The Application and Future of Big Database Studies in Cardiology: A Single-Center Experience

Kuang-Tso Lee,1,2# Ai-Ling Hour,3# Ben-Chang Shia4 and Pao-Hsien Chu1,2,5

As medical research techniques and quality have improved, it is apparent that cardiovascular problems could be better resolved by more strict experiment design. In fact, substantial time and resources should be expended to fulfill the requirements of high quality studies. Many worthy ideas and hypotheses were unable to be verified or proven due to ethical or economic limitations. In recent years, new and various applications and uses of databases have received increasing attention. Important information regarding certain issues such as rare cardiovascular diseases, women’s heart health, post-marketing analysis of different medications, or a combination of clinical and regional cardiac features could be obtained by the use of rigorous statistical methods. However, there are limitations that exist among all databases. One of the key essentials to creating and correctly addressing this research is through reliable processes of analyzing and interpreting these cardiologic databases.

Key Words:  
Big database • Cardiovascular diseases • National Health Insurance Research Database (NHIRD)

INTRODUCTION

The field of medicine is an ancient science. Over the centuries, medicine has developed through careful observation, thoughtful hypotheses, and detailed verification of experiments that in many instances involve the connection of seemingly different factors. In order to apply results to the real world, studies should be rigorously controlled to minimize potential bias. Randomized controlled trials in the cardiovascular field have simplified issues of interest, and have been used to exclude or equalize factors that might interfere with a hypothesis.1,2 Controlled trials can answer specific, selected questions, however there remains some doubt as to how these results can be applied to a real scenario. Moreover, to create the possibility of statistical significance, it is important to minimize potential sampling bias, ensure baseline equalization, and monitor unexpected effects and results. Furthermore, taking the issue of medical ethics into consideration, the sample size of experimental subjects is also a problem. Therefore, considerable funding and time is required for such studies, making it increasingly difficult for single centers and young investigators to afford.

In recent years, many different methods have been developed to analyze different kinds of medical databases. For example, the National Health Insurance Research Database (NHIRD) in Taiwan (http://nhird.nhri.org.tw/en/) consists of comprehensive information about the health of residents in Taiwan, including sex, age, diagnosis of diseases, records of hospitalization, major medical management, and mortality. As of 2014, the NHIRD included 99.9% of the population of Taiwan. The

# Kuang-Tso Lee and Ai-Ling Hour contributed equally to this manuscript.
specific features of the database have been reported in numerous earlier studies. It contains both generalized and specific medical data of substantial importance to medical researchers, reflecting real world medicine in Taiwan. The NHIRD is used for 12.6% (290/2308 as of October 2016) of PubMed publications relating to cardiovascular diseases, and at least 15 manuscripts from Acta Cardiologic Sinica.1-15

The contents of the NHIRD are encrypted to protect personal information, and researchers need to extract interesting content from the database. The major contents include age, sex, diagnoses, admissions, prescriptions, operations, and costs. Because some quantitative data are not available, such as the level of blood hemoglobin, appropriate topics should be chosen to minimize these limitations. Investigators provide the category of encrypted data first, and then targeted information is extracted and “cleaned” to remove unwanted information. Raw data are created which can then be analyzed.

Controlled groups are matched by propensity scoring. Common statistical tools including descriptive statistics, chi-square test, analysis of variance, linear regression, logistics regression, and survival analysis, such as life tables, Kaplan-Meier estimates, and log-rank tests should be chosen on the basis of the study design.

APPLICATION OF DATABASES FOR CARDIOVASCULAR STUDIES FROM TAIWAN

The application of information derived from large databases in the research of cardiovascular diseases can include major adverse cardiovascular events (MACEs) resulting from pregnancy, multidisciplinary diseases, and post-marketing analysis of cardiological treatment. The cardiological studies which have used the NHIRD or single-center databases in our hospital are listed in Table 1. Their conclusions and influences are summarized.

![Table 1](https://example.com/table1.png)

**Table 1.** The published cardiological studies which analyzed NHIRD or single-center based databases in Chang Gung Memorial Hospital, Linkou Medical Center

<table>
<thead>
<tr>
<th>Category</th>
<th>Authors</th>
<th>Disease or intervention</th>
<th>Specific population</th>
<th>Main outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardiology</td>
<td>Li et al. (2015)29</td>
<td>Atrial fibrillation</td>
<td>Heart failure</td>
<td>MACEs</td>
</tr>
<tr>
<td></td>
<td>Tung et al. (2016)</td>
<td>Re-admission</td>
<td>Heart failure</td>
<td>MACEs</td>
</tr>
<tr>
<td></td>
<td>Lin et al. (2014)19</td>
<td>Grown-up congenital heart disease</td>
<td></td>
<td>MACEs</td>
</tr>
<tr>
<td>Woman’s health</td>
<td>Tang et al. (2009)</td>
<td>Pre-eclampsia and eclampsia</td>
<td>Pregnant women</td>
<td>Peripartum stroke</td>
</tr>
<tr>
<td></td>
<td>Lin et al. (2011)16</td>
<td>Pre-eclampsia and eclampsia</td>
<td>Pregnant women</td>
<td>Peripartum MACEs</td>
</tr>
<tr>
<td></td>
<td>Wu et al. (2014)18</td>
<td>Systemic lupus erythematosus</td>
<td>Pregnant women</td>
<td>Postpartum MACEs</td>
</tr>
<tr>
<td>Multidisciplinary</td>
<td>Tsai et al. (2012)20</td>
<td>Acute kidney injury</td>
<td>Post surgeries for aortic dissection</td>
<td>Mortality</td>
</tr>
<tr>
<td></td>
<td>Tsai et al. (2014)23</td>
<td>Acute kidney injury</td>
<td>Acute cardiovascular disease</td>
<td>MACEs</td>
</tr>
<tr>
<td></td>
<td>Chen et al. (2016)21</td>
<td>Pre-operative hemodialysis</td>
<td>Post coronary bypass surgery</td>
<td>MACEs</td>
</tr>
<tr>
<td></td>
<td>Chen et al. (2016)22</td>
<td>Acute respiratory distress syndrome</td>
<td>Post valvular heart surgery</td>
<td>Mortality</td>
</tr>
<tr>
<td></td>
<td>Chen et al. (2015)24</td>
<td>Dental procedures</td>
<td>Incidence of endocarditis</td>
<td>MACEs</td>
</tr>
<tr>
<td></td>
<td>Chou et al. (2015)26</td>
<td>Treated periodontitis</td>
<td>Incidence of atrial fibrillation</td>
<td>Renal dysfunction</td>
</tr>
<tr>
<td></td>
<td>Liu et al. (2016)27</td>
<td>Erectile dysfunction</td>
<td></td>
<td>MACEs</td>
</tr>
<tr>
<td>Post-marketing analysis of medications</td>
<td>Lin et al. (2016)28</td>
<td>Obstructive sleep apnea</td>
<td>Hypertension</td>
<td>Renal dysfunction</td>
</tr>
<tr>
<td></td>
<td>Hsiao et al. (2015)25</td>
<td>Fixed combination of anti-hypertensive medication</td>
<td>Hypertension</td>
<td>MACEs</td>
</tr>
<tr>
<td></td>
<td>Tung et al. (2015)30</td>
<td>Fixed combination of anti-hypertensive medication</td>
<td>Hypertension</td>
<td>Clinical costs</td>
</tr>
<tr>
<td>Biomarkers</td>
<td>Tsai et al. (2015)31</td>
<td>High-sensitivity C-reactive protein</td>
<td>Metabolic syndrome</td>
<td>Arterial stiffness</td>
</tr>
<tr>
<td></td>
<td>Yang et al. (2016)35</td>
<td>Urinary biomarkers</td>
<td>Heart failure</td>
<td>Detection of acute kidney injury</td>
</tr>
</tbody>
</table>
in the following paragraphs.

**Pregnancy medicine and cardiovascular diseases**

Safety is the first priority among most earlier clinical studies. Thus, researchers may often have difficulty in collecting data focusing on women’s health, especially during a woman’s childbearing years. The ultimate experimental design, therefore, should be carefully crafted to avoid or minimize any adverse effects for mothers and their fetuses. Thus, studies enrolling this patient population tend to be retrospective and small in scale. This causes the research to have minimum “power,” thus making hypotheses difficult to confirm. However, with such a comprehensive database as the NHIRD, study “power” is typically not a issue in that a large sample size and real world relevance are apparent.

For example, pre-eclampsia is a disease specific to women. In such cases, the vascular bed of the placenta is abnormal, leading to an elevated systemic concentration of vaso-active cytokines and subsequently the development of hypertension and proteinuria. The underlying environment may be thrombogenic, leading to a higher prevalence of cardiovascular events. Our studies found a strong correlation between pre-eclampsia, stroke, and MACEs. This connection could have been difficult to establish using traditional cohort and single-center studies, particularly when the follow-up period and patient numbers were limited.

Another issue of gestational health is systemic lupus erythematosus. Such patients are typically female, and usually in child-bearing years. In general, little medical evidence is available on pregnant lupus patients, and the cardiovascular risk factors in this group have rarely been given much attention. By using the NHIRD, we found that post-delivery lupus patients had about a 10-fold elevated cardiovascular risk compared to the control group, a finding which helped us to improve clinical decision making.

These two studies highlight that despite the existence of fair cooperation between hospitals, the sample size of patients with certain diseases is relatively small, and the follow-up period is usually insufficiently long. Thus, correlation or causal relationships are usually hard to establish. Fortunately, analysis of a large database can provide an initial impression about possible correlations. With a rigorous research design, database analysis can be used not only for MACEs in gestational women but also for patients with surgically corrected congenital heart disease, or to obtain useful medical information about children, the very elderly, or other vulnerable groups.

**Multidisciplinary research projects in cardiovascular studies**

In the traditional medical research environment, investigators typically focus on research within their specific expertise. In many cases, however, such an approach may not be sufficient to clarify certain issues involving multi-departmental or multi-disciplinary challenges with complicated data. Researchers tend to ignore information that appears unrelated to the specific topic such as geography or regional climate, however these seemingly unrelated factors may be critical to reach a correct result and conclusion. As an example, acute kidney injury (AKI) may complicate the course of aortic dissection. In order to clarify this association, cooperation was necessary between nephrologists and vascular surgeons. The relationships between AKI and bypass surgery, valvular operation, and acute cardiovascular diseases have also been elucidated in this manner. Synchronizing and analyzing databases could facilitate such research; even for apparently distinct sub-special subjects.

Dental disease is considered to be a risk factor contributing to endocarditis. However, establishing and confirming a relationship requires expertise in many subjects, which is a challenge for single-center studies. Dentistry has been incorporated into the NHIRD, and comprehensive cohort data can be obtained. According to our research, severe periodontitis is an adverse prognostic factor for cardiovascular outcomes. Even though annually dental scaling may have a protective role for infective endocarditis and other cardiac events, dental procedures including tooth extraction, periodontal, and endodontic therapies, have been shown to have a neutral effect on the prevalence of subsequent hospitalization for endocarditis. Other factors such as temperature, height, climate, or the economy can also be integrated with medical issues. Therefore, analyzing information from a comprehensive database is valuable for researchers who are dedicated to public health. Additionally, erectile dysfunction, sleep apnea, and new-onset atrial fibrillation have been evaluated to
ascertain their association with MACEs, to provide useful treatment references.

Another category in which the NHIRD is very valuable for research is stroke. Chao et al. studied the use of risk scores and the age threshold for using anticoagulation therapy for Taiwanese patients to prevent stroke.33-37 The population was the largest to date in East Asia, and the results highlighted the importance of racial differences in the era of atrial fibrillation. Importantly, a larger study tried to determine the net effect of using anticoagulants to prevent thromboembolic events for patients with previous intracranial hemorrhage.38 Moreover, Chang et al. documented an increased risk of stroke in patients with atrial fibrillation not recommended for systemic anticoagulants by the current guidelines.39 Researchers can be encouraged to conduct further prospective studies focusing on this group.

Post-marketing analysis of medications

Publication of medication information based upon any study, large or small, requires serious evaluation. Safety, efficacy and adverse events should be confirmed by highly reliable evidence. This evidence can be obtained by academic research, which often requires extensive financial support and labor. However, due to the possibility of uncertain results, many pharmacological companies take great risks to execute and publish post-marketing research, especially when the outcomes do not favor their products. Independent individual clinicians are often interested in differences between similar categories of drugs, to facilitate their practice and improve patient care. Some of this information can be derived from the analysis of large databases. For example, the combined form of anti-hypertensive drugs may be superior in clinical outcomes to individual medications, in that despite similar components, the use of the drugs in combination improves patient compliance and convenience.26,30,40 Such a method could also analyze the real-world effects of some popular categories of medications, for example, novel oral anticoagulants, statins, oral hypoglycemic agents, or even agents used to treat rare diseases.31-44 Results relating to this kind of analysis would establish a basis for further prospective studies.

Build specific databases

The NHIRD comprises many different categories of personal information, events, and procedures, but lacks the specific examination findings such as hemoglobin, ejection fraction of the left ventricle, or quantification of urine protein. Therefore, not all issues can be researched from such a tool, and establishing databases with specific information is essential. Such databases may have been collected for a certain purpose, or may just consist of daily recordings from clinical practice. The principal problem is determining exactly how comprehensive the included information should be. When these databases were being assembled, not all components are considered essential and thus not gathered. Inadequate collection of clinical data can limit later utility. Clinicians often try to obtain as comprehensive laboratory data as possible, but too labile collections may interfere with the quality. Therefore, “bio-banks” containing samples of blood or tissue may be a novel solution to some of these problems. We collect samples of blood for new biomarkers, and compare them to available clinical information. Occasionally, interesting and useful findings have been derived from this information. Although this method cannot replace controlled prospective trials, it can help to prevent undue expenditure of effort, and provide further direction into areas that require additional research.

Databases originate from a variety of different sources, such as a healthcare center to demonstrate in metabolic syndrome whether highly sensitive C-reactive protein can independently predict arterial stiffness,45 a heart failure center studying the obesity and mortality paradox,46,47 a focus on readmission and prognosis,48 applications of urinary biomarkers,49 and a cardiovascular surgery center for the diagnosis of aortic dissection.20 All of these database sources can improve the diagnosis and prognosis of cardiovascular diseases.

LIMITATIONS AND BREAKTHROUGHS OF DATABASES

There are many limitations to the NHIRD. For example, the lack of information about patient lifestyles, personal patient details following decoding, and ultimate mortality due to privacy issues. Correspondingly, the most important database breakthrough should be linked with a laboratory database, and further blood or urine storage samples.
The results of different NHIRD studies, which are based on the same database, are greatly influenced by various exposures, patient selection, matching, statistic methods, and may lead to controversial conclusions. Therefore, we suggest that the characteristics of the targeted patients should be specified. Importantly, the basic study hypothesis should be well studied and considered, not only according to the association, but by means of rational cause-and-effect design.

Large, population-based database inquiries are typically retrospective studies which seek to describe potential associations. The evaluation and investigation which arises from biochemical or molecular samples may offer opportunities to identify disease-causing mechanisms. In addition, appropriate mapping techniques between heterogeneous databases can complement each other.

Other types of databases, such as a climate or air pollution databases can provide various information regarding the environment or natural resources. The spatial and/or temporal models for diseases can be used to predict disease prevalence under specific conditions, and the medical resources can be prepared promptly.

CONCLUSIONS

Analyzing databases is a developing technique with great potential in modern medical research. Similar to all scientific methodology, such database analysis should only be used to test hypotheses with meaningful clinical utilization. However, replication or over explanation of statistical results should be avoided if possible. Under rigorous design study, such database reviews can provide useful information to certain questions, highlight the relationships between components, and provide worthwhile speculation regarding future research.

ACKNOWLEDGEMENT

This work was supported by the National Science Council of Taiwan (Project NMRPG3F0561 and NMRPG 3G6181) and Chang Gung Memorial Hospital (Project CMRPG3E1051). We appreciate Wayne Chu and Yuka Chu’s critical reading.

REFERENCES


