Development of data mining system based on survey on cohort study in the hospital representing the impact which environmental factors give the people
- Mid-and-long term monitoring the relationship between the environmental exposure and the disease structure -
(Abstract of the Final Report)

Contact person Takayuki Hoshino
Information research professional personnel,
Information Systems Division
National Center for Global Health and Medicine
Toyama 1-21-1,Shinjyuku-ku,Tokyo, Japan
Tel:+81-3-5273-5297 Fax:+81-3-3202-4853
E-mail:t-hoshino@it.ncgm.go.jp

Total Budget for FY2012-FY2016 93,266,000Yen
(**FY2015**; 14,265,000Yen)

Key Words EHR, GNU Health, Environmental factor, Air pollutant, Meteorological data, Electronic health record, Data base, Data mining, Preventive medicine,

1. Introduction

This study was carried out under the recent trends of global climate change (e.g., tendency towards global warming), increasing social awareness of air pollutants (particulate matter [PM] 2.5, etc.) and increasing attention to the health impacts of these environmental factors as compared to that in the previous century, in which diseases caused by environmental pollution was a major topic of research. On the basis of the expanded understanding on environmental pollution, individual countries have prepared their own environmental standards. Through the joint efforts of environment-related governmental organs and the private sector to preserve the environment in accordance with appropriate environmental standards, the prevalence of severe diseases caused by environmental pollution has decreased in many developed countries. However, as environmental factors change over time, the features of illnesses affecting people ("disease structure") also change. Monitoring of environmental factors has been actively carried out in individual countries to establish the base for evaluation of the health impacts of environmental changes (e.g., warming) on a global scale, and early stages of research concerning environmental factors and disease structure have started, covering those illnesses on which the impact of environmental factors had not been pointed out, or even if pointed out empirically, evidence for such influence is not yet available. In Japan also, environmental big data have been established and published, and precise monitoring is now under way concerning detailed meteorological data and new environmental factors, such as PM2.5. It has been revealed gradually that new environmental factors affect the structures of illnesses seen in the nation. So that we can analyze how changes in environmental factors (much smaller factors than those covered by past studies) affect our disease structure significantly and also that we may obtain new findings about such impacts, it is indispensable to establish a large database composed of more detailed clinical data collected from a larger population scale than before.

2. Research Objective

In an attempt to combine information science with medicine, this study was designed to conduct analysis of environmental data linked to healthcare information and to identify candidate diseases closely associated with environmental factors, bearing in mind future conduct of data analysis in cooperation with multiple institutions. Through these efforts, the study was aimed at establishing a "common system" enabling integration and analysis of the hospital information system data and the environmental data possessed by the Ministry of Environment (e.g., the data of the Atmospheric Environmental Regional Observation System of the Ministry of Environment) as an infrastructure that would enable continuous monitoring of changes in environmental factors and disease structure.

3. Results

Because the current year is the last year of this study, this report focuses on the following 3 points:

I. Overview of the "common system" created through this study

II. Introduction of concrete examples of advanced analyses conducted by this study group

III. Appeal to institutions where academic publication/introduction is possible (the appeal aimed at facilitating free-of-charge delivery of the "common system" for use in the multi-institution data analyses planned in the future).

I. "Common system" created through this study

"Common system" is a collective term for the system employed for analysis of hospital information system data in combination with environmental big data using the algorithm described below.

(1) Creation of a **"clinical database"** from the hospital information system's big data

Outcome of "clinical database" accumulation

National Center for Global Health and Medicine

2011-2016 Datasets processed for automated analysis

Anonymous patients	572384 cases
Anonymous prescriptions	1290253 cases
Anonymous disease names	239112 cases
Anonymous reservations	608147 cases

Gohyakuyama Clinic

2014

Data on patients with non-allergic rhinitis (NAR)

Markedly improved	130 cases
Improved	88 cases
Unchanged/Discontinued	45 cases
Worsened	1 case

(2) Creation of an "**environmental database**" from environmental big data

Outcome of the "environmental database" accumulation

- Areas for data extraction 1749 area
- Data in units of time 15372000 data (92064000 data, cumulative of 2011-2016 period)
- Data in units of day 640134 data (3833808 data, cumulative of 2011-2016 period)

(3) "Database composed of clinical data combined with environmental data"

(4) Screening/analysis for extraction of candidates of previously unknown new findings

Making use of the "Database composed of clinical data combined with

environmental data," which has been functional after creation based on the results of studies in and before 2015, a system of screening and analysis for extraction of candidates of previously unknown new findings has been developed and put into automated operation for the purpose of allowing extraction of candidate diseases whose structure will change in relation to some particular environmental factors in assumed cases where the candidate diseases to be analyzed are not yet decided (outcome of the study in the current year).

(A) "Quartile odds analysis system (automated operation started in fiscal year [FY] 2016)

(B) Process for selection of candidates from initial candidates for clinico-epidemiological analysis (integrated as software in FY 2016)

(5) Advanced research method making use of database for cohort studies

(A) Process of analyzing the impact of multiple environmental factors related to specific disease

(B) "System for hospital cohort studies of electronic medical records with a mixed model using odds ratio" (automated operation started in FY 2016)

(6) From "**common system**" (a collective term for the series of analytical systems mentioned in (1) through (5)) to "**environmental monitoring system**"

If the common system is applied continuously over the mid- to long-term and is introduced in multiple institutions for cooperative operation, it can be utilized as an **"environmental monitoring system"** applicable to diverse purposes related to analysis of disease structure.

II. Concrete examples of application and analysis by this study group

The "common system" completed during the current year enables extraction of abundant case data from electronic medical records (not possible from conventional paper medical records) and then allows analysis of the data extracted thus. In the present study, the environmental factor data were combined with the electronic medical record big data, including information on diseases, and the combined data were analyzed. This attempt, including the analytical procedure, was reported in a professional journal. Thus, we carried out feedback of the epidemiological data collected by this methodology to the society. Examples of advanced analyses under way to obtain new findings during the current year are given below.

(1) Extraction, using the "common system," of candidate diseases to be analyzed

(2) Analysis of association between summer day PM2.5 exposure and onset of aspiration bronchitis/pneumonia in asthmatic patients

(3) Analysis of PM2.5 exposure under cold environments and the PM2.5 level dependency of the quartile odds ratio for onset of an ischemic attack on day two of exposure in elderly patients with diabetes

(4) Association between peak temperature during a warm period and onset of subarachnoid hemorrhage in patients of all ages without a history of diabetes

III. Appeal to institutions where academic publication/introduction is possible (appeal aimed at facilitating free-of-charge delivery of the "common system" for use in data analysis planned in the future with cooperation of multiple institutions)

As described in the study protocol for the current year, the "common system" developed through this study allows combining the hospital information system data with the

environmental data possessed by the Ministry of Environment (e.g., the data of the Atmospheric Environmental Regional Observation System of the Ministry of Environment) and analyzing the data combined thus, and will serve from now on as a basic system that would enable continuous monitoring of the impacts of environmental factors on disease structure. During the current year, this system was published, and a set of manuals (Japanese version) to be delivered upon request were prepared, and environments facilitating data analysis in cooperation with multiple institutions were arranged. Specifically, the following two methods were used to publish the common system.

(A) Presentation at study meetings

(B) Introduction of the current status of the open source software in the field of healthcare and a data mining system for combining healthcare information with outside data for comparison, using hospital cohorts

In order to further promote the "common system" to the group of hospitals under the National Hospital Organization serving as regional healthcare centers and possessing similar electronic medical record systems, and to publish the system academically, an original paper presenting the system outline, announcing our software free-of-charge delivery program, and inviting institutions interested in introduction of the system will be submitted to the journal "Iryo" (a peer-reviewed professional journal of the National Hospital Organization).

4. Discussion

If the common system is applied continuously over the mid- or long-term and is introduced to multiple institutions for cooperative operation, it can be utilized as an "**environmental monitoring system**" applicable to diverse purposes related to analysis of disease structure. It would additionally allow easy extraction of candidates of epidemiologically useful evidence which can trigger interventional studies and will remarkably improve environmental monitoring and its application, possibly leading to more speedy innovation. In practice, our study repeated pre-study for extraction of candidates for analysis during the course of development using a spiral model (cycles of trial application and reform) towards the goal of developing the above-mentioned "system for hospital cohort studies of electronic medical records with a mixed model using odds ratio (automated operation started in FY 2016)," yielding the numerous results of analysis described in Part II. If these analytical results are reviewed in a multidisciplinary manner by experts in biometeorology, clinical epidemiology and statistics, it will be possible to extract candidates of epidemiologically useful evidence (which can trigger interventional studies) efficiently, and such candidates may be set as objects for mid- or long-term monitoring, as needed, probably serving as a base contributing to innovation in the future, such as mobile terminal-mediated real-time weather forecasts from the viewpoint of health, etc.