# KEVIN LAN

Kevlankl13@gmail.com · LinkedIn

**Summary:** Data Science Engineer with a PharmD and CDC Public Health Informatics Fellowship background, working across real-world evidence (RWE) analytics, clinical data interoperability, and secondary data harmonization (administrative claims, electronic health records) for pharma, healthcare, and federal public health stakeholders. Authors statistical analysis plans (SAPs), builds OMOP CDM- and FHIR-based data models, and maps clinical data to SNOMED CT, RxNorm, and LOINC. Certified in HL7 v2, CDA R2, and FHIR. Has established team-wide QA/validation standards for AI-assisted analytics workflows, pairing rigorous spec-writing and terminology governance with epidemiological framing and output validation.

**Stack:** OMOP CDM, SNOMED CT, RxNorm, FHIR, HL7 v2/CDA R2, LOINC, R, Python, SQL, Airflow, Neo4j/Cypher, Azure (Data Factory, Databricks, Synapse)

## TECHNICAL SKILLS

**Epidemiology & Observational Research:** Observational study design, cohort/inclusion-exclusion criteria definition, disease and outcome characterization, IRB-approved protocol design

**Real-World Data & Interoperability:** Data harmonization, secondary data sources (administrative claims, electronic health records), OMOP CDM, SNOMED CT, RxNorm, LOINC, ICD9/10, HL7 v2, CDA R2 (certified), FHIR, USCDI/USCDI+, PHIN-VAD, HL7/FHIR implementation guide (IG) contributions

**Data Engineering:** R, Python, SQL, Airflow, ETL/ELT pipeline design, Neo4j/Cypher, data validation & QA, pipeline architecture, version control

**Cloud & Tools:** Azure (Data Factory, Databricks, Synapse, DevOps, ML Service), pandas, numpy, sklearn, PySpark, R Shiny

**Analytics & Delivery:** Statistical analysis plans (SAP), data dictionaries, client-facing reporting, AI-assisted development (Cursor, OpenCode; spec-writing, validation-driven review)

## EXPERIENCE

### Cercle AI
August 2025 – Present
*Data Science Engineer*
Remote

- Harmonized multi-source secondary data (2-4 administrative claims and electronic health record feeds, ~2M patient records, multi-TB scale) into an OMOP CDM-style data model, mapping normalized (3NF) clinical fields to SNOMED CT, RxNorm, and LOINC for database development; used AI-assisted development (Cursor, OpenCode) across harmonization, QA, and analytics work while owning spec-writing and output validation
- Created the metadata and data dictionary used to manage terminology and vocabulary mappings across the database, supporting consistent SNOMED CT, RxNorm, and LOINC usage system-wide
- Guide the data science team's solution design for pharma client needs and serve as primary client-facing data lead for the firm's largest (highest-revenue) pharma client, meeting weekly with clinical trial study designers, epidemiologists, clinical managers, and biostatisticians; client subsequently extended the RWD analysis contract and is evaluating a longer-term agreement
- Resolved a previously strained client relationship, rooted in ad hoc solutioning, absence of a shared statistical analysis plan, and limited data transparency for the client's own publication needs, by introducing a standardized SAP template and a client-facing data dictionary (RWD variable definitions, analysis limitations, definition-of-done deliverable criteria), since adopted company-wide; reduced error rates 50%+ and improved turnaround time 15%
- Review SAPs and RWD analyses authored by fellow team members for the same client (4 to date), ensuring methodological consistency and QA rigor prior to client delivery
- Designed and maintain Neo4j/Cypher graph models linking patient-drug relationships across ~2M patients; modeled 180,000+ IUI cycles with stimulation start dates linked to pregnancy and medication outcomes to support observational, epidemiological IUI/IVF treatment-outcome research (pregnancy rate, infertility cause)
- Authored 5 statistical analysis plans (SAPs) over the past year, structuring multi-section deliverables (up to 7 sections per project) defining analysis populations, endpoints, and methodology for client-facing evidence generation, improving project completion rate 60%
- Established team-wide QA/validation standards by building and maintaining a SKILL.md repository (data harmonization QA, cohort-setting/code checking, analysis generation) for Cursor/OpenCode, adopted by other engineers on the team, reducing errors 20% and cutting task completion time 40%
- Co-authored (3rd of 9 authors) "Real-World Insights on Controlled Ovarian Stimulation for IVF/ICSI: Data from a Large Fertility EMR Database in the US (2010-2024)," poster presentation, ISPOR 2026, Philadelphia, PA

**Booz Allen Hamilton** July 2024 – June 2025
*Senior Consultant – Health Data Engineer* Bethesda, MD

- Configured and maintained Microsoft FHIR Server integrations, parsing FHIR resources into a relational structure across 65k+ patient records via an automated harmonization pipeline (MS Fabric, PySpark, OMOP CDM), with custom JSON-parsing scripts that cut processing time 40%
- Served as data governance lead for a CDC-contracted project; working as data modeler, created data models and metadata standards that guided engineering teams' product design decisions, and directly accessed and audited databases to assess data validity and structural integrity
- Designed a data governance and validation workflow standardizing external source integration, cutting integration time from 3 months to 1 month
- Configured FHIR API bulk import of 10k patient records, improving import time 10-fold
- Built semantic metadata models spanning 4 infectious disease domains using GENEPIO ontologies, with outputs ingested into CDC's OAMD system to standardize genomic data representation
- Enriched analytic datasets by integrating external API sources, including CMS insurance data and geographic mapping data, into existing systems
- Engineered a weekly-refreshed RAG pipeline (Azure ML) vectorizing 5GB of policy text into a retrieval store for an AI agent, applying prompt engineering to reduce hallucination and filter irrelevant user queries
- Created a GenAI RAG recommendation engine (cosine similarity, Azure AI Search) surfacing insights from regulatory/biomedical literature, awarded $10,000 from the Fast Prototyping Fund
- Built and demoed a Streamlit/GPT-4 recommendation bot for an HHS tech-innovation project proposal, presenting to the leadership team overseeing the government contract as a candidate solution for improving regulatory policy review

**NIOSH, Centers for Disease Control and Prevention** July 2023 – July 2024
*Health Informatics Scientist – Division of Field Studies and Engineering* Cincinnati, OH

- Served as data architect for NIOSH's Adult Blood Lead Epidemiology Surveillance (ABLES) program, directing a contract engineer and owning the technical roadmap for migrating ABLES to Azure cloud infrastructure, cutting cloud costs 70% and reducing data refresh turnaround from 3 months to as fast as 1 week, enabling states to move to monthly reporting cadence
- Led the OMOP CDM adoption initiative, harmonizing normalized clinical data to SNOMED CT, RxNorm, and LOINC depending on data type, and cross-division data-standard harmonization working sessions for the Division of Field Studies and Engineering (DFSE), representing NIOSH in cross-division study data-capture standards discussions
- Contributed to FHIR implementation guide development and maintained LOINC code sets, the PHIN-VAD value set, and the Message Mapping Guide supporting NIOSH data-capture standards
- Served as subject matter expert in CDC's OHDSI-Vulcan-OMOP work group, contributing to development of the International Patient Access Implementation Guide (IPA IG)
- Coordinated multi-state partner onboarding to the SAMS/SDX health information exchange platform, improving pilot-project data extraction
- Performed governance oversight ensuring contract engineering work and data standards aligned with NIOSH scientific committee study requirements and regulatory expectations
- Authored NORA Fund proposal integrating NIOSH auto-coding with EHR data, accepted, $30,000 granted
- Designed cloud data architecture (Azure Data Factory, Databricks Unity Catalog) and led FHIR API integration/USCDI+ research resulting in accepted submissions to USCDI v3 and USCDI+ v2, cutting state health department ETL/ingestion time from 6 months to 1 week
- Harmonized occupational health data across states with inconsistent units, fields, and coding systems (unit conversion, missingness handling, code standardization) to support ML model development for occupational health analytics

**NCEZID, Centers for Disease Control and Prevention** July 2021 – June 2023
*Public Health Informatics Fellow – Division of Global Migration and Health* Atlanta, GA

- Built statistical compartmental epidemiological models simulating global COVID-19 spread and trend trajectories (chi-square 0.38, p=0.046), informing forecasting and response strategy
- Led manual data labeling (2M tweets, 3-person team) to build a misinformation gold-standard dataset; benchmarked Random Forest against BERT/RoBERTa NLP classifiers (best model: 89% F1, 90%+ accuracy/precision via Databricks fine-tuning), published in an internal report used by CDC leadership
- Applied BERTopic topic modeling to unstructured comment data and built a no/low-code R Shiny interface enabling CDC-wide users to query travel data and review policy comments without writing SQL, cutting policy review time from 3 months to 1 month, validated through A/B testing against the prior manual process
- Engineered a real-time ELT pipeline (Azure Data Factory, Synapse) ingesting 1TB+ per batch of travel and pandemic surveillance data from all US airport and border control stations every 3-7 days, feeding COVID-19 emergency response reports delivered to HHS and the White House emergency health response team
- Synthesized 5-disease forecasting and 15+ surveillance trend-analysis reports informing COVID, Monkeypox, and Ebola response

- Parsed and harmonized terabyte-scale outpatient antibiotics utilization data from HealthVerity's long-term care EHR database into a normalized (3NF) structure, enabling medication-pattern, drug-utilization, and cost analysis for COVID-era antibiotic stewardship reporting
- Published a white paper on USCDI and EHR data standardization and briefed CDC leadership, serving as subject matter expert linking EHR data elements to USCDI use cases and strengthening data elements accepted into USCDI v4
- Standardized travel data capture elements for USCDI+, coordinating with HHS ONC and CMS; submissions accepted and documented in USCDI+ v1 and v2

**University of Virginia Medical Center** Feb 2018 – Jan 2021
*Clinical Pharmacy Intern* Charlottesville, VA

- Designed the UI blueprint for a new EPIC Ambulatory Care patient admission page, adopted by Epic analysts and the UVA pharmacy manager and implemented clinic-wide as the standard ambulatory care module, shaping patient visit workflow and information-capture timing for all clinic visits
- Reviewed medication prescriptions and supported therapy management, improving vaccination rate 150% (2019–2020)
- Co-developed an IRB-approved observational study protocol on inhaled aminoglycoside antibiotic use in outpatient cystic fibrosis patients (~7,000 patient records), owning study design, population/inclusion-exclusion definition, and data/treatment-operationalization, including a double-blind design for downstream analysis; study findings informed a separate published study and an internal cystic fibrosis practice guideline

**Early-Career Pharmacy Informatics Externships** – INOVA Medical Center, FDA, Kit Check Jan – Apr 2021
*Pharmacy & Pharmaceutical Informatics Intern* VA / MD

- Standardized drug administration routes in the EPIC Willow pharmacy EHR environment and built Tableau/SlicerDicer dashboards surfacing drug-administration gaps, reducing non-standard routes 15% and increasing biologic dispensing workflow speed 20% (INOVA Medical Center)
- Analyzed 5 years of Medicare Part D claims data to characterize biologics pricing trends and their projected effect on insurance coverage and policy over a 5-year horizon (FDA)
- Built an automated NDC data-cleaning pipeline standardizing pharmaceutical product codes against RxNorm and First Databank drug data, and designed a drug-shortage data capture method for a medication diversion detection program, improving QA testing speed 15% (Kit Check)

## EDUCATION & CERTIFICATIONS

**CDC Public Health Informatics Fellowship** 2021–2023
(2-year, doctoral-level appointment; travel-data epidemiology & emergency response focus)

**VCU Medical College of Virginia** 2021
Doctor of Pharmacy (PharmD), Emphasis in Informatics & Research

**Penn State University** 2015
BS, Biochemistry

HL7 v2.x, CDA R2 & FHIR Foundation Certification 2022
Certified Scrum Master (CSM) 2023

## PUBLICATIONS & PRESENTATIONS

Yu M, Barletta K, **Lan K**, Field S, Merono C, Goig D, Moegelmose S, Chuderland Ben Arie D, Hariton E, Zhou W. Real-World Insights on Controlled Ovarian Stimulation for IVF/ICSI: Data from a Large Fertility EMR Database in the US (2010-2024). Poster presented at ISPOR 2026; May 2026; Philadelphia, PA.

Seged S, **Lan K**, Zhan S, Eloso J, Biggers B, El Kalach R. The Public Health Informatics Fellowship Program: Training Pharmacists as Data Detectives. *J Am Pharm Assoc.* 2025;65(1):102216.

Katebi V, Panacci A, O'Sullivan T, **Lan K**, Morris J. COVID-19 Vaccine Misinformation Detection Using Classification Models in Cruise-Related Tweets. Internal CDC Report. Accepted.

Chaker S, Henry JA, Hung Y-C, et al. The Kenya Surgical Capacity Study: An Audit of Surgical Equipment/Infrastructure in Level 4 Public Hospitals. *Annals of African Surgery.* 2025;22(1):4-12.

Hung Y-C, Henry JA, Chaker SC, Saad M, Slater ED, Madiraju SK, Waiguru E, Mwai P, Jani P, **Lan K**, Nthumba P. The Kenya Surgical Capacity Study: An Audit of Surgical Workforce in Level Four Hospitals (Part I). *J Surg Res.* D-24-00004.

**Lan K**, Baubly E. DataMosaic Platform Development. Poster/oral presentation, NCEZID, Division of Global Migration and Quarantine, Atlanta, GA, 2022.

**Lan K**, Baubly E. Data Mosaic – SQL-based R Shiny UI Querying Web Application. CDC Public Health Informatics Fellowship Capstone Abstract Presentation, June 2023.

**Lan K**, Wojno A. USCDI Standard on Travel Data Elements Application and Capturing. CDC NCEZID News Update, March 2023.

**Lan K**, Baubly E. Global Travel Order COVID Travel Assessment Page. CDC, November 2021.

Bark A, **Lan K**. Do Pharmacists Speak Well of Pharmacy? A Word Sentiment Analysis of Pharmacy Subreddits. *The Happy PharmD*, September 2020.

**Lan K**, Jian YJ, Chang JH, Yang CM. Retrospective Analysis on Drug Resistance of Clonal Strain Isolated from Emergency Room. Interscience Conference on Infection and Chemotherapy, Korean Society of Infectious Diseases, Seoul, Korea, 2013.

**Lan K**, Vu D. AnyWear VR Startup Product Pitch Deck Competition. Virginia Commonwealth University Da Vinci Center, Richmond, VA, 2018.