Big Data Research for Diabetes-Related Diseases Using the Korean National Health Information Database

Amanuel Tesfaye*

Ethiopian Public Health Institute, Diabetes Research Unit, Ethiopia

Corresponding Author*

Amanuel Tesfaye

Ethiopian Public Health Institute, Diabetes Research Unit, Ethiopia

E-mail: atesfaye@ephi.et

Received: 01-Feb-2025, Manuscript No. jdm-25-37713; **Editor assigned:** 03-Feb-2025, PreQC No. jdm-25-37713; **Reviewed:** 17-Feb-2025, QC No. jdm-25-37713; **Revised:** 22-Feb-2025, Manuscript No. jdm-25-37713; **Published:** 28- Feb-2025, DOI: 10.35248/2155-6156.10001216

Abstract

The advent of big data has transformed healthcare research, particularly in chronic disease epidemiology. In Korea, the National Health Information Database (NHID) provides an extensive, population-wide dataset covering health screenings, claims, and mortality data. This database has become a vital tool for investigating diabetes and its related complications, including cardiovascular disease, renal dysfunction, and cancer. This article explores the structure and capabilities of the NHID, reviews key findings from recent diabetes-related studies, and discusses the strengths, limitations, and future directions of using big data analytics in Korean diabetes research. With proper methodological rigor, NHID-based studies can yield valuable insights for public health interventions and policy development.

Keywords: Big data; Diabetes; Korea; National Health Information Database; NHID; Chronic disease; Epidemiology; Healthcare policy; Real-world data; Population-based research

INTRODUCTION

Diabetes mellitus is a growing public health concern in South Korea, with an estimated adult prevalence of 14.5% as of 2021 [1]. Its associated complications—ranging from cardiovascular disease to end-stage renal disease—represent a significant burden on individuals and the healthcare system. In the era of digital health, big data research offers unprecedented opportunities to analyze trends, risk factors, and outcomes of diabetes on a national scale.

The Korean National Health Information Database (NHID), managed by the National Health Insurance Service (NHIS), integrates claims data, health screening results, demographic information, and mortality data for the entire Korean population [2]. This makes it an ideal resource for large-scale diabetes-related research, enabling both retrospective and prospective cohort studies.

This article outlines the structure of the NHID, highlights its applications in diabetes research, and discusses its impact, limitations, and the future of big data-driven healthcare in Korea.

DESCRIPTION

Structure of the NHID

The NHID consists of several linked datasets:

- **Eligibility database**: Includes demographic characteristics, socioeconomic status, and type of insurance coverage.
- **National health screening database**: Contains anthropometric measurements, laboratory values, and lifestyle questionnaires collected biannually for individuals aged 40 and above.
- **Healthcare claims database**: Records diagnostic codes (based on ICD-10), procedures, prescriptions, and hospitalizations.
- **Mortality database**: Linked with data from Statistics Korea, providing cause-specific death information.

These datasets can be anonymized and longitudinally linked for follow-up studies, making the NHID suitable for identifying disease trends and evaluating health policies over time [3].

Diabetes Identification in the NHID

Patients with diabetes can be identified using:

- ICD-10 codes: E10–E14
- Prescription records for antidiabetic drugs
- Fasting glucose levels (≥126 mg/dL) from health screenings

This multi-criteria approach improves diagnostic accuracy in administrative data research [4].

RESULTS

Key Findings from NHID-Based Diabetes Studies

- **Trends in diabetes incidence and mortality**: A nationwide study using NHID data from 2005 to 2015 revealed a stable diabetes incidence but increasing prevalence due to longer survival. However, **cardiovascular mortality** among diabetic patients has decreased, possibly due to improved management and early detection [5].
- **Diabetes and cardiovascular disease**: NHID data showed that individuals with diabetes had a 1.5–2 times higher risk of myocardial infarction and stroke, and that risk stratification using screening data could predict future events [6].
- **Diabetes and cancer risk**: A cohort study using NHID data found increased risks of liver, pancreas, and colorectal cancers among diabetic patients, especially those with poor glycemic control [7].
- **Health behaviors and outcomes**: Data from the health screening component showed that smoking, alcohol use, and physical inactivity significantly worsened glycemic control and increased complication rates [8].
- **Economic burden**: NHID-based analyses estimated that the annual direct medical cost for diabetes exceeded 2 trillion KRW (~$1.7 billion USD), with costs rising sharply with the presence of complications [9].
- **Medication adherence and outcomes**: Studies leveraging prescription refill data indicated that poor adherence to oral hypoglycemic agents was linked to higher hospitalization and mortality rates [10].

DISCUSSION

Strengths of using the NHID in diabetes research

- **Population-level coverage**: With data from nearly 98% of the population, the NHID provides comprehensive insights into disease burden and outcomes.
- **Longitudinal follow-up**: Enables cohort tracking, time-to-event

analyses, and evaluation of long-term interventions.

- **Integration of health behaviours**: Unique among many national datasets, the health screening component includes data on smoking, alcohol use, and physical activity.

- **Cost-effectiveness**: Secondary data analysis reduces research costs and ethical barriers compared to primary data collection.

Limitations

- **Lack of detailed clinical information**: The NHID does not capture hemoglobin A1c, diet, or continuous glucose monitoring data.

- **Coding bias**: Reliance on ICD codes may misclassify cases if coding is inconsistent.

- **Healthy participant bias**: Health screening data may underrepresent high-risk populations who do not participate in screenings.

- **Limited genetic data**: The NHID is not linked with genomic information, limiting precision medicine research.

Ethical and policy considerations

While the NHID is anonymized, ethical concerns about data privacy and informed consent remain. The Korean government has introduced data access guidelines to ensure secure and ethical use of health data, promoting transparency and public trust [3].

CONCLUSION

The Korean NHID is a powerful resource for advancing research in diabetes and its related diseases. Studies utilizing this big data infrastructure have yielded valuable insights into incidence trends, risk factors, outcomes, and health behaviours. Despite some limitations, continued integration of additional clinical and genomic data will enhance the utility of the NHID for precision public health. Strengthening data governance and fostering interdisciplinary collaboration will be crucial for translating big data findings into actionable health policies and interventions.

References

1. Sahin U (2020) An RNA vaccine drives immunity in checkpoint-inhibitor-treated melanoma. Nature 585: 107-112.
2. Alameh MG (2021) Lipid nanoparticles enhance the efficacy of mRNA and protein subunit vaccines by inducing robust T follicular helper cell and humoral responses. Immunity 54: 2877-2892.
3. Islam MA (2021) Adjuvant-pulsed mRNA vaccine nanoparticle for immunoprophylactic and therapeutic tumor suppression in mice. Biomaterials 266:120431.
4. Van Hoecke L (2021) mRNA in cancer immunotherapy: beyond a source of antigen. Mol. Cancer 20:48.
5. Pulendran B, Arunachalam PS, O'Hagan DT (2021) Emerging concepts in the science of vaccine adjuvants. Nat. Rev. Drug Discov 20: 454-475.
6. Ginn SL, Alexander IE, Edelstein ML, Abedi MR, Wixon J (2013) Gene therapy clinical trials worldwide to an update. Journal of Gene Medicine. 15: 65-77.
7. Allen TM, Cullis PR. (2004) Drug delivery systems: entering the mainstream. Science 303 : 1818-1822.
8. Chakraborty C, Pal S, Doss GP, Wen Z, Lin C (2013) Nanoparticles as “smart” pharmaceutical delivery. Frontiers in Bioscience 18: 1030-1050.
9. Paul D, Sanap G, Shenoy S, Kalyane D, Kalia K, et al. (2021) Artificial intelligence in drug discovery and development. Drug Discov Today 26: 80-93.
10. Sapoval N, Aghazadeh A, Nute MG (2022) Current progress and open challenges for applying deep learning across the biosciences. Nat Commun 13.

Cite this article: Amanuel Tesfaye. Big Data Research for Diabetes-Related Diseases Using the Korean National Health Information Database. J Diabetes Metab, 2025, 16(2): 1216.