Int J Heart Fail. 2021 Jul;3(3):194-196
https://doi.org/10.36628/ijhf.2021.0026
pISSN 2636-154X·eISSN 2636-1558

International Journal of Heart Failure IJHF

Editorial

Check for updates

Using Big Data to Understand Rare Diseases

Jun-Bean Park, MD, PhD[1,2]

[1]Department of Internal Medicine, Seoul National University College of Medicine, Seoul, Korea
[2]Cardiovascular Center, Seoul National University Hospital, Seoul, Korea

▸ See the article "Incidence, Cause of Death, and Survival of Amyloidosis in Korea: A Retrospective Population-Based Study" in volume 3 on page 172.

OPEN ACCESS

Received: Jun 30, 2021
Accepted: Jul 15, 2021

Correspondence to
Jun-Bean Park, MD, PhD
Cardiovascular Center, Seoul National University Hospital and Department of Internal Medicine, Seoul National University College of Medicine, 101 Daehak-ro, Jongno-gu, Seoul 03080, Korea.
E-mail: nanumy1@gmail.com

ORCID iDs
Jun-Bean Park
https://orcid.org/0000-0003-4053-8713

Conflict of Interest
The author has no financial conflicts of interest.

Amyloidosis is a life-threatening systemic disease, with a median survival of less than 1 year for light chain amyloidosis and 4 years for wild-type transthyretin amyloidosis.[1] Since heart is the most commonly involved organ in both types of amyloidosis (although the kidney is the most frequently involved organ for AA amyloidosis),[2] this disease entity has attracted the cardiologists. With the recent emergence of several novel therapies reported to improve prognosis of patients with amyloidosis, the importance and clinical significance of early diagnosis in amyloidosis have been more emphasized than before.[3-5] Despite the need for better understanding the epidemiology of amyloidosis, there has been no population-based epidemiologic data in Korea, yet. One of the major limitations in dealing with rare diseases such as amyloidosis is that the patients are geographically scattered around the country and are usually managed by multiple centers with their own specialized multidisciplinary teams for rare diseases. It is thus not surprising that the primary source of data on epidemiology, natural history, and treatment outcomes of patients with rare diseases is often based on retrospective case reports or series from single or few expert centers,[1] which are subject to referral or selection bias. Furthermore, although case reports and series can provide valuable information, they are also likely to suffer from information bias, since a complete set of clinical data is generally not available in their medical records. A more ideal way of investigating rare disease is to set up robust prospective registries, allowing the collection of a large list of predefined variables and outcomes in a standardized manner. However, the construction and maintenance of prospective registries may take substantial time and effort, which makes it difficult to be started as a first step toward exploring an overview of rare diseases.

Claims data-based observational studies have recently emerged as a novel tool for medical research, which has several advantages over traditional research tools, in particular the virtue of providing real-world evidence. In Korea, the National Health Insurance Service (NHIS) constructed the nationwide claims database, named the Korean National Health Information Database (NHID).[6] This database has many strengths, the first being the large volume of collected data, which is one of the largest claims datasets worldwide, consisting of more than 52 million people. In addition, the Korean NHIS is implementing rare intractable disease (RID) program to support patients with these disease entities, enabling research on rare diseases, with a sufficient sample size allowing a better, comprehensive understanding of the disease.

In this issue of the *International Journal of Heart Failure*, Jang et al.[7] used Korean NHIS data from 2006 through 2017 to study amyloidosis. et al.The nationwide nature of this database

afforded an adequate sample size to analyze the incidence, survival rate and cause of death of amyloidosis, and thus this article can provide a reliable overview of the current status of amyloidosis in Korean population. Considering that this is the first study providing an "overview" of amyloidosis in Korea, the authors deserve kudos for this work. However, this study has important limitations, which are inherent to all research using large administrative databases. The first is the completeness and accuracy of data used for analysis and similar concerns are present in this article. For instance, a considerable portion of diagnoses might be incorrect, given the difficulty in diagnosing amyloidosis and the presence of many photocopies.[8] To enhance the accuracy of diagnosing amyloidosis, it can be recommended to combine RID code (V121 for amyloidosis) to the International Classification of Diseases (ICD) code (E85 for amyloidosis).[9] This is supported by the fact that the accuracy, sensitivity, and specificity of this approach applied to define hypertrophic cardiomyopathy, another rare disease covered by RID program, was reported to be 92.6%, 91.5%, and 100%, respectively.[10]

The study by Jang et al.[7] shows both the opportunities and challenges of using big administrative databases. The data cannot be completely validated at a time and validation over time may be needed, representing an opportunity for future research. Although reviewing whether randomly selected patients with diagnostic codes of interest or operational definitions of interest truly have diseases of interest is a time- and labor-intensive process, it can be the first step in validating the accuracy of the administrative data. As a next step, developing a large-scale biomedical database and research resource by integrating imaging, pathologic, and genomic data with administrative health datasets, such as the U.K. Biobank and The All of Us Research Program, can be a huge leap forward in the field of big data research.[11][12] However, achieving this is a huge challenge as it requires a multifaceted solution for political, legal, and technical issues in data linking. Before completing this challenge, the administrative data will be predominantly used in a stand-alone manner, which may impose limitations to both the analysis and interpretation of research findings. Indeed, Jang et al.[7] aimed to provide data on incidence, survival rate and cause of death of amyloidosis, however, the lack of information on amyloidosis subtypes, which can be readily discerned by using imaging, pathologic, and genomic data, resulted in a heterogeneity of the study population. Since the treatment strategies and outcomes among patients with different subtypes of amyloidosis differ substantially, the impact of the study is limited and should be extrapolated with careful understanding and interpretation of the disease. Furthermore, for the same reason, this study cannot provide a solid basis for planning future research in amyloidosis.

Since we are on the way toward increasing the level of understanding and in-depth knowledge about amyloidosis, the present study by Jang et al.[7] may represent an initial stepping-stone in conducting more accurate and elegant research fed with large datasets. From a broader perspective, this study demonstrates how rare diseases, such as amyloidosis, can be effectively studied on a large-scale using data from the NHID, paving the way for future studies to be performed in this direction. At the heart of the challenge is the need for critical validation of the data from the NHID and its integration with other relevant datasets.

REFERENCES

1. Alexander KM, Orav J, Singh A, et al. Geographic disparities in reported US amyloidosis mortality from 1979 to 2015: potential underdetection of cardiac amyloidosis. JAMA Cardiol 2018;3:865-70.
PUBMED | CROSSREF

2. Falk RH, Comenzo RL, Skinner M. The systemic amyloidoses. N Engl J Med 1997;337:898-909.
PUBMED | CROSSREF
3. Alexander KM, Singh A, Falk RH. Novel pharmacotherapies for cardiac amyloidosis. Pharmacol Ther 2017;180:129-38.
PUBMED | CROSSREF
4. Gertz MA. Immunoglobulin light chain amyloidosis: 2016 update on diagnosis, prognosis, and treatment. Am J Hematol 2016;91:947-56.
PUBMED | CROSSREF
5. Maurer MS, Schwartz JH, Gundapaneni B, et al. Tafamidis treatment for patients with transthyretin amyloid cardiomyopathy. N Engl J Med 2018;379:1007-16.
PUBMED | CROSSREF
6. Park JB, Kim DH, Lee H, et al. Mildly abnormal lipid levels, but not high lipid variability, are associated with increased risk of myocardial infarction and stroke in "Statin-Naive" young population a nationwide cohort study. Circ Res 2020;126:824-35.
PUBMED | CROSSREF
7. Jang SY, Kim D, Choi JO, Jeon ES. Incidence, cause of death, and survival of amyloidosis in Korea: a retrospective population-based study. Int J Heart Fail 2021;3:172-8.
CROSSREF
8. Lee SP, Park JB, Kim HK, Kim YJ, Grogan M, Sohn DW. Contemporary imaging diagnosis of cardiac amyloidosis. J Cardiovasc Imaging 2019;27:1-10.
PUBMED | CROSSREF
9. Choi EK. Cardiovascular research using the Korean National Health Information Database. Korean Circ J 2020;50:754-72.
PUBMED | CROSSREF
10. Park JB, Kim DH, Lee H, et al. Obesity and metabolic health status are determinants for the clinical expression of hypertrophic cardiomyopathy. Eur J Prev Cardiol 2020;27:1849-57.
PUBMED | CROSSREF
11. Sudlow C, Gallacher J, Allen N, et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 2015;12:e1001779.
PUBMED | CROSSREF
12. All of Us Research Program Investigators, Denny JC, Rutter JL, et al. The "All of Us" Research Program. N Engl J Med 2019;381:668-76.
PUBMED | CROSSREF