

Original Article
Medical Informatics



Extracting Structured Genotype Information from Free-Text HLA Reports Using a Rule-Based Approach

Kye Hwa Lee ,¹ Hyo Jung Kim ,² Yi-Jun Kim ,¹ Ju Han Kim ,² and Eun Young Song ³

¹Center for Precision Medicine, Seoul National University Hospital, Seoul, Korea

²Division of Biomedical Informatics, Seoul National University Biomedical Informatics and Systems Biomedical Informatics Research Center, Seoul National University College of Medicine, Seoul, Korea

³Department of Laboratory Medicine, Seoul National University College of Medicine, Seoul, Korea



Received: May 16, 2019

Accepted: Jan 29, 2020

Address for Correspondence:

Eun Young Song, MD, PhD

Division of Biomedical Informatics, Seoul National University Biomedical Informatics (SNUBI) and Systems Biomedical Informatics Research Center, Seoul National University College of Medicine, 103 Daehak-ro, Jongno-gu, Seoul 03080, Korea.
E-mail: eysong1@snu.ac.kr

Kye Hwa Lee, MD, PhD

Center for Precision Medicine, Seoul National University Hospital, 101 Daehak-ro, Jongno-gu, Seoul 03080, Korea.
E-mail: geffa@snu.ac.kr

© 2020 The Korean Academy of Medical Sciences.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

ORCID iDs

Kye Hwa Lee

<https://orcid.org/0000-0002-7593-7020>

Hyo Jung Kim

<https://orcid.org/0000-0001-9555-0926>

Yi-Jun Kim

<https://orcid.org/0000-0002-1763-4267>

Ju Han Kim

<https://orcid.org/0000-0003-1522-9038>

ABSTRACT


Background: Human leukocyte antigen (HLA) typing is important for transplant patients to prevent a severe mismatch reaction, and the result can also support the diagnosis of various disease or prediction of drug side effects. However, such secondary applications of HLA typing results are limited because they are typically provided in free-text format or PDFs on electronic medical records. We here propose a method to convert HLA genotype information stored in an unstructured format into a reusable structured format by extracting serotype/allele information.

Methods: We queried HLA typing reports from the clinical data warehouse of Seoul National University Hospital (SUPPREME) from 2000 to 2018 as a rule-development data set (64,024 reports) and from the most recent year (6,181 reports) as a test set. We used a rule-based natural language approach using a Python regex function to extract the 1) number of patients in the report, 2) clinical characteristics such as indication of the HLA testing, and 3) precise HLA genotypes. The performance of the rules and codes was evaluated by comparison between the extracted results from the test set and a validation set generated by manual curation.

Results: Among 11,287 reports for development set and 1,107 for the test set describing HLA typing for a single patient, iterative rule generation developed 124 extracting rules and 8 cleaning rules for HLA genotypes. Application of these rules extracted HLA genotypes with 0.892–0.999 precision and 0.795–0.998 recall for the five HLA genes. The precision and recall of the extracting rules for the number of patients in a report were 0.997 and 0.994 and those for the clinical variable extraction were 0.997 and 0.992, respectively. All extracted HLA alleles and serotypes were transformed according to formal HLA nomenclature by the cleaning rules.

Conclusion: The rule-based HLA genotype extraction method shows reliable accuracy. We believe that there are significant number of patients who takes profit when this under-used genetic information will be return to them.

Keywords: Major Histocompatibility Complex; HLA Test; Genetic Testing; Electronic Medical Record; Data Sets as Topic

Eun Young Song 

<https://orcid.org/0000-0003-1286-9611>

Funding

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1D1A1A02086109).

Disclosure

The authors have no potential conflicts of interest to disclose.

Author Contributions

Conceptualization: Lee KH, Kim HJ. Funding acquisition: Lee KH. Investigation: Lee KH, Kim HJ. Methodology: Lee KH, Kim YJ. Resources: Song EY. Supervision: Kim JH, Song EY. Writing - original draft: Lee KH, Song EY. Writing - review & editing: Lee KH, Kim HJ.

INTRODUCTION

Human leukocyte antigen (HLA) plays an important role in regulating the immune response.¹ In particular, the self-recognition function is essential when transplanting donor organs that are not consistent with HLA typing in recipients, as it is a major cause of severe transplant rejection. Therefore, transplant recipients should be tested for HLA typing before transplantation to confirm a compatible HLA type with the donor.² Moreover, HLA diversity was recently reported to be associated with more than 100 diseases as well as severe drug hypersensitivity.³⁻⁵ However, the HLA results of transplant patients and donors are rarely used for the diagnosis of HLA-related diseases or for predicting future adverse drug reactions. One of the reasons for this underuse of HLA typing results in clinical settings outside of the context of organ transplantation is the variable level of resolution depending on the testing methods, ranging from a micro-lymphocytotoxicity-based assay to next-generation sequencing.⁶ Developments in next-generation-sequencing technology have now enabled the high-resolution screening for alleles and genotypes as well as serologic typing.⁷ Even in the case of organ transplantation, the resolution of the tests required differs depending on whether the target organ is a solid tissue or blood. However, another, and arguably more important, reason for the underuse of HLA typing is the lack of standardization and structure in the representation format of the HLA typing results in electronic medical records (EMRs).

Standardization with respect to the terminology and format of clinical data has been increasingly emphasized owing to advantages of the continuous utilization of essential clinical information from the EMR while minimizing any loss of data during transformation.³ Because the clinical documents written in a free-texted or semi-structured format frequently contain abundant and/or substantial clinical information, it is vital to transform such records to a structured and standardized format, especially when designing a clinical decision support system (CDSS) to ensure patient safety.⁸ One of the most widely used methods to extract precise data from a free-texted clinical document is natural language processing (NLP)^{9,10} and many tools have been developed for retrieving various types of data from EMR using NLP to enable its secondary use.⁹⁻¹³ In particular, genomics applications have been actively studied for extracting phenotype information using NLP from EMR.^{14,15} However, for the extracting genotype data using these NLP-based methods, most studies have thus far focused on literature databases rather than the EMR itself. In other words, little research has been done on the extraction of genotypes stored in EMR for secondary use.

To overcome these limitations and promote the secondary use of essential and available clinical information, in this study, we evaluated the accuracy of extraction methods which focused on the reports of HLA tests obtained by various typing methods with various degrees of resolution that are stored in a semi-structured format at the clinical data warehouse (CDW) of Seoul National University Hospital (SNUH). We further developed a set of rules for extracting and transforming these data into a structured, standard format. The established rules were evaluated by comparison of a validation set generated by manual curation from the testing set. The findings of this study can expanded the possibility of the secondary use of currently underused genotype data for diverse clinical and medical purposes.

METHODS

Data source

We used the data stored at the CDW of SNUH, SUPREME[®], to retrieve the HLA reports. The test names used to query of the HLA reports from SUPREME were “HLA-[A,B,Cw,DR,DQ] (DNA,[Low,High])”, in which the terms in parentheses were iteratively used. To develop rules and codes for extracting HLA serotype/genotype data from the free-texted HLA reports, we queried HLA typing results obtained between January 1, 2000, and June 30, 2018 at SNUH as a rule-development dataset. To validate and evaluate the performance of the rules and codes developed using the rule-development set, we generated a second HLA data set by querying reports generated from July 1, 2018 to June 30, 2019 as a test set.

The HLA reports were retrieved in Excel files for each rule-development and test set. In both the EMR and CDW, a single HLA report itself was stored in free-texted format and extracted as a single cell of an Excel file for each patient. As shown in Fig. 1, an HLA report resembled the format of a table with spaces and line breaks to express the structure of attribute-value pairs of HLA serotypes, but was not an actual table.

Data extraction method

The extraction process consisted of two steps: 1) develop and apply rules to include reports with only single patient records, and 2) develop and apply rules to extract clinical characteristics and HLA genotypes. We first checked sample HLA reports to determine the

< HLA Typing >

Clinical Indication of study: Kidney transplantation

	Recipient	Donor 1	Donor 2	Donor 3
Name	가나다			
Sex/Age	M/56			
Relation to recipient				
HLA-A(DNA)	A24(*24) A31(*31)			
HLA-B(DNA)	B61(*40:02g) B67(*67)			
HLA-C(DNA)	Not tested			
HLA-DRB1(DNA)	DR4 (*04:01g) DR12(*12:01g)			
HLA-DQB1(DNA)	Not tested			
Sample No.	130751-ABDR			
Date of test	2013.4.29			
보고자(판독의)	마바사 M.D.			

Fig. 1. Representative example of raw data in an HLA report. A real example of an HLA typing report with de-identified patient names is shown. An HLA typing result is represented in one cell in an Excel file. This example includes HLA typing results for three patients because most HLA test subjects were recipients of a transplantation procedure, and the physicians wanted to compare all HLA tests of candidate donors with those of a given recipient on the same page on the electronic medical record. In this cell, the HLA test results were arranged in a tabular form using the space bar and a carriage return but were not structured as actual tables with distinct rows and columns. To improve accuracy, we only focused on reports with the HLA typing results of one patient. HLA = human leukocyte antigen.

pattern of HLA genotypes for developing extraction rules. The iteration processes were applied for 100 reports initially, reaching up to 1,000, and were then applied to the total test set. Based on the first assessment, we developed basic rules and updated codes and then ran the codes to identify missing patterns iteratively. *Rule #1* was first used to identify whether or not there was single patient's record in a report. After excluding the reports with the results from multiple patients, we applied *Rule #2* to extract clinical data (patient's name and indication of HLA testing) and HLA genotypes for the five main HLA genes (A, B, Cw, DR, and DQ). Because it was relatively simple to establish rules for identifying the number of patients or to extract clinical data in the reports, we mainly focused on extracting HLA genotypes accurately in this step. When application of the rules missed some of the genotypes in the report, we printed them on the monitor to visualize the missing pattern and then developed an additional rule or modified existing rules to extract the missed genotypes. This process was performed iteratively until the remaining missing patterns were only simple spelling errors. For example, there was one report with HLA-B genotyping information only; however, this was not extracted correctly because the genotype was recorded as "HLA-B[multiple spaces][line break][multiple spaces]B5-." This extraction process was completed using the regular expression function of Python programming language (version 3; Python Software Foundation; <https://www.python.org/>).

Performance evaluation

The performance of the developed extraction rules was compared with the validation set which was constructed by manual curation (performed by the research nurse and one physician) for the test set. As shown in **Fig. 2**, after building the primary validation set by the first curator, one of the authors reviewed the primary validation set and compared the result with raw data. After completion of the serial curation process, the final validation set was constructed.

Cleaning and converting the HLA typing to nomenclature

In addition to the extracting rules, we developed rules to clean and convert the extracted HLA typing data to a standard HLA nomenclature format as *Rule #3*. The WHO Nomenclature Committee for Factors of the HLA System is responsible for determining nomenclature to standardize the expression of these various HLA test results.^{16,17} Because the data used in this study included a mixture of low-resolution serotyping results and high-resolution allele-typing results, we extracted the allele notations according to the HLA subtypes as well as those based on serological specifications. The cleaning and transforming rules were established after extraction of all HLA alleles.

Ethics statement

The study was approved by the Institutional Review Board (IRB) of SNUH (No. 1811-157-989). The IRB approved the conduction of this study without the informed consent from the participants because this study used anonymized retrospective EMR data.

RESULTS

Summary of HLA typing data

There were 64,023 HLA reports assigned to 16,707 patients in the rule-development set, and there were 6,180 reports for 1,769 patients in the testing set. There were multiple reports for some patients owing to cases in which physicians ordered tests for HLA genes separately for a given patient or when a patient had multiple potential donors. According to *Rule #1*, 52,736

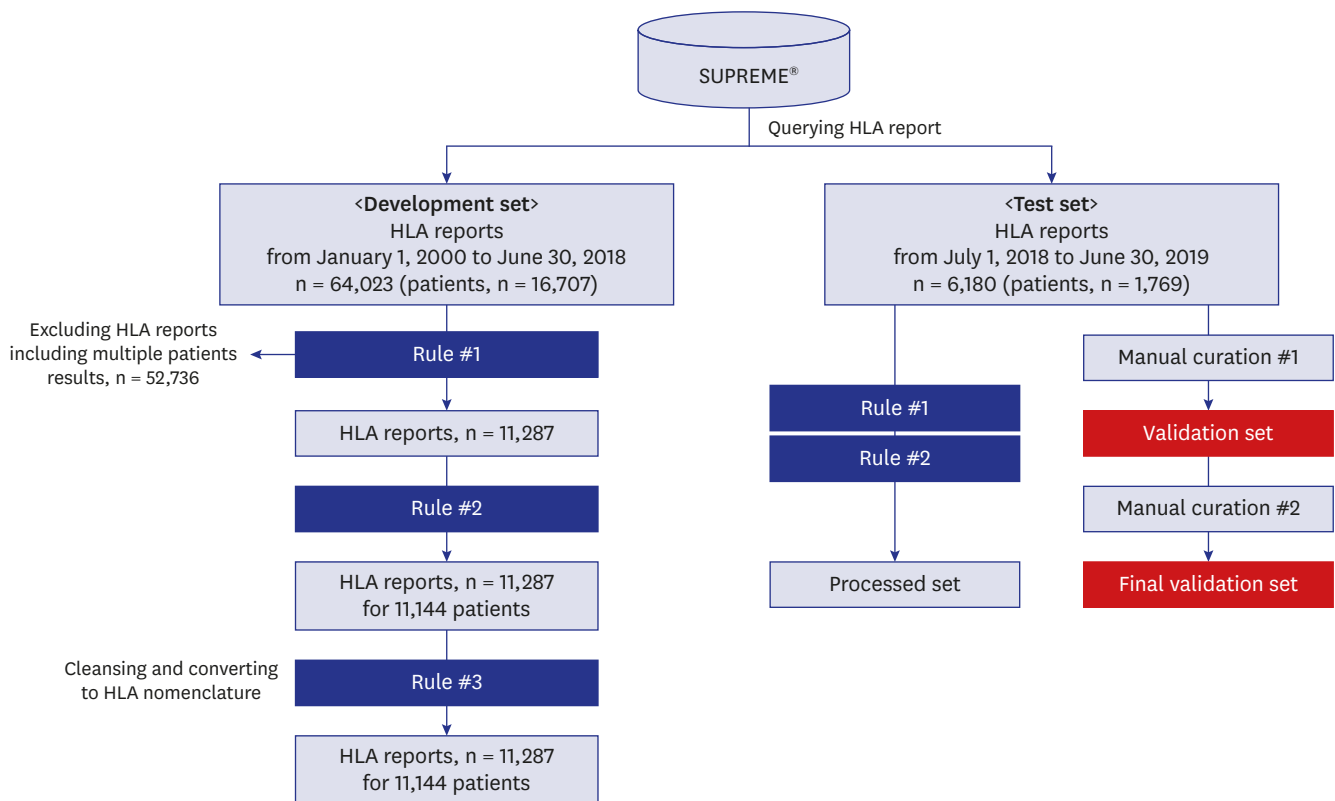


Fig. 2. HLA typing and clinical characteristics extraction pipeline. *Rule #1* was designed to exclude HLA reports with typing results for multiple patients. After excluding these reports, we applied *Rule #2* to extract clinical variables such as name, sex, and indications of HLA typing. *Rule #1* and *Rule #2* are extraction rules, and *Rule #3* is cleaning rule. *Rule #3* was designed to clean the HLA typing results and transform the results to a standard nomenclature. To evaluate the accuracy of the two extraction rules, we applied *Rule #1* and *Rule #2* to the testing set. The rule-based extraction results of the testing set were then compared with the manually curated results of the testing test, as the validation set. The manual curation process was done sequentially by two different investigators. HLA = human leukocyte antigen.

reports and 5,420 patients were excluded in rule-development set. Most of the excluded reports were empty because when tests for multiple HLA genes were ordered for a single patient at the same time, the results recorded in a single report were provided in a combined form. There were 4,039 male and 7,105 female patients, including 2,642 high-resolution tests, 5,835 low-resolution tests, and 2,810 tests with undefined resolution in the final rule-development set.

HLA data extraction rules

There were three types of irregularities detected in the HLA typing results from the raw HLA reports. First, the number of patients with test results was inconsistent. In many cases, only one patient's HLA test results was described in a report, whereas other reports included results for more than two patients'. Second, the resolution of HLA typing methods varied, which also influenced the results. Given the rapid development in HLA typing methods, physicians frequently ordered multiple HLA typing methods with different degrees of resolution at the same time for the same patients. Thus, HLA reports based on low-resolution tests, including serology or analysis of sequence-specific oligonucleotides, only provide information on serotypes (allele groups), whereas the reports based on high-resolution methods, including sequence-based typing, provide information on the specific HLA protein or genotype (DNA substitutions). Third, there was ambiguity in the presentation of the HLA typing result. Because there are two allele types for each HLA locus, if the alleles are homozygous, special

Table 1. Typical Python expressions used to extract the HLA genotype status and clinical variables for each patient

Variables	No. of expressions	Patterns used in regular expression development
Diagnosis	1	^Clinical Indication of study(.+)
Sex/age	1	^((Sex\Age) (Age\Sex))\s+([A-Z]+)\s+(\d+)
Name	1	^Name\s+([\S]+)
HLA genotypes	124	^(HLA-[A B Cw DR DQ])\s+(\DNA)\s+(((A B Cw DR DQ) (\^*\d+:\d+))\s+ +(((A B Cw DR DQ) (\^*\d+:\d+))\s+

HLA = human leukocyte antigen.

characters such as quotes were used to express the second allele. In addition, the use of spaces, line separations, and positions of the line break to obtain a table-like format were also irregular. To minimize problem complexity for standardization, in this study, we focused only on the HLA reports with a single patient result according to *Rule #1* as described above.

After excluding the reports with the results of multiple patients, we applied *Rule #2*, which was developed to extract clinical data (patient's name and indication of HLA testing), with a focus on an accurate determination of HLA genotypes as shown in **Fig. 2**. Through the iteratively developing the rule generation process, including handling missing patterns, we ultimately developed 124 rules to extract HLA genotypes. The patterns of the rules are represented in **Table 1** and the specific expressions used for representation are available at <https://github.com/geffa/HLAgenotypeParsing/blob/master/RegularExpression>. The summary of the initial extraction of the HLA genotypes and clinical characteristics for the rule-development set is provided in **Table 2**. The extraction results demonstrated an uneven distribution in the number of patients that received typing for each of the five HLA gens. Specifically, HLA-B was tested most frequently (ordered for 69.7% of patients in the rule-development set) and the typing of HLA-C was the least frequent (10.4%).

Table 2. HLA genotype frequencies in the test set extracted by *Rule #1* and *Rule #2*

HLA gene	No. of patients who underwent HLA typing	No. of HLA types	Top 5 alleles/frequencies	
			Alleles	Frequencies
HLA-A	4,217	45	*02	0.221
			*24	0.182
			*33	0.130
			*11	0.086
			*24:02	0.068
HLA-B	7,765	101	*15:01	0.072
			*44	0.068
			*51	0.064
			*54	0.048
			*40:02	0.046
HLA-C	1,152	33	*01:02	0.190
			*03:03	0.118
			*03:04	0.098
			*07:02	0.095
			*08:01	0.079
HLA-DR	6,376	65	*04	0.099
			*09:01	0.061
			*12	0.055
			*04:05	0.053
			*13:02	0.052
HLA-DQ	1,631	17	*03:01	0.143
			*03:03	0.125
			*03:02	0.111
			*06:01	0.103
			*05:01	0.090

HLA = human leukocyte antigen.

Validation of NLP accuracy

Evaluation of the extraction rules was based on the standard evaluation metrics of precision and recall. The performance of the extraction rule was evaluated against a validation set generated by manual curation from the test set by a research nurse. Among the 6,180 reports in the testing set, 1,094 reports were identified to contain a single patient result in the validation set. Among these reports, 1,088 reports (99.5%) for 1,039 patients matched with the set extracted by *Rule #1*. The precision of *Rule #1* (designed to extract reports with HLA typing for a single patient) was 0.997 and the recall was 0.994. There were three patients with data that were not extracted by *Rule #1* but that were present in the validation set. All of these cases included names written in English or text with unexpected spaces, and thus were missed by our extraction rules.

The clinical variables including patients' names and indication of HLA typing as well as HLA genotypes, were then extracted by *Rule #2*, and the results are summarized in **Table 3**. For the genotype extraction, we developed rules specific for each HLA gene because the resolution of typing and associated descriptions varied for each gene. The average precision of the determined HLA serotype/alleles was 0.976 and the recall was 0.952. The baseline HLA typing frequencies varied among the five HLA genes, with typing for HLA-C being mostly rare and typing for HLA-B being the most common, as observed in rule-development set. As the coverage and variability of rules should be dependent on the number of tests for each HLA serotype/alleles in the rule-development set, HLA-C typing showed the lowest precision (0.892 for serotype and 0.92 for allele) and the lowest recall (0.795 for serotype and 0.821 for allele). By contrast, the top accuracy

Table 3. Recall and precision for clinical variables and serotype/alleles of HLA genes

Rules	Variables	No. of reports ^a or elements	No. of extracted elements	No. of elements extracted correctly	Precision	Recall
<i>Rule #1</i>	Reports with a single patient record	1,094	1,088	-	-	-
	Patient IDs	1,042	1,039	1,036	0.997	0.994
<i>Rule #2</i>	Names of patients	1,042	1,039	1,036	0.997	0.994
	Indication of HLA testing ^b					
	Bone marrow transplantation	57	55	55	1.000	0.965
	Heart and lung transplantation	1	1	1	1.000	1.000
	Heart recipient	1	0	0	-	0
	Heart transplantation	33	33	33	1.000	1.000
	Kidney & heart transplantation	1	1	1	1.000	1.000
	Kidney transplantation	511	511	511	1.000	1.000
	Liver transplantation	84	83	83	1.000	0.988
	Lung transplantation	26	26	26	1.000	1.000
	PLT refractoriness	4	4	3	1.000	0.750
	NA	376	374	373	0.997	0.992
	Total	1,094	1,088	1,086	0.998	0.993
<i>Rule #3^c</i>	HLA-A					
	Serotype	713	708	705	0.996	0.989
	Allele	693	689	687	0.997	0.991
	HLA-B					
	Serotype	1,008	1,000	996	0.996	0.988
	Allele	989	982	979	0.997	0.989
	HLA-C					
	Serotype	73	65	58	0.892	0.795
	Allele	56	50	46	0.920	0.821
	HLA-DR					
	Serotype	719	716	715	0.999	0.994
	Allele	718	716	690	0.964	0.961
	HLA-DQ					
	Serotype	634	633	632	0.998	0.998
	Allele	634	633	632	0.997	0.997

^aTotal reports, n = 6,180; ^bNumber of reports, ^cNumber of patients.

was obtained for the HLA-DR and DQ genes, which were not the most frequently tested. This is likely related to the fact that these genes were commonly tested using high-resolution techniques, and thus the typing results were represented in a relatively uniform format such as “HLA-gene[Number][Space]*[Number]:[Number]”.

HLA nomenclature mapping

We converted the all of the extracted HLA data into the standard nomenclature format.¹⁴ If there was only serotype information available for a particular type of test with low resolution, only the serotype was saved without proceeding to detailed nomenclature conversion. We also developed rules to clean the genotype results by removing escape characters such as “\n” or empty spaces. We also added asterisk (*) or colon (:) to convert the genotypes according to the standard HLA nomenclature format. When there were additional letters or terms in the result, such as “g” or “group” after 4-digit of genotype to indicate a group of identical nucleotide sequences in a peptide-binding domain, we placed these additional symbols in a separate column. This was done for standardization as well as to avoid confusion with the HLA nomenclature in which the “G” code should follows the first three 3 fields (i.e., six digits) of the allele designation. Instances with other letters that may or may not interfere with the HLA nomenclature, such as “a” or “n”, were subjected to the same process. We used R¹⁸ for the development of Rule #3, and the final rules are listed in Table 3.

Fig. 3 shows an example of a raw HLA report and the corresponding result for the extraction into the database according to our rules so that the input text-based information and the test results were loaded into the HLA type table under the standardized nomenclature.

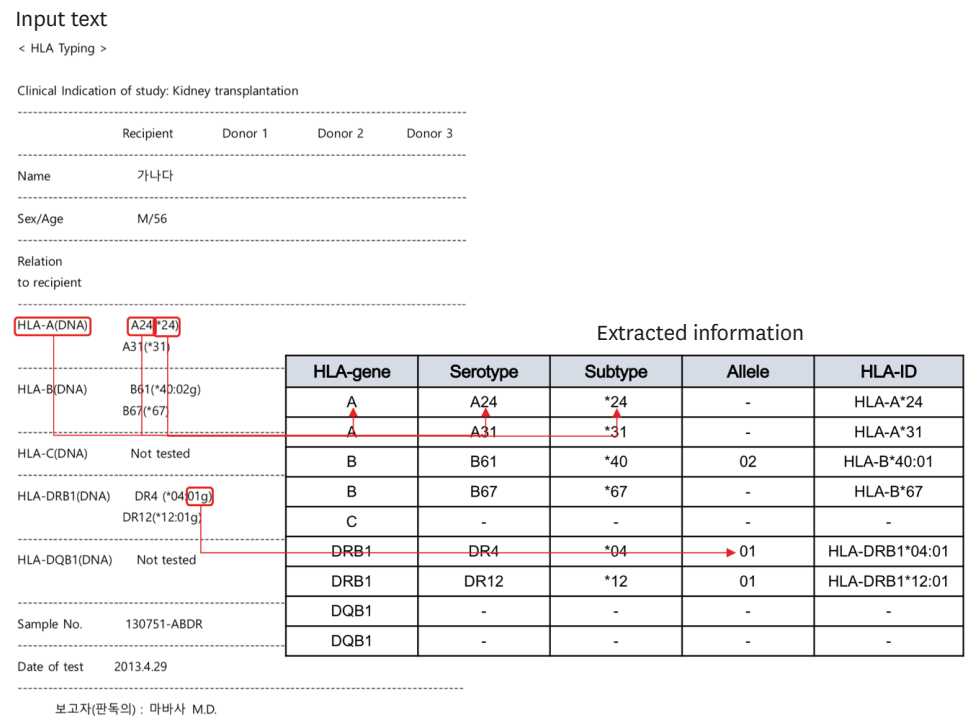


Fig. 3. Example of extracted HLA data from raw data. In this example, the serotype and allele information of the five HLA genes presented in free-text format in the HLA test result report were extracted and stored in the database in a standardized format. HLA = human leukocyte antigen.

DISCUSSION

We have provided a useful method to accurately extract HLA data recorded in a free-text format by applying rule-based NLP methods to construct a standardized HLA genotype database. Application of this method to a clinical database showed that serotype (or locus/protein) and allele frequency statistics—which had previously been stored in the form of unrecognizable free-text data—could be confirmed. Convenient access to such background information could help to support valuable clinical decisions for preventing severe adverse drug reactions related to specific HLA alleles. Therefore, we believe that this HLA database will contribute to the utilization of currently underused HLA genotype information to improve the design and development of CDSS and ensure patient safety.

We reviewed all of the HLA genotype mismatches with validation and automatic extraction. There was a patient data that was not extracted by the rules at all for all of the serotype and genotype of five HLA genes and the validation set of him had the correct result. For this case, the row name of the raw data, which represents HLA gene classes, were differently annotated such as “HLA-A antigen” compare to other HLA reports, “HLA-A.” Another case that was not extracted thoroughly was annotated as a donor and written down with unexpected multiple white spaces (indentation). The other cases were all not extracted because they had a complex representation of HLA serotype and genotype with unexpected symbols. For example, a raw report of a false-negative patient was described as the HLA-A serotype as ‘-’ instead of A2.

Based on this application, we demonstrated that rule-based extraction performed reasonably well for processing HLA data in free-text format. The database use in this study was established from various types of HLA tests with a range of resolution levels according to the testing technique and allelic variability in each patient. In current clinical practice, several types of genetic/genomic testing methods are widely adopted, including polymerase chain reaction, fluorescence in situ hybridization, and next-generation sequencing, which all show a wide range of resolution and data representation. One significant aspect of our developed method is that the results of the HLA tests with various level of resolution could all be transformed into a standardized nomenclature, which allowed for obtaining integrated results about the patient without regard to the specific test methods employed in the pilot test. Moreover, the validation results indicated that our approach yielded accurate results when extracting the data for single patients. However, to accurately extract information from the current free-text HLA data, it is essential that an experienced individual directly conducts a review of the results.

The majority of clinical records mostly comprise unstructured data, because these encompass clinical notes made by physicians as well as device-generated reports or imaging/pathology reports, which are largely input in a free-texted or semi-structured formats.¹⁹ Unstructured data are continuously produced in clinical settings because the test results are typically generated by different devices in different forms, and physicians typically add their final opinions or remarks in free-text format before loading the reports into the EMR system. This situation is further complicated when physicians use unique, self-made abbreviations. Furthermore, most physicians desire a summary view of the raw data to facilitate making a quick decision in busy clinical practices. Although semi-structured and free-text reports are useful for medical practitioners who are already familiar with such formats, their effective utilization in a CDSS and signal detection can be problematic.²⁰ Therefore, our proposed

approach provides a practical solution to convert the underused unstructured genotype data in an EMR to structured and thus reusable data.

The limitations of this study are that we did not perform a utility test of the rule-based named entity recognition system by applying the code to the HLA data generated from other hospitals or organizations. Because the result description pattern of HLA data might vary among institutions, and even between physicians, the performance of the patterns established in this study to extract HLA typing data should be further evaluated in other data sets. We have made this pattern openly available through GitHub and will continuously upbuild the codes. The second limitation is that we only used the regular expression as the NLP method to extract HLA genotype data without considering other trendy methods such as machine learning-based NLP. As shown in Fig. 1, the HLA data in our study did not form a complete sentence or phrase, which was not suitable for traditional text processing. We have adopted regular expressions which have shown that can easily integrate prior knowledge and show reliable performance in clinical text processing, especially in Korean and English mixed EMR data processing.^{21,22} For this reason, we used regular expressions, and the results of the study showed remarkable performance. Lastly, this study process and results is anticipated to proof the long-term reusability, but was not confirmed directly. Because the clinical data generated and stored in EMR is continuously updated and new data are always being added, the extract, transform, and load (ETL) systems should be used to confirm the data reusability and reliability to support the CDSS. However, we constructed the testing set using HLA reports that were recently ordered over one year at SNUH and the performance was found to be reasonable. Therefore, application to newly created data represents the next research challenge. We plan to carry out an advanced modeling study that considers linkages with clinical data and the efficiency of data storage, along with developing data utilization scenarios in the future.

Despite these limitations and remaining challenges, the present study represents the valuable attempt to use data modeling for storing and managing the ever-increasing amount of genomic data contained in EMRs. Although many more developments are expected to complement the process and system, this preliminary applications demonstrates that underused genotype data could be accurately extracted with an NLP method. In particular, for the secondary use of genomic data, it is very important to establish a map according to the resolution of the test and proper nomenclature. Wide adoption of this approach should facilitate access to more reliable data that can be reused for many purposes to the benefit of clinicians in decision-making, and ultimately to the patient.

ACKNOWLEDGMENTS

We are special thanks to JY Lee, a research nurse who helped a lot of HLA genotype curation and built a validation set.

REFERENCES

1. Mosaad YM. Clinical role of human leukocyte antigen in health and disease. *Scand J Immunol* 2015;82(4):283-306.
[PUBMED](#) | [CROSSREF](#)

2. Morishima Y, Sasazuki T, Inoko H, Juji T, Akaza T, Yamamoto K, et al. The clinical significance of human leukocyte antigen (HLA) allele compatibility in patients receiving a marrow transplant from serologically HLA-A, HLA-B, and HLA-DR matched unrelated donors. *Blood* 2002;99(11):4200-6.
[PUBMED](#) | [CROSSREF](#)
3. Hernandez-Boussard T, Kourdis PD, Seto T, Ferrari M, Blayney DW, Rubin D, et al. Mining electronic health records to extract patient-centered outcomes following prostate cancer treatment. *AMIA Annu Symp Proc* 2018;2017:876-82.
[PUBMED](#)
4. Juhn YJ, Kita H, Lee LA, Smith RW, Bagniewski SM, Weaver AL, et al. Childhood asthma and human leukocyte antigen type. *Tissue Antigens* 2007;69(1):38-46.
[PUBMED](#) | [CROSSREF](#)
5. Shiina T, Inoko H, Kulski JK. An update of the HLA genomic region, locus information and disease associations: 2004. *Tissue Antigens* 2004;64(6):631-49.
[PUBMED](#) | [CROSSREF](#)
6. Williams TM. Human leukocyte antigen gene polymorphism and the histocompatibility laboratory. *J Mol Diagn* 2001;3(3):98-104.
[PUBMED](#) | [CROSSREF](#)
7. Choo SY. The HLA system: genetics, immunology, clinical testing, and clinical implications. *Yonsei Med J* 2007;48(1):11-23.
[PUBMED](#) | [CROSSREF](#)
8. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet* 2012;13(6):395-405.
[PUBMED](#) | [CROSSREF](#)
9. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *J Biomed Inform* 2017;73:14-29.
[PUBMED](#) | [CROSSREF](#)
10. Angelino E. Extracting structure from human-readable semistructured text. <https://people.eecs.berkeley.edu/~elaine/pubs/angelino-structure.pdf>. Updated 2012. Accessed March 21, 2019.
11. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. *J Biomed Inform* 2018;77:34-49.
[PUBMED](#) | [CROSSREF](#)
12. Rosier A, Burgun A, Mabo P. Using regular expressions to extract information on pacemaker implantation procedures from clinical reports. *AMIA Annu Symp Proc* 2008;2008:81-5.
[PUBMED](#)
13. Aggarwal A, Garhwal S, Kumar A. HEDEA: a Python tool for extracting and analysing semi-structured information from medical records. *Healthc Inform Res* 2018;24(2):148-53.
[PUBMED](#) | [CROSSREF](#)
14. Denny JC. Chapter 13: Mining electronic health records in the genomics era. *PLOS Comput Biol* 2012;8(12):e1002823.
[PUBMED](#) | [CROSSREF](#)
15. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics* 2011;4(1):13.
[PUBMED](#) | [CROSSREF](#)
16. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SG. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res* 2015;43(D1):D423-31.
[PUBMED](#) | [CROSSREF](#)
17. Marsh SG, Albert ED, Bodmer WF, Bontrop RE, Dupont B, Erlich HA, et al. Nomenclature for factors of the HLA system, 2010. *Tissue Antigens* 2010;75(4):291-455.
[PUBMED](#) | [CROSSREF](#)
18. R Core Team. (2014). R: a language and environment for statistical computing. <http://www.R-project.org/>.
19. Sun W, Cai Z, Li Y, Liu F, Fang S, Wang G. Data processing and text mining technologies on electronic medical records: a review. *J Healthc Eng* 2018;2018:4302425.
[PUBMED](#) | [CROSSREF](#)
20. Scholte M, van Dulmen SA, Neeleman-Van der Steen CW, van der Wees PJ, Nijhuis-van der Sanden MW, Braspenning J. Data extraction from electronic health records (EHRs) for quality measurement of the physical therapy process: comparison between EHR data and survey data. *BMC Med Inform Decis Mak* 2016;16(1):141.
[PUBMED](#) | [CROSSREF](#)

21. Turchin A, Kolatkar NS, Grant RW, Makhni EC, Pendergrass ML, Einbinder JS. Using regular expressions to abstract blood pressure and treatment intensification information from the text of physician notes. *J Am Med Inform Assoc* 2006;13(6):691-5.
[PUBMED](#) | [CROSSREF](#)
22. Shin SY, Park YR, Shin Y, Choi HJ, Park J, Lyu Y, et al. A De-identification method for bilingual clinical texts of various note types. *J Korean Med Sci* 2015;30(1):7-15.
[PUBMED](#) | [CROSSREF](#)