ISSN 1598-0170 (Print)
ISSN 2287-1136 (Online)
http://www.jksii.or.kr

Construction of an PFT database with various clinical information using optical character recognition and regular expression technique☆

Man Young Park[1] Rae Woong Park[2*]

ABSTRACT

The pulmonary function test (PFT) is an essential data source for evaluating the effect of drugs on the lungs or the status of lung function. However, the numeric values of PFT cannot be easily used for clinical studies without labor-intensive manual efforts, because PFTs are usually recorded as image files. This study was aimed at constructing a de-identified, open-access PFT database with various clinical information. For constructing the PFT database, optical character recognition (OCR), regular expression, and the parsing technique were used to extract alphanumeric data from the PFT images in a Korean tertiary teaching hospital. This longitudinal observational database contains 413,000 measurements of PFT from 183,000 patients.

☞ keyword : Chronic Obstructive Pulmonary Disease, FVC, COPD, PFT, database, optical character recognition

1. Introduction

The mortality rate related to respiratory system diseases in Korea is about 39.8 per 0.1 million population, and pneumonia as well as chronic lower airway disease account for as high a proportion of deaths as they are ranked the 6th and 7th among all causes of death.[1] In the United States, chronic lower airway diseases, including Chronic Obstructive Pulmonary Disease (COPD) and asthma, were ranked the 3rd cause of death in 2001, with 42.2 deaths per 0.1 million people[2], and 5.1% of adults have been reported to have COPD.[3] In particular, one of the chronic diseases in adults, asthma, requires long-term medication and treatment.[4,5] Although the study methods that have been conventionally used for COPD, lung cancer, and asthma include case study, clinical trials, and meta-analysis, these previous study methods have limitations such as limited study subjects and a small sample size, and thus, it is challenging to generalize the study results to the entire population. In addition, since a large cohort study takes a substantial amount of time and cost in exchange for solving these problems, it is practically impossible for individual researchers to conduct such studies.

In order to solve these problems, there has been an increase in studies using a nonclinical database, which is open to the public across medical fields, including lung diseases, or using an electronic medical record of the medical facility that the research belongs to. Currently available databases pertaining to lung disease patients are specified only for a certain age group or a certain disease, and thus, there are many limitations in their usefulness. Although the Society of Thoracic Surgeons National Database, a database developed by cardiac surgeons in the U.S., has been constantly collecting and sharing information about basic data, treatment, and tests of patients in thoracic surgery since 1989, this data cannot be compared with other patient groups as the subjects are limited to patients with thoracic surgery.

Since Electronic Medical Record (EMR) data of university level medical facilities include information pertaining to all the medications of patients required for research on lung

1 Mibyeong Research Center, Korea Institute of Oriental Medicine, Daejeon, South Korea
2 Department of Biomedical Informatics, Ajou University School of Medicine, Wonchon-dong, Yeongtong-gu, Suwon, Gyeonggi-do, 442-749, Korea.
* Corresponding author (veritas@ajou.ac.kr)
[Received 2 June 2017, Reviewed 12 June 2017(R2 25 July 2017, R3 17 August 2017), Accepted 23 August 2017]
☆ This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. NRF-2014R1A1A1007557).
☆ A preliminary version of this paper was presented at ICONI 2016 and was selected as an outstanding paper.

diseases, the results of the pulmonary function test (PFT), and clinical information of patients that have been stored for many years, all information required for lung disease research can be sufficiently provided both qualitatively and quantitatively. If a large retrospective cohort study about EMR is utilized, one can perform highly qualified studies on various subjects including prognosis according to the patient's characteristics, disease conditions, effects according to treatments, and changes and improvement in lung functions according to medications. However, as PFT results prior to the introduction of an EMR system were stored as printed papers or scanned images, numeric data cannot be automatically obtained. Moreover, due to the Personal Information Protection Act and ethical issues, it is impossible to access to clinical information of patients except for researchers of a corresponding institute.

Therefore, this study is designed to solve the ethical issues as well as the Personal Information Protection Act issues by removing sensitive information and identifiable data, and establishing a PFT longitudinal observational database containing qualified PFT results, prescribed medications, and patients' basic information.

2. Method

2.1 PFT data source

Since the introduction of the EMR system, PFT data scattered in hospitals are classified into unstandardized data, which are printed papers, and standardized data, which are numeric data stored in the EMR. PFT results have been stored after they are automatically converted to numeric data through an interface between EMR systems, and specific tests which are printed papers are scanned by the medical record team and stored in the file server of the EMR.

2.1.1 Data extraction method for each source Acquisition of a scanned PFT image from a paper

Since PFT data, which used to be printed and filed as papers, has been scanned by the medical record team along with other specific tests and stored in the file system of the EMR since the introduction of EMR system in the corresponding hospital, it was not possible to select PFT image files only. Therefore, the task to extract and classify scanned and stored PFT and all other relevant images (about 900,000 cases) has been performed since the introduction of the EMR. Furthermore, as PFT is usually performed in the department of allergy and clinical Immunology or Pulmonology, test results included different images from the tests conducted in these two departments as well as multiple tests conducted in each of these departments. Thus, it was necessary to classify them by image patterns for optical character recognition (OCR) and parsing.

2.1.2 PFT image extraction

It was almost impossible to manually separate PFTs from about 900,000 image files. In addition, PFT results are large image files. programs providing user interface, such as Windows Explorer or OCR could not be used as the server was down. Therefore, 5,000 image files per folder were stored in 18 folders using JAVA or R programs, and OCR was performed for each folder.

2.1.3 Extraction of required items related to the PFT

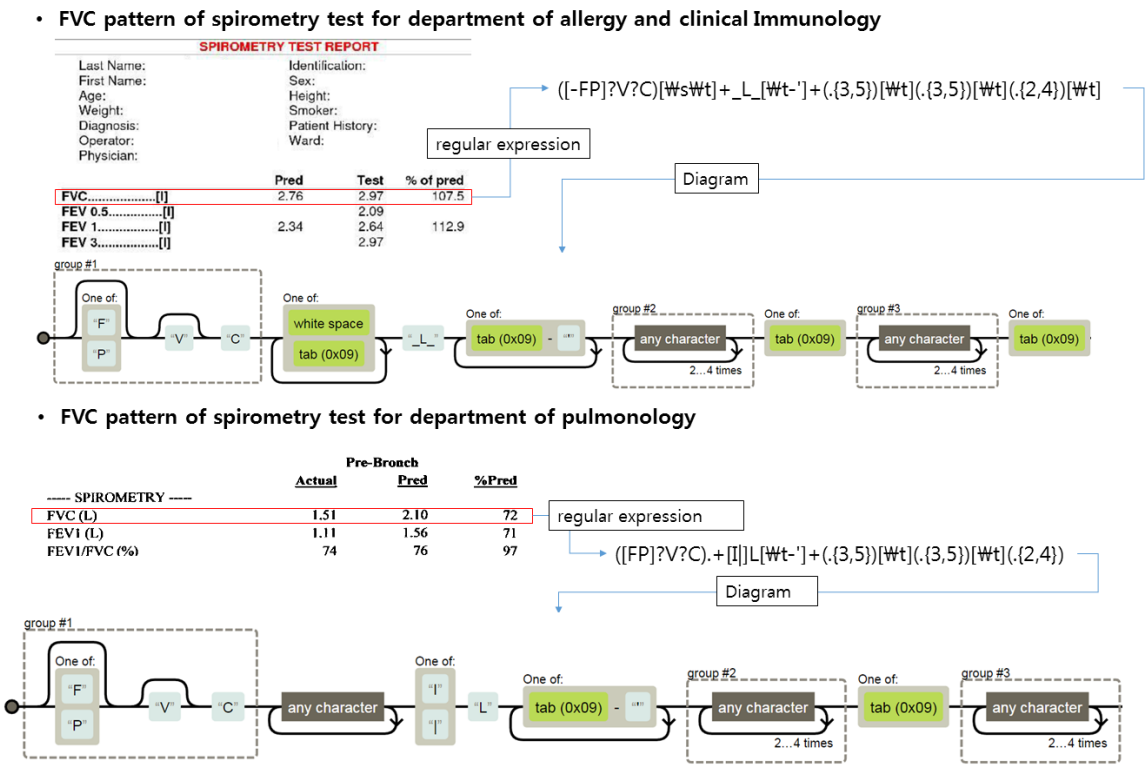
pulmonology involves pre-post data such as results of the Bronchial provocation test and bronchodilator test, or the simple spirometry test. Additionally, the location and items of being extracted data varied according to the department that performed the test. Therefore, parsing was performed by structuring database schema for each images and using regular expressions according to the pattern (Figure 1).

2.14 Verification of the extracted PFT test data

In order to verify if the PFT data extracted through OCR were accurate, a comparison test was performed with the extracted values of FVC, FEV1, and FEV1/FVC and the calculated values (Figure 2).

2.1.5 Data stored in the EMR

Since the introduction of the EMR, the corresponding hospital stored the PFT results in the EMR through an



(Figure 1) regular expression for extracting each PFT test results

interface. The numeric data in the EMR were extracted through a simple Extraction, Transformation, Loading (ETL) procedure.

2.2 De-Identification

all unique identifiers were eliminated. Ages over 80 years were set to 80 years. Highly stigmatized diagnoses such as sexually transmitted disease, abortion, and chromosomal abnormalities were removed from the database. All medication and diagnosis data related to AIDS/HIV infections were also eliminated. A random numeric number from −60 to 60 was generated for each patient, and the number was added to all date data, including birthday for each patient not to be identifiable by that information. the intervals between the dates for each individual patient would be conserved.

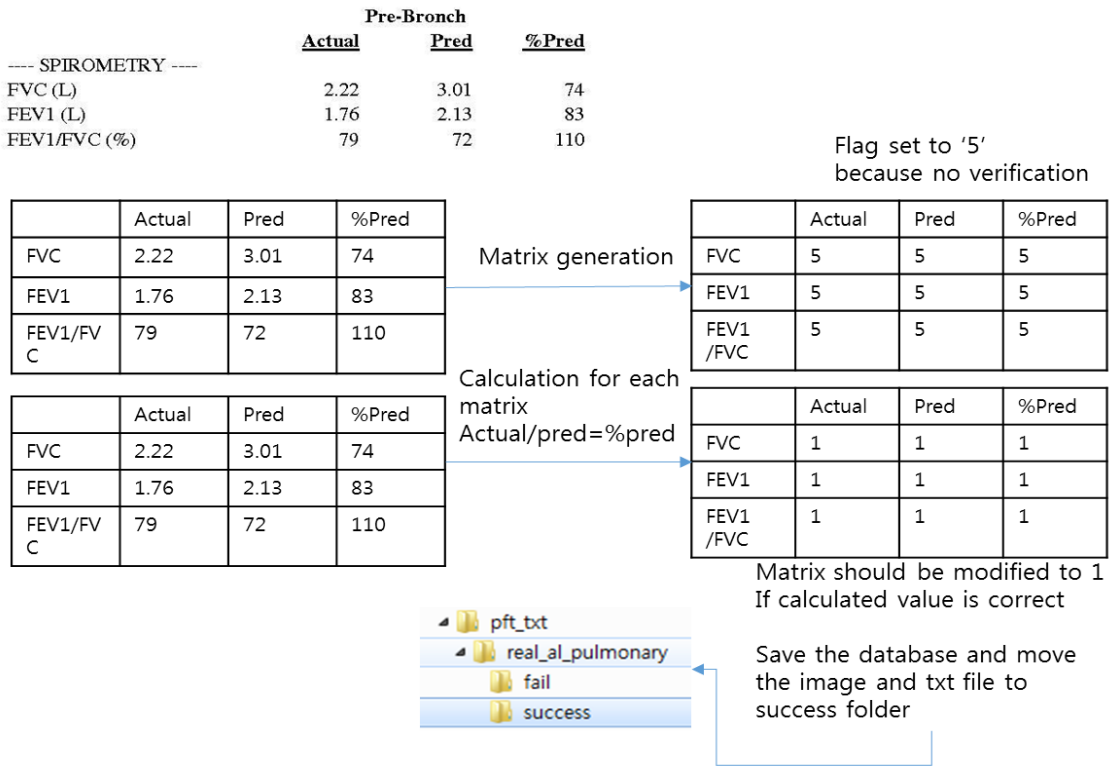
2.3 Study for proof of concept

To prove the usefulness of the database as a surveillance database for detecting the drugs having effect on pulmonary function improvement, we conducted a case-control study against Pethidine, which is most frequently prescribing in the hospital. The cases group was defined as a group with improved pulmonary function. The control group was defined as a group whose lung function was not improved. logistic regression was used for the analysis.

3. Results

3.1 PFT database

The databases consists of seven tables such as PFT result, drug prescription, drug code master, person, diagnosis, diagnosis code master (Figure 3). The PFT sources in the corresponding facilities included health examination,



	Actual	Pred	%Pred
FVC	2.22	3.01	74
FEV1	1.76	2.13	83
FEV1/FVC	79	72	110

	Actual	Pred	%Pred
FVC	5	5	5
FEV1	5	5	5
FEV1/FVC	5	5	5

	Actual	Pred	%Pred
FVC	2.22	3.01	74
FEV1	1.76	2.13	83
FEV1/FVC	79	72	110

	Actual	Pred	%Pred
FVC	1	1	1
FEV1	1	1	1
FEV1/FVC	1	1	1

(Figure 2) Verification of extracted values

industrial medicine, data stored in the EMR, and scanned data. Using these sources, a comprehensive PFT database was established, during which 420,000 PFT results were extracted from about 180,000 patients. In this process, the accuracy of the OCR was about 85.2%, and misunderstood or unrecognized PFT images were manually checked by the researchers and recorded in the database. In reference to Anatomical Therapeutic Chemical (ATC) Level 5, there were 550 types of medications prescribed to patients , and there were about 9 million prescriptions.

3.2 Study for proof of concept

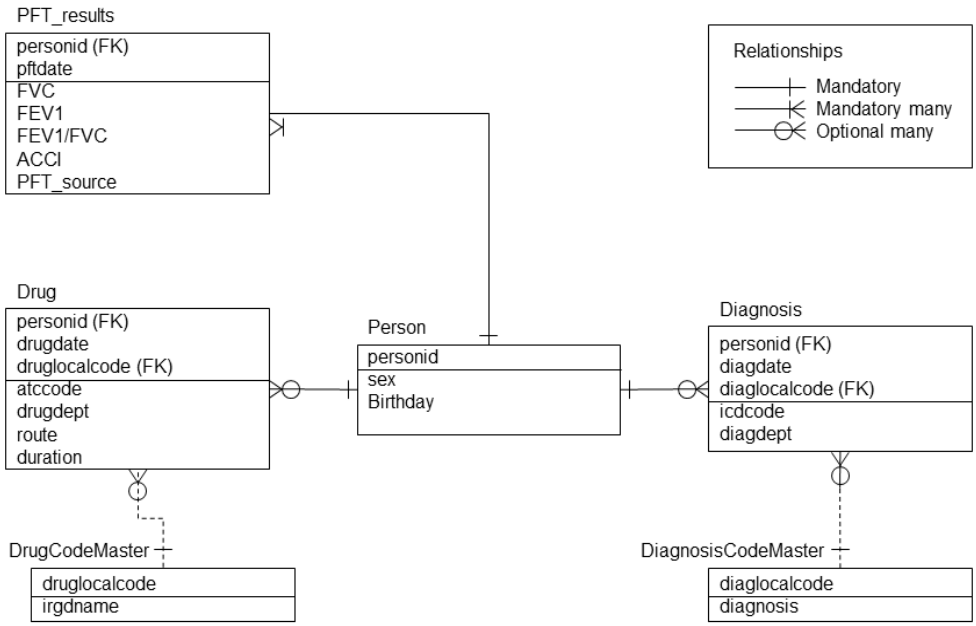
The case group with pulmonary function improved and the control group with no pulmonary function improved were 2,301 and 4,620, respectively (Table 1).

(Table 1) Demographic and clinical characteristics

	Case (N=2,310)	Control (N=4,620)	*p*
SEX			0.604
Female	1,167 (50.5%)	2,366 (51.2%)	
Male	1,143 (49.5%)	2,254 (48.8%)	
AGE	46.5 ± 15.6	46.6 ± 15.4	0.692
Age-adjusted CCI	7.8 ± 2.5	7.8 ± 2.4	0.903
Source			0.871
EMR	80 (3.5%)	150 (3.2%)	
Health Examination	284 (12.3%)	578 (12.5%)	
Scanned PFT	1,946 (84.2%)	3,892 (84.2%)	
DrugYN			<0.001
PethidineYes	2,265 (98.1%)	4,575 (99.0%)	
PethidineNo	45 (1.9%)	45 (1.0%)	

CCI=Charlson comorbidity index

Study has shown that the group having pethidine were improved lung function by about 2.05 times compared to the



(Figure 3) ERD of PFT database

group without pethidine (Table 2).

(Table 2) Adjusted Odds Ratio and 95% Confidence interval for Pethidine Use and effect on pulmonary function improvement

Variable	OR	95% CI	*p*
(Intercept)	0.56	0.4 - 0.77	<0.001
drugYN			
no Pethidine	1	(ref.)	
Pethidine use	2.05	1.35 - 3.12	0.001
Sex2			
Female	1	(ref)	
Male	1.02	0.92 - 1.13	0.752
Age	1	0.99 - 1	0.488
Health Examination	0.87	0.62 - 1.24	0.446
Scan	0.89	0.62 - 1.3	0.548
Age-adjusted CCI	1.01	0.97 - 1.04	0.665

CCI=Charlson comorbidity index

4. Conclusion

A PFT database was created by combining scattered PFT data from a hospital using an automated method. also, we conducted a case-control study against Pethidine, which is most frequently prescribing in the hospital. That drug have an effect on pulmonary function improvement. A drugs were analyzed and the time taken was less than 5 minutes.

The method suggested in the present study can be utilized by anyone to readily establish a database. Additionally, if established databases are consolidated through appropriate de-identification, not only would research on lung function flourish without concerns regarding revealing personal information of patients, but it would also be useful for the screening of drugs for lung function improvement. In conclusion, the findings of the present study are expected to contribute to patients'safety. For anyone who wants to make a database like this, the URL of the source code for OCR, parsing is

'https://figshare.com/articles/_____DB___/5053144'

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. NRF-2014R1A1A1007557).

Reference

[1] The National Statistical Office. 2011 statistic for the cause of death, 2012. http://kostat.go.kr

[2] Kochanek, Kenneth D., et al. "National vital statistics reports." National Vital Statistics Reports, Vol. 59, no. 4, 2011. https://www.cdc.gov/nchs/data/nvsr/nvsr59/nvsr59_04.pdf

[3] Akinbami, Lara J., and Xiang Liu. "Chronic obstructive pulmonary disease among adults aged 18 and over in the United States, 1998–2009." 2011. https://www.cdc.gov/nchs/data/databriefs/db63.pdf

[4] Agertoft, L., and S. Pedersen. "Effects of long-term treatment with an inhaled corticosteroid on growth and pulmonary function in asthmatic children." Respiratory medicine Vol. 88, no. 5, pp. 373-381, 1994.http://www.sciencedirect.com/science/article/pii/0954611194900442

[5] Wallace, Lance A., et al. "Particle concentrations in inner-city homes of children with asthma: the effect of smoking, cooking, and outdoor pollution." Environmental health perspectives, Vol. 111, no. 9, pp. 1265, 2003. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1241585/

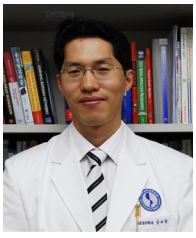
[6] Nathan N, Taam RA, Epaud R, Delacourt C, Deschildre A, Reix P et al. A national internet-linked based database for pediatric interstitial lung diseases: the French network. Orphanet Journal of Rare Diseases, Vol. 7, no. 1, pp. 40, 2012. http://dx.doi.org/10.1186/1750-1172-7-40

[7] St. Antonius Hospital. Database of Interstitial Lung Diseases. Retrieved Jan 18, 2013 from http://clinicaltrials.gov/ct2/show/NCT00267800

[8] The Society of Thoracic Surgeons (STS) National Database. http://www.sts.org/national-database

◐ 저 자 소 개 ◑

박 만 영(Man Young Park)
2005년 한신 대학교 컴퓨터학과(이학사)
2012년 아주대학교 의과 대학원 의료정보학과(이학박사)
2014년 아주대학교 의료정보학과 박사후 연수
2015년 ~현재 한국한의학 연구원 선임연구원
관심분야 : 데이터베이스, 데이터마이닝, 약물역학, 의료정보 etc.
E-mail : pmy10042@kiom.re.kr



박 래 웅(Rae Woong Park)
1995년 아주대병원 수련의, 전공의
2005년 아주대학교 의과대학 조교수
2007년~현재 아주대의료원 유헬스정보연구소 소장
2014년~현재 아주대학교 의과대학 교수
관심분야 : 데이터베이스, 데이터마이닝, 약물역학, 의료정보 etc.
E-mail : veritas@ajou.ac.kr