

Guidelines for the Conduct of Pharmacoepidemiological Studies in
Drug Safety Assessment with Medical Information Databases

Version 1
March 31, 2014

Pharmaceuticals and Medical Devices Agency

Contents

1. Introduction	7
(1) Purpose and object of Guideline	7
(2) Applicability of guideline	7
(3) Background	7
(4) Revision of this guideline	10
2. Preparation of the Study Protocol	11
3. Data sources	16
(1) Characteristics of major data source used in pharmacoepidemiological study and point to consider at the selection	16
① Claims data	16
② HIS/EMR data	17
③ Registry	18
④ Overseas databases	18
(2) Standard coding system	20
① Characteristics of disease name codes	20
② Characteristics of drug codes	20
③ Characteristics of laboratory test item codes	21
④ Points to consider about codes	21
(3) Validation	21
4. Study design and conduct of the study	22
(1) Study design	22
① Cross-sectional study	23
② Active surveillance	23
③ Case series	23
④ Self-controlled case series, case crossover	24
⑤ Cohort study	25
⑥ Case-control study, nested case-control study	25
(2) Points to consider about study designs	26
① Definition of study population	26
② Definition of exposure	26
③ Definition of outcome	27
④ Bias	28
⑤ Adjustment of confounders	29

⑥	Effect modifiers	31
(3)	<i>Data management</i>	31
(4)	<i>Analysis</i>	31
(5)	<i>Quality assurance</i>	31
(6)	<i>Protection of personal information, and ethics</i>	31
5.	Preparation of study report.....	32
6.	Publication of study results	35
7.	References.....	35

Definition of terminology

Term	Definition
HIS/EMR data	This refers to Hospital Information System (HIS)/Electronic Medical Record (EMR) data. HIS is a system for processing information such as orders for medical care at medical institutions. This includes disease names, orders for various tests and drugs, and information about the test results. EMR is a computer-based patient record which includes diagnosis, and the details and course of medical treatment for each patient, etc. Both terms together refer to data on medical treatments in a medical institution.
MIHARI project	A project started in 2009 by the Office of Safety 1 of Pharmaceuticals and Medical Devices Agency (PMDA). It aimed to utilize medical information data by applying pharmacoepidemiological methods for evaluating post-marketing drug safety, with the goal of strengthening post-marketing drug safety measures.
Outcome	All possible consequences resulting from exposure to a cause, preventive or therapeutic intervention. When dealing with a health problem, it refers to all changes occurring as results and changes recognized in the health condition of a patient ⁽¹⁾ .
Medical information data	This refers to data containing medical information such as claims data, HIS/EMR data, and registry, but does not include data on spontaneous reports of adverse drug reactions.
Case	A particular disease or health condition in a study population, or individuals having the disease or condition. In the field of pharmacoepidemiological study, a single patient can be counted as a separate case for each type of disease or injury.
Signal	It is defined by the World Health Organization as “Reported information on a possible causal relationship between an adverse event and a drug, the relationship being unknown or incompletely documented previously.” ⁽²⁾ For clarity, it is also referred to as “safety signal.”
Signal confirmation	Although there has been no international consensus on its definition, this guideline regards it as a quantitative and highly accurate

	estimation of the relationship between a particular adverse event and a drug. This is almost synonymous to hypothesis verification in the pharmacoepidemiological study.
New user	An individual who is newly administered a certain drug. Synonymous includes the term ‘incident user.’
Primary use of data	Use of data collected for the purpose of implementing in that study.
Secondary use of data	Use of data collected for purposes other than implementing in that study.
National health insurance claims database	A database built by the Ministry of Health, Labour, and Welfare (MHLW), by collecting data related to health insurance claims and the specified health checkups and implementation status of the specific health guidance, in accordance with the Act on Assurance of Medical Care for Elderly People. After a review and trial period by an expert meeting set up by the MHLW, since 2013, advancements have been made so that non-governmental bodies can use it for a study that contributes to public interest.
Validation study	An investigation conducted as part of the secondary use of data, where the validity of information contained in the data is checked against another highly reliable source of information. In particular, investigations often concern disease code names or algorithms that extract specific cases from the data.
Pharmacoepidemiological study	Pharmacoepidemiology is defined as “the study of the use and effects of drugs in a population” and includes cohort studies and case-control studies, as well as descriptive studies and active surveillance, etc. It also includes investigations as part of its safety measures conducted by the PMDA, safety-related survey or study by pharmaceutical companies, and studies by researchers. Pharmacoepidemiological studies involve studies that makes primary and secondary use of data.
Clinical event	An event that occurs in clinical settings such as diseases, signs, and symptoms.
E-claim	This is an abbreviation of ‘Electronic health insurance claims processing system.’ Here, an authorized insurance medical institution or pharmacy submits an electronic claim online or by using an

	electronic medium to an examination/payment agency, which examines and settles the bill that is ultimately received by the insurer (3).
--	--

Abbreviations

Abbreviation	Formal name
ATC	Anatomical Therapeutic Chemical Classification System
CCI	Charlson Comorbidity Index
DPC	Diagnosis Procedure Combination
DRS	Disease Risk Score
ENCePP	European Network of Centers for Pharmacoepidemiology and Pharmacovigilance
FDA	U.S. Food and Drug Administration
ICD	International Statistical Classification of Diseases and Related Health Problems
ICH	International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for human use
ISPE	International Society for Pharmacoepidemiology
MedDRA	ICH Pharmaceutical Term List Medical Dictionary for Regulatory Activities Terminology
PMDA	Pharmaceuticals and Medical Devices Agency
PS	Propensity Score
WHO	World Health Organization

1. Introduction

(1) Purpose and object of Guideline

This guideline describes points to consider for the Pharmaceutical and Medical Devices Agency (PMDA), pharmaceutical companies and other relevant organizations to conduct appropriate pharmacoepidemiological studies with secondary use of the medical information database to evaluate the safety of drugs. The guideline is mainly aimed at studies using medical information databases, but some of the guidance is also relevant to pharmacoepidemiological studies that carry out primary data collection rather than secondary use, and can therefore also partially serve as a reference for such studies.

(2) Applicability of guideline

This guideline mainly applies to studies conducted by the PMDA or pharmaceutical companies. As it compiles and summarizes points to consider in conducting pharmacoepidemiological studies from a primarily academic perspective, the guideline may also be useful for researchers affiliated to academic institutions when conducting pharmacoepidemiological studies.

(3) Background

Status of safety evaluation methods for drugs

Until now, spontaneous adverse drug reaction (ADR) reports from medical institutions and pharmaceutical companies, as well as post-marketing studies by pharmaceutical companies have been the main information source for drug safety measures in Japan. A spontaneous ADR report is particularly useful for detecting rare or severe ADRs, and it is an important information source for evaluating the safety of drugs. However, there is a reporting bias arising from the fact that only some of the events recognized as ADRs in the clinical practice are reported, and the proportion of the reporting fluctuates due to various factors. In addition, since the frequency of incidence cannot be determined and it may not be able to quantitatively and appropriately evaluate the risks by setting a comparison group, it may be difficult to make decisions on safety measures. For example, if a single event of a new, previously unknown ADR is reported, it may be difficult to estimate its impact on public health unless the actual frequency of incidence is known. Similarly, if a clinical event with a high background incidence in a certain disease field occurs after use of a drug, or if the effect of the drug on the clinical event is small, it may be difficult to evaluate from spontaneous ADR reports whether the clinical event is due to the underlying disease or it is an ADR of the drug. Moreover, a large number of concomitant drugs make it difficult to properly evaluate the causal relationship between a particular drug reported as a suspected product and an ADR. Normally, post-marketing

studies are conducted by collecting information only about patients exposed to the drug of interest, without a comparison group. It is therefore impossible to quantitatively evaluate relative risk, whereas the frequency of ADRs can be estimated.

In recent years, the context of drug use has been changing due to the increase of patients having multiple comorbidities as the number of elderly patients has increased, and the diversification of prescription drugs owing to the development of new pharmaceutical products. In order to adapt to such changes, there has been increasing demand for a wider variety of information sources and new methods of evaluating drug safety, in addition to spontaneous ADR reports and post-marketing studies.

Introduction of pharmacoepidemiological approaches to safety evaluation

From the examples presented above that illustrate the difficulties in evaluating the safety of drugs based solely on spontaneous ADR reports, pharmacoepidemiological approaches targeting populations may be more useful over the evaluation of individual cases of spontaneous ADR reports. Pharmacoepidemiology is the study of the use and effects of drugs in a population. In addition to the descriptive study focusing on drug usage, pharmacoepidemiology includes an analytical study that quantitatively evaluates the relationship between a drug and an adverse event by statistical analysis, and is closely related to the safety evaluation of drugs. In conventional pharmacoepidemiological studies, the collection of information on the study population involves creating a case report form for each study, requesting the relevant parties to fill them in, and digitizing the obtained information by hand. With primarily manual work being involved, a lot of human and money resources are required, and the cost increased with longer study periods, thus the scope and scale of pharmacoepidemiological studies that could be performed were heavily constrained. However, as medical information databases have gradually become available over the last few years in Japan, it has become possible to conduct pharmacoepidemiological studies that make secondary use of medical information data, in addition to the traditional method of carrying out primary data collection for their studies.

Against this background, governmental bodies have begun using pharmacoepidemiological approaches that utilize medical information databases for safety evaluation of post-marketing drugs. With the proposal on “Review on the Pharmaceutical Administration to Prevent Recurrence of Yakugai (Drug-induced suffering) (final proposal)” (April 28, 2010), by the “Committee for Investigation of Drug-induced Hepatitis Cases and Appropriate Regulatory Administration to Prevent of Yakugai Similar Sufferings,” established by the Ministry of Health, Labour, and Welfare (MHLW), it has been recognized that development of the database of medical information data, establishment of methods for collecting and evaluating information regarding to ADRs are an important issue for post-marketing drug safety measures. It states that “consolidating information infrastructures for understanding the

number of drug users regarding ADRs, calculating the frequency including both of the number of drug exposure and disease incidence information (e.g. ADRs), and evaluating the effect of safety-related regulatory actions, by utilizing the database of medical information data, should proceed.” Furthermore, the proposal titled “Proposal related to safety and reliability of drugs using electronic medical information databases (Japanese sentinel project)” (August 25, 2010) compiled by the “Round-table conference on how to use medical databases for pharmacovigilance” established by MHLW, emphasized the necessity of developing medical information databases to contribute to pharmacovigilance, expectations towards the scale of those databases and the collaboration with the national database of health insurance claims (NDB).

Pharmacoepidemiological studies using medical information databases covering millions and tens of millions of people have already been actively conducted, and many study results have been published in foreign countries. The Food and Drug Administration (FDA) in the United States of America is advancing and implementing Sentinel Initiative, a drug safety evaluation system using pharmacoepidemiological approaches with a large-scale medical information database for pharmacovigilance. Furthermore, the European Medicines Agency (EMA) has been advancing the establishment of an infrastructure named European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP), to conduct a pharmacoepidemiological study to strengthen pharmacoepidemiology and pharmacovigilance in Europe.

In Japan, “Expert meetings on provision of health insurance claim data” has been held since FY 2010 with regards to the use of NDB, where they started to provide claims data recorded on NDB to researchers that made requests during the “Trial Period” between FY 2011 and 2012, and the full operation began from FY 2013. On the basis of the proposals mentioned above, PMDA has launched MIHARI projects in an effort to apply pharmacoepidemiological approaches to medical information databases for pharmacovigilance in FY 2009. Furthermore, Since FY 2011, MHLW and PMDA have collaborated to build a standardized medical information database network for collaboration in 10 healthcare organizations around the country as an initiative to establish a medical information network (designated MID-NET®). In April 2012, the introduction of the Risk Management Plan was announced by a joint notification from the Directors of the Evaluation and Licensing Division and the Safety Division, Pharmaceutical and Food Safety Bureau, MHLW. According to the notification, marketing authorization holders are required to develop and appropriately implement a Risk Management Plan including a pharmacovigilance plan designed to address the risks of concern for new drugs as part of safety measures for the drug, and one of the pharmacovigilance methods is pharmacoepidemiological approaches using medical information databases.

Based on this background, it is expected that the opportunities for pharmacoepidemiological

studies that use medical information database will increase. Due to a lot of considerations when conducting the studies, such as database selection, understanding data characteristics, validation, etc., PMDA established a “Committee for preparation of guidelines on conducting pharmacoepidemiological studies” and summarized the points to consider in this guideline, based on the opinions of experts.

It is hoped that this guideline contributes to promoting and conducting of appropriate pharmacoepidemiological studies.

(4) Revision of this guideline

The points to consider as shown in this guideline reflect the status of the medical information database, as well as general knowledge and methodology in pharmacoepidemiology at the time of writing, and may change with time. This guideline may therefore be revised in the future on the basis of new information.

Committee for preparation of the guideline related to implementation of pharmacoepidemiological study (Phonetic order)

Manabu Akazawa, Professor of Public Health/Epidemiology, Meiji Pharmaceutical University

Mihoko Okada, Professor of Medical Informatics, Department of Medical Welfare Management,
Kawasaki University of Medical Welfare

*Kiyoshi Kubota, Professor of Pharmacoepidemiology Course, Graduate School of Medicine and
Faculty of Medicine, The University of Tokyo

Daisuke Koide, Specially appointed Associate Professor of Clinical epidemiology systems, Graduate
School of Medicine and Faculty of Medicine, The University of Tokyo

Hiroki Sugimori, Professor of Preventive Medicine, School of Graduate Studies, Daito Bunka
University

Takeo Nakayama, Professor of Health Informatics, Graduate School of Medicine and Faculty of
Medicine, Kyoto University

Kunihiko Hayashi, Professor, Graduate School of Health Sciences, Gunma University

Takuhiro Yamaguchi, Professor of Medical Statistics, Graduate School of Medicine, Tohoku
University

(*: Chairman)

2. Preparation of the Study Protocol

A pharmacoepidemiological study must be conducted in accordance with a study protocol created before starting the study. The items which should be included in the study protocol and description procedures are described below. Although this guideline does not indicate the structure of the protocol such as the order of items, the protocol should contain all items listed below. See sections 3, 4 and 5 for points to be considered by the researchers when planning and conducting the study.

- Study title

The title should ideally include the exposure(s) and outcome(s) of interest and study design, so that the content of the study can be easily grasped.

- Study protocol revision history (finalization date, date of update, reason for change, details of change)

In addition to specifying the finalization date of the study protocol, the updated date, the reason and the details of change in order to record the change process if changes are necessary after the finalization data should also be recorded.

- Definition of term

Important terms used in the study protocol should be clearly defined and used consistently.

- Objectives

Describe the following 3 points, with reference to the Good Pharmacoepidemiology Practices Guideline (4) by the International Society for Pharmacoepidemiology.

- 1) Research objectives: Describe the knowledge or information to be gained from the study, or describe the research question by using an interrogative sentence which explains what knowledge or information is to be tried to be gained.
- 2) Specific aims: List exposure(s) and outcome(s) of interest, and hypotheses to be evaluated.
- 3) Rationale: Explain logically how achievement of the specific aims realize the research objectives.

- Background and Previous study

Describe the background to the study and clearly explain why it is being conducted. In addition, if previous studies were used as a reference for the present study, describe the specific points such as problematic points in the results and methodology of those studies. Also, describe the post-marketing

usage of the drug of interest and whether or not there are drugs with the same drug class.

- Organization of study

The organizations and individuals carrying out the study must assume full responsibility for the study. Clearly describe in the study protocol the relationship with the organization or individuals consigning the study, as well as their roles and responsibilities. If all or part of the study will be consigned, describe the names and roles of all persons involved in the study, including those of the consignee.

- Study period

Describe the schedule of the study from the start of protocol preparation to the end of final study report preparation, including each milestone (e.g. completion of study protocol, start of the study, end of the study, completion of the study report etc.)

- Data source selection and data collection methods

List all data sources used in the study and explain how their use is appropriate from the point of view of the study objectives and hypotheses.

In the secondary use of data, briefly describe the original purpose for which the data were collected and its characteristics that could lead to limitations for the study.

Also, describe the results of usage of these data in a previous epidemiological study, as far as can be ascertained, and the data acquisition methods.

If data will be collected for the study in addition to the existing database, describe the method for data collection, and if case report forms for data collection will be created, attach an example form to the study protocol.

- Study design

Describe the reason why it is appropriate to use the selected study design and what limitations it can lead to from the point of view of the study objectives and hypotheses. Examples of study designs include cohort study, case-control study, nested case-control study, case-crossover studies, and self-controlled case series studies. When using a case-crossover or self-controlled case series study design, the potential limitations of the design should be carefully reviewed and described. Many pharmacoepidemiological studies are designed with a control group. If a concurrent comparison/control group is not used or is replaced with a historical control, e.g. because there is no suitable group available, this fact and its justification should be described.

The terms 'prospective' and 'retrospective' are sometimes used in relation to study design but are prone to misunderstanding due to the lack of a universally agreed definition. Their use should therefore

be avoided as far as possible, and if unavoidable, they should be defined before use.

- Sample size, detection power

Describe the detection power provided by the data source together with the method for calculating it, or the sample size required to conduct the study.

- Definition of study population

Clearly define the population being studied. The study population is usually defined as inclusion and exclusion criteria depending on the subject's attributes, location, and time. Clearly describe the conditions for starting and censoring the observation, and explain why these definitions and conditions are appropriate. In cohort studies, describe the period to identify outcome(s) and covariates before the start of observation, and the period to identify the incidence of outcome after starting observation. In case-control and nested case-control studies, describe the period to identify exposure and covariates. Also explain the suitability of the data used in the study in the context of these periods. Use diagrams or figures as necessary to supplement the explanation.

- Definition of exposure

Clearly define the exposure of interest in the study. A list of drug codes should be attached to the study protocol if extracting drug data from a database. Define all exposures if multiple exposures are of interest in the study. The definition of the exposure period includes the date of prescription or dispensing, prescription period, criteria for determining whether to handle as suspend or continue the prescription when the next prescription (dispensing) is later than the date estimated based on the previous prescription period, how to handle overlapping prescription periods, and the end date of the exposure period, etc. If the subjects are limited to those using a drug for the first time (new users), definition of new use should be described. For each definition, explain why the definition is appropriate. Use diagrams or figures as necessary to supplement the explanation.

- Definition of outcome

Clearly define the outcome of interest in the study. A list of diseases or other codes that serve as indicators of outcome should be attached to the study protocol if extracting outcome data from the database. Describe the method of combination when an outcome will be defined by a combination of disease names with treatment drugs or medical procedures, etc. If multiple outcomes are defined, describe each and every outcome. Also describe whether the study will identify outcomes occurring for the first time or those occurring multiple times. Also describe which date is being taken as the date

of occurrence of the outcome (incident date of outcome). For each definition, explain its appropriateness.

- Definition of covariate

Describe the basic information about patient background other than exposure and outcome that will be measured, as well as factors that may act as confounders or effect modifiers, and describe how each covariate will be identified (or a list of codes when they will be defined by codes, as with exposures and outcomes). Furthermore, if there is an important covariate that is difficult to identify in the data source being used, describe methods adopted in the study design or analysis (e.g. performing sensitivity analysis) for dealing with that.

- Validation

Describe whether validation studies have been done, or are planned, for validating the definitions of study population, exposure, outcome, covariates, etc., and attach a study protocol where applicable. If it will not be performed, explain the reason. If definitions of those factors are provided using the results of existing validation studies, explain the reason why it is appropriate to use the results of the validation study, and attach any reference material to the study protocol.

- Data management, statistical analysis

Describe the details and methods of data handling at each stage of data cleaning, data checking, dataset preparation, and up to the final analysis. Describe any analysis for obtaining the results corresponding to the study objectives as the primary analysis, and analyses used to supplement the primary analysis as secondary analyses. In addition, describe the preliminary analysis if it would be performed prior to the primary analysis. It is important to clearly separate the previously planned analysis from an additional interim analysis by documenting the former in an analysis plan. In some cases the statistical analysis protocol may be prepared separately from the study protocol. Describe details of the model used in the analysis, measure of effect, handling of covariates, etc., and describe the statistical software used in the analysis, including version information.

When matching between the exposure group of interest and comparison group will be performed in a cohort study, or matching between cases and control will be performed in a case-control study, describe the matching factors and method of matching, providing specific and detailed procedures, as well as the reasons for using them. Similarly, describe the sampling method when conducting cross-sectional studies.

Describe details of sensitivity analysis that will be performed to confirm the robustness of the

results of analysis. Sensitivity analysis is especially important in pharmacoepidemiological studies with databases because the results of analysis tend to vary significantly depending on study design such as definition of exposure, outcome, covariates, etc. Describe all previously planned sensitivity analyses and ensure that these are differentiated from additional interim sensitivity analyses.

- Study limitations

Even though the appropriate study should be performed in order to obtain results corresponding to the study objectives and hypotheses, if it is foreseeable that the validity of study results cannot be ensured due to reasons such as a limited amount of information or inability to confirm the validity of the information in the data source being used, describe these issues as expected limitations to the study. In particular, the extent to which the limitation of the study affects the internal validity and generalizability (external validity) of the conclusion should be described in detail.

- Publication of results, method of publication

Important results that can affect the safety evaluation of a drug should be publicly disclosed. If it is planned to disclose the results, describe how and where.

- Personal information protection, ethics

Describe whether or not personal information will be used. If personal information will be handled, describe a summary of the security measures that will be taken. If the study will use a data source that removed items related to personal information, describe the fact. If removing items related to personal information from the data source in the course of the study, describe the main details of the removal process. Describe any other ethical matters to keep in mind, or measures that need to be taken. Also, describe whether or not the study is subject to the ethical guidelines of an epidemiological study ⁽⁵⁾, and if the study is to be discussed by an ethics committee, describe the name of the committee and the organization in which this committee has been established.

- Conflicts of interest, preservation of transparency

For all persons involved in the study, describe the status conflict of interest related to all factors that can affect the validity of the study. If there are any conflicts of interest, clearly describe the measures taken to preserve the transparency of the study.

- References

Describe a list of all literature referred to when preparing the study protocol.

3. Data sources

When conducting a pharmacoepidemiological study that makes secondary use of data, it is important to select data sources that are suitable for the study objectives and hypothesis, and to use these sources based on a full understanding of the background that generated the data, and how that background affects the characteristics. The main characteristics and points to consider in relation to the data sources, their selection, standard coding system, and validation are described below. The primary data collection should be adopted if it is deemed appropriate after examining these points.

(1) Characteristics of the major data source used in a pharmacoepidemiological study and points to consider at the selection

Data used in a pharmacoepidemiological study include claims data issued for reimbursement of medical expenses, Hospital Information System / Electronic Medical Record data (HIS/EMR data), and registry, etc. The data source that best suits the objective and hypothesis of the study should be used. Whatever data sources are used, information on the context of the data generation should be obtained (e.g. location of the region, scale and departments of the medical institutions where the data was generated, period which the data was obtained, demographic information such as the age and sex distribution of populations included in the data), and its data characteristics should be checked.

The characteristics of each data source and points to consider on data source selection are given below.

① Claims data

(a) Characteristics

Claims data, a form of invoice data, are prepared by the medical institutions or the pharmacies for requesting the reimbursement of medical or drug dispensing fees, and mainly exist in four types: Medical, dental, drug dispensing, and diagnostic procedure combination (DPC). In Japan, there are several sources of claims data that can be used for the study. Services for providing datasets for analysis and for performing analysis using those data are available. There are also other services that perform analysis by linking claims data with information on Specific Medical Check-ups and Specific Health Guidance. These claims data include not only data collected from multiple medical institutions and pharmacies but also data provided by health insurance societies. The latter covers all claims information for insurance-covered medical services for the society members. Although the scale of the data is usually small, the NDB established by the MHLW is a large-scale database comprising claims information for insurance-covered medical services in the vast majority of the Japanese population.

Such a large-scale database is required when conducting a study for which the outcomes of interest are rare ADRs or diseases, but at the present time, the use of the NDB is limited to academic researchers. In addition, there are many restrictions for its utilization, such as the amount of time required for application and the review process for the data utilization and building a security system for handling the data.

When claims data are aggregated in one place, it is possible to comprehensively follow-up the course from exposure to outcome, since all medical information from the medical institution(s) can be obtained for the existing data period. This guaranteed follow-up period makes this data source suitable for a pharmacoepidemiological study. In other countries, particularly in the United States of America, private insurance companies possess large amounts of claims data concerning tens of millions of people, and many pharmacoepidemiological studies have been conducted using these data.

(b) Points to consider

As claims data do not contain medical practices that are not covered by health insurance (self-paid treatments, treatments insured by worker's accident compensation insurance, motor vehicle liability insurance etc.), it is not possible to obtain information related to health checkups and childbirth for pregnant women, many vaccinations, and injuries from traffic accidents. Therefore, the use of claims data is not suitable for studies that need such information.

② **HIS/EMR data**

(a) Characteristics

HIS/EMR data are held by medical institutions. These contain more informative data items related to medical information than claims data. However, since items and formats of data may differ among medical institutions, standardization of data formats is often a major issue in a study when integrating data from multiple institutions.

Since HIS/EMR data are stored by each medical institution, there are restrictions in utilizing those data in terms of appropriateness on the types and scope of epidemiological studies. If in the future, medical information can be shared among medical institutions with the expansion of regional medical cooperation, the accuracy of patient follow-up including the outpatient treatment period can be improved, resulting in increase of those data values for the epidemiological study. Apart from the NDB, the MHLW and PMDA are constructing a network that can standardize and integrate HIS/EMR data from multiple medical institutions.

(b) Points to consider

When making secondary use of HIS/EMR data from multiple medical institutions, any differences in items and format of those data entered as well as codes used for data such as disease names, drug names, and laboratory test items should be clarified. Analytical strategy in case of differences should also be considered. Since HIS/EMR data are stored by each medical institution, health information such as exposure, confounders, and outcome from other institutions may not be obtained, even if a patient has treatment from several other institutions in parallel. Therefore, in order to obtain reasonable study results, it may be necessary to consider analytical approaches such as limiting the follow-up period to the hospitalization period in the case of acute events. When including outpatients as the study population, some measures should be conducted at least such as limiting the study population to patients continuously being treated in the hospital by using the information like treatment period and number of treatments/consultations and carefully examining its impacts on the study results by estimating the proportion of such patients among all study population.

③ Registry

(a) Characteristics

Registries are created for the purpose of collecting certain specific information and can be divided into the following categories based on type of information: Outcome (disease) registries (cancer registration, etc.), exposure registries (post-marketing studies, etc.), and medical practice registries (collection of information related to specific surgeries, etc.). A registry collects specific data items for each purpose. Thus, on the primary use of registry created for a particular study, collectable information among necessary data for the study should generally be ensured.

(b) Points to consider

In order to make secondary use of an existing registry, if the purpose of creating the registry is different from the purpose of study by secondary use, there is a possibility that information in the registry may be insufficient due to lack of necessary data items. It may be necessary to consider collecting additional information and linking the data with other data (described below).

④ Overseas databases

The use of overseas databases may be an option depending on the study aims and hypothesis. When using it, it is necessary to carefully examine whether the study results are applicable to Japan (generalizability). If there is a clear reason to select an overseas database after comparing the differences in medical insurance systems, medical practices (diagnostic standards and choice of drugs),

racial differences, and method of collecting and coding data being used, the results of examining factors that can affect the study results and reasons for selecting the overseas database should be described in the study protocol and the report.

In addition to the differences in the basic characteristics of each data source described above, when selecting a data source, it is necessary to examine the following aspects of data relevant to the study: 1) Follow-up period, 2) data collection period, 3) population, 4) data items, etc. 1) The follow-up period should clearly be described that it is possible to measure the covariates and outcomes of interest in the study within that period, and the database that sufficiently satisfies these conditions should be selected according to the study objective. For example, if the outcome of interest is the delayed-onset event and the period from exposure to outcome is long, a long follow-up period is necessary. 2) As for the data collection period, when a study conducted concerns a specific period (e.g. epidemic of infectious diseases or revision of medical guidelines, etc.), naturally, the data should include this period. 3) In terms of population, one should confirm that the data being used include the population of interest. For example, if a study concerns pediatric patients, claims data possessed by health insurance societies can be used. On the other hand, if a study concerns elderly patients, as many workers withdraw from the health insurance societies with retirement, claims data from health insurance societies would contain very little data about elderly people, and for this reason it is not suitable for use in the study. Therefore, one should consider the use of data sources, such as HIS/EMR data or registries, other than claims data of health insurance societies. 4) Acquire information on what is held as a data item, and confirm whether information used to define the target population, exposures, outcomes, etc. is available. For example, the dispensing claims data do not include the name of the disease, and DPC claims data do not include the date of disease onset after hospitalization. Therefore, if these items are required, it may be necessary to consider using other data sources.

The number of medical information databases that can be used for a pharmacoepidemiological study in Japan has been gradually increasing, but there are still relatively few options at present. As such, an ideal database suited to the objectives and hypotheses of the study may not always be available; hence, the most suitable one should be selected and the study limitations to the selected database should be clearly described in the study protocol and the report. Furthermore, as shown in the beginning of Section 3, as a result of examining database selection, the primary data collection may be deemed more suitable than secondary use.

If information from a single database is insufficient, multiple databases can be used together by the linkage method. For example, rather than using a single data source, a higher-quality study is expected to be carried out by identifying exposures from prescription information in claims data and

outcomes from the laboratory test values in the HIS/EMR data or disease names in a registry. However, since there is a greater likelihood of identifying individuals from the greater amount of information in linked data than in unlinked databases, careful considerations should be taken when linking data. If linkage will be performed, the target data, the method of linkage and its validity should be described in the study protocol and reports.

(2) *Standard coding system*

When identifying exposures and outcomes in the database study, data encoding these types of information are usually used. When selecting a data source, appropriateness of the coding system for defining the exposures and outcome should be confirmed. The characteristics and points to consider about the standard coding system used in Japan for disease names, drugs and laboratory tests are presented below.

① Characteristics of disease name codes

“The International Classification of Diseases and Related Health Problems Version 10” (ICD-10) codes and the “The Standard Disease Names” (in general an exchange code is used), which are even more detailed disease names compatible to ICD-10 codes, are often available in HIS/EMR data. With respect to claims data, “The Receipt Data Processing System (RESEDEN) Disease Name Master,” which is mapped with “The Standard Disease Names,” is used. Apart from these codes, pharmaceutical companies use “the Medical Dictionary for Regulatory Activities Terminology” (MedDRA) code that mainly expresses ADRs and adverse events to report cases of ADRs and adverse events to the government.

② Characteristics of drug codes

In the HIS/EMR data, a “National Health Insurance Drug Code” or an “Individual Drug Code” (YJ code) is often assigned, and some of medical institutions assign HOT code. With respect to claims data, the RESEDEN drug master is used, and as these RESEDEN drug codes, national health insurance drug codes, and individual drug codes are mapped with the HOT codes. It is therefore possible to cover other coding systems with whichever codes are provided. Furthermore, the Anatomical Therapeutic Chemical classification system (ATC) code, managed by the World Health Organization, has been mapping with individual drug codes. Both ATC codes and individual drug codes are useful for defining the therapeutic classification of drugs.

③ Characteristics of laboratory test item codes

Although JLAC10 managed by the Japanese Society of Laboratory Medicine is available, it is not widespread, and at present, very few medical institutions use JLAC10 in HIS/EMR data. With claims data, the RESEDEN master is also available for medical procedure codes and specific instrument codes that can be used in a study.

④ Points to consider about codes

The above-mentioned codes may be updated and the code itself or names associated with the codes may be changed. The versions of all codes may not be uniform and multiple versions of codes may be used in a study. For this reason, it is necessary to check all versions contained in each code master, and code lists prepared for definitions of exposures and outcomes should be created by using all versions of code masters contained in data of the study.

(3) *Validation*

Validation indicates the procedure of confirming the accuracy of a method ⁽¹⁾. In secondary use of medical information, it is recommended to conduct validation studies for evaluating the accuracy of the definition of exposures and outcomes, etc., used in the study. In particular, if the outcome is defined using only the disease name codes in the claims data, the validation study would be necessary.

Validation studies compare pre-established definitions of exposure and outcome applied to the database against the gold standard information source (data from registry with established reliability, or results of evaluating medical records by proper procedures), in order to evaluate the accuracy of the definitions. An example of procedures using medical records as the gold standard would be to identify all subjects in a particular database, such as subjects corresponding to the exposure or outcome definitions, and to determine whether or not the definition identifies exposure or outcome correctly by reviewing the medical records of these subjects. Judgment criteria for outcomes should be created by referring to any standard diagnostic criteria established in relevant clinical guidelines, etc., if available. Even when there are no relevant clinical guidelines, it is recommended to set criteria based on generally established diagnostic criteria as much as possible. Outcomes are judged according to criteria, but usually judged by clinicians due to the requirement of clinical knowledge. In addition, the judgment should ideally be made by multiple judges to ensure the objectivity of judgment results, and by a small number of judges to minimize the variation in judgment results. Furthermore, in cases requiring the checking of medical records across a number of medical institutions, the methods used in some countries is for a specially-trained person to only extract the necessary information based on a predefined format from each institution's medical records for judging the case of interest, and for

the judge(s) to make their judgment based on the extracted information. Although not well established in Japan, this method would be useful in allowing a small number of judges to judge cases in multiple medical institutions.

When the exposure or outcomes meet the judgment criteria, such data are considered true and accurate. The proportion of subjects judged “true” relative to the subjects extracted from the database for corresponding to the definitions is called the positive predictive value¹ which is an indicator of validity. Other indicators include negative predictive value, sensitivity and specificity, and it is desirable to estimate sensitivity and specificity because they are universal indicators not affected by the composition of the population characteristics in the validation study. However, these estimations require a data source¹ judged based on a gold standard relating to part or all of the target population, such as highly reliable disease registries, but it is often difficult to prepare such a data source. Therefore, only positive a predictive value is usually available from the viewpoint of study feasibility.

Validation studies are originally performed for a specific database and definition for each study. Thus, the positive predictive value is an indicator specific to them used in the study. However, results of validation studies conducted on the same definitions in other databases may be used in case of difficulty evaluating accuracy of the definition based on medical records etc. because study subjects were anonymized without linkage. Furthermore, if creating an appropriate definition of exposure or outcome based on the information in the database is difficult, or if higher validity of study results is necessary, exposure or outcomes should be identified more accurately by reviewing the medical records and other relevant information, as a different approach from the validation study.

4. Study design and conduct of the study

The following is an overview of the main study designs used in a pharmacoepidemiological study as well as points to consider about study designs and conduct of pharmacoepidemiological studies.

(1) Study design

There are many study designs, and the most appropriate one should be selected for each purpose and hypothesis of the study. The characteristics and points to consider about the main study designs used in pharmacoepidemiological studies are discussed below. Among these, designs with a control group are relatively common.

The designs described here are provided for reference. Technical books should be used for

¹ There are various ways to call this, such as positive reaction value and positive hit rate etc., but this guideline follows the “Dictionary of Epidemiology, Version 5.”

considering more details of each design.

① **Cross-sectional study**

(a) Characteristics

This is a study design that collects data for the target study population at a particular point in time, and it is used to investigate the prevalence of diseases and the use/prescription status of drugs, etc. It is also possible to analyze trends over time when conducting cross-sectional studies at multiple time points. This design can be used to evaluate impacts of implementing safety-related regulatory actions through investigation of prescription trends before and after safety-related regulatory actions (e.g. a new warning section in the package insert).

(b) Points to consider

Although control groups can be set in cross-sectional studies, it is difficult to estimate the causal relationship between exposure and outcome because of uncertainties in the temporal relationship between those and as to whether the change in prevalence is due to changes in disease incidence or disease duration, etc.

② **Active surveillance**

(a) Characteristics

According to the ICH E2E Guidelines ⁽⁶⁾, active surveillance is defined as “A method that seeks to ascertain completely the number of adverse events via a continuous pre-organized process.” An example of this is the follow-up of patients treated with a specific drug administration through a risk management program. Patients who fill a prescription for this drug may be asked to complete a brief survey form and give a permit for later contact by the investigator.

The FDA has been advancing the construction of a Sentinel System for monitoring the safety of post-marketing drugs in almost real-time. The sentinel system is a new form of active surveillance that makes secondary use of medical information data. For example, it aims to identify safety signals that have a certain extent validity as quickly as possible with regards to adverse events of concern about new drugs that have just been approved.

③ **Case series**

(a) Characteristics

This is a study design that collects cases having a common outcome, including a collection of cases with a specific outcome regardless of exposure, and a collection of cases with a specific exposure and

outcome ⁽⁷⁾.

(b) Points to consider

Although a case like the latter example discussed in (a) above would lead to generation of a hypothesis related to the relationship between exposure and outcome, it is normally not possible to estimate the absolute risk.

④ Self-controlled case series, case crossover

(a) Characteristics

Both designs use cases as their own control without setting a control group and are able to eliminate the influence of intra-individual confounders that do not change over time. Advantages of these designs are high efficiency, which need only a small number of subjects, because they basically target cases only and uses the follow-up period for each case as the case and control period.

In the Self-Controlled Case Series (SCCS), the incidence rate ratio is obtained as a measure of effect (indicator of action). SCCS was originally designed to evaluate the association between vaccines and acute events, but it has recently been considered for application to other medicines. On the other hand, in Case Cross Over (CCO), the incidence rate ratio is obtained as an odds ratio, as a measure of effect.

(b) Points to consider

These designs have matters to be considered such as: The influence of time-varying confounders cannot be eliminated unless suitable adjustments are made, the definition for setting a risk period (period of risk of outcome occurring after exposure) significantly affects the results of analysis, and the reverse causality is likely to occur because this design is often used for assessing acute events. Therefore, in applying these designs, the validity of using these designs and the necessary conditions for appropriate use should be well considered.

Furthermore, as CCOs usually only set the control period before the case period, it is easily influenced in situations where the prescription trends fluctuate with time. To deal with this issue, it is desirable to consider designs that set a control period after the case period, or case-time-control designs, as alternatives to the standard CCO. The case-time-control design specifies a non-case group (control) matched in time from the same population in which the cases occurred, and it calculates the adjusted odds ratio for the purpose of adjusting the temporal fluctuation in prescription trends ⁽⁸⁾.

⑤ Cohort study

(a) Characteristics

A Cohort is broadly defined as “A group of individuals selected for specific purposes, and followed for a certain period.” Cohort studies measure the incidence of diseases occurring in one or more cohorts, and it is the prototype of all epidemiological studies ⁽⁹⁾. Although it allows the measurement of multiple outcomes from a single cohort, if the outcome is rare, a very large cohort size will be required. When a control group will be provided, the risk difference² and the risk ratio³ or incidence rate difference and incidence rate ratios are used as measure of effect. Although it is possible to measure incidence ratios and incidence rates in cases without a control group, it is not possible to determine the measure of effect.

Data collection for cohort study is labor-intensive, when using the questionnaires as a conventional data collection method, but more efficient when using a large-scale database. For this reason, the cohort study is usually selected as the design of an epidemiological study.

(b) Points to consider

When studying the relationship between a drug of interest and outcome of interest for the purpose of signal evaluation, a control group should be provided for conducting an appropriate evaluation. In addition, a control group is necessary when estimating the relationship between exposures and an outcomes with a high background incidence rate, or between exposures with little drug influence and outcomes. The exposure group and control group populations of interest should be as similar as possible in terms of age, sex, disease, concomitant treatments, etc. Even if there is no proper simultaneously comparable control group, if there is comparable and reliable historical data on populations not receiving the exposure of interest, it is possible to compare with this data (historical control).

⑥ Case-control study, nested case-control study

(a) Characteristics

These designs aim to reach the same goals as cohort studies more efficiently by sampling the controls ⁽⁹⁾. It is possible to estimate the relationship between a single outcome and multiple exposures, and the

² The term “attributable risk” is sometimes used instead of risk difference, but the term attributable risk implies that factors other than exposure that affect outcomes between the two groups are equivalent, and it is sometimes used to mean “attributable fraction,” which is risk difference divided by the incidence in the exposure group, or as other different indicators.

³ The term “relative risk” that encompasses both the risk ratio and the incidence ratio is sometimes used. The term “relative risk” may also be used for odds ratios when the odds ratio in a case-control study can be interpreted as a risk ratio or an incidence rate ratio.

odds ratio can be obtained as a measure of effect.

This design has been relatively common in the conventional data collection approach using questionnaires because this design is more efficient than cohort studies as the amount of data collected is less. When a large-scale database is available for use, nested case-control studies using a cohort are also sometimes conducted.

(b) Points to consider

In general, control should be selected from the source population that gave rise to the cases. If a source population cannot be defined clearly, it is necessary to consider an approach for selecting appropriate controls. The nested case-control design, which is a case control study conducted within the pre-determined cohort, is a useful approach which ensures that controls are selected from the source population.

(2) *Points to consider about study designs*

① Definition of study population

In order to define the study population, it is necessary to check the follow-up period of each individual contained in the study population in the data being used. The definition of the follow-up period, for example, is the period between the start date to end date of data collection in claims data, whereas in HIS/EMR, the follow-up period is from the date of hospitalization to date of discharge. If claims data do not provide such information used to define the follow-up period, it is necessary to consider how the follow-up period should be defined.

The study population needs to be suited to the purpose and hypotheses of the study, and the inclusion and exclusion criteria should be set so as to compose the population. When the study population is identified by applying these criteria to the data during a study, check the information of the people excluded, especially regarding the exclusion criteria, and examine whether or not people have been excluded from the study improperly, from the point of view of the purpose and hypotheses of the study before starting the study. It is also necessary to check that censoring of observation has not occurred in connection with the incidence of specific exposures or outcomes.

A suitable unit of observation should be set, such as person-year, number of people, number of prescriptions, etc.

② Definition of exposure

As information related to drug prescriptions vary by data source in terms of the granularity of information and the types of drugs contained, exposure should be defined by taking these points into

consideration. For example, exposure should be defined by considering the drug coding system being used, whether the prescription information contains the date or just the month, and whether the data contains out-of-hospital prescriptions or in-hospital prescriptions of drugs. List the drug codes of interest exposure and exposures of control groups, and where a drug has multiple dosage forms, clearly indicate the dosage forms used as exposures.

Usually, prescription and dispensing information can be obtained on the secondary use of a medical information database, but information on actual use of the drug is often not available. Therefore, it is necessary to define the exposure period by deducing the actual status of drug use based on the prescription and dispensing information. When defining the exposure period, it is necessary to decide whether the start date of exposure is the date of prescription or the date of dispensing, whether to consider the gap between the previous prescription period and next prescription period as a continuous exposure period, whether to set a grace period after the end of the prescription, and what the conditions are for completion of the exposure period. Furthermore, for chemotherapy with anti-cancer drugs where a prescription cycle is a fixed period of time, and prescriptions that are to be taken as needed, it is necessary to consider the most suitable way to set the exposure period. If information related to the prescription dose is needed, it is necessary to consider how to measure it, and how to calculate the daily dose, cumulative dose, etc. Furthermore, consider the necessity to limit the population to new users of a drug, and if this is necessary, the conditions for defining new users should be set (e.g. no prescription of the drug in 6 months prior to starting the prescription).

③ Definition of outcome

When making secondary use of the data, it is essential to establish a highly valid definition of outcome using information contained in the data. Consider whether or not a highly valid definition of outcome can be achieved by combining drugs, laboratory tests and medical practices used for diagnosis or treatment, but not defining outcome only based on the disease name related to the outcome. List the codes such as disease names, drug names, laboratory test names, and medical practices used to define the outcome. Moreover, it is essential to define the date of outcome. In particular, when the outcome is defined by combining multiple information, consider which date information is defined as the incident date of outcome. When the data contains laboratory test values like HIS/EMR data, it is recommended that the test values are used in defining the outcome. For example, where the outcome is leukopenia or thrombocytopenia, a highly valid definition of outcome is achievable by setting a reference laboratory test value as the outcome. However, it is also necessary to consider the test values before start of exposure and the interval from exposure to testing. Also, when conducting a study using the data from multiple medical institutions, define the outcome considering the difference in the

reference ranges for test values at each medical institution.

When using the disease name codes to define the outcome, it should be aware of the fact that the timing of the data provided with the disease name may differ, depending on the data source. For example, there are situations where the start date of treatment on the claims data and the date of diagnosis on the HIS/EMR data do not match for the same patient. It is also necessary to consider the latent period until the incidence of outcome, as well as whether the study will identify outcomes occurring for the first time or those occurring multiple times.

④ Bias

The errors associated with results and estimation in epidemiological studies are largely divided into 2 types namely, random errors and systematic errors. The former are randomly occurring errors, and they are data variations that cannot be explained by systematic errors. Theoretically, if the size of the study population is made infinite, the variation of mean due to random errors can be reduced to zero. The latter, also called bias, is a systematic deviation from the true value, and the degree of deviation of mean does not change even if the size of the study population becomes infinite ⁽⁹⁾. Although bias can be subdivided into many types, the three broad types are selection bias, information bias, and confounding. Refer to technical texts for the details and strategies for handling each type of bias ⁽¹⁰⁾.

When defining outcomes and exposures using data items included in the medical information database, in particular, it should be noted that information or selection bias may arise when the information contained in the database is different from the facts because of the data deviation from the actual date or missing information. If bias is expected, it is important to conduct a sensitivity analysis in order to ensure the robustness of the results of analysis.

Among confounding, what is particularly problematic in a pharmacoepidemiological study is “confounding by indication,” which is also called “confounding by reason for prescription.” It has been pointed out that in the secondary use of data, the information needed to deal with this issue can be particularly lacking ⁽¹¹⁾. When comparing a particular drug user with a non-user, the proportion of subjects with indicated disease of the drug in the user group is naturally higher than in the non-user group. Even if two drugs with the same indication are compared, if they are prescribed for patients with different severity of the illness and comorbidities, there may be differences in the distribution of severity and comorbidities between the two groups. If the presence or absence of an indication, its severity, or the presence or absence of comorbidity affects the outcome of interest in the study, the incidence of outcome may be also affected by the differences in indication, its severity and the presence of comorbidity, as well as by the presence or absence of exposure, or the difference in exposure. Confounding by indication refers to the effect of the drug itself and some other factors,

that is, the mixing of the effects due to indication, or the results of mixing. It is not possible to correctly estimate the measure of effect without properly adjusting the confounding by indication⁽¹⁰⁾.

It is necessary to sufficiently consider how to deal with every conceivable confounder from the planning stage of the study. In addition, prepare a code list for covariates like confounders, similar to the definition of exposure and outcome, and set the factors that need to be identified. Covariates should generally be identified during the period set before the start of exposure of interest.

⑤ Adjustment of confounders

Confounders can be either measurable or unmeasurable, and if these are likely to exist, efforts should be made to minimize the effect of confounding or to estimate the extent of that effect.

For measurable confounders, there are approaches, such as restricting the study population so that potential confounders do not give rise to confounding, balancing the distribution of confounders between groups by matching, doing stratified analysis by confounders at the analysis stage, equalizing the confounders in each group by standardization, etc.⁽¹²⁾

For unmeasured confounders, there are adjustment approaches such as external adjustment, which measures confounders using external data containing detailed information different from the data being used in the study and uses the results to make adjustments, as well as methods that use the self-control design described in section 4. (1) ④, a method of setting a control group exposed to the drug with the same indication and balancing the distribution of confounders by aligning the characteristics between both groups (active comparator), and sensitivity analysis⁽¹²⁾.

When adjusting for numerous measurable confounders simultaneously, the confounders are often included in a multivariable regression model. However, it is known that if there are too many confounders relative to the number of the outcome incidence, the analysis takes time and the measure of effect cannot be correctly estimated⁽¹³⁾. To deal with this problem, there is a method in which a score summarizing multiple covariates is created, and the number of covariates is reduced by including this score in the multivariable regression model. Such summary scores include the Propensity Score (PS) for predicting exposure from multiple covariates, and the Disease Risk Score (DRS) and the Charlson Comorbidity Index (CCI) for predicting outcome incidences. The details provided below are not to recommend the use of these scores, but should be used according to the purpose of the study.

(a) Propensity Score

PS is a summary score of covariates predicting the probability of receiving exposure, and it is estimated from multiple covariates for each individual, regardless of the relationship with the outcome⁽¹⁴⁾. If there is an exposure of interest (Drug A) and a comparator exposure (Drug B), PS is the

probability of each individual being prescribed Drug A, estimated independently of whether the actual prescribed drug is A or B. If the PS obtained from an individual's covariate is close to 1, Drug A is likely to be prescribed even if Drug B was actually prescribed. Conversely, if the PS is close to 0, the probability of Drug A being prescribed is considered low, even if Drug A was actually prescribed. The PS obtained for each individual can be used for matching, stratification, and regression models.

PS was proposed for the first time in 1983 ⁽¹⁵⁾, and in recent years, it has been used in many studies. In particular, in cohort studies that use databases, PS has been established as a standard method to achieve balance in the covariates between two groups. However, it is important to consider that it is only possible to adjust the observable covariates, and that an accurate PS may not be obtained if the scale of the data source is small. There are various opinions regarding the choice of variables to estimate PS, but at the very least, PS should be estimated by measuring as much as possible all factors that are related to choice of treatment and affect the outcome.

(b) Disease Risk Score

DRS is estimated from multiple factors that predict the outcome and prognosis of a disease, such as CCI, described in the next section. In pharmacoepidemiology, DRS normally mean a summary score estimated from unexposed group data in a study, with the aim of not only simply predicting outcome incidences but also adjusting confounders ⁽¹⁶⁾. This is used as a summary score for confounding factors in stratification and regression models, and it can be used in studies where, for example, cardiovascular disease is the outcome ⁽¹⁷⁾. In pharmacoepidemiological studies, DRS is rarely used compared to PS. However, if the prescribing trend of a drug changes rapidly, such as immediately after the start of new drug sales, it may be appropriate to adjust confounding factors with this DRS rather than PS, and there is also a method of using DRS as an effect modifier. Its usefulness is being reviewed in recent years ^{(18) (19)}.

(c) Charlson Comorbidity Index

There are summary scores directly intended for predicting the incidence of disease outcome. For example, CCI is known as a score that estimates the mortality risk by totaling the scores for specific diseases such as cancer, cardiovascular disease, and diabetes, etc. ⁽²⁰⁾. Although it was originally created using information from medical records, a method of converting the CCI-estimated disease list to ICD-9-CM or ICD-10 has been devised to allow its use in a database study ^{(21) (22)}. Although studies using this score as a covariate have been seen both in Japan and other countries, it is a score created in the United States in the 1980s, and there may be differences in background from current medical practice, such as diagnostic criteria. Therefore, when using CCI, the validity of the CCI

estimation method should be confirmed for each data and disease area used in the study.

⑥ Effect modifiers

If the effect of the exposure on the outcome differs depending on the level of a factor other than the exposure and outcome, the factor is called an effect modifier. For example, if the incidence of ADRs of a drug differs between men and women, sex is the effect modifier. As effect modification vary depending on the measure of effect being used, it is sometimes referred to as the effect-measure modification rather than the effect modification. The presence or absence of effect modifiers should be investigated in the planning stage of the study, and creation of the analytic plan should include estimation of the influence of any important effect modifiers identified.

(3) Data management

Data management, which includes data checking and cleaning performed before analysis, creation of analysis data sets (variable selection and creation), etc., should be recorded in documents, such as reports, well enough to be reproducible by a third party.

(4) Analysis

It is important to make a clear distinction between the analysis prescribed in the study protocol prior to the start of the study and analysis added after confirmation of the protocol. In order to ensure the reproducibility of the analysis, the analytical process should be clearly documented, and the analysis program should be saved with comments so as to be easily understandable by third parties.

(5) Quality assurance

The accuracy of data used for study should be ensured thorough an information security measure, creation of access records to the data, etc. Quality assurance should also be applied to each document and program in the study process.

(6) Protection of personal information, and ethics

Every study that is conducted should be done according to ethical guidelines related to the epidemiological study⁽⁵⁾ and relevant regulations. In addition, regarding the handling of data used for the study, the framework should be implemented for explaining to people outside of the study organization how to deal with security issues such as preventing the leakage of information.

5. Preparation of study report

If submission of a study report is required, all of the items listed below should be included in the report, but do not necessarily specify the structure of the report, such as the order of items. If a study report is not required, a record of these items can be made in any suitable format, containing the following items, in order to ensure the reproducibility of the study.

The study report should include all items mentioned in the study protocol such as the purpose and methods. If the protocol was updated during the study, the study report should reflect the latest version in addition to the sections for results and discussion, etc.

It is also recommended that the contents of the report is checked with the statement of the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) where necessary⁽²³⁾, even aside from the points to consider described below.

- Title of report

The title should ideally correspond to that of the study protocol, for consistency.

- Report update history

As with the study protocol, the update process should be recorded.

- Definition of terminology

Terms used in the report should be added and be described as in the study protocol.

- Purposes

These should be described as in the study protocol. Any additional points should also be stated.

- Background, previous study

These should be described as in the study protocol. Any additional points should also be stated.

- Organization of study

These should be described as in the study protocol. Any additional points should also be stated.

- Study period

The actual period needed for the study should be stated.

- Selection of data source, method for obtaining data

In addition to the contents of the study protocol, any additional points should also be stated, such as the actual methods used to obtain data.

- Study design

This should be described as in the study protocol. Any additional points should also be stated.

- Sample size, detection power

These should be described as in the study protocol. Any additional points should also be stated.

- Definition of study population

This should be described as in the study protocol. Any additional points should also be stated.

- Definition of exposure

This should be described as in the study protocol. Any additional points should also be stated.

- Definition of outcome

This should be described as in the study protocol. Any additional points should also be stated.

- Definition of covariates

This should be described as in the study protocol. Any additional points should also be stated.

- Validation

This should be described as in the study protocol. Any additional points should also be stated.

- Data management, statistical analysis

This should be described as in the study protocol. If the missing values were supplemented or data conversion (coding, categorization, etc.) was performed in the data management process, describe the details of the processes. Also, if there are items that should be noted, such as those with a large proportion of missing values among the measured variables, describe them. Additionally, analyses performed after the confirmation of the study protocol, such as adjustment of confounding factors, presence of effect modifiers, study of interaction, additional sensitivity analysis, etc., should be clearly distinguished from what was mentioned in the protocol. The purpose of these additions should also be described.

- Publication of results and method of publication

The publication schedule at the time of writing the report should be stated as in the study protocol.

- Protection of personal information, and ethics

Describe the method of handling personal information and any other ethical matters. If an ethics committee held discussions on the matter(s), describe the name of the committee, the name of organization in which the committee has been established, the review date, review results, and approval number.

- Conflict of interest, and ensuring transparency

These should be described as in the study protocol. Any additional points should also be stated.

- Study results

The number of people in the study population, number of people corresponding to the inclusion and exclusion criteria, sample sizes of exposure group and control group of interest, the number of people of cases and controls etc. should be stated. A flowchart is useful for showing each stage from identification of the source population to the analysis population.

The distribution of basic measured factors (population statistics, clinical, social factors, etc.) and covariates should be shown, so that it is possible to understand the characteristics of populations identified for analysis, e.g. the exposure group and control group, cases/controls, etc. If matching was performed in a cohort study, show the distribution of populations before and after matching, in order to indicate how the covariates in the two groups were balanced and aligned by matching.

For cohort studies, show the distribution of periods (follow-up period, exposure period) and number of the outcome incidence, and for case control studies, show the frequency of exposure, etc. Before showing the results of analysis using a model, show the results of crude analysis or stratified analysis by the main factors in tables and figures. In addition, show the 95% confidence intervals for the estimated measure of effect.

Clearly and distinctively show the results of primary, secondary, and sensitivity analysis. Additional analyses should be clearly distinguished from what was mentioned in the study protocol.

The report should show not only the results of the final model used for analysis, but also the details of the variable selection process. Describe all candidate variables, approach for selecting the variables, and the variables that were ultimately selected or not selected. If summary scores such as PS and DRS were used, list the variables used to estimate these scores and describe the method for estimating the scores.

- Discussion

Describe how the results obtained are interpreted from the point of view of the purpose of the study. There are no particular restrictions on how to structure this section, but the following sequences is one possibility: 1) Summary of the main results, 2) Comparison with results of previous studies, 3) Interpretation of results (including discussions related to influence of bias and unmeasured confounders), 4) Study limitations, including generalizability of the results, constraints on amount of information obtained and validity of information, and 5) Conclusion. All essential points should be included, not only those listed above.

- References

List the literature referred to in the study.

6. Publication of study results

In principle, the study results should contribute to the public interest, and it is desirable for researchers to explain the results to the general public in order to gain an understanding of the pharmacoepidemiological study. In addition, when publishing the contents of the study in academic journals, the contents of the report should be properly summarized so that there is no discrepancy between it and the publication. In order to avoid publication bias, it is recommended that the results are published even if negative study results are obtained with respect to study objectives or hypotheses.

7. References

- (1) Miquel Porta, Japan Epidemiological Association. Dictionary of Epidemiology, Version 5, 2010, 402p.
- (2) WHO. The Uppsala Monitoring Centre. What Is a Signal? (<http://www.who-umc.org/DynPage.aspx?id=115092&mn1=7347&mn2=7252&mn3=7613&mn4=7614>), (Accessed on November 22, 2013).
- (3) Health Insurance Claims Review & Reimbursement Services. Receipt Processing System. (<http://www.ssk.or.jp/rezept/index.html>), (Accessed on November 22, 2013).
- (4) ISPE. Guidelines for Good Pharmacoepidemiology Practices. 4, 2007. (http://www.pharmacoepi.org/resources/guidelines_08027.cfm), (Accessed on November 22, 2013).
- (5) Ministry of Education, Culture, Sports, Science and Technology, Ministry of Health, Labour and Welfare: Ethical Guidelines on Epidemiological Research.

- (<http://www.mhlw.go.jp/general/seido/kousei/i-kenkyu/sisin2.html>), (Accessed on October 17, 2013).
- (6) ICH E2E Pharmacovigilance Planning. 2004.
 - (7) Olaf M. Dekkers, Matthias Egger, Douglas G. Altman, Jan P. Vandenbroucke. Distinguishing Case Series from Cohort Studies. *Annals of International Medicine*, 2012, vol. 156, p. 37-40.
 - (8) Malcolm Maclure, Bruce Fireman, Jennifer C. Nelson, Wei Hua, Azadeh Shoabi, Antonio Paredes, David Madigan. When Should Case-Only Designs Be Used for Safety Monitoring of Medical Products? *Pharmacoepidemiology and Drug Safety*, 2012, vol. 21(S1), p. 50-61.
 - (9) Rothman, translated by Eiji Yano/Hideki Hashimoto, K J. Rothman's *Epidemiology*. First Edition, Shinoharashinsha Publishing Inc., 2009, 286p.
 - (10) K. J. Rothman, S. Greenland, T. L. Lash. *Modern Epidemiology*. Third edition, Lippincott Williams & Wilkins, 2008, 758p.
 - (11) Til Sturmer, Robert J. Glynn, Kenneth J. Tothman, Jerr Avorn, Sebastian Schneeweiss. Adjustments for Unmeasured Confounders in Pharmacoepidemiologic Database Studies Using External Information. *Medical Care*, 2007, vol. 45 (10 suppl), p. S158-165.
 - (12) Abraham G. Hartzema, Hugh H. Tilson, K. Arnold Chan. *Pharmacoepidemiology and Therapeutic Risk Management*. Harvey Whitney Books Company, 2008, 1050p.
 - (13) P. Peduzzi, J. Concato, E. Kemper, et al. A Simulation Study of the Number of Events Per Variable in Logistic Regression Analysis. *Journal of Clinical Epidemiology*, 1996, vol. 49, p. 1373-1379.
 - (14) Rubin, D. B. Estimating Causal Effects from Large Data Sets Using Propensity Scores. *Annals of International Medicine*, 1997, vol. 127, p. 757-763.
 - (15) Paul R., Donald B. Rubin. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 1983, vol. 70, no. 1, p. 41-55.
 - (16) Miettinen, O. S. Stratification by a Multivariate Confounder Score. *American Journal of Epidemiology*, 1976, vol. 104, no. 6, p. 609-620.
 - (17) Wayne A. Ray, Katherine. T. Murray, Kathi Hall. Azithromycin and the Risk of Cardiovascular Death. *The New England Journal of Medicine*. 2012, vol. 366, p. 1881-1890.
 - (18) Patrick G. Arbogast, Wayne A. Ray. Use of Disease Risk Scores in Pharmacoepidemiologic Studies. *Statistical Methods in Medical Research*, 2009, vol. 18, p. 67-80.
 - (19) Robert J. Glynn, Joshua J. Gagne, Sebastian Schneeweiss. Role of Disease Risk Scores in Comparative Effectiveness Research with Emerging Therapies. *Pharmacoepidemiology and Drug Safety*. 2012, vol. 21(S2), p. 138-147.
 - (20) Mary E. Charlson, Peter Pompei, Kathy L. Ales, C. Ronald MacKenzie. A New Method of Classifying Prognostic Comorbidity in Longitudinal Studies: Development and Validation.

- Journal of Chronic Diseases, 1987, vol. 40, no. 5, p. 373-383.
- (21) Richard A. Deyo, Daniel C. Cherkin, Marcia A. Ciol. Adapting a Clinical Comorbidity Index for Use with ICD-9-CM Administrative Database. *Journal of Clinical Epidemiology*. 1992, vol. 45, no. 6, p. 613-619.
- (22) Vijaya Sundararajan, Toni Henderson, Catherine Perry, Amanda Muggivan, Hude Quan, William Ghali. New ICD-10 Version of the Charlson Comorbidity Index Predicted In-hospital Mortality. *Journal of Clinical Epidemiology*, 2004, vol. 57, p. 1288-1294.
- (23) Erik von Elm, Douglas G. Altman, Matthias Egger, Stuart J. Pocock, Peter C. Gøtzsche, and Jan P. Vandembroucke. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement Guidelines for Reporting Observational Studies. *Epidemiology*, 2007, vol. 18, no. 6, p. 800-804.