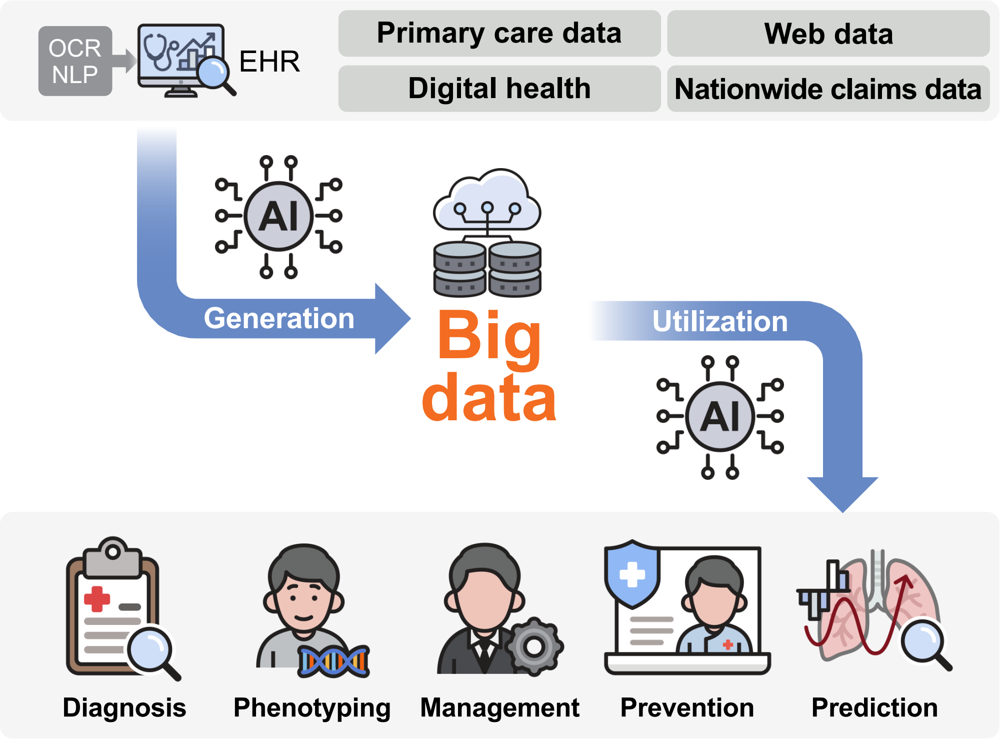
Big Data Research on Severe Asthma

https://doi.org/10.4046/trd.2023.0186
ISSN: 1738-3536(Print)/
2005-6184(Online)
Tuberc Respir Dis 2024;87:213-220

Sang Hyuk Kim, M.D.[1] and Youlim Kim, M.D., Ph.D.[2]

[1]Division of Pulmonary, Allergy, and Critical Care Medicine, Department of Internal Medicine, Dongguk University Gyeongju Hospital, Dongguk University College of Medicine, Gyeongju, [2]Division of Pulmonary and Allergy, Department of Internal Medicine, Konkuk University Hospital, Konkuk University School of Medicine, Seoul, Republic of Korea



Address for correspondence
Youlim Kim, M.D., Ph.D.
Division of Pulmonary and Allergy, Department of Internal Medicine, Konkuk University Hospital, Konkuk University School of Medicine, 120-1 Neungdong-ro, Gwangjin-gu, Seoul 05030, Republic of Korea
Phone 82-2-2030-7524
Fax 82-2-2030-7748
E-mail weilin810707@gmail.com

Received Nov. 18, 2023
Revised Jan. 11, 2024
Accepted Feb. 29, 2024
Published online Mar. 5, 2024

Abstract

The continuously increasing prevalence of severe asthma has imposed an increasing burden worldwide. Despite the emergence of novel therapeutic agents, management of severe asthma remains challenging. Insights garnered from big data may be helpful in the effort to determine the complex nature of severe asthma. In the field of asthma research, a vast amount of big data from various sources, including electronic health records, national claims data, and international cohorts, is now available. However, understanding of the strengths and limitations is required for proper utilization of specific datasets. Use of big data, along with advancements in artificial intelligence techniques, could potentially facilitate the practice of precision medicine in management of severe asthma.

Keywords: Severe Asthma; Big Data; Digital Technology; Multicenter Electronic Health Records; Nationwide Claims Data; International Cohorts

Introduction

Recent advancements in science and technology have led to exponential expansion of the healthcare data[1]. Data reaching complexity and size beyond the capacity of conventional data processing methods is known as big data[2]. Compared with traditional small datasets, the advantages of big data include cost-effectiveness and reduced reliance on statistical estimation, assumptions, and adjustments[3,4]. It may also provide more accurate information regarding patients' real-world experiences[5]. Thus, utilization of big data may lead to conduct of innovative trials and novel discoveries in a variety of research fields, including respiratory medicine[6].

Although only a small portion of patients with asthma are affected by severe asthma, its burden is significant[7,8]. Thanks to the development of newer therapeutic agents, management of severe asthma has shown significant improvement in recent years[9,10]. However, research conducted on severe asthma is still limited, and can be challenging due to the limited number of patients with severe asthma. Fortunately, advances in digital technology have facilitated the collection of big data, providing new opportunities for conduct of research on severe asthma[11]. However, an understanding of the strengths and limitations inherent to each dataset is required for effective utilization of big data. This review will focus on sources of big data as well as relevant research on severe asthma.

Available Dataset

We propose classifying sources of big data for use in research on severe asthma according to their source of generation: electronic health records (EHRs), claims data, and cohort data. These sources can be further classified based on their origin within the category of EHR data, such as single center versus multicenter and primary care versus referral hospital data. In a similar manner, claims and cohort data can be divided according to single-country versus multiple-country datasets. The strengths and weaknesses of the available dataset are summarized in Table 1.

Table 1. Available big data and their strengths and weaknesses

Dataset	Type	Strength	Weakness
Optum	EHR	Longitudinal data Detailed clinical information Linkage to other datasets	High cost for data utilization Lack of data standardization
Optimum Patient Care	EHR	Detailed clinical information Collaborative network	Lack of data standardization Mostly primary care-based population
USA CMS	Claims data	Large number Longitudinal data	Uncertainty in asthma diagnosis Lack of clinical details Mostly elderly (≥65 years) and low-income population
IBM MarketScan	Claims data	Large number Linkage to other datasets	High cost for data utilization Uncertainty in asthma diagnosis Lack of demographic details (e.g., race and income)
Korea NHIS/HIRA	Claims data	Large number Longitudinal data Nationwide data Standardized data Linkage to other datasets	Uncertainty in asthma diagnosis Lack of clinical details
ISAR	Claims	Globalized data Longitudinal data Patients-centered, high-quality, and standardized data Updated quickly	Relatively small number Unequal ethnic registration

EHR: electronic health record; CMS: Centers for Medicare and Medicaid Services; NHIS: Korean National Health Insurance Service; HIRA: Health Insurance Review and Assessment; ISAR: International Severe Asthma Registry.

1. Electronic health record data

Recent advances in artificial intelligence (AI) technology and cloud computing have led to the emergence of commercial big data companies, and several software programs have been developed for extraction of various types of information from EHRs[12]. These technological breakthroughs have enabled the conduct of research on severe asthma utilizing large-scale EHR data. One weakness is that the quality of data can vary across the healthcare providers. In addition, the cost of extracting EHRs data can be substantial. The specific clinical setting from which EHR data originate may affect the representativeness of the data, which could lead to selection bias. However, due to the abundance of clinical information, including pulmonary function tests, large number of patients, and long-term duration of follow-up, this dataset can be considered useful for providing real-world evidence.

Optum, a healthcare services and technology company based in the United States, which offers 'Optum EHR data[13],' is one of the largest healthcare organizations worldwide, with a database of nearly 4 million EHR data, of which 3.2 million are linked to claims data. The dataset includes a wide range of covariates, including pulmonary function tests, laboratory findings, and healthcare costs, thus, it is regarded as a valuable source of medical information on severe asthma. In addition, the Optimum Patient Care (OPC) dataset, a not-for-profit social enterprise based in the United Kingdom (UK) that manages the OPC research database, is also available[14]. This dataset contains de-identified EHR data from general practices across the UK, comprising 22 million patients from over 1,000 general practice sites. The dataset has a well-established collaborative network, which offers a unique advantage. Collection of data by the OPC has expanded beyond the boundaries of UK, even including Australia. The International Severe Asthma Registry (ISAR) is also based on the OPC system for collection of data[15].

2. Claims data

Claims data is automatically generated in the healthcare process, and well-organized nationwide claims datasets have already established in some countries[16]. Despite limitations to accessibility of data, its strengths include size, speed, cost-efficiency, and expandability to other datasets. However, the lack of clinical details may reduce its usefulness, necessitating the use of operational definitions for identifying specific patient groups or events for study. For example, differentiating between difficult-to-treat asthma and severe asthma can be challenging based solely on International Classification of Disease 10th revision codes. Integrating prescription data and healthcare utilization patterns may be helpful in the effort to address this issue, making the data more suitable for use in conducting of robust research.

The USA Centers for Medicare and Medicaid Services (CMS)[17], a federal agency that offers a government-funded healthcare program covering almost 95% of United States citizens aged 65 or older, is a well-known example of claims data. Data collected from this program include healthcare utilization, costs, disease outcomes, and beneficiary demographics. Due to the long history of this system, longitudinal studies have been conducted over extended periods. However, a significant limitation of CMS data is its primary focus on the elderly population, which may not fully represent the broad spectrum of the disease. IBM (International Business Machines Corporation) MarketScan Research Databases, formerly known as Truven Health MarketScan Research Databases, which also provide claims data in the United States[18], includes reimbursed healthcare claims data from more than 250 private health insurance plans for employees, retirees, and their dependents. They are also linked to other useful data, including EHR and mortality. This dataset covers more than 30 million people, excluding those enrolled in medicaid health insurance plans. However, accessing this dataset can be costly, and it raises issues regarding generalizability due to limited coverage from private insurance plans, which may result in selection bias.

The Korean National Health Insurance Service (NHIS) and Health Insurance Review and Assessment (HIRA) database, in South Korea, is a comprehensive claims dataset[19]. This database, which is administered by the Korean government, covers almost 97% of the Korean population through a compulsory universal health insurance program. The greatest strength of these datasets is the universal coverage of Korean citizens, ensuring a high level of representativeness. It also includes regular health check-up data, providing valuable information, including anthropometric measures, changes in personal habits (smoking, alcohol drinking, and physical activity), and laboratory data[20,21]. In addition, it can also be linked to other datasets, including the Korea National Health and Nutrition Examination Survey[22]. However, its historical span is not as long as that of CMS data. Like other claims datasets, the lack of clinical details, including pulmonary function tests, is also a weakness of the NHIS and HIRA databases.

3. Cohort data

Although cohort data are not typically considered big data, due to the relatively lower threshold for what is considered ‘big’ in such diseases, they may be considered in the case of certain rare diseases. Collaboration among multinational cohorts can generate a large dataset comparable to that of conventional big data. Establishment and maintenance of a nationwide severe asthma cohort has been achieved in recent years. Efforts to combine severe asthma cohorts from multiple countries for creation of an international registry, the ISAR, are underway. The goal of the ISAR, which plays an important role in conduct of research on severe asthma[15]. One significant advantage of this international cohort is accuracy in diagnosing severe asthma and the abundance of severe asthma-related data. However, the number of patients may be smaller when compared with EHR data, and primary care patients are generally not included. Nevertheless, during the first three years, more than 5,000 patients with severe asthma were recruited by the ISAR. Because this cohort is ongoing, it is expected that its volume will eventually be comparable to that of other big data. Of particular importance, the ISAR also provides detailed demographics and medication data, including the use of biologics, an important medication in management of severe asthma. Therefore, the results of novel treatment, such as newer biologics, can be rapidly updated in dataset. The remaining challenges is recruitment of patients from diverse ethnicities and regions to ensure acquisition of unbiased data for the specific race or country.

Appropriate Big Data Utilization in Severe Asthma Research

1. Example 1

EHR data can be used for acquisition of real-world data on treatments for severe asthma. Using the Optum dataset, Jeffery et al.[23] attempted to determine the risk of exacerbation in patients with severe asthma following discontinuation of biologics. In the results of the study, which included approximately 2,500 patients with severe asthma who had used biological agents for more than 6 months, a similar risk of asthma exacerbation was observed between stoppers and continuers.

By contrast, the results of a randomized controlled trial (RCT) examining the use of mepolizumab conflicted with findings reported by Jeffery et al.[23], suggesting a potential association between discontinuing mepolizumab and an increased risk of exacerbation[24]. This discrepancy could be due to differences in the study populations, methodologies, or definitions. Despite the possibility of biased outcomes, the use of well-organized study methods using EHR data could potentially lead to replacement of several RCTs[25].

EHR dataset can also be helpful in the conduct of a novel type of research on severe asthma. For example, Ryan et al.[26], examined the treatment patterns and outcomes for patients with the potential for having severe asthma based on their referral status. The study, which combined data from Optimum Patient Care Research Database (OPCRD) and ISAR, enrolled more than 200,000 patients who underwent treatment for asthma in primary care and referral hospitals. Based on the findings of the study, many patients with the potential for having severe asthma in primary care who were not acknowledged could benefit from referral to an asthma specialist. Conduct of similar study, including a large number of patients representing different levels of care, using traditional datasets is not possible.

2. Example 2

Comprehensive information on drug use nationwide from claims data may be useful in the conduct of pharmacoepidemiologic research on severe asthma. Hong et al.[27] used claims data in evaluating the cost-effectiveness of adding tiotropium to the inhaled corticosteroid/long-acting beta-2-agonist in elderly patients with severe asthma. According to their findings, the incremental cost-effectiveness ratio was 60,074 US dollars/quality-adjusted life years, suggesting the cost-effectiveness of add-on tiotropium for treatment of severe asthma. Such findings can provide knowledge that can be used in decision making in treatment of severe asthma, including approval for new drugs and national insurance coverage for costly treatments.

Due to large size and extended follow-up period, claims data can be used for evaluating long-term outcomes, and it can be regarded as an effective and affordable approach to conduct of mortality studies. Lee et al.[28] conducted a study using Korea NHIS data for tracking the long-term outcomes for patients with oral corticosteroid (OCS)-dependent asthma. The result of the study, which included almost 10,000 patients, indicated that death was more common among these patients compared to those with OCS-independent asthma during follow-up periods of more than 10 years. These types of studies are considered more reliable due to their larger sample size and cost-effectiveness compared to the cohort study or RCTs.

3. Example 3

Data from a global cohort can be utilized to study the characteristics of severe asthma worldwide. Wang et al.[29] revealed on the diverse nature of severe asthma across different countries using the ISAR data. The strengths of this study include the diverse registry data, and representative of real-world patients compared to data collected in RCTs. The global cohort also can be used in examination of racial and ethnic differences in severe asthma. Chen et al.[30] reported significant disparities in the use of biologics among different racial groups. The findings of these studies demonstrated that global data on severe asthma can be helpful in the effort to understand how regional and cultural characteristics are compared internationally, which can be regarded as valuable information in the practice of precision medicine[31,32].

Current and Future Directions in Big Data Research on Severe Asthma

1. Current engagement in big data research

The usefulness of data based on the purpose should be determined prior to initiation of research on severe asthma using big data[33], which can be critical in development of a meaningful and effective research plan. Use of prospective data can provide a valuable opportunity for evaluating treatment outcomes, enabling robust examination of causal relationships[34]. However, when the objective is to evaluate epidemiological outcomes, use of large-scale cross-sectional data would be the preferred choice[35]. In addition to data compatibility, close collaboration between clinicians and data scientists is important for optimizing the outcomes. The success of the collaboration will depend on effective communication and mutual understanding between experts. Clinicians provide in-depth knowledge of the domain, insights regarding patients, and medical expertise, while data scientists contribute their analytical knowledge and data manipulation skills. Collaborative synergies of these two different domains could enable a comprehensive approach to treatment of severe asthma, which will lead to determination of the phenotype and endotype of severe asthma, identifi-

Figure 1. Precision medicine for management of severe asthma using big data.

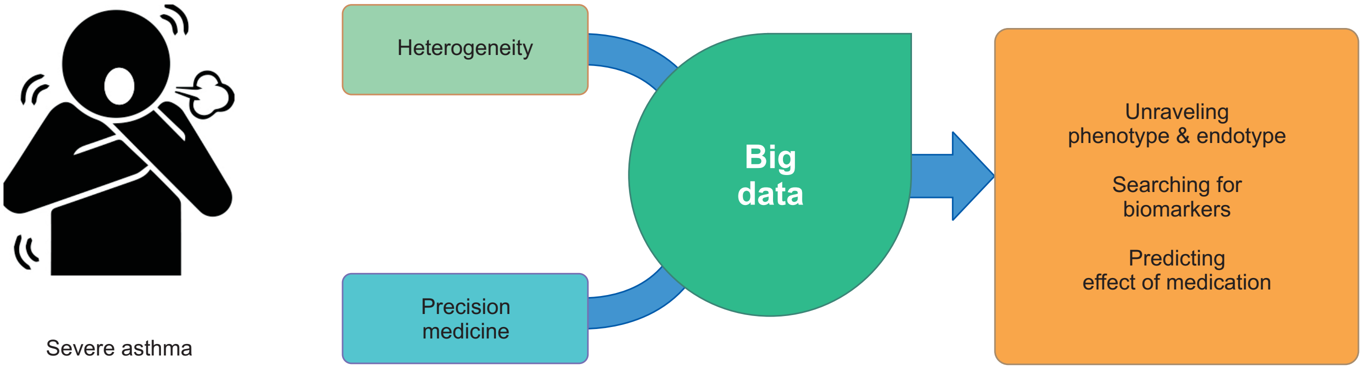
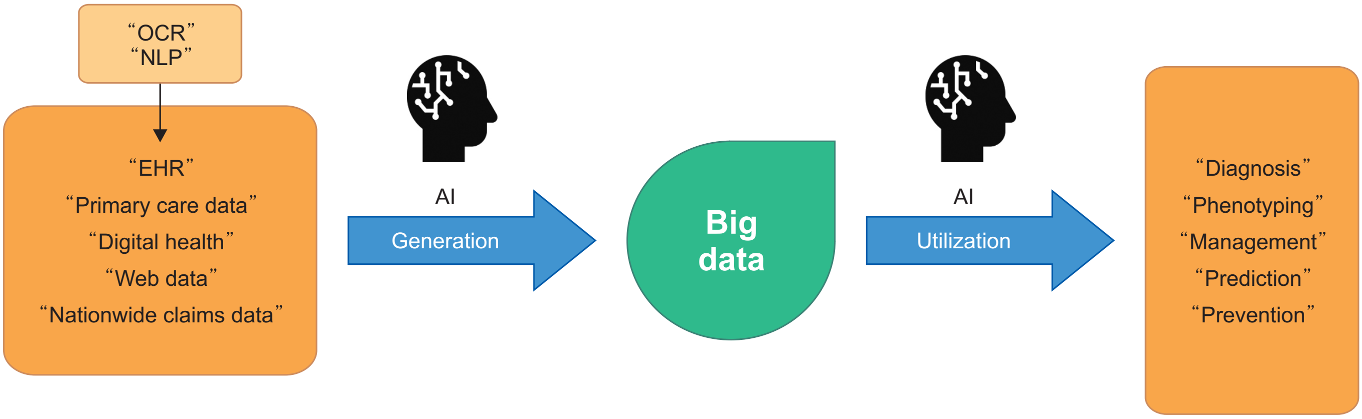


Figure 2. Generation and utilization of big data using artificial intelligence (AI). OCR: optical character recognition; NLP: natural language processing; EHR: electronic health record.



cation of novel biomarkers, and prediction of the medication effects (Figure 1).

2. Artificial intelligence

With the continuing development of AI technology, both the generation and utilization of big data will include appropriate use of AI (Figure 2). Optical character recognition (OCR) can be utilized effectively when using non-digital data. OCR is a technology used to convert different types of documents, including hand-written documents, facilitating the transition of traditional medical records into EHRs[36]. Use of natural language processing, a branch of AI, can enables comprehension of human language based on algorithms by the computer for identification of natural language rules. Use of these new technologies can enable interpretation and organization of unstructured data, including free-text narratives in medical charts, thereby optimizing the use of EHRs in conduct of medical research. This effort may lead to attainment of critical clinical insights from large and non-standardized EHR datasets for improvement of research and clinical practice in treatment of severe asthma[37].

Phenotyping, diagnosis, and management of severe asthma pose numerous challenges, which might be supported with use of AI technologies. Clustering of diseases, extraction of essential features, and prediction of outcomes can be performed without the direction of the researcher using unsupervised AI algorithms[38]. Combining new AI techniques and big data could lead to development of an innovative approach to treatment of severe asthma research[39]. However, because a significant portion of the available data, which contains an abundance of clinical information, is cross-sectional or retrospective, prospective longitudinal data essential for developing robust prediction models of treatment responses is limited. In addition to the longitudinal dataset, concerted efforts toward integration and consolidation of sparse data into a unified framework will also be required. Utilization of big data, along with advancements in AI techniques, will be helpful to clinicians in making prudent decisions in management of severe asthma.

Conclusion

The potential for use of big data as a valuable resource in conduct of research on severe asthma is significant. Thoughtful consideration of the types and design of research questions is required for leveraging the strengths of big data. With the assistance of AI technologies, insights garnered from use of big data can be helpful in the practice of precision medicine for treatment of severe asthma.

Authors' Contributions

Conceptualization: Kim Y. Methodology: Kim SH. Investigation: Kim Y. Writing - original draft preparation: Kim SH. Writing - review and editing: all authors. Approval of final manuscript: all authors.

Conflicts of Interest

Sang Hyuk Kim is an early career editorial board member of the journal, but he was not involved in the peer reviewer selection, evaluation, or decision process of this article. No other potential conflicts of interest relevant to this article were reported.

Funding

No funding to declare.

References

1. Roski J, Bo-Linn GW, Andrews TA. Creating value in health care through big data: opportunities and policy implications. Health Aff (Millwood) 2014;33:1115-22.
2. De Mauro A, Greco M, Grimaldi M. A formal definition of big data based on its essential features. Libr Rev 2016; 65:122-35.
3. Mallappallil M, Sabu J, Gruessner A, Salifu M. A review of big data and medical research. SAGE Open Med 2020;8:2050312120934839.
4. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Inf Sci Syst 2014;2:3.
5. Okada M. Big data and real-world data-based medicine in the management of hypertension. Hypertens Res 2021;44:147-53.
6. Shin SY. Current status and future direction of digital health in Korea. Korean J Physiol Pharmacol 2019;23:311-5.
7. Agache I, Akdis CA, Akdis M, Canonica GW, Casale T, Chivato T, et al. EAACI biologicals guidelines: recommendations for severe asthma. Allergy 2021;76:14-44.
8. Lim GN, Allen JC, Tiew PY, Chen W, Koh MS. Healthcare utilization and health-related quality of life of severe asthma patients in Singapore. J Asthma 2023;60:969-80.
9. Brusselle GG, Koppelman GH. Biologic therapies for severe asthma. N Engl J Med 2022;386:157-71.
10. Kim SH, Kim Y. Tailored biologics selection in severe asthma. Tuberc Respir Dis (Seoul) 2024;87:12-21.
11. Cozzoli N, Salvatore FP, Faccilongo N, Milone M. How

can big data analytics be used for healthcare organization management?: literary framework and future research from a systematic review. BMC Health Serv Res 2022;22:809.

12. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. J Big Data 2019;6:1-25.

13. Wallace PJ, Shah ND, Dennen T, Bleicher PA, Crown WH. Optum labs: building a novel node in the learning health care system. Health Aff (Millwood) 2014;33:1187-94.

14. Lynam A, Curtis C, Stanley B, Heatley H, Worthington C, Roberts EJ, et al. Data-resource profile: United Kingdom Optimum Patient Care Research Database. Pragmat Obs Res 2023;14:39-49.

15. FitzGerald JM, Tran TN, Alacqua M, Altraja A, Backer V, Bjermer L, et al. International severe asthma registry (ISAR): protocol for a global registry. BMC Med Res Methodol 2020;20:212.

16. Kim JA, Yoon S, Kim LY, Kim DS. Towards actualizing the value potential of Korea Health Insurance Review and Assessment (HIRA) data as a resource for health research: strengths, limitations, applications, and strategies for optimal use of HIRA data. J Korean Med Sci 2017;32:718-28.

17. Hennessy S, Leonard CE, Palumbo CM, Newcomb C, Bilker WB. Quality of medicaid and medicare data obtained through Centers for Medicare and Medicaid Services (CMS). Med Care 2007;45:1216-20.

18. Bolognesi MP, Habermann EB. Commercial claims data sources: PearlDiver and individual payer databases. J Bone Joint Surg Am 2022;104(Suppl 3):15-7.

19. Park JS, Lee CH. Clinical study using Healthcare Claims Database. J Rheum Dis 2021;28:119-25.

20. Choi H, Kim SH, Han K, Park TS, Park DW, Moon JY, et al. Association between exercise and risk of cardiovascular diseases in patients with non-cystic fibrosis bronchiectasis. Respir Res 2022;23:288.

21. Kim HK, Song SO, Noh J, Jeong IK, Lee BW. Data configuration and publication trends for the Korean National Health Insurance and Health Insurance Review & Assessment Database. Diabetes Metab J 2020;44:671-8.

22. Kang HS, Kim JY, Park HJ, Jung JW, Choi HS, Park JS, et al. E-cigarette-associated severe pneumonia in Korea using data linkage between the Korea National Health and Nutrition Examination Survey (KNHANES, 2013-2019) and the National Health Insurance Service (NHIS) Claims Database. J Korean Med Sci 2021;36:e331.

23. Jeffery MM, Inselman JW, Maddux JT, Lam RW, Shah ND, Rank MA. Asthma patients who stop asthma biologics have a similar risk of asthma exacerbations as those who continue asthma biologics. J Allergy Clin Immunol Pract 2021;9:2742-50.

24. Moore WC, Kornmann O, Humbert M, Poirier C, Bel EH, Kaneko N, et al. Stopping versus continuing long-term mepolizumab treatment in severe eosinophilic asthma (COMET study). Eur Respir J 2022;59:2100396.

25. Ramagopalan SV, Simpson A, Sammon C. Can real-world data really replace randomised clinical trials? BMC Med 2020;18:13.

26. Ryan D, Heatley H, Heaney LG, Jackson DJ, Pfeffer PE, Busby J, et al. Potential severe asthma hidden in UK primary care. J Allergy Clin Immunol Pract 2021;9:1612-23.

27. Hong SH, Cho JY, Kim TB, Lee EK, Kwon SH, Shin JY. Cost-effectiveness of tiotropium in elderly patients with severe asthma using real-world data. J Allergy Clin Immunol Pract 2021;9:1939-47.

28. Lee H, Ryu J, Nam E, Chung SJ, Yeo Y, Park DW, et al. Increased mortality in patients with corticosteroid-dependent asthma: a nationwide population-based study. Eur Respir J 2019;54:1900804.

29. Wang E, Wechsler ME, Tran TN, Heaney LG, Jones RC, Menzies-Gow AN, et al. Characterization of severe asthma worldwide: data from the International Severe Asthma Registry. Chest 2020;157:790-804.

30. Chen W, Sadatsafavi M, Tran TN, Murray RB, Wong CB, Ali N, et al. Characterization of patients in the International Severe Asthma Registry with high steroid exposure who did or did not initiate biologic therapy. J Asthma Allergy 2022;15:1491-510.

31. Scelo G, Torres-Duque CA, Maspero J, Tran TN, Murray R, Martin N, et al. Analysis of comorbidities and multimorbidity in adult patients in the International Severe Asthma Registry. Ann Allergy Asthma Immunol 2024;132:42-53.

32. Lee JH, Kim HJ, Park CS, Park SY, Park SY, Lee H, et al. Clinical characteristics and disease burden of severe asthma according to oral corticosteroid dependence: real-world assessment from the Korean Severe Asthma Registry (KoSAR). Allergy Asthma Immunol Res 2022;14:412-23.

33. Lee Y, Lee JH, Park SY, Lee JH, Kim JH, Kim HJ, et al. Roles of real-world evidence in severe asthma treatment: challenges and opportunities. ERJ Open Res 2023;9:00248-2022.

34. Grimes DA, Schulz KF. Cohort studies: marching towards outcomes. Lancet 2002;359:341-5.

35. Wang X, Cheng Z. Cross-sectional studies: strengths, weaknesses, and recommendations. Chest 2020;158(1S):S65-71.

36. Goodrum H, Roberts K, Bernstam EV. Automatic classification of scanned electronic health record documents. Int J Med Inform 2020;144:104302.

37. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health

records. NPJ Digit Med 2022;5:194.

38. Exarchos KP, Beltsiou M, Votti CA, Kostikas K. Artificial intelligence techniques in asthma: a systematic review and critical appraisal of the existing literature. Eur Respir J 2020;56:2000521.

39. Inselman JW, Jeffery MM, Maddux JT, Lam RW, Shah ND, Rank MA, et al. A prediction model for asthma exacerbations after stopping asthma biologics. Ann Allergy Asthma Immunol 2023;130:305-11.