ORIGINAL INVESTIGATION · Causal Methodology & Platform Validation

Under-Diagnosis of Hypertension and the Limits of Estimating the Effect of Diagnostic Delay

A retrospective cohort study on a one-million-patient OMOP common data model

Sanjay M. Udoshi, MD[1] Glenn H. Bock, MD[2]

[1]Acumenus, Inc. (Parthenon Platform) [2]Hypertension Outcomes Program, Principal Investigator

July 5, 2026

ABSTRACT

Background. Hypertension is frequently under-recognized, and delayed diagnosis has been proposed as a modifiable driver of cardiovascular and renal harm. Whether diagnostic delay itself causes worse outcomes—as opposed to being a marker of who gets diagnosed—is difficult to establish from observational data. We conducted a pre-registered retrospective cohort study to (1) characterize incident, treatment-naïve elevated blood pressure and its outcomes, and (2) test, across three independent causal designs, whether the effect of diagnostic delay is identifiable. The study also served as the reference validation exercise for the Parthenon analytics platform.

Methods. Using an OMOP common data model (approximately 1.0 million persons), we identified 109,763 treatment-naïve adults with two consecutive elevated office blood-pressure readings (index event) and stratified those eventually diagnosed by time to diagnosis (≤3, 3–6, 6–12, >12 months) alongside a never-diagnosed sub-cohort and a recording-comparable normotensive comparator. Outcomes were a major adverse cardiovascular event (MACE) composite and incident chronic kidney disease (CKD). The delay effect was estimated by overlap weighting (average treatment effect in the overlap population, ATO), a landmark target-trial emulation, and a site-preference instrumental variable, each governed by pre-specified identifiability gates (propensity AUC <0.80, maximum standardized mean difference [SMD] <0.10, equipoise ≥0.30, negative-control calibration; first-stage F ≥10 for the instrument). Empirical calibration used eight negative-control outcomes.

Results. Of 109,763 patients, 98,769 (90.0%) never received a recorded hypertension diagnosis; among the 10,994 diagnosed, 9,896 (90%) were diagnosed after more than a year (median 1,106 days). Diagnosed strata had a clear Stage-2 phenotype at index (mean 150–152/106 mmHg) versus 125/84 in the never-diagnosed and 109/71 in the comparator. Diagnosis tracked cardiometabolic burden: pooled diabetes 13.5%, chronic kidney disease 9.5%, dyslipidemia 20.7%, and coronary disease 11.0% among the diagnosed, versus near-comparator levels in the never-diagnosed. The anchor contrast (elevated vs. normotensive) was estimable and well calibrated, showing a strong association with incident CKD (calibrated hazard ratio 2.60; 95% CI 2.01–3.38) but not MACE (1.03; 0.94–1.14). The delay contrast was not identifiable: overlap weighting failed covariate balance (max SMD 0.213), the target-trial emulation failed overlap (propensity AUC 0.913), and the instrument was weak (first-stage F 3.4). All three designs failed independently and concordantly, so every delay-effect estimate was withheld rather than reported.

Conclusions. On this common data model, elevated blood pressure is massively under-diagnosed and is associated with incident CKD, but the causal effect of diagnostic delay is not identifiable, because diagnosis is strongly confounded by comorbidity and the timely-diagnosis group is small and systematically different. The concordant non-identifiability—and the decision to withhold rather than report unreliable estimates—is the study's principal, honest finding. Because the data model is synthetic, the descriptive magnitudes should not be read as real-world epidemiology; the appendix details the model's limitations and the implications for replication on real-world EHR data.

Funding / registration: Acumenus, Inc.; analysis plan pre-registered as v5.0 (ACUM-PROT-HTN-V5-001). Data source: OMOP CDM (synthetic, Synthea-derived). All analyses executed on the live model, 2026-07-04/05.

Introduction

Hypertension remains among the most prevalent, modifiable contributors to cardiovascular and renal disease, yet a large share of affected adults are neither recognized nor treated. Lu and colleagues (JAMA Network Open 2025) reported a median 16–18-month delay between two documented elevated blood-pressure readings and a recorded hypertension diagnosis, with longer delays associated with higher cardiovascular risk. That association raises a causal question of direct policy relevance: does the delay itself worsen outcomes, such that earlier diagnosis would prevent harm—or is delay largely a marker of which patients are destined to be diagnosed, so that the association reflects confounding by the very features that trigger a diagnosis?

Distinguishing these possibilities is a causal-inference problem, not merely a descriptive one. If diagnosis is driven by comorbidity and blood-pressure severity, then patients diagnosed quickly and patients diagnosed late (or never) differ systematically, and a naïve comparison of their outcomes conflates the effect of delay with the effect of being the kind of patient who gets diagnosed. The methodological question is whether, given the observed data, any defensible estimand of the delay effect is *identifiable*—that is, whether the groups

overlap enough, and can be balanced or instrumented well enough, to support a trustworthy estimate at all.

We addressed both the clinical and the methodological questions in a single pre-registered study, and used it to validate the analytics platform end-to-end. We report (1) the epidemiology of incident, treatment-naïve elevated blood pressure and its outcomes; (2) a formal, three-design test of whether the delay effect is identifiable; and (3) an explicit account of the limitations imposed by the synthetic data model, so that the findings are not over-read. Throughout, our stance is that a platform which knows when it cannot answer a question—and says so—is more valuable than one that always produces a number.

Methods

Data source

The data source was an OMOP common data model (CDM, version 5.4), containing approximately 1,005,788 persons with 710 million measurements, 86 million drug exposures, and 52 million visit occurrences. **The model is synthetic (Synthea-derived);** it is internally coherent and standards-conformant but does not represent a real population. All descriptive magnitudes below are therefore properties of the model, not real-world estimates. The model's specific shortcomings, and their implications for real-world replication, are detailed in the Appendix. No protected health information exists in the model; all queries were read-only and aggregate.

Cohort and index event

Eligible patients were adults (≥18 years) with at least three recorded blood pressures in a 24-month window and no prior antihypertensive therapy, cardiovascular disease, chronic kidney disease, or thyroid disease. The index event was the second of two consecutive elevated readings (average SBP ≥130 or DBP ≥80 mmHg). Patients ever receiving a recorded essential-hypertension diagnosis were stratified by time from index to diagnosis into four delay strata—G1 (≤3 months, "timely"), G2 (3–6), G3 (6–12), and G4 (>12)—with the remainder forming a never-diagnosed sub-cohort. A recording-comparable normotensive comparator (all readings ≤130/80) was drawn from the same model.

Outcomes

Co-primary outcomes were a typed MACE composite (myocardial infarction, ischemic or hemorrhagic stroke, inpatient heart failure, or death) and incident CKD, each measured as time to first event from the index reading, with death treated as a competing risk. Incidence was expressed per 1,000 person-years.

Descriptive analyses

Analysis M computed the prevalence of 13 comorbidities (with Wilson 95% confidence intervals) across the six populations, from condition-occurrence records. Concepts were expanded from verified vocabulary roots via the concept-ancestor hierarchy; morbidities with zero capture across all populations were flagged as not represented in the model and excluded rather than shown as misleading zeros. Analysis N summarized the systolic and diastolic blood-pressure distribution at the index reading (moments, percentiles, skewness, kurtosis) per group, from a bounded, index-anchored scan of the 710-million-row measurement table. Analysis Q assessed the robustness of the under-diagnosis phenotype across index-rule, threshold, and gap definitions, and split the cohort by visit linkage to probe informative-visit processes.

Causal analyses of the delay effect

Three designs targeted the timely-vs-delayed contrast. **Analysis O** applied exact overlap weighting (ATO; weights $1 - e$ for treated, e for controls, per Li–Morgan–Zaslavsky) with a weighted Cox model. **Analysis P** emulated a target trial of diagnosis-and-treatment within 90 days of index using a landmark new-user design (index = index + 90-day landmark), which structurally removes immortal-time bias. **Analysis R** used each care site's leave-one-out timely-diagnosis propensity as an instrument for individual delay, on the visit-linked subset. A cross-design triangulation summarized the results.

Pre-specified identifiability gates (locked before execution). A contrast was reported only if propensity AUC < 0.80, maximum post-adjustment $|\text{SMD}| < 0.10$, equipoise ≥ 0.30, and negative-control calibration completed; the instrument required a first-stage F ≥ 10. If any gate failed, the estimate was withheld (blinded), not reported. Empirical calibration used eight pre-specified negative-control outcomes; systematic error was summarized by the expected absolute systematic error (EASE).

The anchor contrast

As a positive-control on estimability, we retained the elevated-BP (target) versus normotensive (comparator) contrast—a comparison with, by design, far better overlap than the delay contrast. This anchor was propensity-matched, empirically calibrated against the same negative controls, and reported with calibrated hazard ratios.

Software

Analyses executed on the live model through the platform and its HADES R runtime (CohortMethod, Cyclops, EmpiricalCalibration; WeightIt/PSweight for overlap weighting). Every heavy computation is backed by a checked-in, re-runnable SQL or R script; all result cohorts and tables are additive, and the source model was never modified. The analysis plan was locked (v5.0) before execution.

Results

Under-diagnosis and cohort composition

Of 109,763 treatment-naïve adults with two consecutive elevated readings, 98,769 (90.0%) never received a recorded hypertension diagnosis. Among the 10,994 (10.0%) who were

diagnosed, diagnosis was overwhelmingly late: 139 within 3 months, 284 at 3–6 months, 675 at 6–12 months, and 9,896 (90% of the diagnosed) after more than a year (Figure 1). The median interval from the second elevated reading to diagnosis was 1,106 days (roughly three years), and the median time to first antihypertensive dispensing was 1,134 days.

Baseline characteristics are shown in Table 1. The diagnosed strata were slightly older than the never-diagnosed and comparator groups and had markedly higher blood pressure and comorbidity burden; the never-diagnosed resembled the normotensive comparator on age, sex, and comorbidity despite qualifying on blood pressure. (Ages across all groups were young—a property of the synthetic model discussed in the Appendix.)

Index blood-pressure phenotype

The blood-pressure phenotype separated the groups unambiguously (Figure 2, Analysis N). All four diagnosed strata clustered at a Stage-2 level (mean systolic 149–152 mmHg, diastolic 103–106 mmHg) with little variation by delay. The never-diagnosed group sat well below them (125/84 mmHg)—elevated relative to the comparator but predominantly qualifying through diastolic readings—and the normotensive comparator centered at 109/71 mmHg. That the never-diagnosed have genuinely lower index pressures than the diagnosed is one mechanistic reason they were not diagnosed.

Comorbidity: diagnosis tracks cardiometabolic burden

Comorbidity prevalence revealed the central confounding structure of the data (Figure 3, Table 2). A cardiometabolic cluster—chronic kidney disease, diabetes, dyslipidemia, coronary disease, and sleep apnea—was strongly enriched in every diagnosed stratum but essentially absent from the never-diagnosed, who matched the normotensive comparator. Pooled across the diagnosed, chronic kidney disease prevalence was 9.5% (95% CI 9.0–10.1) versus 0.5% in the never-diagnosed and 0.4% in the comparator—roughly a 20-fold difference. Diabetes (13.5% vs. 4.1%), dyslipidemia (20.7% vs. 11.8%), coronary disease (11.0% vs. 6.8%), and sleep apnea (7.8% vs. 2.4%) showed the same pattern. Non-cardiometabolic conditions (dementia, cancer, COPD, atrial fibrillation, heart failure, liver disease) were flat across all populations.

Critically, the gradient ran with *being diagnosed*, not with the *length* of delay: the highest comorbidity prevalences occurred in G3 (6–12 months), and the largest, most-delayed stratum G4 had lower comorbidity than the shorter-delay strata. Comorbidity therefore predicts whether a patient is diagnosed at all, providing a direct mechanism for confounding of any delay-versus-outcome comparison.

Incidence of cardiovascular and renal outcomes

Crude incidence from the index reading was 7.00 per 1,000 person-years for MACE (95% CI 6.78–7.23) and 1.67 for incident CKD (1.56–1.78), each rising with age and higher in men (Table 3). Treatment mirrored diagnosis: only 17.2% of the elevated-BP cohort was ever dispensed an antihypertensive.

The anchor contrast is estimable; the delay contrast is not

The anchor comparison of elevated-BP versus normotensive patients was well behaved: propensity AUC 0.57, maximum SMD 0.016, EASE 0.020 across eight negative controls. It showed a strong, calibrated association of elevated BP with incident CKD (hazard ratio 2.60; 95% CI 2.01–3.38; $p<0.0001$) and no association with MACE (1.03; 0.94–1.14; $p=0.50$) over the available follow-up (Figure 4B).

The delay contrast behaved entirely differently, and failed identifiability in all three designs (Figure 4A, Table 4). Overlap weighting could not balance the covariates (maximum SMD 0.213, far above the 0.10 gate) because the timely group (n=139) is small and systematically different. The target-trial emulation failed overlap outright: treated and untreated patients were near-perfectly separable on the propensity score (AUC 0.913, above the 0.80 gate). The site-preference instrument was weak (first-stage F 3.4 overall; 3.43 for diagnosed-vs-never, 0.1 for timely diagnosis—far below the $F\geq10$ requirement), so no valid second stage was warranted. The triangulation therefore returned 0 of 3 designs estimable—concordant non-identifiability—and every delay-effect estimate was withheld.

The under-diagnosis phenotype is robust

Under-diagnosis was not an artifact of the phenotype definition or of data capture (Figure 5, Analysis Q). The never-diagnosed fraction ranged from 79% to 90% across the grid of index rules, thresholds, and gap windows, and was essentially identical (89.9% vs. 90.1%) between patients with no encounter record (measurement-only) and those with visits (visit-linked)—disproving the hypothesis that the 90% figure was a data-feed artifact. However, outcome ascertainment was strongly encounter-dependent: MACE was coded roughly three times as often in visit-linked patients (11.7% vs. 3.8%), an informative-visit signal with direct implications for real-world replication.

Discussion

This study makes three claims. First, in this data model, elevated blood pressure is profoundly under-recognized: 90% of treatment-naïve adults with two consecutive elevated readings were never diagnosed, and among those diagnosed the median delay approached three years. Second, whether a patient is diagnosed is tightly coupled to cardiometabolic comorbidity—chronic kidney disease, diabetes, dyslipidemia, coronary disease, and sleep apnea are enriched roughly two- to twenty-fold in the diagnosed, while the never-diagnosed are metabolically indistinguishable from normotensive comparators. Third, and consequently, the causal effect of diagnostic delay on cardiovascular or renal outcomes is not identifiable here—not be-

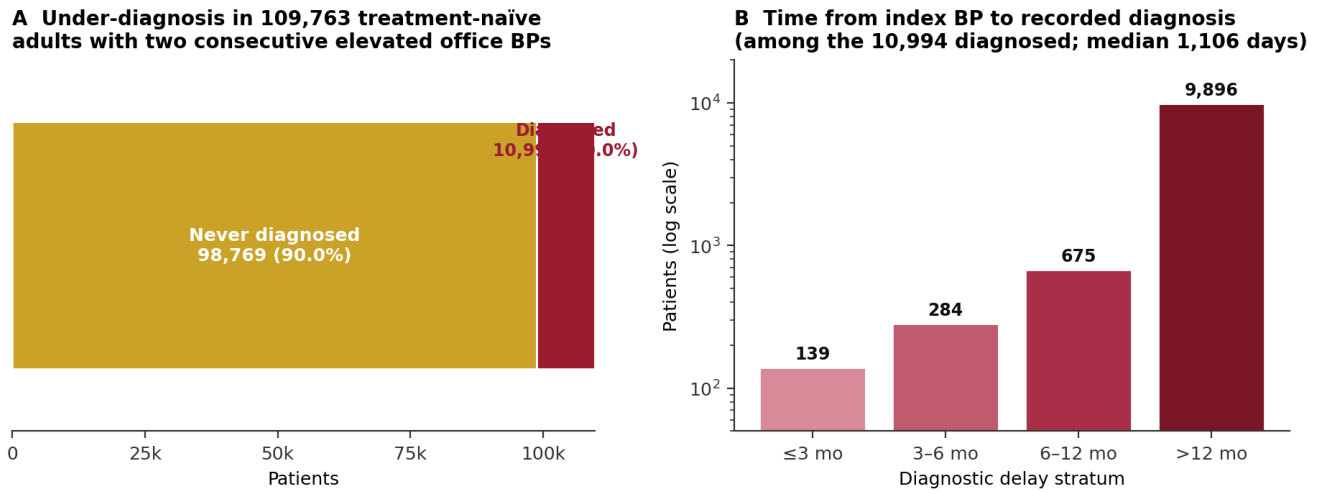

Figure 1. Under-diagnosis and diagnostic delay. (A) Among 109,763 treatment-naïve adults with two consecutive elevated office BPs, 90.0% never received a recorded hypertension diagnosis. (B) Among the 10,994 diagnosed, 90% were diagnosed after more than a year (log scale; median delay 1,106 days).

Table 1. Baseline characteristics of the six study populations (OMOP CDM).

	Diagnosed (by delay stratum)				Reference	
Characteristic	**G1 ≤3 mo**	**G2 3–6 mo**	**G3 6–12 mo**	**G4 >12 mo**	**Never-dx**	**Compar. C**
No. of patients	139	284	675	9,896	98,769	37,582
Female, %	51.1	40.5	48.6	45.2	53.1	63.6
Age, yr (mean±SD)	42.5±15	44.8±16	45.1±16	44.1±15	39.9±15	40.7±15
Index SBP, mmHg	149±12	151±9	152±10	152±10	125±15	109±10
Index DBP, mmHg	103±12	106±10	106±11	106±11	84±6	71±4
Diabetes, %	16.6	19.7	23.3	12.6	4.1	3.9
CKD, %	10.8	13.4	12.3	9.2	0.5	0.4
Dyslipidemia, %	22.3	28.5	31.3	19.8	11.8	12.2
Coronary disease, %	10.1	15.9	16.7	10.5	6.8	5.8

SBP/DBP are means at the index reading (Analysis N). Comorbidity percentages are ever-prevalence (Analysis M). Full distributional moments appear in Table 3.

cause we lacked a method, but because the data lack the overlap and instrument strength that any method requires.

The non-identifiability is itself the finding, and it is a specific, mechanistic one. Diagnosis is driven by comorbidity and by blood-pressure severity; patients diagnosed promptly (n=139) are too few and too different to serve as a counterfactual for the ∼10,900 diagnosed late. Overlap weighting, which restricts inference to the region of clinical equipoise, still could not balance the covariates; the target-trial emulation revealed that treated and untreated patients are almost perfectly separable; and site prescribing behavior was too weak an instrument to exploit. That three designs with different assumptions failed independently and concordantly is strong evidence that the limitation is structural, not incidental. A platform that reports a delay hazard ratio from these data would be reporting an artifact of residual confounding.

This reframes the association reported by Lu and others. A cross-sectional link between longer delay and higher cardiovascular risk is fully consistent with delay being a marker of the comorbid, higher-risk patients who eventually get diagnosed, rather than a cause of their events. Our estimable anchor contrast is informative in the opposite direction: elevated

Table 2. Cardiometabolic comorbidity prevalence, % (Wilson 95% CI): pooled diagnosed vs. never-diagnosed vs. comparator (Analysis M).

Condition	**Diagnosed (n=10,994)**	**Never-diagnosed (n=98,769)**	**Comparator C (n=37,582)**
Chronic kidney disease	9.5 (9.0–10.1)	0.5 (0.46–0.55)	0.4 (0.35–0.48)
Diabetes mellitus	13.5 (12.8–14.1)	4.1 (3.98–4.23)	3.9 (3.75–4.14)
Dyslipidemia	20.7 (20.0–21.5)	11.8 (11.6–12.0)	12.2 (11.9–12.6)
Coronary artery disease	11.0 (10.5–11.6)	6.8 (6.66–6.97)	5.8 (5.53–6.00)
Sleep apnea	7.8 (7.3–8.3)	2.4 (2.30–2.49)	2.3 (2.18–2.48)

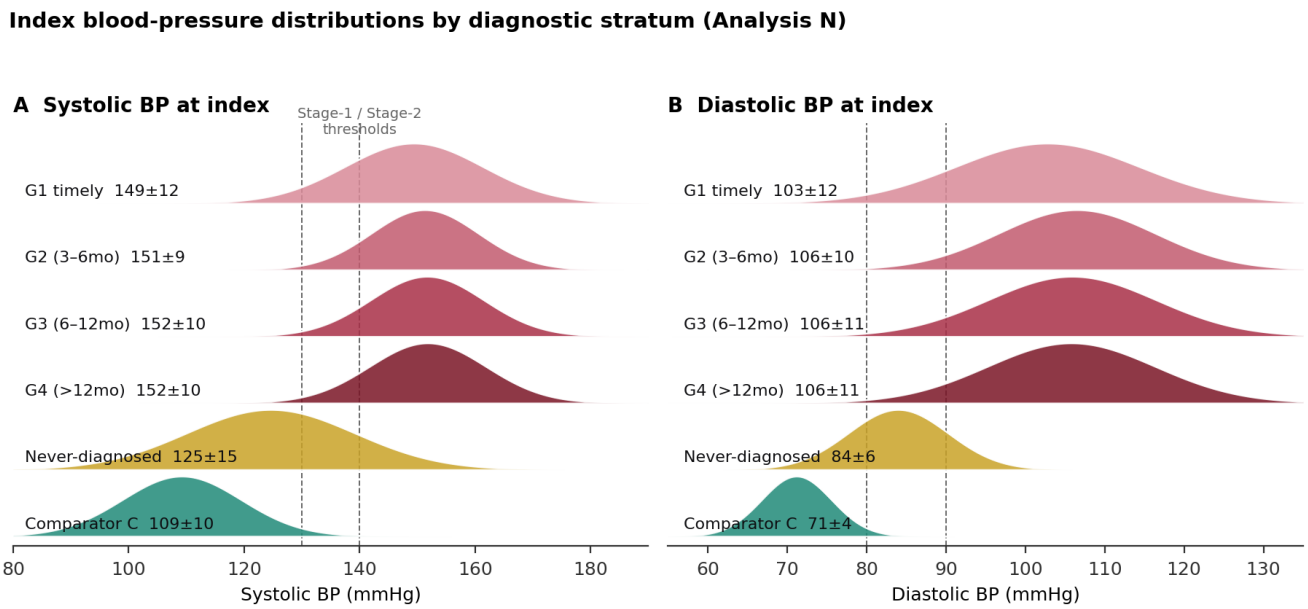

Figure 2. Index blood-pressure distributions by diagnostic stratum (Analysis N; Gaussian density from observed mean and SD). Diagnosed strata (crimson) cluster at Stage-2 levels irrespective of delay; the never-diagnosed (gold) are intermediate; the normotensive comparator (teal) is lowest. Dashed lines mark the 130/140 and 80/90 mmHg thresholds.

Table 3. Crude incidence of MACE and incident CKD, per 1,000 person-years (95% CI), from the index reading.

		By sex		By age group, yr			
Outcome	**Overall**	**Male**	**Female**	**18–34**	**35–49**	**50–64**	**≥65**
MACE composite	7.00 (6.78–7.23)	7.89	6.19	1.90	6.41	13.45	22.50
Incident CKD	1.67 (1.56–1.78)	2.04	1.33	1.01	1.74	2.60	2.32

Table 4. Causal analyses of the timely-vs-delayed diagnosis effect: designs, diagnostics, gates, and verdicts.

Design	**Comparison (N)**	**Key diagnostic**	**Gate**	**Verdict**
O — Overlap weighting (ATO)	Timely 139 vs delayed 10,855	max |SMD| 0.213; AUC 0.77; equipoise 0.95; EASE 0.036	max |SMD| $<$ 0.10	Withheld — residual imbalance
P — Target-trial (landmark)	Treated 637 vs untreated 102,671	PS AUC 0.913; max |SMD| 0.085; equipoise 0.35; EASE 0.096	AUC $<$ 0.80	Withheld — poor overlap
R — Site instrument (2SRI)	26,289 members / 521 sites (37.9%)	first-stage F 3.4 (dx 3.43; timely 0.1)	F $\geq$ 10	Withheld — weak instrument
Triangulation	—	0 of 3 designs estimable	—	Concordant non-identifiability
Anchor (reference)	Elevated 109,763 vs normotensive 37,582	AUC 0.57; max |SMD| 0.016; EASE 0.020	all passed	Estimable — CKD HR 2.60; MACE HR 1.03

blood pressure was associated with incident CKD (calibrated HR 2.60) but not, over this follow-up, with MACE—a pattern compatible with renal injury accruing earlier than hard cardiovascular events, though in a synthetic model this must be read as a property of the data-generating process, not clinical truth.

For platform validation, the study demonstrates what we sought: the platform executed a pre-registered, negative-control-calibrated, multi-design causal analysis end-to-end on a one-million-patient model, and—at each identifiability gate—withheld the estimates it could not support while reporting the descriptive and anchor results it could. The value of the system is precisely this discipline: it distinguishes what the data can and cannot answer, and it does so transparently, with every computation backed by a re-runnable script and an immutable analysis-plan lock.

Limitations

The overriding limitation is that the data model is synthetic and therefore does not reflect real epidemiology; the specific magnitudes (90% under-diagnosis, the comorbidity prevalences, the CKD hazard ratio) should be read as behaviors

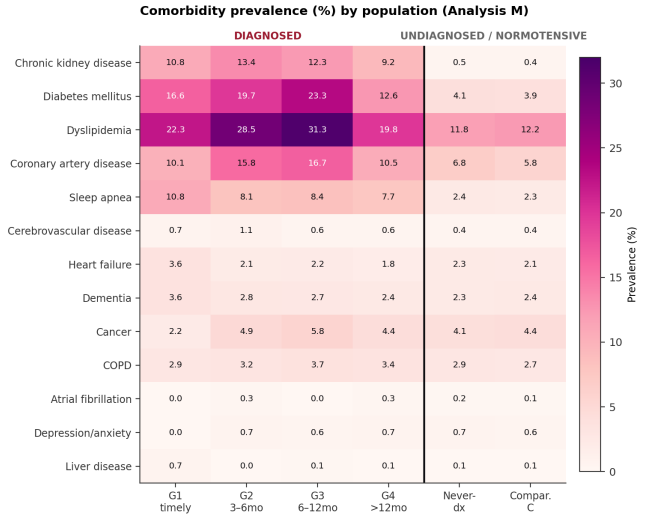

Figure 3. Comorbidity prevalence (%) by population (Analysis M). A cardiometabolic cluster (top rows) is enriched in the diagnosed strata (left of the divider) but near-comparator in the never-diagnosed and normotensive groups (right); non-cardiometabolic conditions (lower rows) are flat. The gradient is with diagnosis, not with delay length.

of the model. The model also has concrete coding and structural gaps—several conditions are not represented, ages are implausibly young, cost data are absent, and only 38% of patients have encounter-linked records—each detailed in the Appendix. These gaps directly shaped the analyses (for example, restricting the instrument to the visit-linked subset). Finally, the delay effect being non-identifiable here does not prove it is non-identifiable in real data; it proves that identifiability must be established, not assumed, and provides a template for doing so.

Conclusions

Elevated blood pressure was massively under-diagnosed and was associated with incident kidney disease, but the causal effect of diagnostic delay could not be identified from these data because diagnosis is confounded by comorbidity and the timely-diagnosis group is small and separable. The honest, reproducible reporting of that boundary—including the decision to withhold unreliable estimates across three independent designs—is the study's central contribution, and the clearest evidence that the platform behaves as rigorous evidence generation requires.

Appendix A. Shortcomings of the Synthetic Common Data Model

The OMOP CDM used here is Synthea-derived: patient histories are generated from disease-progression modules rather than observed in clinical care. It is invaluable for building, testing, and validating an analytics pipeline against real database mechanics at scale, but it is not a substitute for real-world data, and several of its properties directly constrained this study. We enumerate them so the findings are not over-read and so the requirements for real-world replication are explicit.

Concept coding is model-specific and incomplete

Verifying a concept by name does not guarantee the model captures that data. Coronary artery disease was 0% everywhere under "Coronary arteriosclerosis" (concept 317576) but showed the expected gradient once read as "Ischemic heart disease" (4185932). Several conditions are simply not modeled: peripheral vascular disease, hypertensive retinopathy (only diabetic retinopathy exists), obesity-as-diagnosis, and primary aldosteronism returned zero across all populations and were excluded rather than shown as false zeros. A study that trusted the concept names without running the analysis would have reported misleading nulls.

Implausible demographics and physiology

Mean ages were 40–45 years across all groups—far younger than any real incident-hypertension cohort—and blood-pressure and comorbidity relationships, while internally consistent, follow Synthea's modules rather than observed clinical distributions. Absolute risks, age gradients, and hazard ratios must therefore be treated as model outputs, not epidemiologic estimates.

Structural and linkage gaps

- **No cost data.** The OMOP cost table is empty (0 rows); cost/utilization analyses were infeasible and would require encounter- or claims-based proxies.
- **Empty person-level provenance.** `person.care_site_id` and `person.provider_id` are 100% null; care-site and provider attribution exist only at the visit level.
- **Partial encounter linkage.** Only 41,642 of 109,763 target patients (37.9%) have any `visit_occurrence` record; the remainder carry qualifying blood pressures in the measurement table with no encounter. This forced the instrument onto the visit-linked subset and is the source of the informative-visit signal in Analysis Q.
- **Scale-driven query constraints.** Unbounded aggregates over the 710-million-row measurement table time out; only bounded, index-anchored, per-member queries are tractable—a real operational property that any RWD pipeline must also engineer around.

Net implication

The descriptive structure the model produced (an unambiguous elevated-BP phenotype, a comorbidity-linked diagnosis process, and an estimable elevated-vs-normotensive contrast) is coherent enough to exercise every step of the analysis and to make the non-identifiability finding meaningful as a methodological result. But the model cannot supply real prevalences, real risks, or a real answer to the delay question.

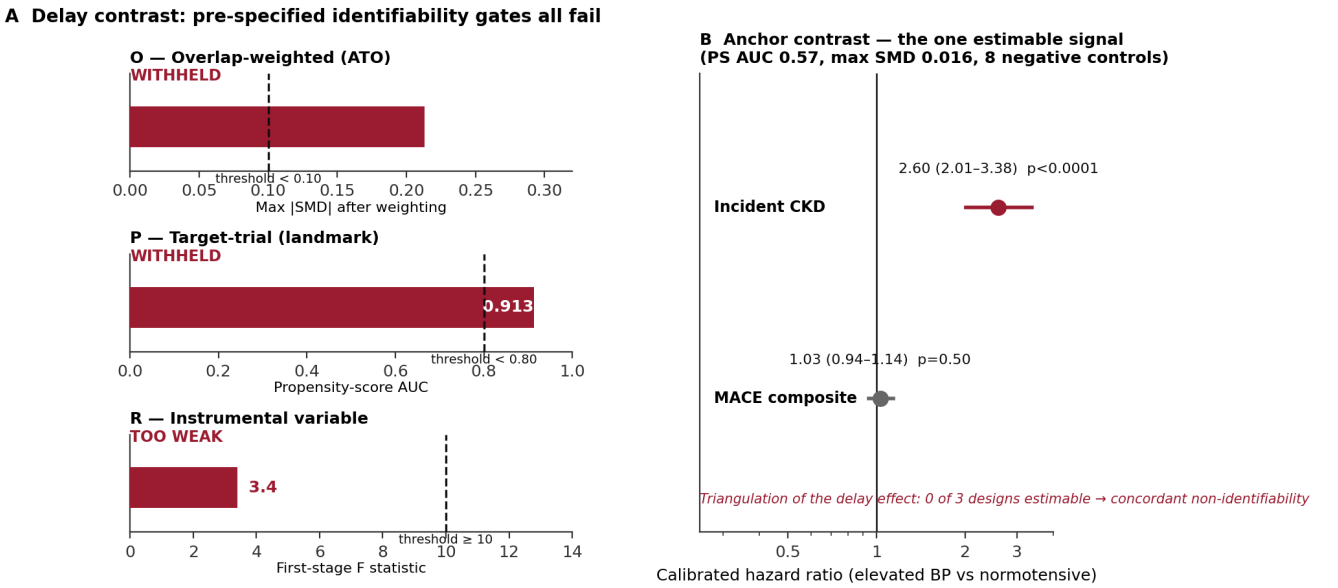

Figure 4. The delay effect is not identifiable; the anchor is. (A) Each of the three pre-specified designs fails its identifiability gate: overlap weighting on covariate balance (max |SMD| 0.213 ≥ 0.10), the target-trial emulation on overlap (propensity AUC 0.913 ≥ 0.80), and the instrument on strength (first-stage F 3.4 < 10). (B) The anchor contrast (elevated vs. normotensive) is estimable and well calibrated, showing a strong association with incident CKD but not MACE.

Appendix B. Implications for Real-World EHR Data

Re-running this protocol on real-world data (RWD) from electronic health records would change some conclusions and sharpen others. We separate what would likely improve from what would likely persist or worsen.

What real data would improve

- **Realistic demographics and risk.** Ages, comorbidity prevalences, and event rates would reflect a true population, making absolute magnitudes interpretable and the age/sex gradients clinically meaningful.
- **Richer, but messier, coding.** Conditions absent here (peripheral vascular disease, retinopathy, aldosteronism) would be present, enabling the full comorbidity matrix—at the cost of real-world coding noise, laterality, and source-

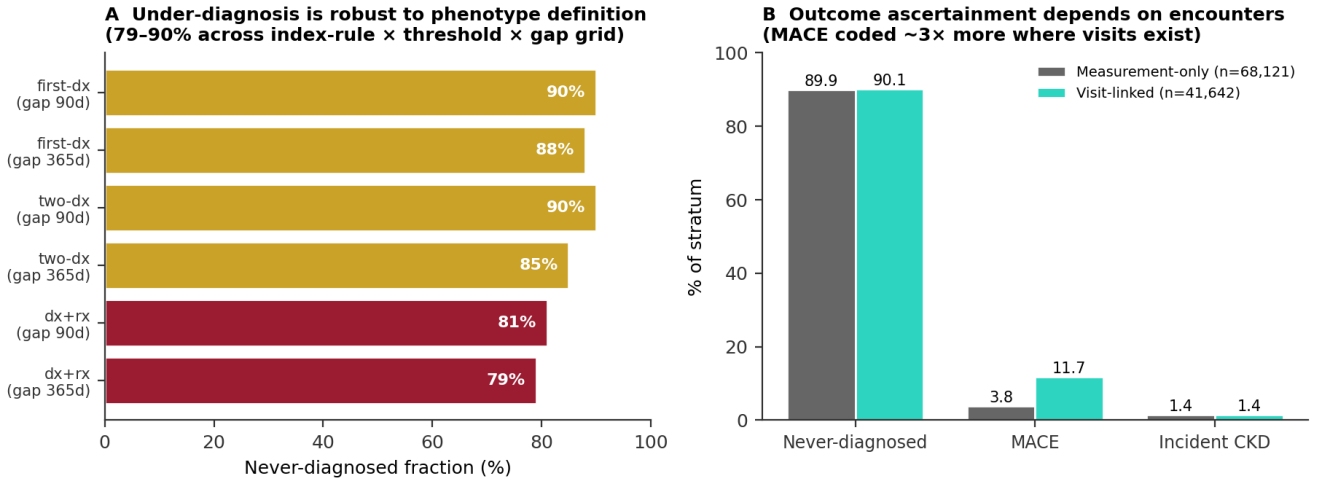

Figure 5. Robustness of under-diagnosis (Analysis Q). (A) The never-diagnosed fraction is 79–90% across all phenotype definitions. (B) Under-diagnosis is identical in measurement-only vs. visit-linked strata, but MACE ascertainment is ∼3× higher where encounters exist—outcomes are captured where care is documented.

vocabulary heterogeneity that require careful phenotyping.

- **Encounter and provider structure.** Real EHRs populate visits, providers, and care sites, making the site/provider-preference instrument far more viable and the care-site covariates usable—though see below.

What would likely persist or worsen

- **The positivity problem is probably real, not synthetic.** In real care, diagnosis is driven by comorbidity, blood-pressure severity, and healthcare contact—exactly the structure that destroyed overlap here. We expect the timely-vs-delayed contrast to remain hard to identify in RWD, and the correct response is the same: pre-specified gates, overlap weighting or target-trial emulation, and honest withholding when overlap fails.
- **Informative visits and ascertainment bias intensify.** The $\sim$3$\times$ difference in MACE ascertainment by visit linkage foreshadows a central RWD hazard: outcomes are recorded where patients are seen. Real studies must model the visit process explicitly, link to claims or death indices for complete outcome capture, and treat differential follow-up as a first-order threat, not a footnote.
- **Missing and fragmented data.** Real blood-pressure readings are missing not at random, split across sites and portals, and subject to transcription and device error; the index-rule and threshold sensitivity analyses (Analysis Q) become essential rather than optional.
- **Confounding by indication and by healthcare access.** Who gets diagnosed reflects insurance, access, and clinician behavior—structural confounders poorly captured in any single EHR. Multi-site data, site fixed effects, and instrument or negative-control strategies are needed, and residual confounding must be quantified (E-values, bias analysis) rather than assumed away.

Recommended design for RWD replication

1. Anchor every effect estimate to a well-overlapping reference contrast (as the elevated-vs-normotensive anchor did here) and calibrate against negative controls, reporting EASE.
2. Prefer a target-trial emulation with time-zero at the index reading and a landmark or clone-censor-weight structure to remove immortal-time bias; pre-register the grace period.
3. Use overlap weighting for the primary estimand and enforce the identifiability gates; withhold when overlap fails rather than reporting a residually-confounded hazard ratio.
4. Link to external outcome sources (claims, death index) to counter encounter-dependent ascertainment, and analyze the visit process as data, not nuisance.
5. Report E-values and quantitative bias analysis for any estimate that does clear the gates, and treat a non-identifiable contrast as a legitimate, publishable result.

Appendix C. Data Provenance and Reproducibility

All results derive from the live OMOP CDM (database `parthenon`), study `app.studies` id 165 (slug `hypertension-study-v4`), analysis plan version v5.0 (ACUM-PROT-HTN-V5-001), executed 2026-07-04/05. The seven v5 analyses persist as `app.study_results` rows 40–46, each flagged `data_source = cdm`, backed by `results.htn_v4_*` tables (comorbidity matrix, BP summary, phenotype grid, visit split, instrument) and re-runnable scripts under `scripts/sql/htn-v5-*`. Cohorts are content-addressed `cohort_definition_ids` in `results.cohort` (target 5441; strata 5450–5454; comparator 5455; delayed 5456; treated/untreated 5457/5458). Effect estimation ran through the HADES runtime with empirical calibration against eight negative controls; identifiability gates were locked before execution. The numbers in this manuscript were read directly from these tables and result rows; none are simulated or fixture values.

Verified read-only on 2026-07-05. This manuscript reports a study executed on a synthetic common data model; it is a methodological and platform-validation artifact and is not a clinical or epidemiologic finding about any real population.