An Agentic AI Framework Integrating Data-Driven Hypothesis Generation with Traditional Validation for Real-World Evaluation of Emerging Therapies

Peichang Shi, Kevin Shang, Jia Zhao, Catherine Major
Contact: Shi_peichang@bah.com

MSR139

Background & Objectives

Real-world evidence (RWE) is increasingly required to complement RCTs, but existing methods force investigators to specify outcomes in advance. This introduces selection bias for new therapies (where little is known) and emerging health threats (e.g., COVID-19), where no predefined hypotheses exist. Drug effectiveness is a systemic outcome that traditional hypothesis-driven approaches cannot fully capture.

To address this gap, we propose an agentic AI framework for hypothesis-free exploratory pattern detection in claims data, enabling unbiased discovery of unanticipated drug effects without prior assumptions, followed by traditional statistical validation of identified signals.

Framework

This study proposes an agentic AI framework for data-driven hypothesis generation using longitudinal claims data, which leverages BioClinicalBERT to autonomously detect temporal shifts in ICD/CPT coding patterns, identify clinically analogous patient trajectories and control cohorts, and generate literature-informed mechanistic hypotheses. Candidate hypotheses are prioritized through human-in-the-loop review before undergoing formal causal validation using quasi-experimental methods such as Difference-in-Differences (DiD). By integrating representation learning, pattern discovery, clinical reasoning, and econometric validation, the framework bridges exploratory AI-driven discovery with hypothesis-driven causal inference.

Case Study

Early Adoption of Apixaban vs. Warfarin (2016)
We demonstrate the framework using a retrospective claims-based study of early Apixaban adoption compared with Warfarin in 2016. Although Apixaban was approved by the FDA in late 2012, real-world population-level effects and utilization patterns were still emerging during the study period, providing a suitable setting for AI-assisted hypothesis generation and causal validation.
Data: 2016 Synthetic Healthcare Database (~2M patients)
Exposure: First Apixaban or Warfarin prescription in April 2016
Pre-index: 90-day ICD/CPT history before first fill
Follow-up: May–December 2016
Comparator: 1:1 propensity score matched Warfarin initiators

Agentic AI Workflow

Synthetic Healthcare Database for Research

↓

Cohort Selection
First Apixaban or Warfarin fill in April 2016 | Apixaban: n=431 | Warfarin: n=977

↓

Pre-index BERT Embedding (Bio_ClinicalBERT)
90-day lookback ICD/CPT codes as clinical sentence -> 768-dim [CLS] embedding per patient

↓

Propensity Score Matching
Logistic Regression on 768-dim embeddings -> PS | Greedy 1:1 caliper -> 295 matched pairs

↓

Post-index Code Collection (May-Dec 2016)
Inpatient + Outpatient ICD/CPT codes | Aggregate per patient -> BERT embedding

↓

Unsupervised Clustering
UMAP 768->20 dims | HDBSCAN | Chi-square per cluster | ICD co-occurrence networks

↓

AI-Driven Hypothesis Generation
Rank codes by OR | Group by ICD chapter | Map to clinical hypotheses | Literature check

↓

Human-in-the-Loop Review 2 | Clinical plausibility review | Hypothesis prioritization | Approve for DiD testing

↓

Difference-in-Differences Validation
outcome ~ group + period + group x period | Cluster-robust SE | Confirm/reject hypotheses

↓

Validated Clinical Evidence
Transparent | Reproducible | Regulatory-ready

Research Objectives

1. Develop BERT-based PSM using ICD/CPT sequences as patient representations
2. Apply HDBSCAN clustering to identify post-index clinical subgroups without predefined outcomes
3. Generate structured clinical hypotheses from AI-detected patterns and validate via DiD
4. Demonstrate a generalizable agentic AI framework for early evidence generation

Data & Patient Selection

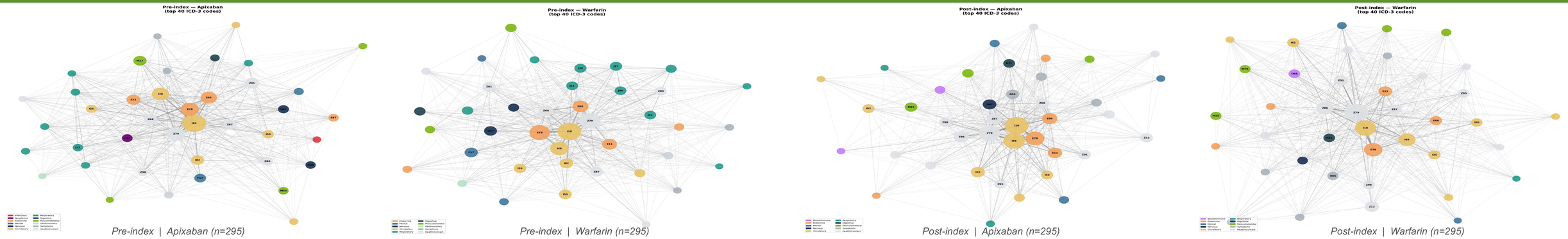
The Synthetic Healthcare Database for Research is publicly available synthetic claims (inpatient, outpatient, pharmacy) for ~2 million patients in 2016, structured to replicate distributional properties of real-world administrative data.

Selection Step	N
Synthetic Healthcare Database (2016)	~2,000,000
First anticoagulant fill in April 2016	1,408
-- Apixaban initiators	431
-- Warfarin initiators	977
With >=1 pre-index ICD/CPT code	910
After BERT-PSM (matched pairs)	**590**

Data set Limitations

- **Synthetic dataset:** Findings are based on synthetic claims data and may not fully represent real-world populations.
- **Sample size:** Smaller cohorts may reduce statistical power.
- **Covariate adjustment:** Analyses primarily used ICD/CPT codes without other demographic covariates
- **Single database/year:** Findings are based on a single 2016 dataset with a pre–post design rather than a longitudinal follow-up, which may limit generalizability..

ICD Co-occurrence Networks: Pre-index vs. Post-index (Apixaban and Warfarin)



Pre-index | Apixaban (n=295) *Pre-index | Warfarin (n=295)* *Post-index | Apixaban (n=295)* *Post-index | Warfarin (n=295)*

Nodes = ICD 3-character categories | Node size = patient frequency | Edge weight = co-occurrence within same patient | Color = ICD chapter | Top 40 codes per panel

AI-Generated Hypotheses

Apixaban Cluster -- Proposed Hypotheses

H1 -- Arrhythmia management: Greater post-index increases in cardioversion and catheter ablation, consistent with DOAC preference in AF rhythm-control strategies.
H2 -- Cardiac complexity: More post-index documentation of prior MI and cardiac surgery (CABG) history despite PSM

Warfarin Cluster -- Proposed Hypotheses

H3 -- Bleeding signal: Greater post-index increase in anemia codes, consistent with warfarin's known hemorrhagic risk.
H4 -- Long-term drug therapy coding
H5 -- Monitoring burden: Higher PT/INR test rates post-index reflecting mandatory anticoagulation monitoring absent with apixaban.

DiD Validation Results

Model: outcome ~ group + period + group x period | Cluster-robust SE

Code	Outcome	Pre APX	Post APX	Pre WAR	Post WAR	DiD	p-value
92960	Electrical cardioversion	0.7%	5.1%	0.7%	1.0%	+0.041	* 0.010
Z981	Cardiac surgery hx (CABG)	0.7%	3.1%	0.3%	0.0%	+0.027	* 0.011
--	**-- Warfarin increase --**						
Z79	Long-term drug therapy coding	0.043	0.071	0.032	0.193	-0.1321429	*<0.001
D6859	Anemia (other/unspecified)	0.3%	0.0%	0.3%	3.7%	-0.037	* 0.002
85610	PT/INR coagulation test	7.5%	8.1%	7.5%	14.6%	-0.064	* 0.048

All hypotheses H1-H5 confirmed. Apixaban (blue) | Warfarin (pink)

Clinical Interpretation

Apixaban – Arrhythmia & Cardiac Management

- 5x more cardioversions post-index vs. warfarin (5.1% vs 1.0%, DiD +4.1%, p=0.010) – consistent with DOAC periprocedural convenience (no INR monitoring or bridging required), not necessarily higher arrhythmia burden
- Higher post-index capture of prior CABG history (Z981: 0.7%→3.1% vs. 0.3%→0.0%, DiD +2.7%, p=0.011) – likely reflects more frequent follow-up and historical code completion in the apixaban group rather than new cardiac events

Warfarin – Monitoring Burden & Long-Term Therapy Coding

- PT/INR testing doubled post-index (85610: 7.5%→14.6% vs. 7.5%→8.1%, DiD –6.4%, p=0.048) – serves as a positive control, confirming expected warfarin monitoring requirements
- Long-term drug therapy coding (Z79) increased sharply in warfarin (3.2%→19.3% vs. 4.3%→7.1%, DiD –13.2%, p<0.001) – may reflect both more intensive follow-up and potentially a higher proportion of patients requiring indefinite anticoagulation (e.g., mechanical valves, recurrent VTE)

Warfarin – Bleeding Safety Signal

- Anemia emerged exclusively in warfarin (D6859: 0.3%→3.7% vs. 0.3%→0.0%, DiD –3.7%, p=0.002) – consistent with warfarin's known hemorrhagic risk; the absence of new anemia codes in apixaban is directionally reassuring

Conclusions

1. Agentic AI generated clinically plausible hypotheses without predefined outcomes—all 6 validated (DiD, p<0.05)
2. Apixaban – higher rates of cardioversion and historical cardiac code capture, consistent with DOAC convenience and more frequent outpatient encounters
3. Warfarin – expected increases in PT/INR monitoring and long-term therapy coding (both positive controls) plus a reproducible bleeding signal (anemia)
4. Human-in-the-loop ensured clinical validity with scalable, regulatory-grade rigor

Future Work

1. Evaluate larger real-world datasets and more robust causal inference methods.
2. Extend to multi-year longitudinal studies of disease trajectories and care pathways.
3. Incorporate additional covariates such as age, sex/gender etc.
4. Expand applications to fraud detection, inappropriate utilization, and anomalous billing patterns.

Key References

- Alsentzer E, et al. Publicly available clinical BERT embeddings. NAACL Clinical NLP Workshop. 2019.
- Granger CB, et al. Apixaban versus warfarin in AF (ARISTOTLE). N Engl J Med. 2011
- Hylek EM, et al. Major hemorrhage and tolerability of warfarin in the first year of therapy among elderly patients with atrial fibrillation. *Circulation.* 2007

Acknowledgement

We thank Christina Hedge and Christopher Brossart for their leadership.