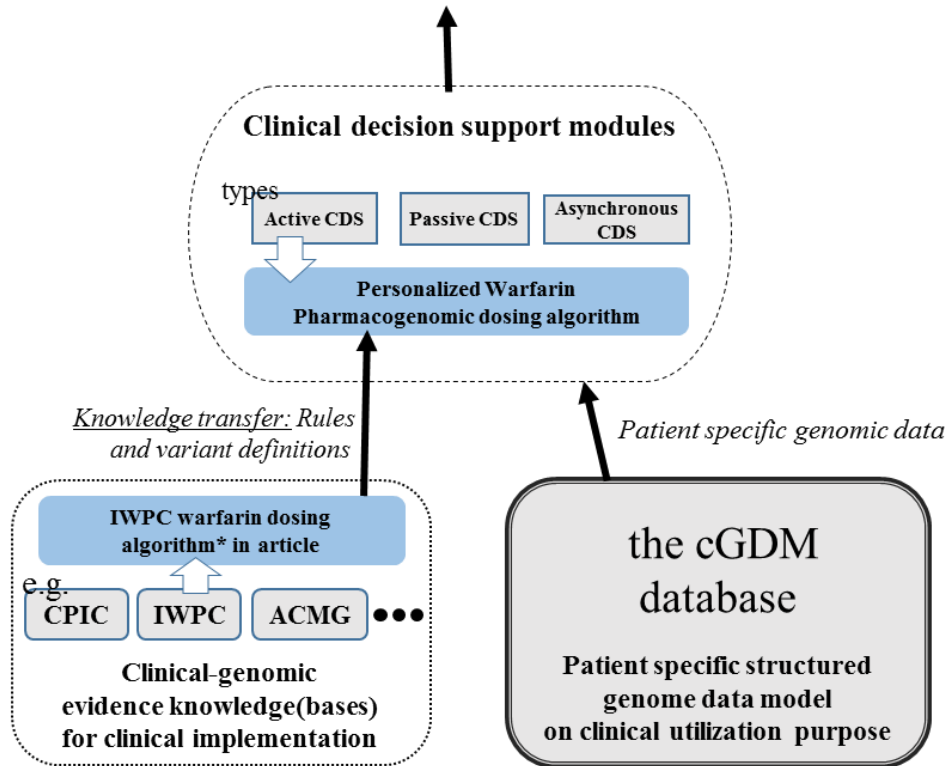


Figure 1.7 The conceptual map of genomic decision support system based on the cGDM

While the accumulation of confirmatory knowledge could seem relatively slow compared to the speed of the vast discovery of the bioinformatics field, the benefits and impacts the two will have on patients when they are seamlessly connected are evident. The cGDM brings this process into computational space.



# **Chapter 2. Pharmacogenomic Clinical Decision Support: Modular Implementation of CPIC Guideline**

## **2.1. Introduction**

As the development of sequencing technology and the results of research on pharmacogenomics (PGx) accumulate, efforts are being made to apply personalized drug prescriptions and dose adjustments in the clinical field. The same drug may cause adverse reactions due to congenital or acquired causes, and drug adverse reactions are a major obstacle to the safe and effective use of drugs. “The social costs and health disadvantages of these adverse drug reactions are well known. PGx use cases are of particular interest because over half of all primary care patients are exposed to PGx relevant drugs. Studies have found that 7% of U.S. Food and Drug Administration (FDA)-approved medications and 18% of the 4 billion prescriptions written in the United States per year are affected by actionable PGx variants that nearly all individuals (98%) have at least one known, actionable variant by current Clinical Pharmacogenetics Implementation Consortium (CPIC) guidelines and that when pharmacogenes with at least one known, actionable, inherited variant are considered, over 97% of the U.S. population has at least one high-risk diplotype with an estimated impact on nearly 75 million prescriptions”<sup>41</sup>. Therefore, it is expected that

applying knowledge about the drug genome to avoid predictable adverse reactions to patients and maximizing the effects of drug treatment prior to drug prescriptions would improve patient safety and quality of treatment.

Various efforts are being made to establish a knowledge resource of pharmacogenomic knowledge that can be applied in clinical practice and to connect it to clinical information systems. Representatives are as follows; Clinical Pharmacogenetics Implementation Consortium (CPIC) of the Pharmacogenomics Research Network (PharmGKB)<sup>42</sup> and the Dutch Pharmacogenetics Working Group (DPWG)<sup>43</sup>, International warfarin pharmacogenetics consortium (IWPC)<sup>44</sup>, Canadian Pharmacogenomics Network for Drug Safety (CPNDS)<sup>45</sup>. Efforts have been made to implement informed decision making using pharmacogenomic information in clinical settings based on these refined knowledge resources. In particular, recent attempts at systematic clinical implementation have been reported by the European Consortium<sup>46</sup>, the IGNITE Network Pharmacogenetics Working Group<sup>47,48</sup>, and the United Kingdom<sup>49</sup>. In order for PGx to become routine in practice, attention has been paid to establishing a PGx decision support system integrated with EHR.

However, it has not been proposed as a sustainable, scalable, and interoperable design among different sites. When considering the complexity of dealing with the volatility of PGx knowledge and the considerable amount of information in patient-specific genomic data as an

extension of the clinical context, PGx clinical decision support pipeline focused on knowledge representation is needed. Moreover, data processing methods is needed to provide PGx test result on demands. Clinical decision support (CDS) holds great promise for genomics but has had limited utility because executing CDS has required manual entry of genetic conditions into the problem list for decision support<sup>50</sup>.

In the study, we aim to develop a PGx CDS pipeline linking between clinical actionable drug-gene interaction knowledge and personal genomic data. First of all, we transform CPIC guideline knowledge resources into a machine-readable structured database. Finally, we suggest a PGx CDS service design based on the data model layer, both on CPIC guideline knowledge resources and personal genomic data.

## 2.2. Purpose of Research

We propose PGx CDS that enables modular implementation between heterogeneous existing clinical information systems. Modeling of medical knowledge and representation of and reasoning about medical knowledge are the significant steps of the construction of CDS tool<sup>70</sup>. Although CPIC guidelines supporting the clinical application of pharmacogenomics knowledge provide reliable content, considerable modeling activities are required to transform knowledge from human-interpretable form to a machine-readable form for consistent application.

Thus, we firstly collected, integrated CPIC guideline contents. Data integration gives a unified landscape by combining data from disconnected resources<sup>51</sup> In this process, modeling the relationship between the sources and the global schema is, therefore, a crucial aspect. Then, we transform CPIC guideline knowledge resource to the machine-readable structured database along with content analysis. Exploratory analysis of the collected dataset reveals the rules or properties that the content implicitly implied. Finally, we propose a modular PGx CDS service by capturing the explicit and implicit knowledge flow of the CPIC knowledge resource through the modeling process and seamlessly unites actionable drug-gene interaction knowledge with patient genomic information on computational space.

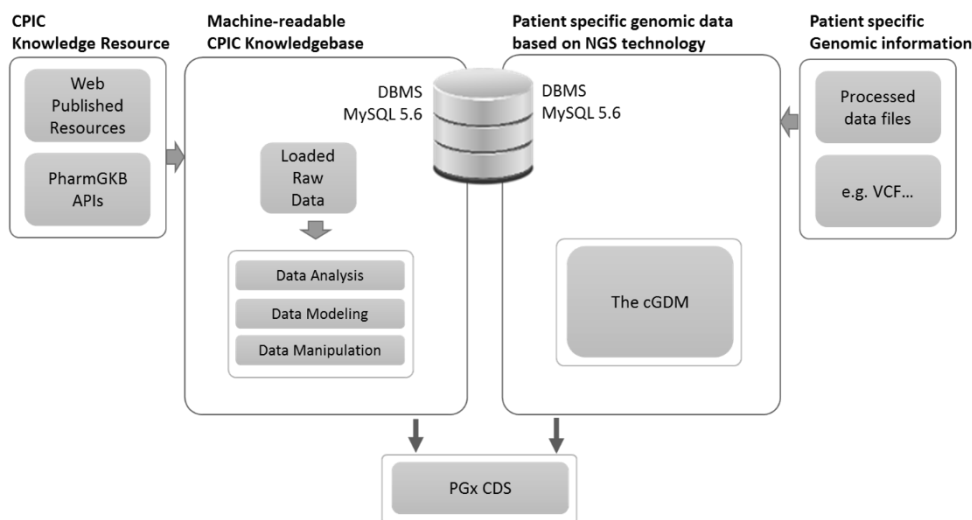
## **2.3. Material and Methods**

### **2.3.1 Material: CPIC guideline as knowledge resource**

The CPIC was formed in 2009 as a shared project between PharmGKB (<https://www.pharmgkb.org>) and the Pharmacogenomics Research Network (PGRN) (<http://www.pgrn.org>). One of the goals of CPIC is to provide peer-reviewed, updated, evidence-based, freely accessible guidelines for gene-drug pairs<sup>6</sup>. All CPIC guidelines adhere to a standard format, and the terms used in CPIC guidelines to describe allele function and phenotype are standardized<sup>7,52</sup>. An ultimate goal for CPIC guidelines is to provide actionable guidelines for clinicians to make more precision decisions for specific drugs when genetic results are available. As a result of the admirable contribution of the consortium, it provides the most world-widely adoptable clinical pharmacogenomic implementation knowledge base. Efforts are underway to make CPIC guidelines more machine-readable, including making the guidelines available in various file formats<sup>53</sup>.

## 2.3.2. Data Collection

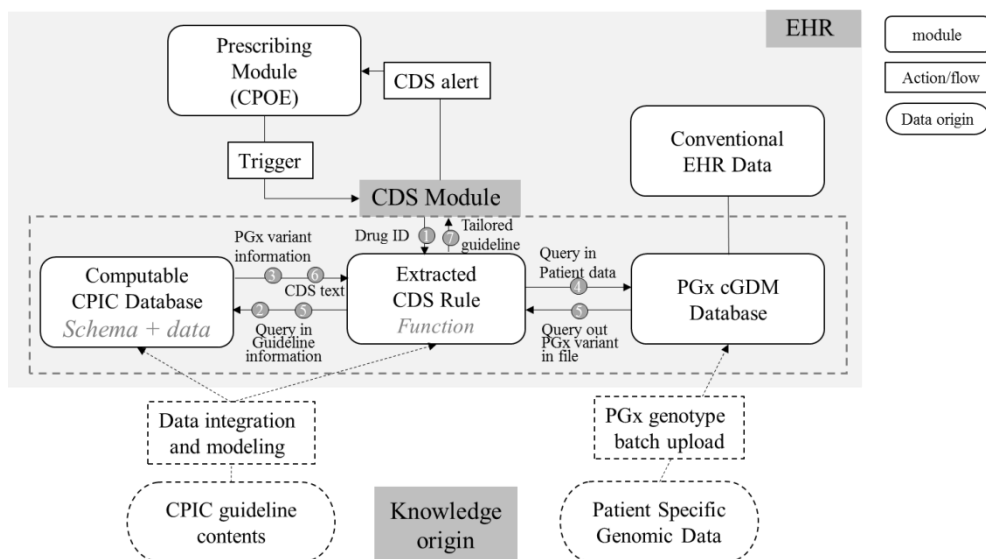
CPIC guideline datasets are first collected between July 10th and August 30th, and updated between 2019 March 15th and March 30th in 2020, via open assessed CPIC webpages and PharmGKB APIs. Collected data items are as follows; guideline list (drug-gene pair information included), drug resource mapping, gene resource mapping, gene allele definition, gene diplotype phenotype, clinical decision support guidelines. Except for the guideline list, other data formats are downloaded in comma-separated values form. Collected datasets are imported to a relational database management system (MySQL 5.6) for exploratory analysis and data-driven restructuring.



\*CPIC: The Clinical Pharmacogenetics Implementation Consortium; DBMS: database management system; PGx CDS: pharmacogenomics clinical decision support system; VCF: variant call format

**Figure 2.1. The configuration of the study environment**

### 2.3.3. Clinical decision support service architecture



**Figure 2.2. Modular implementation of PGx CDS overview**

As discussed in Chapter 1, we perceive patient-specific genomic information as a sub-dimension of representation that reflects the patient's health status. Therefore, we consider the data level integration so that the service architecture ensures agile combined and computation with other sub-dimensional information.

Among collected 6 CPIC content categories, guideline title, drug resource mapping, gene resource mapping, and gene allele definition are used to construct a computable CPIC database (Figure 2.2, middle-left). Others, gene diplotype-phenotype and clinical decision support guideline categories, are applied to CDS rule function that matches PGx variant definition and patient genomic information and selects a personalized PGx CDS to alert given drug prescribing condition. The cGDM is adopted as a

patient-specific genome data model, developed in Chapter 1, to serve as a data layer infrastructure supporting the intellectual interplay between medical experts and informed decision-making.



## **2.4. Results**

### **2.4.1. Collected CPIC guideline and exploratory analysis**

The CPIC guidelines reviewed for machine-readable data conversion are a total of 24 guideline entries (Table 2.1) published to date on the official website<sup>42,54</sup>. Each guideline contains specific information related to certain gene-drug pairs; unique 20 genes and 62 drugs. Each guideline gives well-curated knowledge in forms of procedural subcategories such as drug resource mapping, gene resource mapping, gene allele definition, gene diplotype-phenotype, allele frequency, clinical decision support guidelines. However, mainly due to differences in how each gene affects the drug efficacy or biological characters, the composition of the provided items are varied.

Table 2.2 shows representative CPIC content items and their dataset availability according to each guideline. In the case of drug and gene resource mapping, every dataset is available. HLA-A and HLA-B gene allele definitions are not defined in CPIC standard format due to its unique biological character and high complexity. Gene diplotype-phenotype tables are not provided when the former form of information is not describable, or the only haplotype is existed (G6PD), or the structural variants have a meaningful proportion in the PGx gene. When the items reflect the PGx

drug-gene interpretation process, ensuring the entire item shows the feasibility of building a seamless digitalized pipeline. To explicit clinical decision support workflow and recommendation text files, guidelines that have complete data items are 10; 1) CYP2D6, CYP2C19 and Tricyclic Antidepressants (for 2 of 7 drugs), 2) CYP2D6 and Atomoxetine, 3) TPMT, NUDT15 and Thiopurines, 4) DPYD and Fluoropyrimidines, 5) CYP2D6, CYP2C19 and Selective Serotonin Reuptake Inhibitors, 6) RYR1, CACNA1S and Volatile anesthetic agents and Succinylcholine, 7) CYP2B6 and efavirenz, 8) CYP2D6 and Ondansetron and Tropisetron, 9) CYP2D6 and Tamoxifen, CYP2C19 and Voriconazole, 10) CYP2C9 and NSAIDs (for 7 of 15 drugs).

**Table 2.1. The collected CPIC guideline overview**

CPIC Guideline Title	Drug or Ingredient (unique n = 62)	Gene (n = 20)
HLA-B and Abacavir	abacavir	HLA-B
HLA-B and Allopurinol	allopurinol	HLA-B
CYP2D6, CYP2C19 and Tricyclic Antidepressants	amitriptyline, clomipramine, desipramine, doxepin, imipramine, nortriptyline, trimipramine	CYP2C19, CYP2D6
UGT1A1 and Atazanavir	atazanavir	UGT1A1
CYP2D6 and Atomoxetine	atomoxetine	CYP2D6
TPMT, NUDT15 and Thiopurines	azathioprine, mercaptopurine, thioguanine	TPMT, NUDT15
DPYD and Fluoropyrimidines	capecitabine, fluorouracil, tegafur	DPYD
HLA-A, HLA-B and Carbamazepine and Oxcarbazepine	carbamazepine, oxcarbazepine	HLA-A, HLA-B
CYP2D6, CYP2C19 and Selective Serotonin Reuptake Inhibitors	citalopram, escitalopram, fluvoxamine, paroxetine, sertraline	CYP2D6, CYP2C19
CYP2C19 and Clopidogrel	clopidogrel	CYP2C19
CYP2D6 and Codeine	codeine	CYP2D6
RYR1, CACNA1S and Volatile anesthetic agents and Succinylcholine	desflurane, enflurane, halothane, methoxyflurane, isoflurane, sevoflurane, succinylcholine	RYR1, CACNA1S
CYP2B6 and efavirenz	efavirenz	CYP2B6
CFTR and Ivacaftor	ivacaftor	CFTR
CYP2D6 and Ondansetron and Tropisetron	ondansetron, tropisetron	CYP2D6
IFNL3 and Peginterferon-alpha-based Regimens	peginterferon alfa-2a, peginterferon alfa-2b, ribavirin	IFNL3
CYP2C9, HLA-B and Phenytoin	phenytoin	CYP2C9, HLA-B
G6PD and Rasburicase	rasburicase	G6PD
SLCO1B1 and Simvastatin	simvastatin	SLCO1B1
CYP3A5 and Tacrolimus	tacrolimus	CYP3A5
CYP2D6 and Tamoxifen	tamoxifen	CYP2D6
CYP2C19 and Voriconazole	voriconazole	CYP2C19
CYP2C9, VKORC1, CYP4F2 and Warfarin	warfarin	CYP2C9, VKORC1, CYP4F2
CYP2C0 and NSAIDs	aspirin, diclofenac, celecoxib, flurbiprofen, aceclofenac, ibuprofen, indomethacin, lornoxicam, lumiracoxib, meloxicam, metamizole, nabumetone, naproxen, piroxicam, tenoxicam	CYP2C8, CYP2C9

**Table 2.2. Dataset list and its availability over guidelines**

CPIC Guideline Title	Original Publication Date	Most Recent Update Date	Drug Resource Mapping	Gene Resource Mapping	Gene Allele Definition	Gene Diplotype-phenotype	Clinical Decision Support
HLA-B and Abacavir	April 2012	May 2014			Not available	Not available	Not available
HLA-B and Allopurinol	February 2013	June 2015			Not available	Not available	Not available
CYP2D6, CYP2C19 and Tricyclic Antidepressants	May 2013	October 2019					(2/7)
UGT1A1 and Atazanavir	September 2015	November 2017					Not available
CYP2D6 and Atomoxetine	February 2019	October 2019					
TPMT, NUDT15 and Thiopurines	March 2011	February 2019					
DPYD and Fluoropyrimidines	December 2013	January 2020					
HLA-A, HLA-B and Carbamazepine and Oxcarbazepine	September 2013	December 2017			Not available	Not available	
CYP2D6, CYP2C19 and Selective Serotonin Reuptake Inhibitors	August 2015	October 2019					
CYP2C19 and Clopidogrel	August 2011	March 2017					Not available
CYP2D6 and Codeine	February 2012	October 2019					Not available
RYR1, CACNA1S and Volatile anesthetic agents and Succinylcholine	November 2018	September 2019				Not applicable*	
CYP2B6 and efavirenz	April 2019	No updates					
CFTR and Ivacaftor	March 2014	May 2019				Not available	Not available
CYP2D6 and Ondansetron and Tropisetron	December 2016	October 2019					
IFNL3 and Peginterferon-alpha-based Regimens	February 2014	No updates				Not available	Not available
CYP2C9, HLA-B and Phenytoin	November 2014	No updates			Not available	Not available	Not available
G6PD and Rasburicase	August 2014	September 2018				Not available	Not available
SLCO1B1 and Simvastatin	October 2014	No updates					Not available
CYP3A5 and Tacrolimus	July 2015	No updates					Not available
CYP2D6 and Tamoxifen	January 2018	October 2019					
CYP2C19 and Voriconazole	December 2016	No updates					
CYP2C9, VKORC1, CYP4F2 and Warfarin	December 2016	No updates				Not applicable*	Not available
CYP2C9 and NSAIDs	March 2020	No updates	(7/15)	(1/2)	(1/2)	(1/2)	(7/15)
Number of available files grouped by guidelines			23	23	20	15	11

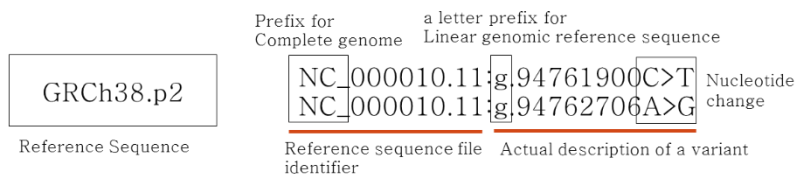
## 2.4.2. Data integration and modeling

In this section, we briefly examine each CPIC content item in terms of its attribute and value set. On top of that, *CPIC guideline title* contains drug-gene pair information at ingredient or drug class level. *Drug resource mapping* file provides for each drug of ingredient, respectively, which has four attributes; ‘Drug or Ingredient,’ ‘Source,’ ‘Code Type,’ ‘Code.’ Source attribute has a member of RxNorm, DrugBank, ATC, PharmGKB. In summary, this item provides definitions of drugs that can be identified in four representative external drug knowledge bases. *Gene resource mapping* file is also expressed in the same attribute set, and provides unique indexes of 4 different external genome knowledge bases for each gene; PharmGKB, Ensembl, NCBI, HGNC.

*The Gene allele definition table* can be divided into four districts when clustered with similar value properties as below (Figure 2.3). This table is a collection of PGx variant information in a gene. For example, we can start \*4 in the C district. At the same line in D district, we can find the alternative allele Y and G. In the first line of those, reference allele C and A are shown. We could make the exact HGVS nomenclature when combine assigned A+B district. In this case, CYP2C19 \*4 consists of two variants; NC\_000010.11:g.94761900C>T and NC\_000010.11:g.94762706A>G. This expression is interoperable with any line of A+B, for example, rs12248560 and rs28399504 in terms of rsID from NCBI dbSNP. The machine cannot

interpret the table, evidently. We naturally extracted codified token from part A. As a consequence, we abstracted each value pattern and named its properties. As a consequence of data modeling and reconstruction, district A of gene allele definition table over 17 gene files results in Table 2.3.

(a) HGVS nomenclature for CYO2C19 \*4 variant



(b) CYP2C19 allele definition table from CPIC (and PharmGKB)

	A	B	C	D	E	F	G	H	I	J	K	L
1	GENE: CYP2C19											
2	Nucleotide change to gene from <a href="http://www.pharmvar.org">http://www.pharmvar.org</a>											
3	Effect on protein (NP_000760.1)	-806C>T	1A>G	7C>T	10T>C	50T>C	55A>C	83A>T	151A>G	12401C>T	12416C>T	12455G>C
4	Position at NC_000010.11 (Homo sapiens chromosome 10, GRCh38.p2)	5' region	M1V	P3S	F4L	L17P	I19L	K28I	S51G	R73C	H78Y	G91R
5	Position at NC_008384.3 (CYP2C19 RefSeqGene, forward relative to chromosome)	g.4220C>T	g.5026A>G	g.5032C>T	g.5035T>C	g.5075T>C	g.5080A>C	g.5108A>T	g.5176A>G	g.17426C>T	g.17441C>T	g.17480G>C
6	rsID	rs12248560	rs28399504	rs367543002	rs367543003	rs55752064	rs17882687			rs145328984		rs118203756
7	CYP2C19 Allele											
8	*1		A	C	T	T	A	A	A	C	C	G
9	*2											
10	*3											
11	*4	Y	G									
12	*5											
13	*6											
14	*7											
15	*8											
16	*9											
17	*10											
18	*11											
19	*12											
20	*13											
21	*14					C						
22	*15						C					
23	*16											
24	*17	T										
25	**											

Figure 2.3. Gene allele definition table example

- (a) Variant expression in HGVS nomenclature and its meaning.
- (b) Gene allele definition table collected from CPIC guideline contents. File has for distinctive areas; A) Reference Sequence level related values; B) Detail location and variant information given A; C) Star allele nomenclature; D) actual variant information at locus A+B

**Table 2.3. Reference Sequence Information for Locus assignment**

HGNC_Gene_Symbol	Chromosome	Reference_Sequence_Source	Reference_Assembly	Complete Genomic Molecule ID	Genomic Region ID	Protein ID
CACNA1S	1	NCBI RefSeq	GRCh38.p7	NC_000001.11	NG_009816.1	NP_000060.2
CFTR	7	NCBI RefSeq	GRCh38.p2	NC_000007.14	NG_016465.3	NP_000483.3
CYP2B6	19	NCBI RefSeq	GRCh38.p2	NC_000019.10	NG_007929.1	NP_000758.1
CYP2C19	10	NCBI RefSeq	GRCh38.p2	NC_000010.11	NG_008384.3	NP_000760.1
CYP2C9	10	NCBI RefSeq	GRCh38.p2	NC_000010.11	NG_008385.1	NP_000762.2
CYP2D6	22	NCBI RefSeq	GRCh38.p2	NC_000022.11	NG_008376.3	NP_000097.3
CYP3A5	7	NCBI RefSeq	GRCh38.p2	NC_000007.14	NG_007938.1	NP_000768.1
CYP4F2	19	NCBI RefSeq	GRCh38.p2	NC_000019.10	NG_007971.2	NP_001073.3
DPYD <sup>+</sup>	1	NCBI RefSeq	GRCh38.p2	NC_000001.11	NG_008807.2	NP_000101.2
G6PD	X	NCBI RefSeq	GRCh38.p2	NC_000023.11	NG_009015.2	
IFNL3 <sup>+</sup>	19	NCBI RefSeq	GRCh38.p2	NC_000019.10	NG_042193.1	
NUDT15	13	NCBI RefSeq	GRCh38.p7	NC_000013.11	NG_047021.1	NP_060753.1
RYR1	19	NCBI RefSeq	GRCh38.p2	NC_000019.10	NG_008866.1	NP_000531.2
SLCO1B1	12	NCBI RefSeq	GRCh38.p2	NC_000012.12	NG_011745.1	NP_006437.3
TPMT	6	NCBI RefSeq	GRCh38.p2	NC_000006.12	NG_012137.2	NP_000358.1
UGT1A1	2	NCBI RefSeq	GRCh38.p2	NC_000002.12	NG_002601.2	NP_000454.1
VKORC1	16	NCBI RefSeq	GRCh38.p2	NC_000016.10	NG_011564.1	

\* HLA-A, HLA-B, CYP2C8 Allele Definition Tables are not available

<sup>+</sup> source - <https://www.pharmgkb.org/page/pgxGeneRef>

Table 2.4 shows information density and terminology variation in the value field of the gene allele definition table. Among 17 available PGx gene variant information, 11 genes adopted star allele nomenclature<sup>55</sup>, and G6PD has its own nomenclature, and WHO class to designate distinctive functions on drug reaction mechanism<sup>56</sup>, two genes have a single PGx variant. Almost of PGx variant over 17 genes are single nucleotide variant (SNV) or insertion/deletion (InDel), but CYP2B6 and CYP2D6 include 14 and 4 copy number variants respectively. The number of different loci that appear in CPIC guideline contents is 702.

**Table 2.4. Gene allele definition table data profiles**

HGNC Gene Symbol (n=20)	No of Loci	No of assigned designation	Matrix size	Example values	
CACNA1S	2	2	4	Reference	c.520C>T
CFTR	40	42	1,640	2789+5G->A	S977F
CYP2B6 <sup>*</sup>	38	38	1,444	*1	*38
CYP2C19	34	34	1,156	*1	*37
CYP2C8				not available	
CYP2C9	58	61	3,538	*1	*61
CYP2D6 <sup>+</sup>	128	146	18,560	*1	*9xN, *139
CYP3A5	8	8	64	*1	*9
CYP4F2	2	2	4	*1	*3
DPYD	15	93	1395	Reference	c.1003G>T (*11)
G6PD	173	187	32,351	202G>A_376A>G_1264C>G	Yunan <sup>++</sup>
HLA-A				not available	
HLA-B				not available	
IFNL3	single variant(g.39248147C>T)			rs12979860 reference (C)	rs12979860 variant (T)
NUDT15	17	19	323	*1	*19
RYR1	43	48	2,064	Reference	c.1021G>A
SLCO1B1	29	37	1,073	*10	*9
TPMT	39	43	1,677	*1	*9
UGT1A1	5	10	50	*1	*80+*37
VKORC1	single variant(g.3109638C>T)			rs9923231 reference (C)	rs9923231 variant (T)

\* Star allele available gene count: N=11 (CYP2B6; CYP2C19; CYP2C9; CYP2D6; CYP3A5; CYP4F2; DPYD; NUDT15; SLCO1B1; TPMT; UGT1A1)

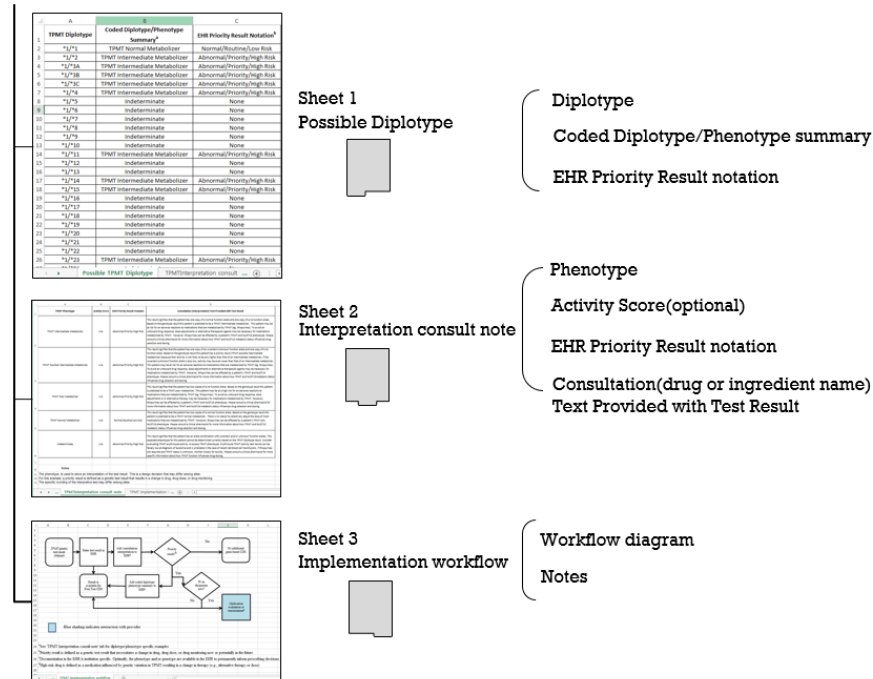
<sup>+</sup> CYP2D6 and CYP2B6 include 14 and 4 copy number variants respectively

<sup>++</sup> G6PD Genetic Variant Nomenclature and WHO Class



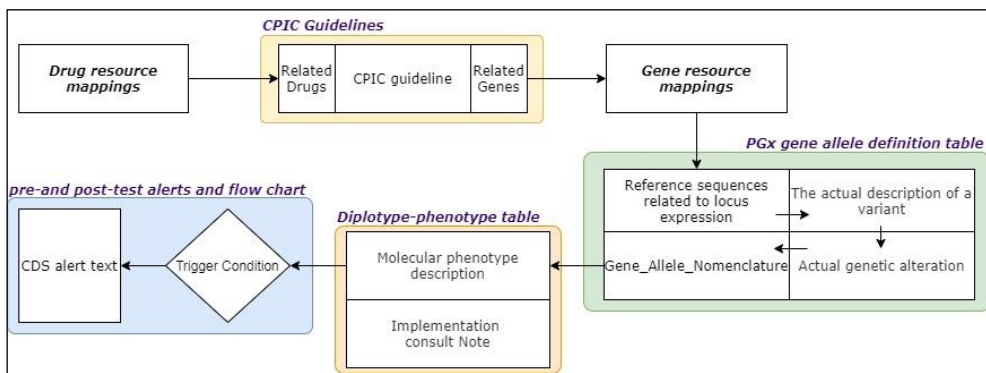
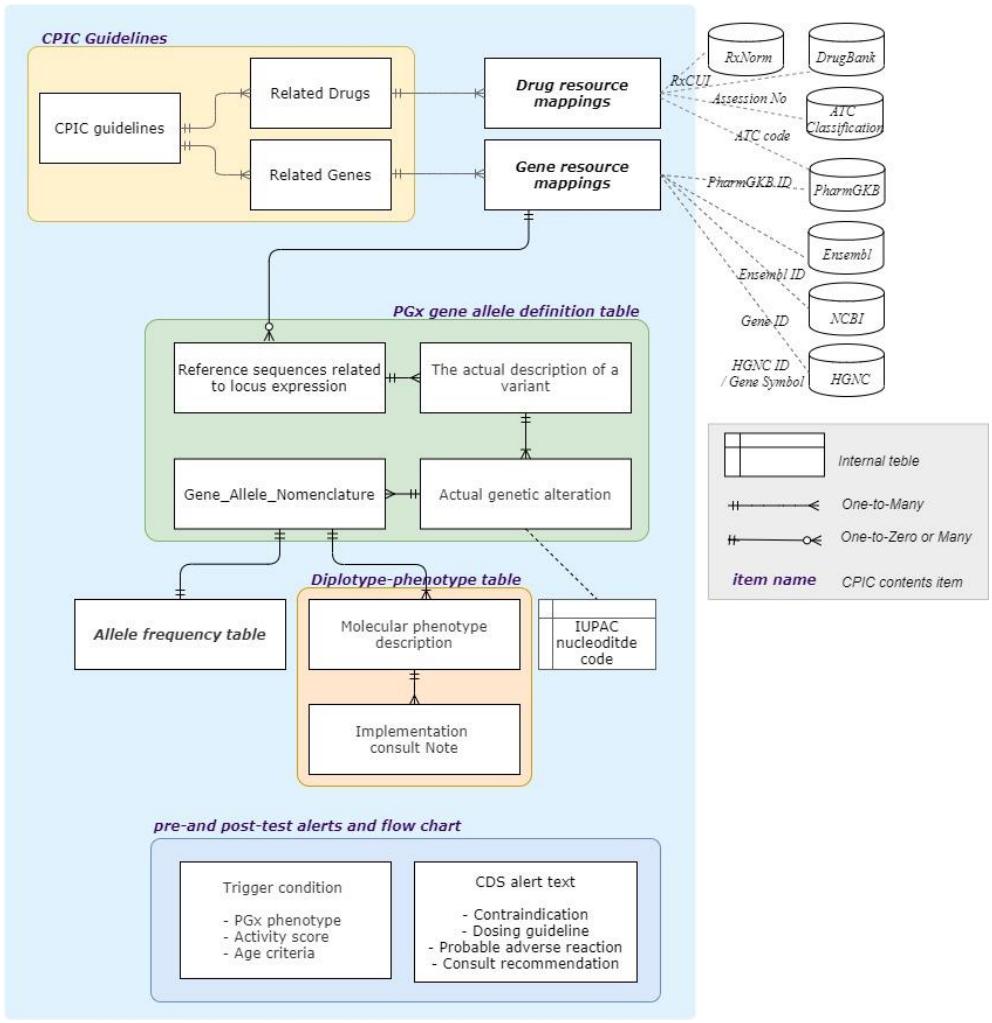
The *Diplotype-Phenotype table* consists of 3 sheets, each of which is a ‘possible Diplotype,’ ‘Interpretation consult note,’ and ‘Implementation workflow.’

Item: Diplotype-Phenotype table



**Figure 2.4. Diplotype-Phenotype table example and its meta-data structure**

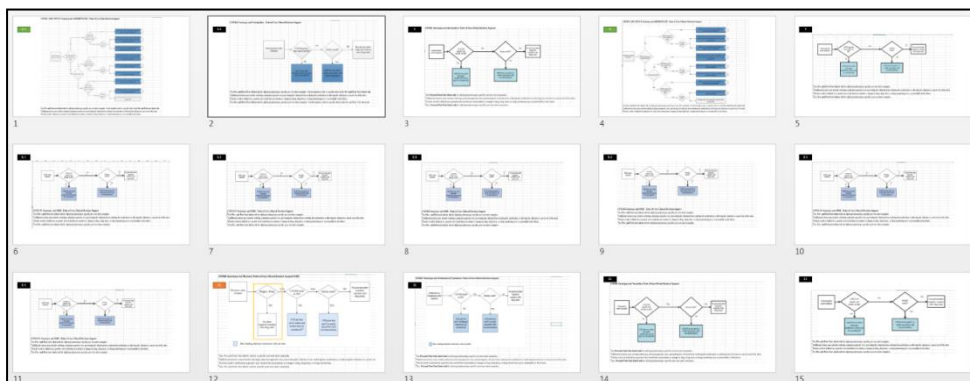
Data model construction was conducted within these multilayer data analysis results. Each rough data structure embedded in original contents has been reclassified into atomic level attributes, a group of entities, and data flow according to the CDS service scheme of this study. Data flow is declared as relations in the constructed data model. Normalization and abstraction were applied until the ambiguity of overlapping properties, and abnormal cardinality disappeared for the design of the entity-relationship model for the CDS service. Computable conversion of the CPIC knowledgebase and linking scheme in PGx CDS to patient genomic data based on knowledge representation is shown in Figure 2.5.



**Figure 2.5. Snapshot of CPIC guidelines content structure converted to be computable**

### 2.4.3. CDS Rule Extraction

The pre-and post-test alert file consists of two sheets; 'Pre- and post-test alerts,' 'Flow Chart.' Flow chart helps end-user's understanding also easily convert to a conditional phrase in computer language. However, the trigger condition, a particular exact subset, is offered by the 'Pre- and post-test alerts' sheet. In other words, conditional trigger information for CDS function is distributed in two sheets. Firstly, 'Flow Chart' has one common condition whether the patient's genomic information is available or not. There are two exceptions over three guidelines; one is filtering weight over 40 kg criteria in case of 'CYP2B6 and efavirenz', the other has branched alert message between for pediatrics and adults in case of 'CYP2D6 and Atomoxetine' and 'CYP2C19 and Voriconazole'. The latter type of exception does not appear in 'Flow chart' but implied to provide two alert text message columns in 'Pre- and post-test alerts.' Through this separation and regrouping process, we constructed trigger condition, alert message, and trigger condition-alert message relation.

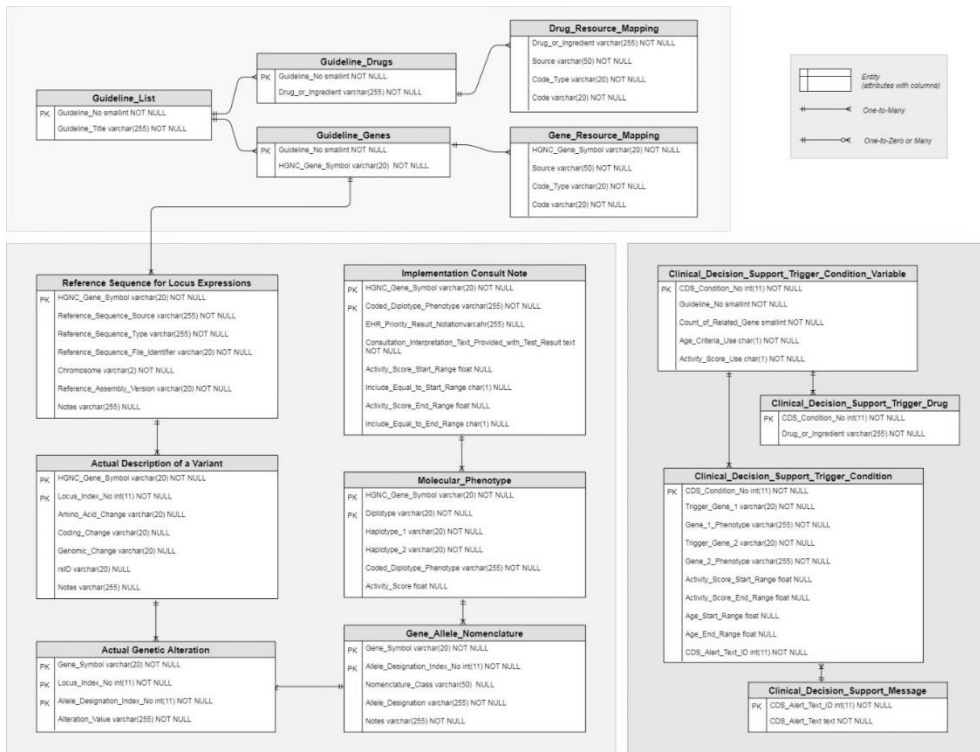


**Figure 2.6. Collection of 'Flow chart' over available 15 guidelines**

#### **2.4.4. Structured database construction**

Finally, we have constructed a machine-readable CPIC guideline database in the form of a relational database. The database includes 15 tables and 46 unique attributes (Figure 2.7). Interestingly, the left and right parts of the ERD are separated.

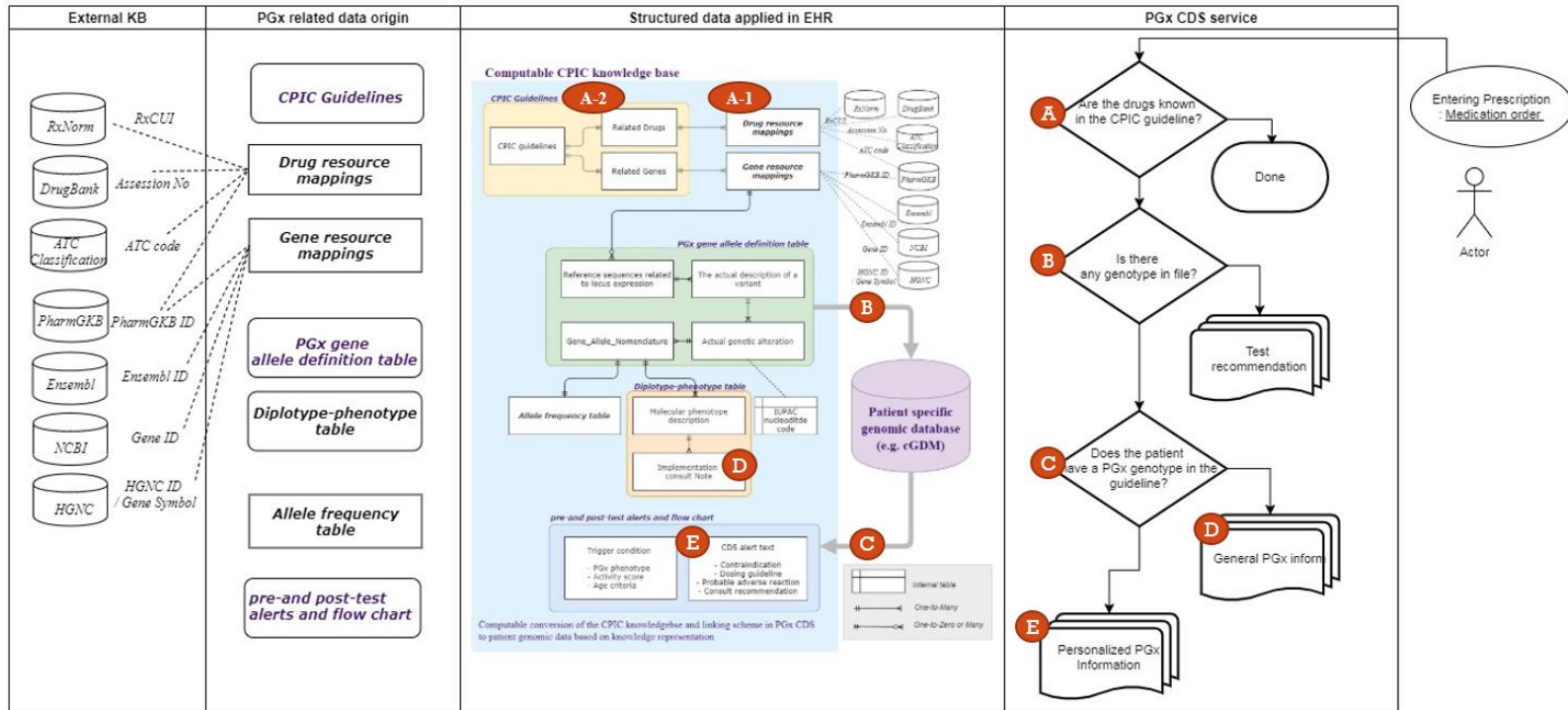
The left side represents the knowledge that declares PGx related variant definition and converts those findings into interpretable codified phenotypes for each drug-gene pair for which the guideline is targeted. The right part is a guide that provides a tailored CDS message when an individual's codified phenotype and prescribing drug ingredient is known. The CDS message contents could break down a set of properties comprised of contraindication, dose adjustment guidelines, probable adverse reactions, and consult recommendations to the clinical pharmacist for further consideration. However, in this study, the CDS alert text was not structured because the distribution of the corresponding attributes when segmented by sentence was irregular.



**Figure 2.7. Entity-relationship diagram of reconstructed database based on CPIC contents**

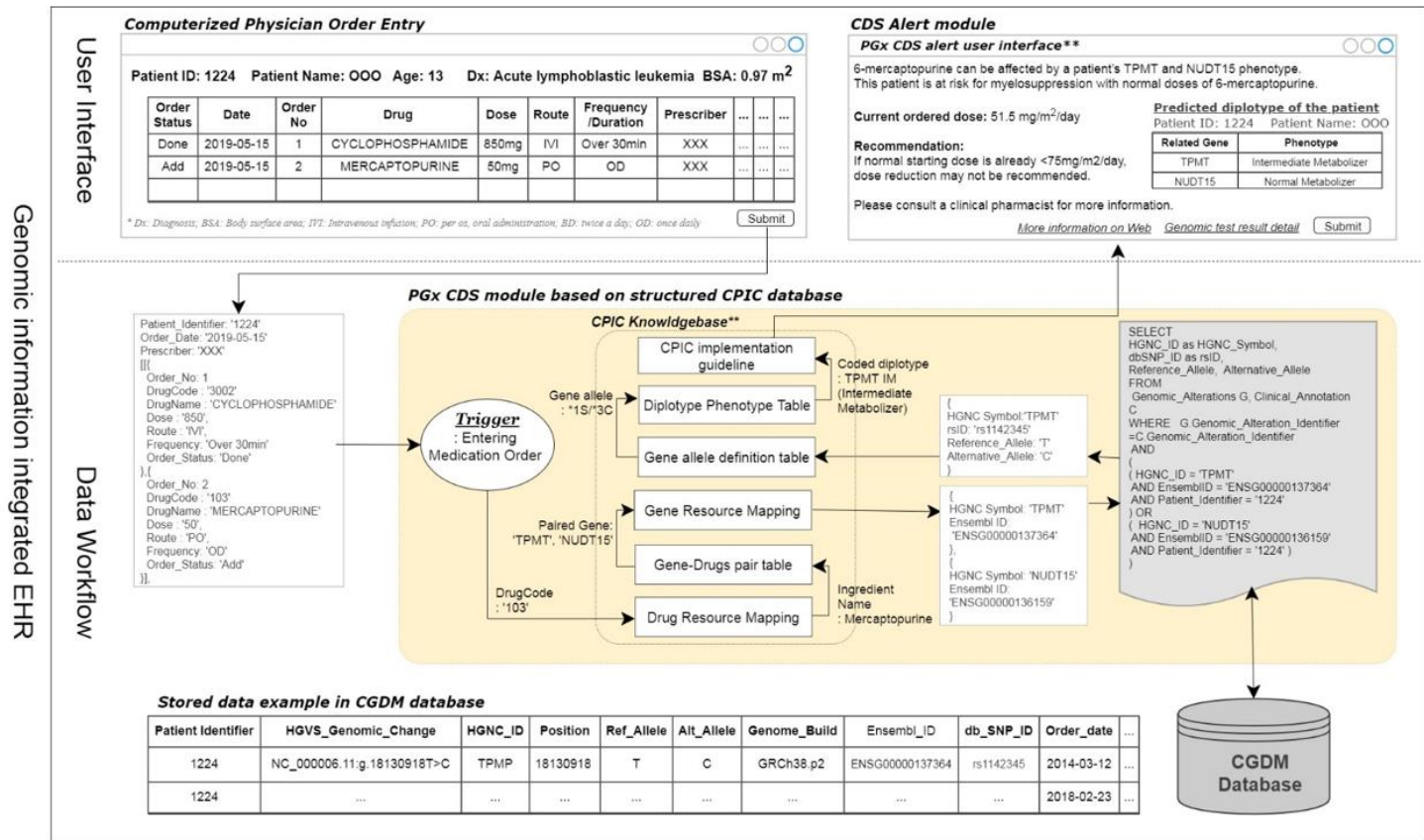
## 2.4.5. PGx CDS service module

Figure 2.8 shows the developed PGx CDS service module. When the system evoked, the CDS module looks at patient genome data stored in the EHR server and returns potential phenotype according to the CPIC PGx variant definition. Also, the module query out individualized recommendations for the prescriber. The novelty of this modular solution is the machine-readable conversion of the CPIC guideline and seamless function execution in a single EHR system. Data modeling reveals four components of the CPIC knowledge resource. The first is targeted phenomena identifier, CPIC guideline title, and drug-gene pair information. The effort to provide curated and filtered PGx variant definition list with expert knowledge with clinical relevance. Then, they try to capture related annotation systems for interpretation, such as the star allele system. This information is presented in the nomenclature field in the Gene-allele definition table and codified data field in the Diplotype-phenotype table. Final CDS alert texts are given with the assumption that a person who looks at guidelines knows the specific genotype information. Data flow crack is found in here, but we could bridge this gap with the patient-specific genome database proposed in Chapter 1. Finally, seamless PGx CDS are enabled shown in Fig. 2.9. Through the data collection and reconstruction process, we could briefly explore the colossal landscape of their accomplishment. For enhancing usability, CPIC does process standardization along with the development of new guidelines.



**Figure 2.8. PGx CDS module architecture**

(A) In this step, service refers to the data in (A-1) and (A-2) to check if the prescribed drug has relevance to the pharmacogenomic guideline. (B) Execute a query into a patient-specific genome database by referring to pharmacogenomic variant information declared in the CPIC knowledge base. (C) The search result includes the possession of genomic information of the patient is returns in the form of a phenotype. (D) Provide general guidance on the drug-dielectric guidelines. (E) Provide individualized PGx CDS alert message



\*\* Reference : Relling, M. V., Schwab, M., Whirl-Carrillo, M., Suarez-Kurtz, G., Pui, C. H., Stein, C. M., ... & Caudle, K. E. (2019). Clinical Pharmacogenetics Implementation Consortium Guideline for Thiopurine Dosing Based on TPMT and NUDT 15 Genotypes : 2018 Update *Clinical Pharmacology & Therapeutics*,105(5), 1095-1105.

Figure 2.9. PGx CDS module integration scenario with dataflow



# Chapter 3. Clinical Application of Clinical Genome Data Model: Integrating Star Allele and HLA Data Models

“An ideal nomenclature would be one that is entirely unambiguous. One might hope that a geneticist of the year 2493 could pick up a 1993 copy of *The American Journal of Human Genetics* and quickly understand, from the designation of a mutation and without extensive study of other sources, the location of a nucleotide change. However, the complexity of the genome and its functions is such that a perfect nomenclature is unachievable.<sup>57</sup>” (Ernest Beutler, 1993)

## 3.1. Introduction

As Beutler envisioned, the perceived complexity of the genomics is expanding, and a perfect nomenclature is not achieved yet. However, there is some accomplishment, such as the HGVS nomenclature and star allele system that helps effective communication between scientists. The HGVS<sup>58</sup> nomenclature has advantages in figure out a specific locus from the nomenclature. Nevertheless, it does not specify a specific reference sequence. Thus the same variant could be described using different reference sequences, which might cause confusion. Furthermore, the expression is not scalable enough to express functional combinations. Thus star allele nomenclature was introduced in 2006<sup>55</sup>. The star allele nomenclature could contain multiple-locus in one name (so-called star), and one locus could be placed in redundant stars. The star-allele nomenclature is the result of efforts to standardize genetic polymorphism annotation for the cytochrome P450 genes. As clinical pharmacogenetic testing becomes widespread, this system

has played a vital role in effectively delivering the patient's genotype and predicted clinical phenotype. As genomics research expands, the system remains a valuable tool for the broader community of genetic researchers to exploit our ever-improving ability to catalog variability in the human genome<sup>55</sup>. However, as scientific discoveries accumulate, the number of assigned stars is increased, and the complexity of the naming system itself is also expanded. For example, \*1 is mostly accepted as a reference sequence functionality, but a few exceptions occur as known population distribution of the variants are changed. In addition, there are highly curated representative registries according to research interest so we could use those naming system as an auxiliary identifier. We prove the concept in Chapter 2 using PGx variant definition construction and interoperable interpretation in the data of the patient-specific genomic information in cGDM.

Furthermore, there are independent nomenclatures such as the human leukocyte antigen (HLA) system. The HLA system<sup>59</sup> is known to be the most polymorphic in human. The HLA polymorphism is not evenly spread throughout the molecule but is clustered in the antigen-binding groove<sup>60</sup>. HLA is a protein that plays a vital role in our body's immune function with a wide variety of allele types.<sup>61</sup> HLA diversity is particularly important in organ transplantation because transplant recipients and donors with different serological HLA proteins will exhibit organ transplant rejection<sup>62</sup>. Therefore, transplant recipients must perform HLA screening

before transplantation. Recently, HLA diversity has been reported to cause severe drug hypersensitivity as well as organ transplantation<sup>63</sup>. However, the HLA results of transplant patients and donors have not been used to predict future adverse drug reactions. This is because the HLA test is performed in various ways, from a simple serological test to an NGS test. Besides, while the nomenclature that represents the HLA test results is continuously updated, the test results simply have been stored in free text in the electronic medical record (EMR) <sup>64</sup>.

## **3.2. Purpose of Research**

Firstly, the HLA database is designed to be used in clinical practice with data-driven approach. Construction of HLA DB linked in hospital information system could bring clinical pharmacogenomics information to physicians. Secondly, the HLA database is covering multiple test methods enable to protect from the harm due to the non-use of health-related data<sup>65</sup>. Ultimately, we try to validate the model consistency to cope with the evolving annotation systems by construction of HLA database.

### **3.3. Material and Methods**

We used the dataset extracted the results of the HLA test performed and demographics of patients using SUPREME® between February 2002 and June 2018, a clinical data warehouse of Seoul National University Hospital<sup>66</sup>. With a data-driven approach, we could extract clinical context enriched entities and attributes. Also, HLA nomenclature has been adopted as the primary material for designing and elaborating the HLA entity.

We designed the cGDM HLA as a physical data model in a relational database on MySQL 5.6 in an agile manner. Data-driven modeling is comprised of data mining and clarification of implicit properties and relations<sup>67</sup>.

## **3.4. Results**

### **3.4.1. Summary of collected dataset**

Collected dataset from SUPREME<sup>®</sup> has 11,287 records for 11,144 patients; 4,039 male and 7,105 female patients, including 2,642 high-resolution tests, 5,835 low-resolution tests, and 2,810 tests. Gathered data fields are shown in Table 3.1 below. We filtered these fields with data existence, and remove its redundancy. Then, the reclassification of each field was conducted compared to the cGDM schema. Unlike the expectation that it will be a true subset of the existing cGDM schema, except for the HLA nomenclature, unique properties remain that called 'related patient.' This is caused by a unique clinical context when the HLA test ordered, organ transplantation. In this case, donor-recipient tag information or family relationship information has significant meaning for test result application. For internal integrity, we decide to capture this information with the appended entity for further use.

**Table 3.1. Extracted field list gathered from the EHR records**

<b>Document item name</b>	<b>full name or example data</b>
MRN	patient identification no
PatientDOB	Birthdate
PatientName	patient name
PatientSex	patient sex
TestCode	test code
TestDate	test date
TestName	test name
Name	name (data not found)
PatientType	donor/recipient
diagnosis	dx (data not found)
RelatedPatientsNo	relatives(data not found)
A1_gene	A11
A1_allele	*11
A2_gene	A24
A2_allele	*24
B1_gene	B7
B1_allele	*07
B2_gene	B62
B2_allele	*15
C1_gene	Not tested
C1_allele	Not tested
C2_gene	Not tested
C2_allele	Not tested
DR1_gene	DR1
DR1_allele	*01:01g
DR2_gene	DR4
DR2_allele	*04:03g
DQ1_gene	Not tested
DQ1_allele	Not tested
DQ2_gene	Not tested
DQ2_allele	Not tested
RelatedPatientName	NA

### 3.4.2. HLA data model

HLA entity is added in forms of tokenized HLA nomenclature. HLA gene classes and its subtypes are represented in Supplementary information 2. Because this nomenclature is logically well developed, one of the major challenges was in its version control. Opportunely, the HLA community provides a version conversion tool and table as a text file. We parsed the HLA test results from the dataset with nomenclature logic and normalized its values with mass conversion when we uploaded the dataset to the DBMS table.

Common entities with the cGDM

Extension entity represent HLA nomenclature

Subject		
1 Subject Identification Number	Subject_ID	GDM generate
2 Subject name	Subject_Name	Name
3 Patient Identification Number	Patient_ID	MRN
4 Birth date	Birth_Date	PatientDOB
5 Gender	Gender	PatientSex
6 Race	Race	
7 Ethnicity	Ethnicity	
8 Institution code	Institution_Code	GDM generate
9 Register Identification Number	Register_ID	GDM generate
10 Submission date	Submission_Date	GDM generate

Specimen		
1 Specimen Identification Number	Specimen_ID	GDM generate
2 Subject Identification Number	Subject_ID	GDM generate
3 Specimen origin type	Origin_Type	
4 Body site	Body_Site	
5 Body site code	Body_Site_Code	
6 Physical type	Physical_Type	
7 Physical type code	Physical_Type_Code	
8 Specimen type	Specimen_Type	
9 Specimen block Identification Number	Block_ID	SampleNo1
10 Specimen accession Identification Number	Accession_ID	
11 Collection date	Collection_Date	DateOfTest1
12 Received date	Received_Date	
13 Differentiation state	Differ_State	

Protocol		
1 Protocol Identification Number	Protocol_ID	GDM generate
2 Specimen Identification Number	Specimen_ID	GDM generate
3 Test name	Test_Name	TestName
4 Type of sequencing	Sequencing_Type	GDM generate
5 Ordered date	Order_Date	
6 Order Identification Number	Order_ID	
7 Lab name	Lab_Name	
8 Reagent	Reagent	
9 Received Date	Receive_Date	
10 Bioinformatician	Bioinformatician	
11 Analytics institution	Analytics_Institution	
12 Sequencer Identification Number	Sequencer_ID	
13 Panel Identification Number	Panel_ID	
14 Pipeline Identification Number	Pipeline_ID	
15 SNV/InDel pipeline detail iden	SNV_InDel_PD_ID	
16 Copy Number Variation pipeline	CNV_PD_ID	
17 Translocation pipeline detail	Translocation_PD_ID	
18 Microsatellite Instability Alterat	MSI_PD_ID	
19 Tumor mutation burden pipeline	TMB_PD_ID	
20 Document creation date	Docu_Creation_Date	
21 Document version	Docu_Version	

HLA Star Allele		
1 Star Allele Identification Number	Star_Allele_ID	GDM generate
2 Protocol Identification Number	Protocol_ID	
3 A1_gene	A1_gene	A11
4 A1_allele	A1_allele	*11
5 A2_gene	A2_gene	A24
6 A2_allele	A2_allele	*24
7 B1_gene	B1_gene	B7
8 B1_allele	B1_allele	*07
9 B2_gene	B2_gene	B62
10 B2_allele	B2_allele	*15
11 C1_gene	C1_gene	Not tested
12 C1_allele	C1_allele	Not tested
13 C2_gene	C2_gene	Not tested
14 C2_allele	C2_allele	Not tested
15 DR1_gene	DR1_gene	DR1
16 DR1_allele	DR1_allele	*01:01g
17 DR2_gene	DR2_gene	DR4
18 DR2_allele	DR2_allele	*04:03g
19 DQ1_gene	DQ1_gene	Not tested
20 DQ1_allele	DQ1_allele	Not tested
21 DQ2_gene	DQ2_gene	Not tested
22 DQ2_allele	DQ2_allele	Not tested

Figure 3.1. HLA Database design merged in the cGDM schema



## General Discussion<sup>†</sup>

The rapid accumulation of genome information has led to a paradigm shift in medicine. Nevertheless, significant barriers remain to overcome inflection points. Through multi-disciplinary analysis and consideration of this phenomenon, we determined two main causes: 1) reliability-related result variance among numerous pipelines and processes, and 2) the unique data structure of genome information. Since these two causes have mutual influences, an integrative solution may be more effective than a point solution. Moreover, we foresee that GIS will become an essential component of an integrated clinical information system in the precision medicine era. In this context, this cGDM could serve as a genomic information representation scheme enabling the intellectual interaction between medical experts and informed decision making, ultimately contributing to the enhancement of personal genomic data utilization at the point of care.

---

<sup>†</sup> The part of the dissertation general discussion published in following paper: Kim, H. J., Kim, H. J., Park, Y., Lee, W. S., Lim, Y., & Kim, J. H. (2020). clinical Genome Data Model (cGDM) provides interactive clinical Decision Support for precision Medicine. *Scientific reports*, 10(1), 1-13.

## **The GDM as an Infrastructure for a GIS**

We recommend the GDM as a genomic information representation scheme for clinical purposes. To ensure the convenient and appropriate clinical use of genomic data, medical informatics technology is needed as part of the infrastructure supporting the integration of clinic and genomic layers of information<sup>68,69</sup>. Given the multi-level and multi-dimensional nature of health, clinicians must perform decision-making for a given case based on a collection of segmented data representing a person's health, including laboratory data, imaging, and observation data assessed by experts. Currently, a clinical information system is typically used as a core tool for supporting this knowledge in a management process. To broaden perspectives in the era of precision medicine, we propose a concept of genome information system (GIS) as an integral component of an expected clinical information system for precision medicine (Fig. 1.1).

The cGDM can serve as a data-level infrastructure for implementation of the GIS. When decision makers face unfamiliar health-status measurements, determining clinical significance and meaning is challenging<sup>69,70</sup>. The cGDM was designed to preserve genomic information at an appropriate information scale and granularity covering the procedural dimension, which is related to the confidence level as a clinical measurement for clinical application. The design of the cGDM allows processed genomic data for a general purpose to be stored and merged with

existing clinical data, providing outputs in an interoperable data format. Likewise, sequencing analysis, data processing, and presentation of processed information can be managed in a form that can be explicitly confirmed. Once data are uploaded to the cGDM-based database, they serve as a supportive backbone for any downstream functional applications such as report generation or a clinical decision support system. (e.g., Fig 8; Fig 3)

To develop a system for the systematic management of genomic data, it is necessary to unify its data structure with that of other existing components of clinical information systems, ensuring sufficient reliability for identifying the original data generation process<sup>71</sup>

## **Current Approach to Genomic Data Management**

The Health Level 7 (HL7) clinical genomics working group provided a model for health information exchange and Fast Health Interoperability Resources (FHIR) genomics, a model that integrates genetic and clinical information via the HL7 interfacing standard<sup>70,72</sup>. FHIR provides standards for medical and genomic information exchange and offers open-source and open application programming interfaces (APIs) that can easily be applied in clinical fields among heterogeneous data sources. FHIR and FHIR genomics have made substantial contributions toward the implementation of medical information exchange and are drawing electronic health records vendors' attention in this respect.

The Global Alliance for Genomics & Health (GA4GH) was established in 2013 to develop public tools that enable the responsible, voluntary, and secure sharing of clinical and genomic data<sup>73</sup>. The federated approach of GA4GH does not involve the storage and management of data in centralized data repositories. Instead, it provides an API that enables users to request and share data while holding data for institutions<sup>74</sup>.

The FHIR and GA4GH consortium of HL7 were developed with the intention to facilitate the exchange of genomic and clinical data among multiple sites. Both resources have a common character as a form of information exchange at the communication level. These systems use the latest web technologies such as the representational state transfer (REST)

API to make it easier for developers to implement clinical applications or information systems in the healthcare industry.

The International Organization for Standardization (ISO) Technical Committee 215 (Medical Information) has proposed genomic information standards. ISO 27720:2009 (GSVML; General Sequence Variation Markup Language) is a standard that defines how genetic sequencing variation information is exchanged based on XML. The scope of this standard is in the data exchange format and does not include the database schema. Although all genetic sequencing is within the standard's scope, the SNP is the main target of this standard. Another standard for more specific clinical utilization of genomic information is ISO/TS 20428 Health information - Data elements and their metadata for describing structure information in electronic health records established in 2017. Additionally, ISO/CD TS 23357 Genomic informatics – clinical genomics data sharing specification for next generation sequencing is under development state.

**Table 4.1 Comparison table of characteristics of related resources**

Resource	Publication (year)	Data management scope			Computability			Purpose	Organization
		Storage	Exchange	Clinical data linkage	Patient identification	for CDS rule	for report generation		
cGDM	2020	O	X	O	O	O	O	Data level EHR integration	SNUBI
OMOP G-CDM	2019	O	X	O	X	X	X	Federated Research Network	OHDSI
FHIR Genomics	2020 (2015~)	X	O	O	O	O *	O	Information Exchange	HL7
GA4GH Genomics API	in progress (2015~)	X	O	X	X	X	X	Data interchange for bioinformatics research	GA4GH
ISO/TS 20428:2017	2017	X	O	O	O	X	O	Structuring sequencing report	ISO/TC215 (Health Informatics)
ISO/TS 25720:2009	2009	X	O	X	X	X	O	SNP data exchange	ISO/TC215 (Health Informatics)
GDC	2017	X	O	X	X	X	X	Cancer related genomic data sharing	NIH NCI

\*via SMART on FHIR, CDS Hooks, HL7 Inforbutton

cGDM: clinical Genome Data Model; OMOP G-CDM: Observational Medical Outcomes Partnership Genome Common Data Model; FHIR: Fast Healthcare Interoperability Resources; GA4GH: Global Alliance for Genomics and Health; ISO/TS 20428:2017: Health informatics - Data elements and their metadata for describing structured clinical genomic sequence information in electronic health records; ISO/TS 25720:2009: Health informatics - Genomic Sequence Variation Markup Language(GSVML); API: Application Programming Interface; GDC: Genomic Data Common; SNP: Single Nucleotide Polymorphism; SNUBI: Seoul National University Biomedical Informatics; OHDSI: Observational Health Data Science and Informatics; HL7: Health Level Seven; NIH: National Institutes of Health; NCI: National Cancer Institute

Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM) aims to conduct distributed research across observational databases in multiple institutions using a common data model approach. Genomic Common Data Model (G-CDM) proposed as an extension part of OMOP-CDM represents genomic information<sup>75</sup>. Focused on research purposes, the granularity and scale of knowledge representation have limited for multifaceted clinical application.

The almost resources discussed earlier focus on data exchange formats for utilization rather than on EHR integration of genomic information. Therefore, the system is being developed by designing functions first rather than expressing knowledge of the genomic information itself, and by further defining the element whenever the function is added. This development methodology has strength for easy and fast software function development. On the one hand, however, all of reviewed resources are on a separate layer from the ground level schema in data management.

## **The cGDM: A Step beyond the Capabilities of the Existing Systems**

To develop a system for the systematic management of genomic data, it is necessary to unify the data structure with that of other components of clinical information systems, and to ensure sufficient reliability for identifying the data generation process<sup>18</sup>. Conventional systems have focussed on data structure unification issues first, to harmonise heterogeneous systems among separate institutions<sup>76</sup>. By contrast, our model was designed to achieve both clinico-genomic knowledge representation accompanied by traceability of the genomic data, to enable determination the clinical significance of a genomic test result provided to a clinician.

Through the developed cGDM, standardization and integration of the structure of genomic data can be realized, along with tracing of the information in a step-by-step manner until the data related to the target are extracted according to clinical or research requirements. To secure the clarity of genomic information, we defined the basis for each attribute and focused on designing an entity set that can accurately represent the genomic data to be delivered to the target user, without information distortion, through composition of the basis.

To allow better assessment of the meaningfulness of genomic information, we defined the basis for each attribute and focused on



designing an entity set that accurately represents the genomic data that are delivered to the target user, without information distortion. Furthermore, the cGDM is adaptable as a data-level extension to any existing information system, regardless of database system or application platform. Effectiveness and feasibility of genomic data management in the computational environment in terms of the data-level EHR integration approach by the cGDM were also broadly evaluated in Chapter 2 and 3.

## **Unrecognized Ambiguity in the Interdisciplinary Knowledge Interplay**

Accumulation of basic, translational, and regulatory science is a prerequisite to implementing personalized medicine in routine care<sup>22</sup>. As a basic science, bioinformatics has witnessed explosive and rapid progress since the completion of the Human Genome Project. In the context of regulatory science, there are currently several ongoing efforts within the bioinformatics and molecular biology domains,<sup>10,11,77</sup> with great maturation in the body of knowledge during the last decade, including principles and recommendations related to NGS technology. These efforts have focussed primarily on the standardization of bioinformatics protocols and the file structures for intra- or interlaboratory communication.

Translational science represents the next challenge for the realization of actual health promotion with personalized medicine<sup>78</sup>. In the context of clinico-genomics, translational approaches ultimately target the syntactic and semantic interoperability between genomics and clinical practice, to ensure business continuity in terms of knowledge management<sup>23,24,79</sup>. Previous approaches have stressed a need for structural transformation to overcome the currently low adaptation of genomic information for clinical decision-making. However, the other major cause, the knowledge gap, has yet to be seriously considered because the solution appears obvious: the education of medical experts in bioinformatics

principles.

Nevertheless, this raises the question of the specific level of bioinformatics knowledge required in clinical practice. Our working group agreed that clinicians do not need to be bioinformatics experts to implement precision medicine. Preferably, the key is education on how to understand genomic data and confidence levels, and then be provided with sufficient information to make clinical decisions. Based on this perspective, we identified a previously unrecognised ambiguity related to the knowledge interplay between bioinformatics and medical practices (Fig. 3). Although the genome is the most concrete type of observational data representing an individual's inheritance, the genomic information delivered to clinicians is rarely transformed to a human-readable form and is also rarely a direct representation of the genomic sequence. Instead, this information is more of an intellectual product, processed in a purpose-weighted result file structure. Thus, the question of reliability of the genomic information must be addressed before it is adopted by the physician, similar to other types of conventional observational data.

Considering the knowledge gap in this clinico-genomic context, unrecognised ambiguities may occur on each side. For example, when linking the outputs of bioinformatics to clinical fields, the indicator of information quality moves from internal consistency within the same protocol to external consistency between different protocols. Thus, to

accomplish the final goal of precision medicine, more discussion is needed about how data will cross this intermediate space, then about how to best represent and deliver crossover information.

## **Adoption of FMEA to Information Processing**

To best of our knowledge, the methodology proposed herein has not yet been applied in the field of genetic information processing. FMEA is the most commonly used methodology for determining reliability of manufacturing and design processes<sup>17,20,21,80,81</sup>. We perceive the result of genetic testing not as an output of static measurement, but rather as an output of an intellectual production process. When conducting bioinformatics analyses, there is no requirement for unification among the processes, since the internal consistency within each process guarantees scientific rigour. Moreover, the flexible data specifications used in the bioinformatics field have the advantage of supporting various research applications<sup>7</sup>, but that advantage becomes an obstacle to data integration for comprehensive clinical decision making. In addition, relevant external knowledge, tools, platforms, and analytical techniques cannot be unified because they are still under development. Considering this large interdisciplinary hyperspace, our approach aims to improve the quality of information delivery while responding to an enormous, growing body of knowledge that has yet to be integrated within its own basic-science field. Therefore, the FMEA was adopted to derive and clarify a set of metadata designed to prevent information from being distorted.

To facilitate the use of genomic test results in clinical practice, it is

essential to integrate genomic data into clinical decision support systems regarding data volume and knowledge management<sup>6,34,37,82</sup>. Data modeling is the first and most crucial step in the multi-tiered design of information systems. The final product reliability, for example specific clinical decision support algorithms or integrated information systems, is hardly improved over the designed reliability on the lower level of architecture (data-level)<sup>20</sup>. This viewpoint was projected to the study design. An important consideration is that the analytic scheme presented here can help to enhance clinico-genomic understanding for experts on both the medical and bioinformatics sides of the workflow. (see Methods Section) Throughout the development of this method, we focussed on equally weighting the clinical perspective and bioinformatics process analysis in the context of business continuity, starting from our initial clinical intention through bioinformatics information processing by a knowledge-based protocol, finally offering a deliverable and interpretable form to the point-of-care clinician.

## Limitations

Multi-omics data have a fundamental limitation of unification, which is derived from the difference of knowledge expression forms related to the processing methodology, final processed data depending on the target layer, and its biological characteristics. In addition, prior to NGS, there were already several structured models according to differences in data scale and technical maturity. The entity and attribute set defined in the GDM is derived from analysis of the workflow of NGS. Therefore, we do not consider the elements of other technology-based workflows in multi-omics layers.

The methods, equipment, data processing and analytical techniques for extracting data from targets in nature will continue to evolve and accumulate. The cGDM was designed to be flexible and able to readily adapt to technological changes. However, an eventual failure in responding to these changes cannot be excluded and represents a potential limitation of this study.

Several standard models have been generated, based on differences in data scale and technical maturity, prior to the development of NGS technology. Thus, we have not considered multi-omics data. Focussing on NGS technology-based workflow helped us to determine an optimized information scale and granularity for the clinical level, and to design a model to generalise and process genomic data based on individual patients.

The cGDM could be extended to be a part of technology-wide data model integration for multi-omics data management.

The data model proposed in this study aims to clarify blind points within the interdisciplinary genomic-clinical interface, connecting separated expertise within a single platform to provide a broad perspective that covers the information reliability required for clinical evidence. In particular, we have made a novel attempt to adopt the FMEA method for a systematic meta-level data design process. Future work will focus on the development of functional systems to conduct real-world validation, including a data-upload pipeline from processed genome data files, as well as a clinical decision support tools based on the cGDM.



# Supplementary Information

## Supplementary Figure S1. PGx CDS mock-up application based on the cGDM architecture

**Stored data example in the cGDM database**

Patient Identifier	HGVS_Genomic_Change	HGNC_Gene_Symbol	Position	Ref_Allele	Alt_Allele	Genotype	Biomarker_Resource	dbSNP_ID	Order_date
1224	NC_000010.10:g.96702047C>T	CYP2C9	96702047	C	T	1 0	IWPC	rs1799853	2018-08-23
1224	NC_000010.10:g.96741053A>C	CYP2C9	96741053	A	C	0 1	IWPC	rs1057910	2018-08-23
1224	...	...	...	...	...	...	...	...	...

← → ↻

1000 genome phase 3 CEU

Pt Name (Pt No)  
Gender-Age-Race-Height(cm)-Weight(kg)

Mario Speedwagon (PA06984)  
M-21-Unknown-187-79

Petey Cruiser (PA06985)  
F-22-Unknown-189-61

Anna Sthesia (PA06986)  
M-23-Unknown-161-97

Paul Molve (PA06989)  
F-24-Unknown-156-45

Anna Mull (PA06994)  
M-25-Unknown-194-65

Gail Forcewind (PA07000)  
F-26-Unknown-168-81

Paige Turner (PA07037)  
F-27-Unknown-185-63

Bob Frapples (PA07048)  
M-28-Unknown-172-95

iaahav@at.ac.nubl.org:8080/admin/map/menu.jsp#

Patient No: PA06989    Name: Paul Molve    F / 24  
Dx: Deep vein thrombosis    156cm / 45kg   

Date	Order No	Drug	Dose	Dose Unit	Route	Frequency/Duration	Prescriber
2020-07-10	1	warfarin	6	mg	PO	OD	Dr. Kim

PGx CDS message

Warfarin dosing can be modified with a patient's VKORC1 and CYP2C9 phenotype. IWPC warfarin pharmacogenetics dosing estimation applied to the patient's case display below.

Current ordered dose: 6 mg/day (= 42.0 mg/week)

Recommendation: 3.7 mg/day = 26.0 mg/week

Related Gene	PGx Genotype
VKORC1	A/G
CYP2C9	*2/*2

Please consult a clinical pharmacist for further considerations.

[More evidence information on Web](#)

### Supplementary Table S1. Table Specification of the cGDM

The logical entities and attributes expressed in Figure 1.5 were converted into physical entities and attributes. Here, we provided our physical data model as the following table. The required data type, description, and example value for each attribute defined are described. All of the logical entities and attributes in Figure 1.5 have been transformed and defined in the physical model presented here. So, by applying this sort of conversion to physical model, each researchers can construct a genomic database according to the environment of the existing information system.

#### CLINICAL IDENTIFIER Table specification

#	Logical Name	Physical Name	PK	Required	Data Type	Description	Example
1	Subject Identifier	Subject_Identifier	PK	Yes	int(11)	Arbitrary person identifier defined in the CGDM database	1
2	Patient Number	Patient_Number		Yes	varchar(20)	Patient number of existing HIS database used to link with the CGDM database	12345678
3	Medical Institution Identifier	Institution_Identifier		Yes	varchar(20)	An abbreviation of the hospital name where the patient data linked with the CGDM database	SNUH
4	Order Identifier	Order_Identifier		Yes	varchar(20)	Unique key value represents an order of existing HIS database used to link with the CGDM database	602489471
5	Clinician Identifier	Clinicain_Identifier		Yes	varchar(20)	Unique key value represents a physician of existing HIS database used to link with the CGDM database	A2068494
6	Submission Date	Submission_Date		Yes	datetime	Date of the beginning of the data production period (e.g. ordered date)	2018-08-17 13:44

**EXPERIMENT RELATED INFORMATION Table specification**

#	Logical Name	Physical Name	PK	Required	Data Type	Description	Example
1	Experiment Identifier	Experiment_Identifier	PK	Yes	int(11)	Arbitrary identifier of the experiment defined in the CGDM database	11
2	Subject Identifier	Subject_Identifier	FK	Yes	int(11)	Arbitrary person identifier defined in the CGDM database	1
3	Test Description	Test_Description		No	text	Detailed description for ordered test	
4	Type of sequencing	Sequencing_Type		Yes	varchar(50)	Library strategy for genome sequencing	{WGS, WES, Targeted sequencing, etc.} <sup>72</sup>
5	Platform technology	Platform_Technology		Yes	varchar(20)	The technology platform used to identify the variant	NGS
6	Sequencer	Sequencer		Yes	varchar(50)	Sequencing equipment	Illumina Hiseq 2500
7	Sequencing Institution	Sequencing_Institution		Yes	varchar(50)	Name of sequencing institution	SNUBI
8	Experimenter	Experimenter		Yes	varchar(50)	Name of the primary experimenter	BJ Min
9	Collection Date	Collection_Date		Yes	datetime	Date of the sample collection	2018-09-03 11:00

**BIOINFORMATICS PROTOCOL RELATED INFORMATION Table specification**

#	Logical Name	Physical Name	PK	Required	Data Type	Description	Example
1	Bioinformatics Protocol Identifier	BI_Protocol_Identifier	PK	Yes	int(11)	Arbitrary identifier of the bioinformatics protocol defined in the CGDM database	121

2	Experiment Identifier	Experiment_Identifier	FK	Yes	int(11)	Arbitrary identifier of the experiment defined in the CGDM database	11
3	Pipeline Name	Pipeline_Name		Yes	varchar(50)	Name of the pipeline	SNUBI WXS data pipeline
4	Step (of the pipeline)	Step		Yes	int(3)	The order in which the steps are executed	1
5	Tool (of the pipeline)	Tool		Yes	varchar(50)	Procedure description	(alignment, sort, deduplication, variant calling, etc.)
6	Parameter (of the pipeline)	Parameter		Yes	varchar(50)	The name of tools	GATK
7	Datasource origin (used in the pipeline)	Datasource_Origin		Yes	varchar(50)	The version of tools	v2.5-2
8	Datasource version (used in the pipeline)	Datasource_Version		No	varchar(50)	Preset parameters used for the step	stand_call_conf=30,stand_emit_conf=10
9	Datasource Build (used in the pipeline)	Datasource_Build		No	varchar(50)	The source of databases	1kG, Mills, dbSNP137
10	Analytics Institution	Analytics_Institution		Yes	varchar(50)	Name of the bioinformatics analytics institution	SNUBI
11	Bioinformatician	Bioinformatician		Yes	varchar(50)	Name of the primary bioinformatician	YM Park
12	Received Date	Received_Date		Yes	datetime	Date of the raw data file (eg. BAM file) received	2018-09-15 17:35
13	Documentation Date	Documentation_Date		Yes	datetime	Date of the processed data stored in the CGDM database	2018-09-22 11:22

**QUALITY CHECK Table specification**

#	Logical Name	Physical Name	PK	Required	Data Type	Description	Example
1	Quality Check Identifier	QC_Identifier	PK	Yes	int(11)	Arbitrary identifier of the quality check matrix in the CGDM database	123
2	Bioinformatics Protocol Identifier	BI_Protocol_Identifier	FK	Yes	int(11)	Arbitrary identifier of the bioinformatics protocol in the CGDM database	121
3	Total Reads	Total_Reads		Yes	bigint	Total number of reads	100720000
4	Total Aligned Reads	Total_Aligned_Reads		No	bigint	Total number of aligned reads	99168912
5	% Reads Aligned	Reads_Aligned_Percent		No	float	Percentage of reads aligned	98.46 (= 4/3)
6	Total Bases	Total_Bases		No	bigint	Total number of bases	7260000
7	Total Mapped Bases	Mapped_Bases		No	bigint	Total number of mapped bases	7050000
8	Average on target depth	Depth_Mean		No	float	Mean on target depth	71.94
9	Standard deviation on target depth	Depth_SD		No	float	Standard deviation of on target depth	16.54
10	On Target Bases	Target_Bases		No	bigint	On target bases	2640000

**GENOMIC ALTERATION Table specification**

#	Logical Name	Physical Name	PK	Required	Data Type	Description	Example
1	Genomic Alteration Identifier	Genomic_Alteration_Identifier	PK	Yes	int(11)	Arbitrary identifier of the genomic alteration defined in the CGDM database	14009

2	Bioinformatics Protocol Identifier	BI_Protocol_Identifier	FK	Yes	int(11)	Arbitrary identifier of the bioinformatics protocol defined in the CGDM database	121
3	Position	Position		Yes	varchar(255)	The genomic position where the alteration occurs	180888597
4	Reference allele	Reference_Allele		Yes	varchar(255)	The base found in the reference genome	A
5	Alternative allele	Alternative_Allele		Yes	varchar(255)	Any base other than the reference	T
6	Chromosome	Chromosome		Yes	varchar(2)	The chromosome where the alteration occurs	7
7	Cytogenetic location	Cytogenetic_Location		No	text	Cytogenetic band that the location of the alteration maps to	17q12
8	Codon	Codon		No	text	The codon where the alteration is identified	12
9	Exon	Exon		No	varchar(10)	The exonic location where the alteration is identified	19
10	HGVS genomic change	HGVS_Genomic_Change		Yes	text	Description of the nucleotide change for a genomic sequence (supplied by HGVS)	NG_007873.3:g.176429T>A
11	HGVS coding change	HGVS_Coding_Change		No	text	Description of the nucleotide change for a coding DNA sequence (supplied by HGVS)	NM_004333.4:c.1799T>A

12	HGVS protein change	HGVS_Protein_Change	No	text	Description of the nucleotide change for a protein sequence (supplied by HGVS)	NP_004324.2:p.Val600Glu
13	HGVS version	HGVS_Version	Yes	varchar(20)	The version number of HGVS	HGVS version 15.11
14	dbSNP Identification Number	dbSNP_ID	No	varchar(20)	The identification tag (supplied by NCBI dbSNP)	rs56046546
15	dbVar Identification Number	dbVar_ID	No	varchar(20)	The identification tag (supplied by NCBI dbVar)	nsv1123397
16	Genome build	Genome_Build	No	varchar(20)	Genomic coordinates of the reference	GRCh37/hg19
17	Genomic source	Genomic_Source	Yes	varchar(10)	Class of genomic source	{Somatic, Germline, Unknown, etc.}
18	HGNC gene symbol	HGNC_Gene_Symbol	No	varchar(20)	The official gene symbol approved by the HGNC	ALK, JMJD7-PAL2G4B
19	Entrez gene ID	Entrez_ID	No	integer	Entrez Gene ID (supplied by NCBI)	238
20	Ensembl gene ID	Ensembl_ID	No	char(15)	Ensembl Gene ID (supplied by Ensembl)	ENSG00000171094
21	Genotype	Genotype	No	char(3)	Allelic state of the given variant	0 1, 0 0, . ., etc
22	clinVar Variation Identification Number	clinVar_Variant_ID	No	varchar(20)	The identification tag (supplied by clinVar)	188275
23	COSMIC Identification Number	COSMIC_ID	No	varchar(10)	The identification tag (supplied by COSMIC)	COSM476
24	Molecular Effects	Molecular_Effect	No	varchar(50)	Effects of mutations on protein function	{Missense, Nonsense, Frameshift, Promoter, etc}

25	Variant type	Variant_Type		Yes	varchar(20)	The type of variant in a sequence of DNA	{ Substitution, Deletion, Duplication, Insertion, InDel, Inversion, Conversion, etc. }
26	Functional Domain	Functional_Domain		No	varchar(50)	The functional domain where the alteration occurs	ATP-binding domain

**CLINICAL ANNOTATION Table specification**

#	Logical Name	Physical Name	PK	Required	Data Type	Description	Example
1	Clinical Annotation Identifier	Clinical_Annotation_Identifier	PK	Yes	int(11)	Arbitrary identifier of the clinical annotation defined in the CGDM database	22
2	Genomic Alteration Identifier	Genomic_Alteration_Identifier	FK	Yes	int(11)	Arbitrary identifier of the genomic alteration defined in the CGDM database	14009
3	Biomarker Datasource	Biomarker_Datasource		Yes	varchar(255)	Name of datasource for biomarkers of genomic data	ACMG actionable genes
4	Biomarker Name	Biomarker_Name		Yes	varchar(50)	Name of predictive indicator from biomarker datasource	EGFR Exon 19 Deletion

**MICROSATELLITE INSTABILITY Table specification**

#	Logical Name	Physical Name	PK	Required	Data Type	Description	Example
1	Microsatellite Instability Identifier	MSI_Identifier	PK	Yes	int(11)	Arbitrary identifier of microsatellite instability defined in the CGDM database	14



2	Bioinformatics Protocol Identifier	BI_Protocol_Identifier	FK	Yes	int(11)	Arbitrary identifier of the bioinformatics protocol defined in the CGDM database	121
3	MSI phenotype	MSI_Phenotype		Yes	varchar(50)	Distinct phenotype of the microsatellite instability	{Microsatellite Stable (MSS), MSI-Low (MSI-L), MSI-High (MSI-H), Indeterminate MSI}
4	MSI marker name	MSI_Marker_Name		Yes	varchar(20)	Name of the MSI marker	BAT26
5	MSI marker status	MSI_Marker_Status		Yes	varchar(20)	Determined MSI status	Positive

**Supplementary Table S2. IUPAC nucleotide code table for processing double/triple based code**

Symbol	Meaning
a	a; adenine
c	c; cytosine
g	g; guanine
t	t; thymine in DNA; uracil in RNA
m	a or c
r	a or g
w	a or t
s	c or g
y	c or t
k	g or t
v	a or c or g; not t
h	a or c or t; not g
d	a or g or t; not c
b	c or g or t; not a
n	a or c or g or t

\*reference: Cornish-Bowden, A. Nucl Acid Res 13, 3021-3030 (1985)  
[https://www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/MolBioReview/iupac\\_nt\\_abbreviations.html](https://www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/MolBioReview/iupac_nt_abbreviations.html) 에서 재인용

**Supplementary Table S3. Number of HLA alleles**

<i>Category</i>	<i>Locus</i>	<i>Allele number</i>	<i>Protein number</i>	<i>Null allele number</i>
<i>Class I</i>	HLA-A	673	527	46
	HLA-B	1077	911	38
	HLA-C	360	283	8
	HLA-E	9	3	0
	HLA-F	21	4	0
	HLA-G	36	14	1
	Pseudogenes	39		
	Total	2215	1742	93
<i>Class II</i>	HLA-DRA	3	2	0
	HLA-DRB	669	546	8
	HLA-DQA1	34	25	1
	HLA-DQB1	93	68	1
	HLA-DPA1	27	16	0
	HLA-DPB1	128	114	2
	HLA-DMA	4	4	0
	HLA-DMB	7	7	0
	HLA-DOA	12	3	1
	HLA-DOB	9	4	0
	Total	986	789	13
<i>MHC-like</i>	MICA	64	54	0
	MICB	30	19	2
	Total	94	73	2

\* reference: Shiina, T., Hosomichi, K., Inoko, H., & Kulski, J. K. (2009). The HLA genomic loci map: expression, interaction, diversity and disease. *Journal of human genetics*, 54(1), 15-39. Table 4. Number of HLA alleles

# Bibliography

- 1 Ginsburg, G. S. & Willard, H. F. Genomic and personalized medicine: foundations and applications. *Translational research* **154**, 277-287 (2009).
- 2 Downing, G. J., Boyle, S. N., Brinner, K. M. & Osheroff, J. A. Information management to enable personalized medicine: stakeholder roles in building clinical decision support. *BMC medical informatics and decision making* **9**, 44 (2009).
- 3 Collins, F. S. & Varmus, H. A new initiative on precision medicine. *New England Journal of Medicine* **372**, 793-795 (2015).
- 4 Dewey, F. E. *et al.* Clinical interpretation and implications of whole-genome sequencing. *Jama* **311**, 1035-1045 (2014).
- 5 McCarty, C. A. *et al.* The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC medical genomics* **4**, 13 (2011).
- 6 Masys, D. R. *et al.* Technical desiderata for the integration of genomic data into Electronic Health Records. *J Biomed Inform* **45**, 419-422, doi:10.1016/j.jbi.2011.12.005 (2012).
- 7 Lubin, I. M. *et al.* Principles and Recommendations for Standardizing the Use of the Next-Generation Sequencing Variant File in Clinical Settings. *J Mol Diagn* **19**, 417-426, doi:10.1016/j.jmoldx.2016.12.001 (2017).
- 8 Kho, A. N. *et al.* Practical challenges in integrating genomic data into the electronic health record. *Genet Med* **15**, 772-778, doi:10.1038/gim.2013.131 (2013).
- 9 Kassakian, S. Z., Yackel, T. R., Gorman, P. N. & Dorr, D. A. Clinical decisions support malfunctions in a commercial electronic health record. *Applied clinical informatics* **8**, 910-923 (2017).
- 10 Yang, W. *et al.* Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic acids research* **41**, D955-D961 (2012).
- 11 Roukos, D. H. Next-generation, genome sequencing-based biomarkers: concerns and challenges for medical practice. *Biomarkers in medicine* **4**, 583-586 (2010).
- 12 Roy, S. *et al.* Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists. *J Mol Diagn* **20**, 4-27, doi:10.1016/j.jmoldx.2017.11.003 (2018).
- 13 Oliver, G. R., Hart, S. N. & Klee, E. W. Bioinformatics for clinical next

- generation sequencing. *Clin Chem* **61**, 124-135, doi:10.1373/clinchem.2014.224360 (2015).
- 14 Gargis, A. S. *et al.* Good laboratory practice for clinical next-generation sequencing informatics pipelines. *Nat Biotechnol* **33**, 689-693, doi:10.1038/nbt.3237 (2015).
  - 15 Han, P. K. J. *et al.* A taxonomy of medical uncertainties in clinical genome sequencing. *Genet Med* **19**, 918-925, doi:10.1038/gim.2016.212 (2017).
  - 16 Simianu, V. V. *et al.* Understanding clinical and non-clinical decisions under uncertainty: a scenario-based survey. *BMC medical informatics and decision making* **16**, 153 (2016).
  - 17 Shebl, N. A., Franklin, B. D. & Barber, N. Is failure mode and effect analysis reliable? *Journal of patient safety* **5**, 86-94 (2009).
  - 18 Singh, V., Pungotra, H., Singh, S. & Gill, S. S. Prioritization of Failure Modes in Process FMEA using Fuzzy Logic. *International Journal Of Enhanced Research In Science Technology & Engineering* **2** (2013).
  - 19 Certa, A., Hopps, F., Inghilleri, R. & La Fata, C. M. A Dempster-Shafer Theory-based approach to the Failure Mode, Effects and Criticality Analysis (FMECA) under epistemic uncertainty: application to the propulsion system of a fishing vessel. *Reliability Engineering & System Safety* **159**, 69-79 (2017).
  - 20 Teng, S.-H. & Ho, S.-Y. Failure mode and effects analysis: an integrated approach for product design and process control. *International journal of quality & reliability management* **13**, 8-26 (1996).
  - 21 Gilchrist, W. Modelling Failure Modes and Effects Analysis. *International Journal of Quality & Reliability Management* **10**, doi:10.1108/02656719310040105 (1993).
  - 22 Eubanks, C. F., Kmenta, S. & Ishii, K. in *ASME Design Engineering Technical Conferences*. 14-17.
  - 23 Reifer, D. J. Software failure modes and effects analysis. *IEEE Transactions on reliability* **28**, 247-249 (1979).
  - 24 Vajna, S. in *ASME 2003 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. 375-382 (American Society of Mechanical Engineers).
  - 25 Sayyadi Tooranloo, H., Ayatollah, A. S. & Alboghobish, S. Evaluating knowledge management failure factors using intuitionistic fuzzy FMEA approach. *Knowledge and Information Systems* **57**, 183-205, doi:10.1007/s10115-018-1172-3 (2018).
  - 26 Cabanes, B., Hubac, S., Le Masson, P. & Weil, B. in *14th International Design Conference (DESIGN 2016)*.

- 27 Chandrasegaran, S. K. *et al.* The evolution, challenges, and future of knowledge representation in product design systems. *Computer-aided design* **45**, 204-228 (2013).
- 28 Blount, G., Kneebone, S. & Kingston, M. Selection of knowledge-based engineering design applications. *Journal of Engineering Design* **6**, 31-38 (1995).
- 29 Tamisier, T. & Feltz, F. A Data Model for Knowledge Representation in Collaborative Systems. *Data Science Journal* **6**, S225-S233 (2007).
- 30 Navathe, S. B. & Schkolnick, M. in *Proceedings of the 1978 ACM SIGMOD international conference on management of data.* 144-156 (ACM).
- 31 Smith, J. M. & Smith, D. C. Database abstractions: aggregation and generalization. *ACM Transactions on Database Systems (TODS)* **2**, 105-133 (1977).
- 32 Consortium, G. P. A global reference for human genetic variation. *Nature* **526**, 68 (2015).
- 33 National Cancer Institute GDC Data Portal TCGA PAAD dataset. at <https://portal.gdc.cancer.gov/projects/TCGA-PAAD> (2017)
- 34 Sen, A., Al Kawam, A. & Datta, A. Emergence of DSS efforts in genomics: Past contributions and challenges. *Decision Support Systems* **116**, 77-90 (2019).
- 35 Overby, C. L., Tarczy-Hornoch, P., Hoath, J. I., Kalet, I. J. & Veenstra, D. L. in *BMC bioinformatics.* S10 (BioMed Central).
- 36 Hoffman, M. A. & Williams, M. S. Electronic medical records and personalized medicine. *Human genetics* **130**, 33-39 (2011).
- 37 Castaneda, C. *et al.* Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. *Journal of clinical bioinformatics* **5**, 4 (2015).
- 38 Dinu, V. & Nadkarni, P. Guidelines for the effective use of entity-attribute-value modeling for biomedical databases. *Int J Med Inform* **76**, 769-779, doi:10.1016/j.ijmedinf.2006.09.023 (2007).
- 39 Peleg, M. The Role of Modeling in Clinical Information System Development Life Cycle. *Methods of information in medicine* **50**, 7-10 (2011).
- 40 Williams, M. S. *et al.* Genomic Information for Clinicians in the Electronic Health Record: Lessons Learned from ClinGen and eMERGE. (2019).
- 41 Dolin, R. H., Boxwala, A. & Shalaby, J. A Pharmacogenomics Clinical Decision Support Service Based on FHIR and CDS Hooks. *Methods Inf Med* **57**, e115-e123, doi:10.1055/s-0038-1676466 (2018).

- 42 Relling, M. V. & Klein, T. E. CPIC: Clinical Pharmacogenetics Implementation Consortium of the Pharmacogenomics Research Network. *Clin Pharmacol Ther* **89**, 464-467, doi:10.1038/clpt.2010.279 (2011).
- 43 Swen, J. *et al.* Pharmacogenetics: from bench to byte. *Clinical Pharmacology & Therapeutics* **83**, 781-787 (2008).
- 44 Consortium, I. W. P. Estimation of the warfarin dose with clinical and pharmacogenetic data. *New England Journal of Medicine* **360**, 753-764 (2009).
- 45 Ross, C. J. *et al.* The Canadian Pharmacogenomics Network for Drug Safety: a model for safety pharmacology. *Thyroid* **20**, 681-687 (2010).
- 46 Blagec, K. *et al.* Implementing pharmacogenomics decision support across seven European countries: The Ubiquitous Pharmacogenomics (U-PGx) project. *J Am Med Inform Assoc* **25**, 893-898, doi:10.1093/jamia/ocy005 (2018).
- 47 Cavallari, L. H. *et al.* Multi-site investigation of strategies for the clinical implementation of CYP2D6 genotyping to guide drug prescribing. *Genet Med*, doi:10.1038/s41436-019-0484-3 (2019).
- 48 Cicali, E. J. *et al.* Challenges and lessons learned from clinical pharmacogenetic implementation of multiple gene-drug pairs across ambulatory care settings. *Genet Med*, doi:10.1038/s41436-019-0500-7 (2019).
- 49 Pearce, C. *et al.* Delivering genomic medicine in the United Kingdom National Health Service: a systematic review and narrative synthesis. *Genet Med*, doi:10.1038/s41436-019-0579-x (2019).
- 50 Walton, N. A., Johnson, D. K., Person, T. N. & Chamala, S. Genomic Data in the Electronic Health Record. *Advances in Molecular Pathology* **2**, 21-33, doi:10.1016/j.yamp.2019.07.001 (2019).
- 51 Lenzerini, M. in *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems.* 233-246.
- 52 Kho, A. N. *et al.* Practical challenges in integrating genomic data into the electronic health record. *Genetics in Medicine* **15**, 772 (2013).
- 53 Caudle, K. E. *et al.* Standardizing terms for clinical pharmacogenetic test results: consensus terms from the Clinical Pharmacogenetics Implementation Consortium (CPIC). *Genet Med* **19**, 215-223, doi:10.1038/gim.2016.87 (2017).
- 54 The Clinical Pharmacogenetics Implementation Consortium (CPIC®), CPIC guidelines. <https://cpicpgx.org/guidelines/> [accessed 2020-05-10]
- 55 Robarge, J., Li, L., Desta, Z., Nguyen, A. & Flockhart, D. The star-allele nomenclature: retooling for translational genomics. *Clinical Pharmacology*

- & *Therapeutics* **82**, 244-248 (2007).
- 56 Minucci, A. *et al.* Glucose-6-phosphate dehydrogenase (G6PD) mutations database: review of the "old" and update of the new mutations. *Blood Cells Mol Dis* **48**, 154-165, doi:10.1016/j.bcmd.2012.01.001 (2012).
- 57 Beutler, E. The designation of mutations. *American journal of human genetics* **53**, 783 (1993).
- 58 den Dunnen, J. T. & Antonarakis, S. E. Nomenclature for the description of human sequence variations. *Hum Genet* **109**, 121-124, doi:10.1007/s004390100505 (2001).
- 59 Bodmer, J. G. *et al.* Nomenclature for factors of the HLA system, 1989. *Immunobiology* **180**, 278-292 (1990).
- 60 Klein, J. & Sato, A. The HLA system. *New England Journal of Medicine* **343**, 702-709 (2000).
- 61 Mosaad, Y. Clinical role of human leukocyte antigen in health and disease. *Scandinavian journal of immunology* **82**, 283-306 (2015).
- 62 Mahdi, B. M. A glow of HLA typing in organ transplantation. *Clinical and translational medicine* **2**, 6 (2013).
- 63 Fan, W.-L. *et al.* HLA association with drug-induced adverse reactions. *Journal of immunology research* **2017** (2017).
- 64 Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in medicine* **17**, 405-423 (2015).
- 65 Jones, K. H. *et al.* The other side of the coin: Harm due to the non-use of health-related data. *Int J Med Inform* **97**, 43-51, doi:10.1016/j.ijmedinf.2016.09.010 (2017).
- 66 Lee, K. H., Kim, H. J., Kim, Y.-J., Kim, J. H. & Song, E. Y. Extracting Structured Genotype Information from Free-Text HLA Reports Using a Rule-Based Approach. *Journal of Korean Medical Science* **35** (2020).
- 67 Solomatine, D., See, L. M. & Abrahart, R. in *Practical hydroinformatics* 17-30 (Springer, 2009).
- 68 Warner, J. L., Jain, S. K. & Levy, M. A. Integrating cancer genomic data into electronic health records. *Genome medicine* **8**, 113 (2016).
- 69 Pennington, J. W. *et al.* Genomic decision support needs in pediatric primary care. *Journal of the American Medical Informatics Association* **24**, 851-856 (2017).
- 70 Heale, B. S. *et al.* Integrating genomic resources with electronic health records using the HL7 Infobutton standard. *Applied clinical informatics* **7**,



817-831 (2016).

- 71 Hamburg, M. A. & Collins, F. S. The path to personalized medicine. *New England Journal of Medicine* **363**, 301-304 (2010).
- 72 Alterovitz, G. *et al.* FHIR Genomics: enabling standardization for precision medicine use cases. *NPJ genomic medicine* **5**, 1-4 (2020).
- 73 Page, A. *et al.* Genomics. A federated ecosystem for sharing genomic, clinical data. Global Alliance for Genomics and Health. *Science* **352**, 1278-1280 (2016).
- 74 Lawler, M. *et al.* All the world's a stage: facilitating discovery science and improved cancer care through the Global Alliance for Genomics and Health. *Cancer discovery* **5**, 1133-1136 (2015).
- 75 Shin, S. J. *et al.* Genomic Common Data Model for Seamless Interoperation of Biomedical Data in Clinical Practice: Retrospective Study. *J Med Internet Res* **21**, e13249, doi:10.2196/13249 (2019).
- 76 Haarbrandt, B. *et al.* HiGHmed—an open platform approach to enhance care and research across institutional boundaries. *Methods of information in medicine* **57**, e66-e81 (2018).
- 77 Rector, A. L. Thesauri and formal classifications: terminologies for people and machines. *Methods of information in medicine* **37**, 501-509 (1998).
- 78 Barile, S., Polese, F., Saviano, M. & Carrubbo, L. in *Innovating in Practice* 417-438 (Springer, 2017).
- 79 Tooranloo, H. S., Ayatollah, A. S. & Alboghobish, S. Evaluating knowledge management failure factors using intuitionistic fuzzy FMEA approach. *Knowledge and Information Systems*, 1-23 (2018).
- 80 DeRosier, J., Stalhandske, E., Bagian, J. P. & Nudell, T. Using health care failure mode and effect analysis™: the VA National Center for Patient Safety's prospective risk analysis system. *The Joint Commission journal on quality improvement* **28**, 248-267 (2002).
- 81 Deandrea, S. *et al.* Implementation of Failure Mode and Effects Analysis to the specimens flow in a population-based colorectal cancer screening programme using immunochemical faecal occult blood tests: a quality improvement project in the Milan colorectal cancer screening programme. *BMJ Open Qual* **7**, e000299 (2018).
- 82 Overby, C. L. *et al.* Developing a prototype system for integrating pharmacogenomics findings into clinical practice. *Journal of personalized medicine* **2**, 241-256 (2012).
- \* Kim, H. J., Kim, H. J., Park, Y., Lee, W. S., Lim, Y., & Kim, J. H. (2020). clinical Genome Data Model (cGDM) provides interactive clinical Decision Support for precision Medicine. *Scientific reports*, 10(1), 1-13.

## [Website]

The Clinical Pharmacogenetics Implementation Consortium (CPIC®), CPIC guidelines. <https://cpicpgx.org/guidelines/> [accessed 2020-05-10]

Health Level Seven, FHIR Genomics, <https://www.hl7.org/fhir/genomics.html> [accessed 2020-07-03]

The Global Alliance for Genomics and Health alliance, GA4GH Genomics API, <https://ga4gh-schemas.readthedocs.io/en/latest/> [accessed 2020-07-03]

International Organization for Standardization, IOS 25720:2009 Genomic Sequence Variation Markup Language(GSVML) <https://www.iso.org/standard/43182.html> [accessed 2020-07-03]

International Organization for Standardization, ISO/TS 20428:2017 Data elements and their metadata for describing structured clinical genomic sequence information in electronic health records <https://www.iso.org/standard/67981.html> [accessed 2020-07-03]

National Institution NHI NCI GDC <https://gdc.cancer.gov/developers/gdc-data-model> [accessed 2020-07-03]

## [Acknowledgement]

The main body of the dissertation chapter 1 and part of the general discussion has been published as the following paper: Kim, H. J., Kim, H. J., Park, Y., Lee, W. S., Lim, Y., & Kim, J. H. (2020). clinical Genome Data Model (cGDM) provides interactive clinical Decision Support for precision Medicine. Scientific reports, 10(1), 1-13.

# 국문 초록

## 정밀의학을 위한 임상유전체데이터모델

김 효 정

서울대학교 의과대학

의료정보학 협동과정

진료 현장에서 의사결정을 내려야 하는 임상에게 개인 유전체 정보를 다른 임상 근거들과 통합하여 보다 쉽게 다룰 수 있도록 구조화하여 지원하는 것은 정밀의학 구현을 위한 의료정보학의 주요 과제 중 하나이다. 차세대 염기서열 분석법과 같은 대량신속처리 유전체 기술의 등장과 그에 따른 해석정보의 축적으로 정밀 의학 및 개인 맞춤형 의학으로의 전환이 가시화 되는 듯 보였으나, 차세대염기서열 분석 기술 기반의 개인유전체 정보의 임상 활용은 여전히 제한적이다. 선행연구에서는 임상현장에서 유전체정보의 활용이 더딘 이유로 의료 전문가와 생물정보학자들 사이의 지식 격차, 진료 현장과 생물정보학 작업절차 간의 분리, 유전체 데이터만의 독특한 양적, 질적 자료구조의 특성과 같은 복합적인 원인을 제시하고 있다. 이러한 문제를 해결하고자 하는 시도로서 개인유전체정보를 병원정보시스템에 통합해야 한다는 요구가 높아지고 있으나 임상현장에서 활용하는 것을 목적으로 하는 지속가능하고 상호운용가능한 저장, 관리, 처리 방식에 대한 구체적인 논의는 부족한 실정이다.

본 연구에서는 임상정보시스템에 개인 유전체 정보가 통합되어 임상에 적용되기까지 현재의 장벽들을 문헌고찰을 통해 재탐색하고 관련된 개념과 방법들을 고찰하였다. 그리고 차세대 염기서열 분석방법을 기반으로 한 데이터를 어떻게 임상에서 활용하기 쉽도록 저장하고 처리하고 전달할 것인가 하는 당면한 과제에 단계적으로 접근하였다. 정보시스템 설계에 있어 데이터 모델의 설계는 최종시스템의 기능이 데이터 모델에 표현된 정보량 안에서 제한된다는 점에서 가장 일차적이며 중요한 단계이다. 따라서 1장에서는 다학제적 논의를 통해 임상 의사결정에 활용할 수 있는 유전체 지식표현을 논리적 데이터모델의 형태로 도출하여 차세대염기서열분석기술 기반의 임상유전체데이터모델(cGDM; clinical Genome Data Model)을 제안하였다. 2장에서는 약물치료를 개인별로 최적화하기 위해 이용 가능한 유전체검사결과를 사용하는 방법에 대한 지식체인 CPIC guideline을 구조화하여 1장에서 구축한 환자의 유전체 정보와 데이터 레벨의 정보흐름을 구현함으로써 모듈 방식의 약물유전체 임상 의사결정지원시스템을 제시한다. 3장에서는 생명정보학에서 임상적 의미를 드러내는 발견들이 지속됨에 따른 명명체계의 다양함을 수용하는 확장 체계의 하나로서 HLA gene에 대한 구조화된 정보 설계와 구현을 다루었다. 즉, HLA nomenclature를 대상으로 지식표현을 설계, 확장하여 임상유전체데이터모델의 지속가능성과 확장성을 검증하였다.

본 연구에서는 중개과학으로서 의료정보학이 정밀의료에 기여할 수 있는 다학제적공간을 탐색하고 정보시스템의 지식표현, 기능구현, 사용성 측면을 포괄하는 접근을 시도하였다. 본 연구의 결과로 제시된 임상유전체데이터모델은 논리적인 데이터모델 수준에서 설계되어 기존 병원정보시스템에 사용된 개발 언어에 제약을 받지 않고 데이터 수준의 확장체계로 활용할 수 있다. 즉, 정형화된 데이터를 기반으로

임상정보를 처리하는 기존의 다양한 정보시스템 아키텍처의 설계에 통합되어 각 기관 혹은 사용자의 필요에 맞게 CDSS나 서식에 연결하는 등 다양한 기능의 구현을 지원할 수 있다. 또한 연구용 데이터의 수집과 분석에 사용될 수도 있어 개인유전체분석결과를 실질적인 데이터 순환 사이클에 연결하는 데 기여할 수 있다. 궁극적으로, 의료전문가와 정보를 활용한 임상 의사결정간의 지적상호작용을 지원하는 데이터 계층 인프라를 제공한다.

**Keyword:** 정밀의료, 지식중개, 지식공학, 통합병원정보시스템, 유전체데이터모델, 약물유전체정보를 활용한 임상 의사결정지원시스템

**Student Number:** 2015-30615