

김지오 | Zio Kim

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

Clinical/RWE Data Scientist

RWE, OMOP CDM 기반의 의료 데이터에 대한 End-to-End 분석에 전문성을 가진 Clinical Data Scientist입니다. 서울대학교병원과 에비드넷 (EvidNet) 데이터사이언스팀을 거치며 건강보험공단, 심사평가원, 상급종합병원, 다기관 CDM 등 대규모 의료 데이터를 활용해 인과추론, Causal Inference, 생존분석 등 통계기법 중심의 임상 연구를 주도하였습니다.

현재는 AI Product 팀에서 임상 분석 워크플로우를 실제 서비스로 구현하고 검증하는 Product engineer의 역할을 맡고 있습니다. 임상 데이터 분석이라는 핵심 역량에 프로덕트 관점을 더해, 데이터의 가치를 소비자의 시스템에 유연하게 연결하는 데 집중하고 있습니다.

Research Interests

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

Work Experience | 경력

에비드넷 EvidNet | AI Product Team

Seoul, Korea

26.04 - Currently Data Scientist / Clinical AI Product Engineer

— 신사업 R&D 및 분석 플랫폼 구현

— 임상 연구 경험을 바탕으로 플랫폼의 임상적 타당성 (Clinical validity)를 QC하고, 검증된 수정사항을 백엔드/서버 등 시스템 레벨에 직접 반영하여 분석 워크플로우 고도화

에비드넷 EvidNet | Data Science Team

Seoul, Korea

25.07 - 26.03 (9M) Data Scientist

— 글로벌 제약사/보험사 대상 다기관 OMOP CDM 기반 RWE 연구의 설계, 분석, 모델링, 프로토콜 작성까지 전 과정 주도

— 환자 Trajectory 분석, Unsupervised Learning 기반 클러스터링, 생존분석 (Cox) 및 AutoML 등을 활용한 모델링

— 신규 과제 수주를 위한 다수의 Feasibility 분석 및 신사업 기획

서울대학교병원 Seoul National University Hospital

Seoul, Korea

21.09 - 25.07 (3Y 10M) 의공학과/신장내과/산부인과 임상통계연구원

— 서울대병원, 국민건강보험공단, 심사평가원, UKBiobank 등 의료빅데이터를 사용한 다양한 인과추론 및 헬스케어 데이터 분석 연구 (학사/석사 과정과 병행)

(PI: Hyung-Jin Yoon, M.D, Ph.D, Seung-mi Lee, M.D, Ph.D, Hajeong Lee, M.D, Ph.D)

Education | 교육

서울대학교 Seoul National University

Seoul, Korea

23.03 - 25.02 (2Y 0M) 공과대학 바이오엔지니어링학과 석사 (임상빅데이터연구실)

— Thesis: Estimating the Heterogeneous Effect of Maternal Renal Disorders on Neonatal Outcomes: A Causal Machine Learning Approach (Prof. 윤형진 교수님)

세종대학교 Sejong University

Seoul, Korea

18.03 - 23.02 (5Y 0M) 소프트웨어융합대학 데이터사이언스학과 학사 [공학인증]

— 졸업 프로젝트: [Natural Language Processing-based AI Chatbot for Diagnosing Pet Diseases](#) (Prof. 이승원 교수님)

— 20.01 - 20.12 (1Y 0M) 부학생회장, 소프트웨어융합대학 학생회

* # 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, Jeeseun Lee, Sehoon Park, Min Hyuk Lim, Joong Shin Park, Hyung-Jin Yoon, Seung Mi Lee(+), Hajeong Lee(+)

- [연구 내용] 신장질환 산모에게서 태어난 신생아의 선천성 기형 발병에 대한 risk estimation
- [공동 1저자/주 분석자] 국민건강보험공단 맞춤형 DB 사용, 전체 분석 과정 설계 및 진행, Raw table에서의 신생아 기형, 산모 신장질환, 기타 약제 및 기저질환 등의 변수 정의, 통계 모델링, 데이터 클리닝 및 전처리 (used: 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(+)

- [연구 내용] Thermal Inversion을 도구변수로 활용하여, 대기오염이 만성 비염 환자의 외래 방문에 미치는 인과적 영향 규명
- [분석자] Instrumental variable analysis 모델링 및 수행, Subgroup analysis, Thermal inversion 데이터 크롤링 (used: 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(+)

- [연구 내용] 항히스타민제의 복용기간별 정신질환 발병에 미치는 risk estimation
- [공동 1저자/주 분석자] 데이터 클리닝 및 전처리, Raw table에서의 정신질환 및 약제 정의, 통계 모델링 (used: R, SAS, SQL)
- 국민건강보험공단 맞춤형DB 데이터 구축 설계

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(+)

- [연구 내용] 인과추론 방법론의 일종인 IV 방법론을 사용하여 천식 외래 환자수와 대기오염물질의 인과 관계 규명
- [주 분석자] Causal Inference 방법론인 Instrumental variable analysis 모델링 및 수행 (used: R)
- [주 분석자] Subgroup analysis, Negative control outcome and exposure analysis 수행 (used: R)
- [주 분석자] Thermal inversion 데이터 크롤링, Main figure 작업 (used: Python)
- 환경성질환/대기오염 데이터 클리닝 및 전처리 (used: 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(+)

- 서울대학교병원 건강검진 데이터 수집, 전처리 및 클리닝 (used: Python)
- IRB 승인 신청서 작성 및 기타 데이터 접근 권한 획득을 위한 행정 절차 수행

Conferences | 학회

International 국제

25.02 **HealthInf 2025, Oral presentation, Causal Machine Learning Approach for Quantifying Heterogeneous Effects of Maternal Renal Disorders on Adverse Pregnancy Outcomes** (포르투, 포르투갈)

Zio Kim(#), Hyung-Jin Yoon(+)

- [연구 내용] Causal Forest 방법론을 사용하여 모성 신장 질환이 임신 예후에 미치는 영향을 개인별로 heterogeneous 하게 규명함
- [주 분석자] Causal Forest modeling, GLM/Cox modeling, 질환에 대한 조작적 정의, 데이터 클리닝 및 전처리 (used: R, SAS)

24.02 **HealthINF 2024**, Poster presentation, The efficient removal of spatial autocorrelation in environmental epidemiological data through eigenvector spatial filtering (로마, 이탈리아)

Zio Kim(#), Su Hwan Kim(#), Jin Youp Kim, Hyung-Jin Yoon(+)

- [연구 내용] ESF 방법론을 사용하여 SA를 제거하고 천식, 비염, 아토피의 발병과 대기 오염 물질의 인과관계 규명
- [주 분석자] Eigenvector Spatial Filtering 모델링, Fixed effect/Random effect 모델링 및 비교 (used: Python, R)

Domestic 국내

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 (서울, 대한민국)

Gun Ahn(#), Won Chang Choi(#), **Zio Kim(#)**, Jiwon Ryu(+), Hyung-Jin Yoon(+)

- [연구 내용] 조영제를 사용하지 않은 CT에 GAN 모델을 적용하여 조영제를 사용한 CT를 생성
- CNN 3D mass classification 모델링, 종양 추출 (mass extraction), CT 데이터 클리닝 (used: Python, Tensorflow)

서울대학교 Seoul National University

Seoul, Korea

21.09 - 23.02 (1Y 5M) 임상빅데이터연구실 (Clinical Big Data Lab, CBDL) 학부연구생

— 서울대병원/국민건강보험공단/심사평가원 의료데이터를 사용한 다양한 인과추론 및 헬스케어 데이터 분석 연구에 인턴으로 참여

Projects | 프로젝트

25.07 - 26.03 [글로벌 A 제약사] 희귀질환 치료 Trajectory 및 Disease Burden 연구

— **Role:** Lead Data Scientist (연구 설계, 전처리, 모델링, 프로토콜 및 최종 보고서 작성)

— **Description:** OMOP-CDM 다기관 데이터 기반 비지도 학습 (K-means, Sequence analysis, Hierarchical clustering)과 통계 모델 (IR, IRR, Cox)을 병행 비교하여 임상적 활용도가 높은 최적의 분석 전략 수립 및 의뢰사의 가설에 부합하는 분석 결과 도출

— **Key Results:** 초기 설계된 통계 모델 (GBTM)의 구조적 한계점을 선제적으로 파악하고, 의뢰사의 목적에 부합하는 새로운 분석 프레임워크를 구조화하여 역제안함. 추가 수정 없는 긍정적 피드백을 이끌어내며 후속 (Wave 2) 프로젝트 계약 체결에 성공함.

25.10 - 25.12 [국내 B 헬스케어/소비재 기업] 대사/근골격계 질환 환자 클러스터링 분석

— **Role:** Lead Data Scientist (중도 투입되어 메인 분석 리드 및 최종 보고서 작성)

— **Description:** 대사/근골격계 질환 환자의 특색별 sub-grouping을 위한 OMOP-CDM 다기관 데이터 기반 클러스터링 모델링

— **Key Results:** Multi-task LASSO, Hierarchical clustering 등 다양한 방법론을 적용하여 다각도로 sensitivity analysis를 수행하였으며 2달의 프로젝트 기간 동안 주 3~4회 발생한 의뢰사의 추가분석 요청을 기한 내 100% 완수하였음. 전임자의 프로젝트를 중도 인수인계 받아 투입되었음에도 빠르게 선행 내용을 파악하고 의뢰사의 니즈에 맞는 분석 결과물을 제시하였으며, 최종적으로 100p 내외 분량의 최종 보고서를 단독 작성하여 프로젝트를 원활하게 마무리하였음.

26.01 - 26.04 [국내 C 보험사] 급성감염 질환 환자군 위험률 및 동반질환 분석

— **Role:** Main Data Analyst (코호트 정의, 전처리 및 자동화 파이프라인 구축)

— **Description:** 다기관 OMOP-CDM 데이터를 활용하여 임상/배양검사 기반의 급성감염 코호트를 정의하고, 계층별 (연도/성별/나이대별) 위험률 및 동반질환 통계 산출

— **Key Results:** CRP (C-Reactive Protein) 수치와 배양검사 양성 판정 시점 간의 시간차를 고려한 복잡한 임상 조건의 코호트 추출 로직을 정확히 구현하고, 16개 기관 대규모 데이터 처리를 위해 Python 기반의 데이터 병합 및 EDA 자동화 파이프라인을 구축하였음. 배양검사 검체별 concept 포함, 항생제 처방 확인 등 임상적 디테일이 요구되는 QC 과정을 주도하여 신뢰도 높은 통계 결과를 도출함.

26.03 - [국내 D 병원] 제2형 당뇨병 (T2DM) 환자 응급실 방문 위험 요인 분석

— **Role:** Main Data Scientist (논문 리비전 추가분석 및 Rebuttal 대응)

— **Description:** OMOP-CDM 데이터를 활용한 제2형 당뇨병 환자의 응급실 방문 예측 모델 (AutoML 기반)의 심층 성능 검증 및 논문 리비전 수행

— **Key Results:** 전임자의 연구를 인수인계받아 단기간 내에 데이터 파이프라인과 ML 모델링 구조를 파악하고, 저널의 Major revision 요구사항에 대한 모든 추가 분석을 완수함 (Decision Curve Analysis, Youden's Index 기반의 임계값 재설정, Site-stratified validation 등)

26.01 - 26.02 [국내 E 보험사] 주요 장기 (간/폐/신장) 질환 RWE 통계 분석

— **Role:** Main Data Analyst (주요 치료 정의 수립, 코호트 추출 및 통계 산출)

— **Description:** 다기관 OMOP-CDM 데이터를 활용하여 간/폐/신장질환 코호트를 구축하고, 질환별 주요 치료 (약물/수술/시술) 패턴 및 의료비용 (급여액) 통계 산출

— **Key Results:** 질환별 임상 가이드라인 및 레퍼런스를 검토하여 의뢰사가 요청한 주요 치료 (약물, 수술, 시술)에 대한 조작적 정의를 수립하고, 이를 다기관 concept/수가코드로 매핑하여 분석 기준을 명확화 함. KCD 주진단 코드 기반의 정밀한 환자 추출과 더불어, '진단일 기준 go일 이내 시행된 치료' 만 집계하는 시간 조건 (Time-window) 로직을 적용하여 다기관 환자수 및 치료 시행 건수를 정확하게 산출함. 의료비용 데이터가 확보된 기관을 대상으로 질환 및 치료별 급여액 통계 파이프라인을 구축 및 QC하여, 보험사 신규 상품 기획을 위한 핵심 기초 RWD (Real-World Data)를 기한 내 성공적으로 납품함.

25.07 - 26.03 [PoC] 신규 과제 수주를 위한 Feasibility check

— **Description:** 제약/보험사 신규 RWE 프로젝트 수주를 위한 다기관 OMOP-CDM 기반 코호트 추출 및 사전 타당성 (Feasibility) 분석

— **Key Results:** 통풍, 희귀질환, 피부질환(Keloid) 등 특정 질환 환자군 분석과 심혈관계 주요 약제들의 단독/병용 처방 패턴 분석 등 다수의 PoC 프로젝트를 성공적으로 완수하여 신규 과제 수주에 기여함. 평균 2~3일 내에 코호트 설계부터 데이터 추출, 결과 리포팅까지 신속하게 완료해야 하는 타이트한 PoC 업무 환경에서 모든 프로젝트의 데드라인을 100% 준수함.

22.03 - 22.06 **반려견 증상 기반 질환 진단 챗봇 <의사소통>** Natural Language Processing-based AI Chatbot for Diagnosing Pet Diseases

Zio Kim, Seeun Choi, Jaesuk Kim, Yoomin Lee

— 학부 Capstone 졸업 프로젝트

— [프로젝트 팀 리더] SBert 기반의 문장 유사도 검사 알고리즘 구축 (semantic text similarity modeling) / 백엔드 및 웹 구현 (used: Python, JS, Flask)

— 데이터사이언스분과 1위 (금상) 수상

Honors & Awards | 수상

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

Skills and Techniques | 기술 및 자격증

Programming

- **Python: [주언어]** 데이터 전처리 및 클리닝, 결측/이상치 처리, Feature engineering, EDA, 통계/ML 모델링 등에 활용. 데이터에 대한 이해를 기반으로 한 전처리 및 EDA가 가능함. 데이터의 특성을 파악하고 알맞은 모형을 선택하여 분석 수행, 결과 해석을 할 수 있으며 결과를 원하는 방향으로 시각화 할 수 있음. 10억+행 이상의 정형데이터 처리 경험 다수와 1만장+ 이상의 이미지 데이터 처리 경험이 있음.
- **R: [주언어]** Cox/PSM/IPTW/IRR/시각화 등의 통계 모델링 및 전처리 작업 수행 가능.
- **SQL:** 건강보험공단 분석센터에서 데이터 추출 및 처리를 하는 데 활용하였음. 쿼리를 사용하여 원하는 데이터를 추출할 수 있고, 관계형 데이터베이스를 이해하고 설계할 수 있음. JOIN, MERGE 등을 사용하여 대규모 데이터를 결합하고 정제할 수 있음. 속도를 고려하여 쿼리를 최적화 할 수 있음.
- **C:** 학부 1~4학기의 전공 강의 수강 중 습득. 자료구조/알고리즘의 기본적인 지식을 가지고 있으며 C를 사용하여 알고리즘 문제를 해결할 수 있음.

Tools

- **SAS:** 건강보험공단 분석센터에서 활용하였음. SAS 내에서 SQL을 활용하여 데이터 처리를 할 수 있으며, 이외 SAS 문법으로 데이터 EDA 및 클리닝, PSM, 회귀분석/생존분석 등 대부분의 통계 모델링을 할 수 있음.
- **Docker:** 에비드넷의 CDM 플랫폼 '피더넷(FeederNet)'을 활용한 다기관 임상 데이터 추출 프로세스 수행 과정에서 경험하였으며, 컨테이너 기반 환경에서 SQL 쿼리를 실행하여 여러 의료 기관의 대규모 실사용 데이터(RWD)를 수집 및 처리하였음.

Certificate

- **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 | 학업

- **Data Science & Analytics:** 데이터 문제 해결 실습, 데이터 시각화, 데이터 분석 및 심화 데이터 처리, 의사결정분석 등을 통해 데이터 기반의 통찰 도출 역량을 다지고 기초를 쌓았음.
- **Computer Science & Math:** 캡스톤 디자인, 자료구조, 알고리즘, 데이터베이스 등을 통해 CS 기초 역량을 다졌으며 선형대수학, 통계학 등을 수강하여 수리적 분석 토대를 마련하였음.