All blue links are clickable for further information.

Zio Kim (김지오)

+82) 10-2964-9967 · ziokim@snu.ac.kr / zio0323gm@gmail.com · github: ZIO-KIM

Clinical/RWE Data Scientist
I am a **Clinical Data Scientist specializing in end-to-end medical data analysis based on RWE and OMOP CDM**. Throughout my career at Seoul National University Hospital and the Data Science Team at EvidNet, I have spearheaded clinical research focusing on statistical methodologies such as Causal Inference and Survival Analysis, utilizing large-scale healthcare data from the National Health Insurance Service (NHIS), Health Insurance Review and Assessment Service (HIRA), tertiary hospitals, and multi-site CDM.

Currently, I serve as a **Product Engineer within the AI Product Team**, where I implement and validate clinical analysis workflows into real-world services. *By integrating a product-oriented perspective into my core competency in clinical data analysis, I focus on seamlessly bridging the gap between data value and consumer systems.*

Research Interests
· Clinical & RWE Data Science · Survival Analysis & Statistical Modeling · Causal ML / Causal Inference · Multi-Site RWD & OMOP CDM · Patient Trajectory

Work Experience

EvidNet | AI Product Team **Seoul, Korea**
26.04 - Currently **Data Scientist / Clinical AI Product Engineer**
— R&D for new business initiatives and implementation of clinical analytics platforms.
— Conduct Quality Control (QC) for the clinical validity of platforms based on clinical research expertise, and enhance analysis workflows by directly implementing validated modifications at the system level (backend/server).

EvidNet | Data Science Team **Seoul, Korea**
25.07 - 26.03 (9M) **Data Scientist**
— Spearheaded the entire lifecycle of multi-site OMOP CDM-based RWE research for global pharmaceutical and insurance companies, covering study design, analysis, modeling, and protocol development.
— Performed advanced modeling utilizing patient trajectory analysis, unsupervised learning-based clustering, survival analysis (Cox), and AutoML.
— Conducted numerous feasibility analyses to secure new projects and participated in new business strategy planning.

Seoul National University Hospital **Seoul, Korea**
21.09 - 25.07 (3Y 10M) **Clinical Statistics Researcher (Depts. of Biomedical Engineering, Nephrology, and OB-GYN)**
— Conducted diverse research in causal inference and healthcare data analysis using large-scale medical big data (SNUH, NHIS, HIRA, UK Biobank, etc.) concurrently with academic programs.
(PI: Hyung-Jin Yoon, M.D, Ph.D, Seung-mi Lee, M.D, Ph.D, Hajeong Lee, M.D, Ph.D)

Education

Seoul National University **Seoul, Korea**
23.03 - 25.02 (2Y 0M) **Master of Engineering in Bioengineering (Clinical Big Data Lab)**
— Thesis: Estimating the Heterogeneous Effect of Maternal Renal Disorders on Neonatal Outcomes: A Causal Machine Learning Approach (Advisor: Prof. Hyung-Jin Yoon)

Sejong University **Seoul, Korea**
18.03 - 23.02 (5Y 0M) **Bachelor of Engineering in Data Science** [ABEEK Accredited]
— Graduation Project: Natural Language Processing-based AI Chatbot for Diagnosing Pet Diseases 🔗 (Advisor: Prof. Seung-won Lee)
— 20.01 - 20.12 (1Y 0M) Leadership: Vice President of the Student Council, College of Software Convergence

Publications (SCI/SCIE)

** # denotes first author or co-first author*

2026 **Risk of Congenital Malformation in Newborns from Mothers with Kidney Diseases: A Nationwide Population-based Study** 🔗 ***(Communications Medicine, IF=6.3)*** ***(Co-first authored)***
Seung Hyun Han(#), **Zio Kim(#)**, Subin Jeong, Seungyeon Kim, Jeongin Song, Jeesun Lee, Sehoon Park, Min Hyuk Lim, Joong Shin Park, Hyung-Jin Yoon, Seung Mi Lee(†), Hajeong Lee(†)
— **Description:** Estimated the risk of congenital malformations in newborns born to mothers with kidney diseases.
— **Role (Co-First Author/Lead Analyst):** Designed and executed the entire analysis workflow using the NHIS (National Health Insurance Service) customized database. Defined operational variables for neonatal malformations, maternal kidney diseases, concomitant medications, and underlying conditions from raw tables. Performed extensive data cleaning, preprocessing, and statistical modeling. *(Tech Stack: SAS, R, SQL)*

2026 **Causal Effect of Air Pollution on Outpatient Visits for Chronic Rhinitis** 🔗 ***(Laryngoscope, IF=2.5)***
Su Hwan Kim(#), Seong Pyo Kim, **Zio Kim**, Heung-Woo Park, Jin Youp Kim, Hyung-Jin Yoon(†)
— **Description:** Investigated the causal effect of air pollution on outpatient visits among patients with chronic rhinitis utilizing thermal inversion as an instrumental variable (IV).
— **Role (Analyst):** Modeled and conducted Instrumental Variable (IV) analysis and subgroup analysis. Scraped and processed thermal inversion data. *(Tech Stack: Python, R)*

2025 **Neuropsychiatric adverse effects of antihistamine: A nationwide data based epidemiological study** 🔗
(Asian Pacific Journal of Allergy and Immunology, IF=2.3) ***(Co-first authored)***
Zio Kim(#), Jin Youp Kim (#), Su Hwan Kim(#), Kyung-Lak Son, Chae-Seo Rhee(†), Hyung-Jin Yoon(†)
— **Description:** Estimated the risk of neuropsychiatric disorder onset based on the duration of antihistamine usage.
— **Role (Co-First Author/Lead Analyst):** Designed the data architecture for the NHIS customized dataset. Defined operational definitions for psychiatric disorders and medications from raw tables. Conducted data cleaning, preprocessing, and statistical modeling. *(Tech Stack: R, SAS, SQL)*

2023 **Causal relationship between asthma outpatient visits and air pollution with instrumental variable approach** 🔗 ***(Allergy, IF=12.6)***
Su Hwan Kim(#), Seong Pyo Kim(#), Jae-In Song, **Zio Kim**, Jin Youp Kim, Hyung-Jin Yoon(†)
— **Description:** Identified the causal relationship between asthma outpatient visits and air pollutants using an Instrumental Variable (IV) approach.
— **Role (Lead Analyst):** Modeled and executed causal inference analyses, including IV analysis, subgroup analysis, and negative control outcome/exposure analysis. Crawled thermal inversion data and generated main figures. Handled data cleaning and preprocessing for environmental disease and air pollution datasets. *(Tech Stack: Python, R)*

2024 **Deep learning-based long-term risk evaluation of incident type 2 diabetes using electrocardiogram in a non-diabetic population: a retrospective, multicentre study** 🔗 ***(EClinicalMedicine, IF=9.6)***
Junmo Kim(#), Hyun-Lim Yang, Su Hwan Kim, Siun Kim, Jisoo Lee, Jiwon Ryu, Kwangsoo Kim, **Zio Kim**, Gun Ahn, Doyun Kwon, Hyung-Jin Yoon(†)
— **Role (Analyst):** Collected, cleaned, and preprocessed health screening data from Seoul National University Hospital *(Tech Stack: Python)*. Prepared Institutional Review Board (IRB) applications and managed administrative procedures to acquire data access permissions.

Conferences

International

25.02 **HealthInf 2025, Oral presentation, Causal Machine Learning Approach for Quantifying Heteregeneous Effects of Maternal Renal Disorders on Adverse Pregnancy Outcomes *(Porto, Portugal)***
Zio Kim(#), Hyung-Jin Yoon(†)
— **Description:** Quantified the individualized, heterogeneous effects of maternal renal disorders on pregnancy outcomes utilizing a Causal Forest approach.
— **Role (Lead Analyst):** Developed Causal Forest, Generalized Linear Models (GLM), and Cox proportional-hazards models. Established operational definitions for diseases and conducted extensive data cleaning and preprocessing. *(Tech Stack: R, SAS)*

24.02 **HealthINF 2024** 🔗 **, Poster presentation, The efficient removal of spatial autocorrelation in environmental epidemiological data through eigenvector spatial filtering *(Rome, Italy)***
Zio Kim(#), Su Hwan Kim(#), Jin Youp Kim, Hyung-Jin Yoon(†)
— **Description:** Investigated the causal relationship between air pollutants and the onset of asthma, allergic rhinitis, and atopic dermatitis by eliminating spatial autocorrelation (SA) using the Eigenvector Spatial Filtering (ESF) methodology.
— **Role (Lead Analyst):** Modeled Eigenvector Spatial Filtering (ESF), compared Fixed-Effect and Random-Effect models 🔗, and performed data cleaning and preprocessing. *(Tech Stack: Python, R)*

Domestic

24.10 **Korean Society of Artificial Intelligence in Medicine (KoSAIM) 2024** 🔗 **, Poster presentation, Advances in Renal Cancer CT Imaging: GAN-Enhanced Contrast Visualization and U-Net3+ Based Tumor Detection Techniques *(Seoul, South Korea)***
Gun Ahn(#), Won Chang Choi(#), **Zio Kim(#)**, Jiwon Ryu(†), Hyung-Jin Yoon(†)
— **Description:** Generated contrast-enhanced CT images from non-contrast CT scans by applying Generative Adversarial Network (GAN) models.
— **Role (Co-First Author/Lead Analyst):** Modeled CNN 3D mass classification, performed tumor (mass) extraction, and conducted CT data cleaning 🔗. *(Tech Stack: Python, TensorFlow)*

Research Experiences

Seoul National University **Seoul, Korea**
21.09 - 23.02 (1Y 5M) **Undergraduate Research Assistant (Intern), Clinical Big Data Lab (CBDL)**
— Participated as a research intern in various studies focused on causal inference and healthcare data analytics.
— Conducted data analysis using large-scale medical big data from Seoul National University Hospital (SNUH), National Health Insurance Service (NHIS), and Health Insurance Review and Assessment Service (HIRA).

Projects

25.07 - 26.03 **[Global Pharmaceutical Company A] Rare Disease Treatment Trajectory and Disease Burden Study**
— **Role:** Lead Data Scientist (Study design, preprocessing, modeling, protocol, and final report writing)
— **Description:** Established an optimal and clinically applicable analysis strategy by comparing unsupervised learning methods (K-means, Sequence analysis, Hierarchical clustering) with statistical models (IR, IRR, Cox) using multi-site OMOP-CDM data to validate the client's hypotheses.
— **Key Results:** Proactively identified the structural limitations of the initially designed statistical model (GBTM) and proposed a novel analysis framework tailored to the client's objectives. Delivered highly satisfactory results requiring zero additional revisions, successfully securing the contract for the subsequent Wave 2 project.

25.10 - 25.12 **[Domestic Healthcare/Consumer Goods Company B] Metabolic/Musculoskeletal Disease Patient Clustering Analysis**
— **Role:** Lead Data Scientist (Onboarded mid-project to lead main analysis and write the final report)
— **Description:** Conducted OMOP-CDM multi-site data-based clustering modeling to subgroup patients based on metabolic and musculoskeletal disease characteristics.
— **Key Results:** Performed extensive sensitivity analyses utilizing various methodologies, including Multi-task LASSO and Hierarchical clustering. Successfully met 100% of the deadlines for the client's frequent ad-hoc requests (3-4 times a week) over the 2-month period. Despite taking over the project mid-stream from a predecessor, rapidly acquired domain context, delivered outputs precisely meeting the client's needs, and successfully concluded the project by independently writing a comprehensive ~100-page final report.

26.01 - 26.04 **[Domestic Insurance Company C] Acute Infection Cohort Risk Rate and Comorbidity Analysis**
— **Role:** Main Data Analyst (Cohort definition, preprocessing, and automated pipeline construction)
— **Description:** Defined acute infection cohorts based on clinical and culture test results using multi-site OMOP-CDM data, and calculated risk rates and comorbidity statistics stratified by year, sex, and age group.
— **Key Results:** Accurately implemented complex cohort extraction logic accounting for the specific time window between elevated CRP (C-Reactive Protein) levels and positive culture test dates. Built a Python-based data merging and EDA automation pipeline to process large-scale data across 16 institutions. Led rigorous Quality Control (QC) processes requiring high clinical attention to detail, such as integrating culture specimen concepts and flagging antibiotic prescriptions, to derive highly reliable statistical outcomes.

26.03 - **[Domestic Hospital D] Risk Factor Analysis for Emergency Department (ED) Visits in Patients with Type 2 Diabetes Mellitus (T2DM)**
— **Role:** Main Data Scientist (Additional analysis for manuscript revision and rebuttal response)
— **Description:** Conducted in-depth performance validation and manuscript revision for an AutoML-based ED visit prediction model for T2DM patients using OMOP-CDM data.
— **Key Results:** Assumed responsibility for a predecessor's study, rapidly understanding the data pipeline and ML modeling architecture. Successfully fulfilled all major revision requirements requested by the journal, including Decision Curve Analysis (DCA), Youden's Index-based threshold resetting, and site-stratified validation.

26.01 - 26.02 **[Domestic Insurance Company E] Major Organ (Liver/Lung/Kidney) Diseases RWE Statistical Analysis**
— **Role:** Main Data Analyst (Operational definition of major treatments, cohort extraction, and statistical calculation)
— **Description:** Built cohorts for liver, lung, and kidney diseases using multi-site OMOP-CDM data, and calculated statistics on major treatment patterns (medications/surgeries/procedures) and medical costs (reimbursements).
— **Key Results:** Reviewed clinical guidelines and references to establish operational definitions for the major treatments requested by the client, mapping them to multi-site OMOP concepts and reimbursement fee codes to clarify analysis criteria. Accurately calculated the number of patients and treatment instances across multiple institutions by combining KCD primary diagnosis codes with strict time-window logic (capturing treatments administered strictly within 90 days of diagnosis). Built and QC'd a reimbursement cost statistics pipeline for institutions with available financial data, successfully delivering core Real-World Data (RWD) on time for the client's new insurance product planning.

25.07 - 26.03 **[PoC] Feasibility Checks for Securing New Projects**
— **Description:** Conducted multi-site OMOP-CDM-based cohort extraction and feasibility analyses to secure new RWE projects from pharmaceutical and insurance clients.
— **Key Results:** Contributed to securing new project contracts by successfully completing numerous Proof of Concept (PoC) analyses, including specific patient cohort analyses (e.g., gout, rare diseases, keloid) and prescription pattern analyses for cardiovascular medications. Maintained a 100% on-time delivery rate in a highly demanding PoC environment, requiring rapid cohort design, data extraction, and result reporting within tight 2 to 3-day turnaround times.

22.03 - 22.06 **Natural Language Processing-based AI Chatbot for Diagnosing Pet Diseases**
Zio Kim, Seeun Choi, Jaesuk Kim, Yoomin Lee
— **Description:** Undergraduate Capstone Project.
— **Role (Project Team Leader):** Built a sentence similarity checking algorithm utilizing SBert (semantic text similarity modeling) and implemented the backend and web application . *(Tech Stack: Python, JS, Flask)*
— **Key Results:** Won 1st Place (Gold Prize) in the Data Science Division.

Honors & Awards

2022 **1st Place (Gold Prize)** , Capstone Design Competition, Sejong University - College of Software Convergence
2022 **Scholarship**, Academic Grade-enhancing Scholarship, Sejong University
2018 **5th Place**, SW algorithm competition,Sejong University
2018 **Scholarship**, Scholarship for Excellence in Language Proficiency, Sejong University

Skills and Techniques

Programming
— **Python (Primary):** Proficient in end-to-end data pipelines, including preprocessing, feature engineering, and statistical/ML modeling. Skilled in model selection, result interpretation, and customized data visualization. Extensive experience processing 1B+ rows of structured data and 10K+ image datasets.
— **R (Primary):** Expert in advanced statistical modeling (Cox Proportional Hazards, PSM, IPTW, IRR), data visualization, and complex clinical data preprocessing.
— **SQL**: Experienced in designing complex queries and managing RDBMS for large-scale data extraction (NHIS/HIRA databases). Proficient in optimizing queries (JOIN, MERGE) for high-performance data integration and refining.
— **C**: Strong foundation in data structures and algorithms; proficient in algorithmic problem-solving.

Tools
— **SAS**: Experienced in utilizing SAS (including SAS-SQL) for large-scale clinical data management, EDA, and advanced statistical modeling, including PSM, Regression, and Survival Analysis.
— **Docker**: Leveraged for multi-site clinical data extraction via FeederNet (EvidNet's proprietary CDM platform). Experienced in executing SQL queries within containerized environments to retrieve and process large-scale RWD across multiple medical institutions.

Certifications
— **SQLD** (2021.10)

Languages
— **Fluent in English**
  — **OPIC**: AL (Advanced Low) | Sep, 2024
  — **New TEPS**: 456/600 | Feb, 2022
  — **TOEFL**: 99/120 | May, 2019
  — **TOEIC**: 945/990 | Aug, 2018

Relevant Courseworks

— **Data Science & Analytics:** Data Problem Solving and Practice, Data Visualization, Introduction to Data Analytics, Advanced Data Processing, Decision Analysis

— **Computer Science & Math:** Capstone Design, Data Structures and Lab, Algorithms and Lab, Database Systems, Linear Algebra and Programming, Introduction to Statistics