Comparing pooled data, meta-analysis, and federated learning approaches to emulate EHR-based target trials across health systems

Max Sunog[1], **Marie-Laure Charpignon** [234], **Colin Magdamo**[4], **Jeff Klann**[1], **Youssef Hbid**[5], **Haoyang Li** [6], **Chengxi Zang** [6], **Mohit Sharma** [6], **Anish Kasam**[7], **Antonio Surcel**[8], **Ran Abuhasira**[1], **Thibaut Horel**[9], **Munther Dahleh** [9], **Deborah Blacker** [14], **Fei Wang** [6], **Joanna Tzoulaki**[5], **Michael Hogarth**[7], **Bella Vakulenko-Lagun**[8], **Mark Albers**[14]

[1] **Massachusetts General Hospital,** [2] **University of California, Berkeley,**
[3] **Kaiser Permanente,** [4] **Harvard Medical School,** [5] **Imperial College London**
[6] **Weill Cornell Medical School,** [7] **University of California, San Diego**
[8] **University of Haifa,** [9] **Massachusetts Institute of Technology**

Abstract

Many forms of healthcare research, including studies using the target trial emulation (TTE) framework, rely on data from the electronic health record (EHR). Because data from multiple EHRs often cannot be combined, studies that require multiple data sources often collaborate by combining independent analyses via meta-analysis (MA). However, MA can be ineffective when there is high heterogeneity among sites or when the outcome of interest is rare, two common scenarios in TTE. Alternatively, de-identified data from multiple health systems have been aggregated in secure enclaves. While valuable, the setup and maintenance of these platforms present significant administrative challenges, and restrictions on uploading sensitive data hinders their utility. To address these limitations, federated learning (FL) methods facilitate collaboration when MA and data enclaves are insufficient. To illustrate the advantages of FL, we used data in the ENACT enclave from EHRs in Massachusetts and California to empirically compare TTE results obtained via MA and FL to reference results derived from pooled data. FL consistently produced results closer to the reference than MA, with a larger effect for rarer outcomes. These findings motivate the creation of DRIAD-FL, our platform to expand methods for federating TTEs across five diverse health systems located in the US, the United Kingdom, and Israel.

Keywords: Federated learning, target trial emulation, meta-analysis, dementia

Data and Code Availability The study uses EHR data from the Research Patient Data Registry (RPDR) in Massachusetts, the University of California Health Data Warehouse (UCHDW) in San Diego, INSIGHT Clinical Research Network (ICRN) in New York, the Clinical Practice Research Datalink (CPRD) in the United Kingdom, and Clalit in Israel. Because the data contain patient information, they cannot be made available.

Institutional Review Board (IRB) This research was performed under MGB IRB approval (protocol 2023P000604).

1. Introduction

In drug repurposing research, the decades of medical history available for millions of patients within the EHR enable target trial emulation (TTE), the retrospective emulation of a randomized controlled trial (RCT) (Hernan and Robins, 2016). To execute a TTE, potential participants are identified using prescription records of the treatment of interest or the control drug/drug class. Exclusion criteria, based on prior drug prescriptions, diagnoses, and sociodemographics, are used to restrict the pool of patients to those who would be eligible for the RCT. In the resulting cohort, follow-up times are determined using

records indicative of the outcome(s) of interest or their last recorded healthcare encounter.

Although RCTs remain the gold standard to evaluate the clinical effects of a treatment, they can be infeasible due to cost considerations, recruitment challenges, or the required length of follow-up. For instance, Alzheimer's disease and related dementias (ADRD) have a prodromal phase of up to 15 years before clinical symptoms manifest (Bateman et al., 2012). For RCTs that measure the effect of repurposing treatment candidates, this extensive pre-clinical phase poses a challenge; in such cases, TTE is a valuable tool.

However, EHR-based TTE is often limited by the scale of the source health system, regardless of the total number of patients in the underlying EHR. For example, Massachusetts General Brigham (MGB) is a reference center for the treatment of amyotrophic lateral sclerosis (ALS) and the associated data warehouse comprises over 13 million patients, but has data for only ∼10k patients with ALS (1991-2025). Additionally, strict eligibility criteria may reduce cohort size. For example, removing patients without a primary care provider in the considered health system to maximize data quality can reduce cohort size by one third (Sunog et al., 2025). Further, a TTE in a single EHR may not be representative of the target population, sociodemographically or clinically, as providers in a given system are likely to make similar clinical and charting decisions (e.g., based on internal protocols).

To address these challenges, aggregating data from many EHRs would be ideal. However, important concerns about patient data privacy and security, as well as intellectual property considerations, prohibit or drastically encumber the sharing of data across health systems. In light of these restrictions, meta-analysis (MA) has traditionally been a convenient method to leverage data from multiple EHRs without sharing individual patient information. In MA, TTEs are first executed independently at each site, and the resulting estimated treatment effects are combined (e.g., using a weighted average, with weights inversely proportional to the variance of the estimated effect) (Borenstein et al., 2007).

Despite the flexibility granted by minimal data sharing requirements, MA has limited applications. For example, a traditional MA assumes that each contributing study used the same set of variables (Qin et al., 2022), which requires significant data harmonization among sites a priori. Moreover, MA is typically highly biased when heterogeneity across sites is high or when the event rate at a site is too small (1% is often used as a heuristic, but the necessary event rate depends on many factors and may not be evident) (Efthimiou, 2018). While established methods can mitigate such bias in certain cases (discussed in **Appendix A**), strict constraints can limit their utility in a drug repurposing TTE. Thus, MA is often ineffective for TTE when the outcome is rare or when the estimation of heterogeneous treatment effects requires granular stratification. Further, a study with zero observed events will generally be discarded from a MA. TTEs with a long maximum follow-up period face similar challenges, as the size of the risk set diminishes at later time points because participants either experience the outcome(s) of interest or are censored due to loss to follow-up.

Beyond statistical innovation, the construction of enclaves – secure, centralized databases of cohorts aggregated from many health systems – allows for the implementation of TTEs using the gold standard of pooled EHR data. Despite this attractive feature, setting up enclaves is often administratively burdensome. In fact, adding a single site to an existing enclave requires complying with their security protocols and drafting new data use agreements with all currently participating institutions. These steps become increasingly demanding as more sites join (particularly international partners), thereby limiting the scalability of enclaves. Enclaves are also limited by the de-identification process, which typically involves removing sensitive data such as zip codes and provider notes. Although such information is not strictly necessary for TTE, it can enhance emulation quality; zip codes can reduce unmeasured confounding via mapping to socioeconomic status, and provider notes can complement structured data to refine the timing and ascertainment of clinical events. Further, the variety of EHR data schemas (e.g., i2b2, OMOP, and PCORnet), coding systems (e.g., ICD9, ICD10, SNOMED), and site-specific implementations can disrupt data aggregation and increase the extent of missing data within an enclave (Cook et al., 2022). Despite these limitations, enclaves uniquely enable pooling EHR data

and are a critical tool to assess the utility of collaborative strategies through the comparison of methods using pooled data with distributed ones.

More recently, federated learning (FL) has been proposed as an alternative strategy to leverage multiple EHRs in TTE. In FL, each site trains a model locally and distributes model summary data (e.g., gradients) to others to learn a shared model incorporating information from all sites. Because the shared data are not patient-level information, FL has fewer barriers to entry than enclave participation. It also does not restrict the use of sensitive attributes. Infrastructure needs vary, but all FL methods require harmonizing trial specifications (e.g., exclusion criteria, outcome definition) and at least one round of communication during model training (i.e., sharing gradients). However, harmonization requirements can be less restrictive in FL than in MA (Zeng et al., 2024) (Han et al., 2023).

In contrast to MA, FL approaches allow the inclusion of studies with zero events (Schuemie et al., 2021). Moreover, in simulation studies, FL treatment effect estimates are closer to those from pooled data than from MA, especially in the presence of high heterogeneity or rare outcomes. An evaluation of federated ODACoR methods (One-shot Distributed Algorithms for Competing Risk) found that the bias of the subdistribution HR estimate, relative to that obtained using pooled data, was much lower (.2% to 8.5%) than with MA (39.7%) when the outcome incidence rate was set to .5% (Zhang et al., 2024). These results promote FL for TTEs with rare outcomes (e.g., ALS, which affects only .0052% of the US population (Mehta et al., 2021)).

To address this clinical reality, our team with access to structurally diverse EHRs located in Massachusetts, New York, California, the United Kingdom, and Israel, developed Drug Repurposing in AD-Federated Learning (DRIAD-FL), a platform for TTE federation. To demonstrate the utility of our platform, a pilot TTE was performed using data from two of these health systems: MGB in Massachusetts and the University of California Health Data Warehouse (UCHDW) in San Diego. Datasets were compiled in the ENACT enclave to compare a reference pooled analysis with MA and FL. Although this study is limited to the two participating sites with data in ENACT, it illustrates the importance of FL for future collaborative TTEs in the full group.

2. Methods

We compared the effect of the antidiabetic drugs metformin and sulfonylureas on ADRD onset, accounting for the competing risk of death. Our TTE specification followed Charpignon et al. (2022), with deviations caused by limitations of the enclave (full details in **Appendix B**).

The federated TTE was executed using a modified version of ODACoR-O (Zhang et al., 2024). In this few-shot method, each site initially executes a local TTE to produce a Cox proportional hazards model for the subdistribution hazard ratio. Then, every site shares the set of local time-to-event values for patients with the primary outcome, as well as point estimates and variances for coefficients of outcome models, which are used to calculate the corresponding MA estimates. Next, each site calculates and shares summary-level statistics characterizing their local risk sets at all times represented in the pooled set of time-to-event values. These statistics are collectively sufficient to produce the first and second derivatives of the pooled likelihood. Finally, these derivatives are applied to produce the final, federated coefficients, via the Newton method.

In this study, we introduced two modifications tailored to our causal setting. First, we altered the definition of the risk set. In ODACoR, it is defined at a given time t as the union of two groups: (1) participants i with time-to-event values $t_i \geq t$ and (2) participants who experienced the competing event. The use of this risk set produces a subdistribution hazard; alternatively, estimating the causal effect of treatment on outcome incidence requires using the cause-specific hazard instead. This entails defining the risk set as only the former group (1) in the initial local TTEs, as well as in calculations of summary-level statistics and pooled likelihood gradients. Next, we introduced inverse propensity of treatment weighting to emulate the randomization characteristic of a RCT and isolate the causal effect of treatment. Logistic regression models for the propensity score were trained separately at each site and stabilized average treatment effect weights were derived accordingly. These weights were used in both initial local TTEs

and summary-level statistic calculations, requiring the shared sets of time-to-event values to be augmented with corresponding sets of weights.

With this modified federation method, the TTE was executed analogously as a pooled TTE to produce a gold standard HR. The pooled HR was used to evaluate the accuracy of HRs generated via MA of local TTEs and federated TTE.

3. Results

Table 1: Abridged summary table stratified by cohort (A-MGB, B-UCHDW) and treatment arm. Full table in **Appendix C**. HTN: hypertension, CVD: cardiovascular disease. Comorbidities identified via diagnosis records.

Feature	Met-A	Sulf-A	Met-B	Sulf-B
Total	43655	5240	4491	372
% of Site	89.3%	10.7%	92.4%	7.6%
ADRD	6.9%	9.0%	5.6%	7.0%
Death	6.8%	15.9%	5.8%	10.2%
Age	65.7	70.4	66.1	68.6
Sex F	49.8%	45.9%	52.2%	49.2%
HTN	53.9%	45.5%	39.6%	37.9%
Obesity	16.3%	6.1%	7.4%	5.4%
CVD	16.8%	16.9%	11.4%	12.9%
Cancer	17.7%	17.8%	17.5%	13.2%

The cohorts sourced from the MGB and UCHDW databases had important differences (Table 1). After applying eligibility criteria, the MGB and UCHDW cohorts comprised 47,895 and 4,863 patients, respectively. While the relative proportions of treatment arms were similar, the cumulative incidence rates of ADRD and ACM by the end of follow-up were higher in the MGB cohort (ADRD: 7.2%, ACM: 7.8%) than in the UCHDW cohort (ADRD: 5.7%, ACM: 6.1%). Notably, at baseline, MGB patients had a 30.7% longer history in the EHR system than their UCHDW counterparts, on average, allowing a more complete capture of conditions in the patient's EHR. This difference may have contributed to the higher baseline prevalence rates of each comorbidity and to the lower share of patients with missing HbA1C and BMI values in the MGB cohort (**Appendix C**). However, these contrasts could also reflect differences in the composition of the distinct underlying populations, which may ultimately result in a heterogeneous baseline risk of ADRD between the two systems.

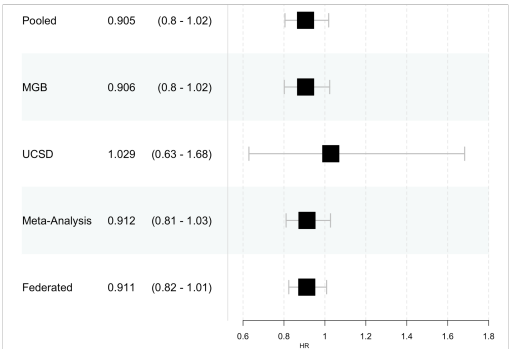

Figure 1: HR of initiating metformin vs. sulfonylureas on ADRD using pooled data, MA, and FL. Point estimates and 95% CIs.

The estimated HRs resulting from every analysis revealed a similar nearly significant protective effect of metformin, except for the single-site UCHDW trial (1.029; CI: .63-1.68), which yielded a very wide CI due to a small cohort size (Figure 1). The MGB trial produced a HR (.906; .8-1.02) within the CI in Charpignon et al. (2022) (.81; .69-.94), although the effect size was affected by the modified trial specification. The FL HR (.911; .82-1.01) was very close to that of the reference pooled analysis (.905; .79-1.02). MA also produced a HR (.912; .81-1.03) close to the reference, albeit less precise and slightly less accurate.

To evaluate more challenging scenarios, three sensitivity analyses were conducted, resulting in increasingly lower outcome incidence rates: (1) memory loss diagnoses removed from the outcome definition (4.4% outcome incidence), (2) only dementia diagnoses kept in the outcome (3.3%), and (3) outcome defined as in (2) and patients aged 70+ at baseline excluded (1.6%). (2) and (3) also excluded the BMI and HbA1C covariates. Results appear in **Appendix D**.

MA produced more accurate HRs in (1); because the diagnosis of memory loss is handled differently across health systems, removing such diagnoses from the outcome definition may have reduced heterogeneity. Otherwise, the relative

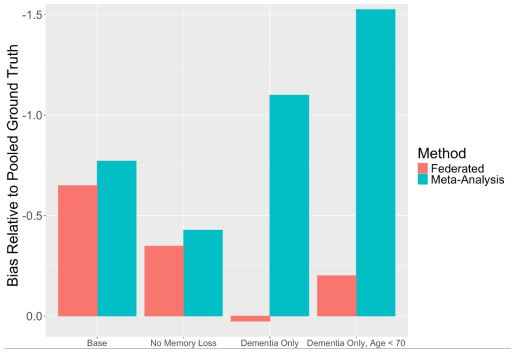

Figure 2: Bias of MA and FL HRs of initiating metformin vs. sulfonylureas on ADRD, relative to pooled analysis (100 · [Pooled HR-Aggregated HR]/[Pooled HR]), in primary and sensitivity analyses.

bias was inversely related to cumulative outcome incidence, as expected. However, the bias of FL HRs was smaller in sensitivity analyses, especially when considering fewer covariates. Even in (3), FL produced 86.8% less bias than MA.

4. Discussion

This study compared two approaches to a TTE conducted jointly in two health systems. FL consistently produced an accurate HR, while MA generally resulted in more bias when cumulative outcome incidence was lower, suggesting that FL may be especially useful in TTE for rare diseases.

While the MGB and UCHDW health systems serve distinct populations and exhibit varying clinical practices, health systems located in countries with national health systems are likely to differ even further, and may result in MAs with more bias than we could test in ENACT. In the future, we plan to introduce data from additional sites into ENACT, which will enable evaluations of FL in a more heterogeneous data environment, closer to that of our international team.

In follow-up studies, we will test FL in a range of scenarios to characterize how study parameters (e.g., outcome rarity, imbalance between treatment arms, missing data patterns) impact the reduction in bias, relative to MA. This large-scale assessment of FL's comparative performance across data settings will help determine when FL benefits outweigh the additional costs incurred, including in terms of logistics (e.g., setting up data use agreements, elaborating joint study protocols, updating IRB language) and computational time. Using a computer with an 8-core 2.5GHz CPU and 32GB of RAM, training both single-site models and generating the MA estimate took 12s, while the overall federation took 3,168s. Such a 264-fold increase in computational time may be a bottleneck for studies with very large datasets. Moreover, real-world applications require extensive communication among sites, which may incur additional delays.

This study aimed to emulate a real-world use of federation, but the enclave environment precluded an exact replication of a federated TTE using local data from multiple EHRs. The de-identification of UCHDW data prevented the inclusion of zip codes used to adjust for social vulnerability, yielding residual unmeasured confounding. The MGB data did not contain a comprehensive list of healthcare encounter dates, which may have altered some participants' censoring times. Beyond these known sources of divergence, data transfer often carries a risk of degradation. Further, the conversion of encodings used in the OMOP schema into their equivalents in the i2b2 schema was required to align UCHDW and MGB data but may have been imperfect. Nonetheless, these costs are necessary to evaluate FL in real-world healthcare data.

Among the additional sites in the FL platform, two comprise records from closed health care systems covering different populations. Preliminary data suggest greater heterogeneity between these two sites and MGB and UCHDW (**Appendix E**). Overall ADRD outcome incidence rates range from 8.3% to 17.5%. Further, one cohort has 60% more ADRD events among sulfonylurea initiators, while another cohort has 10% fewer. The distributions of covariates such as obesity and stroke also vary widely across cohorts. These differences indicate that FL will be critical for TTEs in which the full consortium contributes data. In future work, DRIAD-FL will replicate this TTE with FL and MA, as well as a TTE with a rare outcome (i.e., ALS) to evaluate the benefits of FL when addressing other limitations of MA.

References

RJ Bateman, C Xiong, TLS Benzinger, AM Fagan, A Goate, NC Fox, DS Marcus, NJ Cairns, X Xie, TM Blazey, DM Holtzman, V Buckles A Santacruz, A Oliver, K Moulder, PS Aisen, B Ghetti, WE Klunk, E McDade, RN Martins, CL Masters, R Mayeux, JM Ringman, MN Rossor, PR Schofield, RA Sperling, S Salloway, and JC Morris. Clinical and biomarker changes in dominantly inherited alzheimer's disease. *New England Journal of Medicine*, 367:795–804, August 2012.

M Borenstein, L Hedges, and H Rothstein. Meta-analysis: Fixed effect vs. random effects. 2007. URL https://meta-analysis.com/download/Meta-analysis_fixed_effect_vs_random_effects%20072607.pdf.

ML Charpignon, B Vakulenko-Lagun, B Zheng, C Magdamo, B Su, K Evans, S Rodriguez, A Sokolov, S Boswell, YH Sheu, M Somai, L Middleton, B T Hyman, R A Betensky, S N Finkelstein, R E Welsch, I Tzoulaki, D Blacker, S Das, and M W Albers. Causal inference in medical records and complementary systems pharmacology for metformin drug repurposing towards dementia. *Nature Communications*, 13 (7652), December 2022.

L Cook, J Espinoza, N Weiskopf, N Mathews, D Dorr, K Gonzales, A Wilcox, and C Madlock-Brown. Issues with variability in electronic health record data about race and ethnicity: Descriptive analysis of the national covid cohort collaborative data enclave. *Journal of Medical Internet Reserach*, 10, September 2022.

O Efthimiou. Practical guide to the meta-analysis of rare events. *Evidence Based Mental Health*, 21, April 2018.

L Han, Z Shen, and J Zubizarreta. Multiply robust federated estimation of targeted average treatment effects, 2023. URL https://arxiv.org/abs/2309.12600.

M Hernan and J Robins. Using big data to emulate a target trial when a randomized trial is not available. *American journal of epidemiology*, 183(8):758–764, APR 2016.

PB Imrey. Limitations of meta-analyses of studies with high heterogeneity. *JAMA Network Open*, 3, January 2020.

P Mehta, J Raymond, R Punjani, T Larson, F Bove, W Kay, and K Horton. Prevalence of amyotrophic lateral sclerosis (als), united states, 2016. amyotrophic lateral sclerosis and frontotemporal degeneration. *Taylor and Francis Online*, 23, May 2021.

J Qin, Y Liu, and P Li. A selective review of statistical methods using calibration information from similar studies. *Statistical Theory and Related Fields*, January 2022.

MJ Schuemie, Y Chen, and MA Suchard. Combining cox regressions across a heterogeneous distributed research network facing small and zero counts. *Statistical Methods in Medical Research*, 31:438–450, November 2021.

M Sunog, C Magdamo, ML Charpignon, and M Albers. Investigating primary care indications to improve the quality of electronic health record data in target trial emulation for dementia. *medRxiv*, April 2025.

S Yusuf, R Peto, J Lewis, R Collins, and P Sleight. Beta blockade during and after myocardial infarction: An overview of the randomized trials. *Progress in Cardiovascular Diseases*, 27(5):335–371, 1985.

Z Zeng, EH Kennedy, LM Bodnar, and AI Naimi. Efficient generalization and transportation, 2024. URL https://arxiv.org/abs/2302.00092.

D Zhang, J Tong, N Jing, Y Yang, C Luo, Y Lu, DA Christakis, D Guthe, M Hornig, KJ Kelleher, KE Morse, CM Rogerson, J Divers, RJ Carroll, CB Forrest, and Y Chen. Learning competing risks across multiple hospitals: one-shot distributed algorithms. *Journal of the American Medical Informatics Association*, 31: 1102–1112, April 2024.

Appendix A. Meta-analysis Methods for Addressing Heterogeneity and Rare Outcomes

To reduce the bias of meta-analysis for studies with rare outcomes and enable meta-analysis involving single-zero or double-zero studies (those were one or both treatment arms have zero event occurrences), a number of methods have been developed with different applications. Still, many TTEs are unable to utilize any of these methods because of their various limitations. Here we discuss three such methods:

Continuity correction can be used to include single-zero and double-zero studies by adding a small constant to the recorded count of each treatment-outcome pair. This correction is simple to implement and enables the computation of meta-analysis, but it often yields results that are highly biased (Efthimiou, 2018). Therefore, it is preferable to avoid continuity correction if possible.

Peto's method consists in performing meta-analysis with a modified odds ratio as the estimand, which can allow for the inclusion of single-zero studies (but not double-zero studies) and produces less biased results than a meta-analysis of the classical odds ratio (Yusuf et al., 1985). However, this method cannot benefit a study using any other estimand such as a hazard ratio, which is often the desired metric in TTE. Also, the reduction in bias is only effective in settings where the treatment effect is not large and the sizes of all treatment arms are similar (Efthimiou, 2018). For many TTEs, the size of the trial arms at any given health system are significantly unbalanced because national guidelines and/or hospital-wide practices will tend to favor one treatment for the relevant indication.

Bayesian meta-analysis is a more flexible method in which individual studies do not produce independent results, but rather are used to the prior estimands. This strategy can target various metrics, including hazard ratios, and it does not disregard single-zero or double-zero results. By design, its implementation requires access to prior estimates of the target metrics (Efthimiou, 2018); however, TTE for drug discovery or repurposing frequently investigates treatment effects for which there are no published data, and therefore these analysis are unsuitable for Bayesian meta-analysis.

The established method for performing meta-analysis when there is heterogeneity among studies is to substitute the FE model (which weights the results of each study inversely to the study's variance) with an RE model (which weights the results of each study inversely to the study's variance *and* to the variance between studies). RE meta-analysis typically performs better than FE meta-analysis in such settings and it can be sufficient in certain cases, but it isn't always preferable. In general, the results of RE meta-analysis have less precision than FE meta-analysis, so the meta-analysis of studies with low heterogeneity are better suited by a FE model. When heterogeneity is very high, the variance between studies can dominate the in-study variance, inducing disproportionately high weights for small studies that can make meta-analysis results difficult to interpret. Given these risks, it is critical to understand the extent of heterogeneity when deciding between FE and RE, but such an assessment can be challenging and may require statistical techniques in conjunction with a careful evaluation of potential differences in design across studies (Imrey, 2020).

Appendix B. TTE Specifications

Patients from Site 1 and Site 2 were selecting via queries requiring that they had:

1. A record of a metformin or sulfonylurea prescription at age 50 or older
2. No record of a metformin or sulfonylurea prescription before age 50

For all patients meeting these criteria, demographic, prescription, diagnosis, and procedure data were uploaded into ENACT. With these data, the trial was executed as specified by Charpignon et al. (2022) with the following modifications:

1. The follow-up time was extended to December 2024 to utilize the extent of available data.
2. The covariates for outcome adjustment did not include features reliant on sensitive data

or lab results (e.g., socioeconomic variables) because the data were not available for UCHDW patients in the enclave.

3. The eligibility criterion requiring a visit within the 18 months prior to baseline was removed because the dates for all visits were not available for MGB patients in the enclave. Instead, a criterion that patients have at least 18 months of history in the EHR prior to baseline was used as a substitute.

To identify ADRD outcomes, we used a set of ICD codes and medication RxNorm codes that construct a broad definition including not only dementia diagnoses, but memory loss and mild cognitive impairment diagnoses as well. The inclusion of these codes is informed by expert consultation, and attempts to account for 1) the common, severe delays in ADRD diagnosis, as well as 2) the widespread practice among primary care physicians and other clinicians not specialized in neurology to use memory loss broadly for patients with progressed symptoms.

The covariates were also selected based on expert consultation. In both the propensity and outcome models, the following covariates were used: Age at baseline, Sex at birth, Hypertension diagnosis prior to baseline, Stroke diagnosis prior to baseline, Chronic obstructive pulmonary disease diagnosis prior to baseline, Obesity diagnosis prior to baseline, Cardiovascular disease diagnosis prior to baseline, Cancer (broad definition) diagnosis prior to baseline, Cancer (selective definition) diagnosis prior to baseline, HbA1C record prior to baseline (categorical: missing, ¡7, 7-10, ¿10), BMI record prior to baseline (categorical: missing, ¡25, 25-¡30, $\geq$ 30)

ADRD indication codes:

Memory Loss:

F04, 780.93, R41.2, R41.3, I69.211, I69.311, I69.811, I69.911, 294.0, G30, G31.0, G31.8, F01.5, F02.8, R41.81, 797, 797.0

Mild cognitive impairment and general neurodegeneration:

G31.1, G31.9, 331.2, 331.8, 331.9, 438.0, G31.84, G31.81, G31.85, G31.89, 331.89, 331.83, 331, 290.9, 290.8, 294.9, 294.8, 294, F01.A, F01.B, F01.C, F02.A, F02.B, F02.C, F03.A, F03.B, F03.C

Dementia:

F03.9, 331.1, F02.80, F03.90, G31.83, F01.50, F03.91, F02.81, G31.09, G31.01, F03, F01.50, F01.51, 331.19, 331.11, 331.82, G30.8, G30.9, G30.0, G30.1, 290.0, 290.20, 290.40, 290.3, 290.43, 290.10, 290.41, 290.21, 290.12, 290.13, 290.11, 290, 290.42, 290.4, 290.2, 290.1, 294.10, 294.20, 294.21, 294.11, 294.1, 331.0, 331.00, F02, 294.2, F01

RxNorms:

135446, 1602583, 2597448, 1858971, 1858970, 135447, 1430990, 236559, 1100187, 1602588, 1602594, 1805422, 1805427, 2597453, 2597459, 997224, 997230, 1602584, 1602593, 1805421, 1805426, 2597449, 2597458, 997221, 997227, 998582, 1602585, 2597452, 367663, 2654337, 1170743, 1170744, 1602586, 1602587, 2597450, 2597451, 1100184, 1599803, 1599805, 1805420, 1805425, 2597446, 2597456, 997216, 997220, 997223, 997226, 997229, 2597442, 2597455, 997215, 997219, 997225, 998579, 1599802, 2597445, 371957, 483068, 483071, 1160637, 1160638, 1160639, 1295290, 1599800, 1599801, 2597443, 2597444, 583099, 4637, 860693, 602734, 602737, 860697, 860709, 860717, 583133, 602732, 860696, 860708, 860716, 583101, 602733, 2654831, 1178299, 1178300, 310436, 310437, 579148, 860695, 860707, 860715, 860901, 330343, 330344, 330345, 860694, 860706, 860714, 860900, 384641, 384642, 583097, 1163353, 1163354, 1163355, 405206, 996633, 996634, 996572, 996624, 6719, 236685, 996563, 996574, 996597, 996605, 996611, 996617, 996742, 996562, 996573, 996595, 996604, 996610, 996616, 996741, 406108, 607609, 996596, 2655351, 2658253, 2658795, 1178753, 1178754, 1178755, 996561, 996571, 996594, 996603, 996609, 996615, 996740, 996752, 996560, 996570, 996592, 996602, 996608, 996614, 996739, 996751, 372757, 577156, 996593, 1159791, 1159792, 1159793, 225807, 183379, 994808, 1308571, 226665, 226666, 226667, 226668, 725105, 751302, 1308570, 574012, 574013, 574014, 574015, 725104, 751301, 1805980, 366553, 1173232, 1173233, 1173234, 1296125, 1308569, 312834, 312835, 312836, 314214, 314215, 725021, 725023, 1308568, 331507, 331508, 331509, 331510, 331511, 725019, 725022, 1805978, 373797, 374628, 1157970, 1157971, 1157972, 1157973, 1295358, 997218, 997222, 997228, 998581, 998584, 998585, 998586, 1599804, 1599806, 1602589, 1602595, 429251, 602736, 860698, 860699, 860710, 860711, 860718,

860719, 860903, 413274, 996598, 996599, 996606, 996607, 996612, 996613, 996618, 996619, 996748, 996750, 996754, 1359577, 1359895, 1360107, 1360122, 1360267, 1360423, 285017

To identify covariates, we used the following ICD9 and ICD10 codes as indications:

Hypertension:

I27.24, I87.323, I87.322, I87.329, I87.393, I87.333, I87.313, I87.312, I87.319, I87.303, I87.309, I10, I15.2, I15.1, I15.8, I27.29, I27.22, I27.23, I15.0, I27.21, 459.31, 459.33, 459.30, 401, d401.9.3, 405.99, 572.3, 416.0, 401.9, 405.91, 401.1, 642.03, 642.01, 401.0, 405.01, 405.09

Stroke:

G46.4, G46.3, I67.89, I67.848, I67.81, I68.8, G46.8, G45.8, 997.02, 437.1, 437.9

COPD:

J44.1, J44.0, J84.17, J84.89

Obesity:

E66.01, E66.2, E66.8, E66.09, E66.1, 278.01, 278, 278.03

CVD:

I50.41, I50.31, I50.43, I50.33, I50.813, I50.23, I50.21, I25.760, I25.812, I25.758, I25.759, I25.750, I25.811, I25.111, I25.118, I25.119, I25.110, I25.10, I50.82, I50.42, I50.32, I50.812, I50.22, Z48.21, I50.84, T86.23, T86.21, Z94.1, I11.0, I11.9, T86.298, I24.8, I25.89, Q24.8, I45.5, I27.89, I09.89, I97.130, Z95.812, Z95.811, Z95.4, Z95.2, Z95.3, I27.22, I09.81, M05.30, I50.814, I50.40, T86.20, I50.30, I50.20, I50.811, I50.83, M05.312, M05.39, B57.0, I25.768, I25.769, B57.2, I39, T86.32, T86.31, T86.22, T86.39, I97.131, I25.751, A18.84, 415, 414.1, 414.10, 402.11, 996.83, 861.11, 861.01, 414.07, 414.06, V42.1, V43.3, V42.2, 429.82, 428.1, 402.01, 996.02, 411.89, 414.19, 416.8, 996.71, 426.6, 429.89, 398.99, 746.89, 414.8, V13.65, 398.91, 746.9, 402.91, 402.90, 416, 416.1, 391.8, 398.9, 086.0, 746.86, V43.21, V43.22, V43.2, 402.00, 164.1, 392.0, 392.9, 402.9, 746.7

Cancer:

V10.83, Z85.048, V10.06, 238.2, V10.46, 239.0, 162.9, C34.90, D49.0, 238.0, 189.0, C64.9, D43.2, D37.5, V10.3, 153.9, D49.7, C50.919, C77.0, Z85.79, V10.79, Z85.238, C50.411, C50.811, C02.1, 191.9, C08.9, C79.51, Z85.528, D37.9, D44.0, 174.9, C79.31, C54.1, C56.1, C77.4, 162.5, 162.3, C34.11, C34.31, C67.9, C67.8, 153.6, 197.5, 153.0, 157.9, 197.7, C25.9, C78.7, C62.90, C62.10, 186.9, 197.0, C78.00, C44.90, V10.47, C50.911, 180.0, 180.9, C50.912, 182.0, C54.3, C54.9, C54.2, Z08, C77.3, C69.20, D48.9, 190.5, C96, C64.2, C77.8, Z86.008, Z85.118, 338.3, C77.9, Z86.03, C50.211, C34.91, C34.30, C53.8, C50.119, 188.9, 198.5, C79.52, 154.1, C71.8, 191.0, V10.87, 198.81, C34.82, D70.1, 155.0, C80.1, 198.4, C56.9, 238.8, 199.1, C44, C34.32, C44.99, C44.390, 173.39, C44.399, C44.391, C21.8, C79.9, C78.89, C78.02, C78.01, C79.00, 164.0, C50.412, 198.3, C50.011, C96.9, Z85.038, 154.0, 174.4, C34, C34.92, C34.00, C50, 159.9, C26.9, 156.0, C15.8, V10.81, C34.10, Z86.000, C04.9, 144.9, C49.4, C49.9, 171.5, 171.9, 196.3, D49.6, C69.02, 196.2, V10.51, C67.2, V10.11, 191.1, Z85.9, D49.4, C67.1, 197.6, V10.07, 155.2, 173.90, V10.43, 183.0, C09.9, C02.8, C22.8, C75.9, C69.31, 155.1, C24.1, Z85.09, V10.05, D49.2, 239.2, D44.3, 239.6, 188.8, Z85.068, Z85.07, V10.09, D49.9, D47.9, 239.9, 238.79, 173.30, C44.301, C44.309, C44.300, 196.9, C64.1, C67.0, C50.512, Z85.89, Z85.831, V10.89, C34.12, 145.9, Z85.028, D37.6, D44.2, 237.4, D44.9, V10.52, D49.89, D48.3, C18.1, 239.7, D37.4, 198.82, 151.9, 156.2, 162.8, C34.80, Z85.818, 161.9, C44.602, 170.7, C40.20, C41.9, D44.4, C50.111, 189.2, C66.9, C66.2, C50.112, 174.8, 195.5, C76.51, 239.89, 173.60, C44.601, C68.9, C04.1, C74.02, C74.90, C44.599, 174.1, 184.4, D48.4, 235.4, 238.9, 160.2, 160.9, D44.10, 198.89, C67.4, 188.4, C49.0, D49.5, 239.5, D49.1, 239.1, 142.0, 238.1, C49.21, Z86.001, C79.70, 157.0, 198.7, 196.0, D37.8, C50.812, Z85.819, 190.0, C69.40, 152.9, 152.3, 195.2, C79.71, C50.212, C74.00, 194.0, 195.0, 158.9, C50.929, V10.04, C39.9, 157.1, 153.4, C24.0, 156.1, C44.40, C69.30, C47.9, D43.0, 146.0, 189.8, C65.2, 239.4, 170.4, C40.00, C72.9, 171.6, C49.6, 171.7, C49.5, V10.01, D47.Z9, R97.21, 153.5, 161.0, C80.0, 199.0, 171.3, C49.20, V10.44, C34.81, C34.01, 145.3, 148.9, D41.00, C16.8, 146.8, 173.8, 152.0, 188.2, C50.219, D44.7, 197.8, 150.9, 151.0, 147.9, V10.03, C60.9, C79.72, 151.5, C56.2, 197.1, 196.1, 154.3, 196.8, V10.85, 162.0, 149.0, C79.01, C50.012, C50.819, C79.32, V10.42, D43.4, C25.4, C24.9, C65.9, 189.1, C65.1, 153.3, V10.02, 236.5, 150.4, C49.22, 170.9, 192.0, C72.50, C50.311, 174.3, 237.3, D44.6, C41.4, 170.6, D42.0, 202.90, C96.Z, 141.9, C40.22, C40.21, C78.4, C79.11, 173.9, C69.01, 190.3, 141.4, C14.8, 173.3, C49.3, 142.9, C25.7, 161.1, D38.5, 235.9, D38.6, D48.60,

238.3, D49.81, 239.81, 153.7, 191.2, C24.8, 237.0, C50.312, 174.0, 151.4, Z85.59, 163.9, 163.8, D37.01, 237.5, D43.8, D37.1, D37.2, 235.2, 157.2, 194.4, 173.40, 146.9, Z85.00, C66.1, C69.42, C44.9, C18.8, D37.09, V10.21, C50.419, 184.0, C10.8, 192.2, C50.511, C32.8, 141.0, C41.0, C44.299, D42.9, 237.6, D42.1, 235.3, 162.2, C34.02, 197.4, C09.1, C00.2, 140.9, C30.0, 160.0, 173.0, C44.0, 238.6, C57.00, C44.509, C25.8, 157.8, C44.609, D39.11, 173.5, 141.2, D43.9, D43.3, 237.9, C44.49, D37.05, D37.030, 198.2, 174.5, C50.519, 197.2, V10.22, 173.2, C44.101, V10.41, C79.62, C79.60, C79.61, 203.80, V10.90, C44.709, 198.1, 170.2, C41.2, 235.1, V10.84, 190.6, 170.1, V58.42, C76.40, Z85.858, C44.702, C44.209, C40.10, 170.5, C44.202, D41.01, D39.8, 145.5, 145.2, C62.92, 196.6, C69.32, 189.9, 235.5, C50.611, 162.4, D38.1, C50.319, 191.7, 196.5, 235.6, 147.8, C11.8, C57.7, C72.42, V10.88, C50.922, D40.9, 236.6, C63.9, 191.8, C50.019, 154.8, C21.2, 183.2, 141.6, D49.3, C44.590, 173.59, C44.591, 161.2, C41.3, 170.3, 174.2, D41.02, C50.122, 154.2, 192.1, C70.9, 158.0, Z85.53, 173.79, C44.791, C69.90, 190.9, C49.11, 171.2, 191.6, 173.00, D39.10, D48.62, 142.8, D37.3, D48.2, 143.9, 187.9, C54.8, 182.8, C00.1, V10.29, 148.1, C44.699, 156.9, 188.3, C67.3, 198.6, 188.1, C00.9, C50.021, 180.8, C10.1, C51.8, C51.1, 152.8, C44.500, 173.50, C44.501, C50.921, 180.1, C79.40, D48.61, C49.10, 171.4, 171.0, 235.7, C17.8, D38.2, 153.2, 236.7, 173.10, C64, C50.612, 174.6, C44.201, 173.20, Z85.848, 151.8, C69.00, 236.91, D41.20, 173.99, 159.0, 170.0, 148.8, C13.8, C44.80, D47.09, C62.11, D37.03, 236.3, D39.9, C76.50, D43.1, C57.9, 194.3, 173.4, 188.0, C54, 145.0, C74.91, 150.5, D40.8, 236.2, C44.89, 173.89, 153.8, C54.0, 182.1, 173.7, 152.1, 175.9, C79.10, D49.519, 157.4, 142.1, 189.3, C69.92, 184.2, 192.9, Z85.54, C04.8, C69.91, C44.4, C44.20, 187.4, D49.59, 191.3, C79.02, C50.121, C40.01, C44.292, Z85.50, V10.50, C62.12, D44.11, C48.8, C69.62, 184.8, C72.41, 148.3, C13.2, C57.01, C69.60, C69.50, C69.52, 190.7, 239.3, C62.02, 175.0, C50.029, C50.022, C49.12, 153.1, 149.9, 191.5, C71.5, D41.21, D38.3, 235.8, D38.4, 195.1, C50.222, 187.7, 197, C78, C76.8, 195.8, 236.0, 157.3, 238.7, C78.80, D37.031, 235.0, D37.032, D37.039, 141.1, 236.4, D40.10, 152.2, V10.59, 182, 150.8, 237.2, D37.02, D37.04, C31.8, 144.0, C04.0, C50.829, 151.2, Z86.007, 173.70, C44.701, 173.19, C68.8, D00-D09, C49.8, C74.01, C62.91, D40.11, C44.102, 195.3, 184.9, 146.6, C69.61, C69.41, 191.4, C10.4, 198.0, H47.42, 183.4, C40.11, C09.8, C74.92, V10.53, 173.49, C57.8, C50.522, C69.82, C69.10, 190.4, D39.12, 150.1, C76.52, D44.12, 159.8, C50.619, C63.7, V10.00, C00.0, 173.69, C44.691, 238.5, C50.822, 160.1, C30.1, 377.52, C06.89, 160.8, 146.1, 173.6, 150.3, 190.1, 377.51, 144.8, C57.02, 192.3, 197.3, C74.10, 150.2, C44.109, D41.11, 183.8, C44.692, 188.5, 184.1, C51.0, C44.799, C47.0, 164.2, C44.792, 165.9, 141.8, 140.5, 173.1, C44.10, 173.09, D41.12, 158.8, C76.42, C13.0, 148.0, 145.6, D40.12, 190.2, C38.8, C47.21, C72.30, C00.6, C40.31, C69.21, 187.2, C05.8, C76.41, 154, D49.511, 147.1, 194.6, C50.821, 164.9, C72.59, 173.80, V10.40, 194.1, 141.3, 189.4, C72.21, C40.02, 183.9, C31.3, C72.32, C50.221, 140.4, 143.1, 194.9, C44.191, 160.5, 140.1, 160.3, C72.31, 140.8, C00.8, 147.0, 156.8, C50.421, 173.29, C44.291, 187.1, 164.1, 159.1, C72.40, 181, D41.3, C96.20, C32.3, C62.01, C40.12, C47.8, 171.8, C60.8, C78.30, V10.49, C80.2, 236.90, D41.9, C62.00, 143.0, C50.622, 161.8, C44.192, 151.3, C69.51, 145.1, 161.3, C40.32, C63.12, C74.11, C44.59, C40.90, C63.11, 151.1, C47.3, C69.22, C40.81, 165.0, 186.0, C69.12, D41.8, 202.96, C69.81, 203.81, 145.8, C06.80, V10.20, 149.8, D41.22, C69.11, C50.422, 150.0

Appendix C. Full Summary Table for the Primary Cohort

Table 2: Summary table for the MGB cohort, stratified by treatment arm. HTN: hypertension, CVD: cardiovascular disease. HTN, Stroke, COPD, Obesity, CVD, and Cancer identified via diagnosis records.

Feature	Met	Sulf
Total	43655	5240
% of Site	89.3%	10.7%
ADRD	6.9%	9.0%
Death	6.8%	15.9%
Age	65.7	70.4
Sex F	49.8%	45.9%
HTN	53.9%	45.5%
Stroke	0.7%	0.5%
COPD	2.3%	2.4%
Obesity	16.3%	6.1%
CVD	16.8%	16.9%
Cancer	17.7%	17.8%
HbA1C 7	35.3%	14.1%
7-10	14.9%	20.8%
$\geq$ 10	2.5%	2.8%
Missing	47.3%	62.3%
BMI 25	0.7%	0.9%
25-29	2.0%	1.1%
$\geq$ 30	16.5%	5.9%
Missing	80.8%	92.1%
Years in EHR	5.8	4.7

Table 3: Summary table for the UCHDW cohort, stratified by treatment arm. HTN: hypertension, CVD: cardiovascular disease. HTN, Stroke, COPD, Obesity, CVD, and Cancer identified via diagnosis records.

Feature	Met	Sulf
Total	4491	372
% of Site 10.7%	92.4%	7.6%
ADRD	5.6%	7.0%
Death 5.8%	10.2%	
Age	66.1	68.6
Sex F	52.2%	49.2%
HTN	39.6%	37.9%
Stroke	0.4%	0.3%
COPD	1.2%	1.1%
Obesity	7.4%	5.4%
CVD	11.4%	12.9%
Cancer	17.5%	13.2%
HbA1C 7	30.4%	11.0%
7-10	13.0%	13.7%
$\geq$ 10	2.7%	2.4%
Missing	53.9%	72.9%
BMI 25	0.5%	0.3%
25-29	2%	0.3%
$\geq$ 30	13.0%	5.1%
Missing	84.5%	94.4%
Years in EHR	4.4	3.6

Appendix D. Hazard Ratios Comparison for Sensitivity Analyses

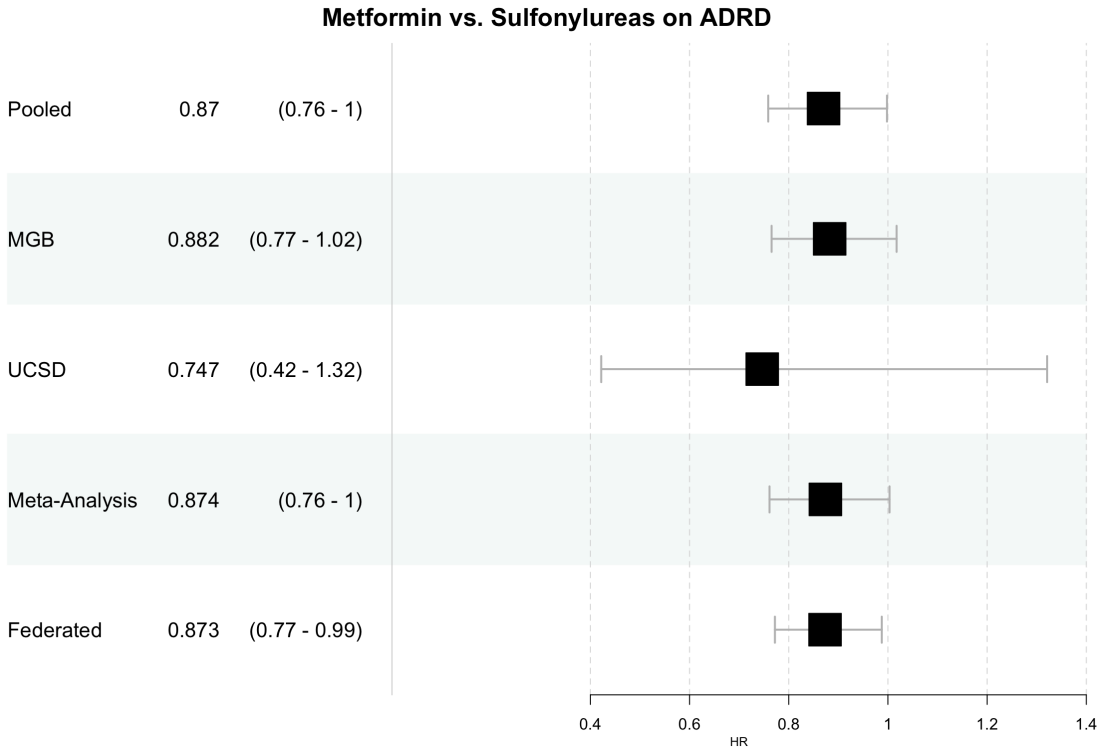

Figure 3: Sensitivity analysis 1: HR of initiating metformin vs. sulfonylureas on ADRD using pooled data, MA, and FL (excluding memory loss diagnoses) using pooled data, MA, and FL.

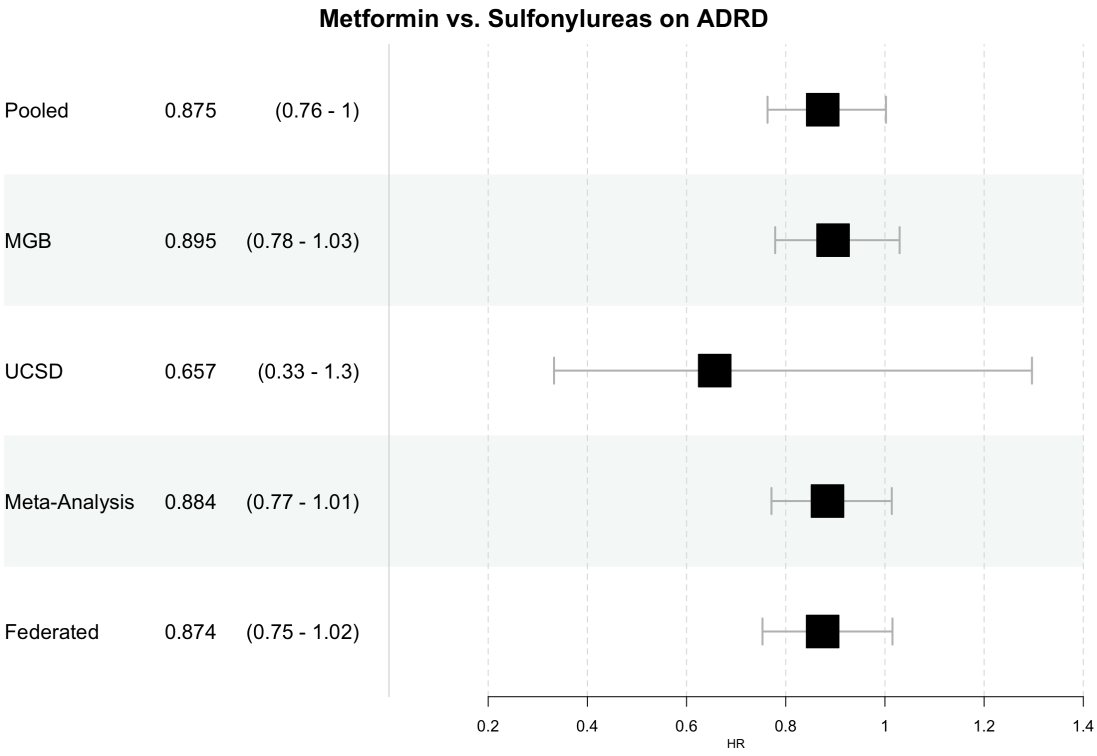

Figure 4: Sensitivity analysis 2: HR of initiating metformin vs. sulfonylureas on ADRD using pooled data, MA, and FL (excluding memory loss, mild cognitive impairment, and general neurodegeneration diagnoses) using pooled data, MA, and FL.

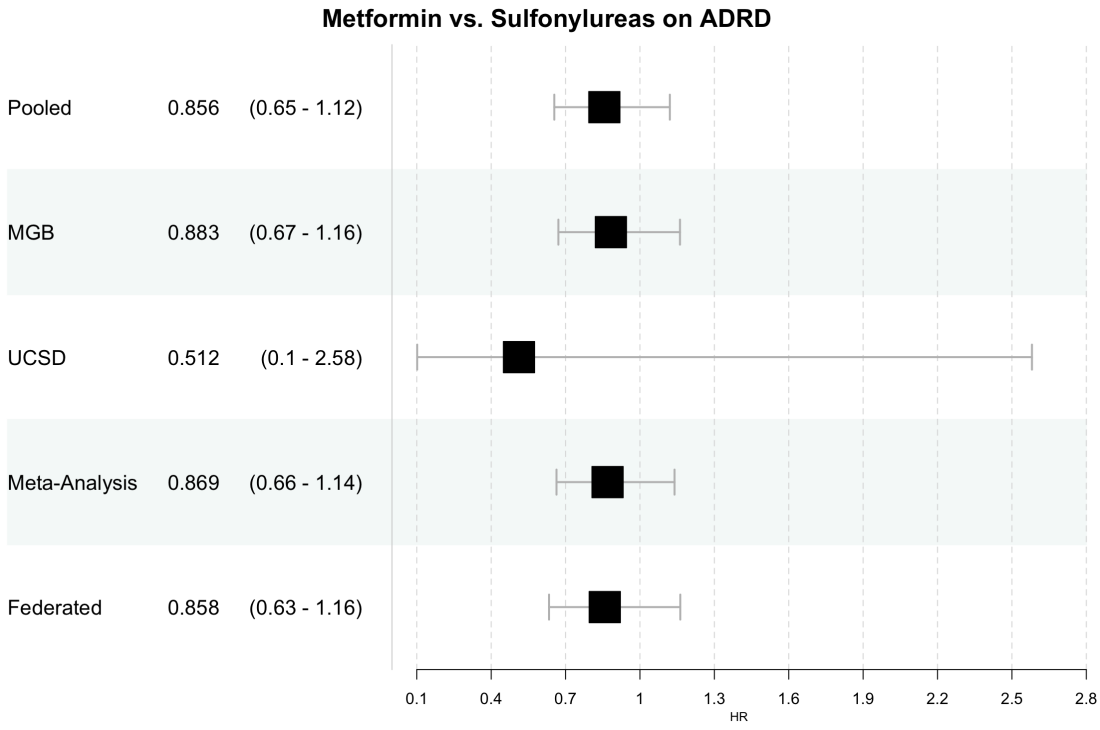

Figure 5: Sensitivity analysis 3: comparison of the HR on ADRD (excluding memory loss, mild cognitive impairment, and general neurodegeneration diagnoses) among patients initiating metformin vs. sulfonylureas at ages 50-70 using pooled data, MA, and FL.

Appendix E. Summary Statistics for Cohorts from Additional Sites in DRIAD-FL

Site 3 (INSIGHT CRN) is an EHR database containing healthcare data from multiple neighboring health systems within New York City for over 20,000,000 patients. In preparation for a real-world federated trial across all five sites, we have built the framework for and executed a single-site TTE at INSIGHT CRN. The estimated HR of initiating metformin vs. sulfonylureas on ADRD was .955, with 95% CI .914 - .997.

Table 4: Summary table for INSIGHT CRN, stratified by treatment arm. FuT is follow-up time, HTN is hypertension, and CVD is cardiovascular disease.

Feature	Met	Sulf
Total	124,830	31,263
Trt	78.0%	20.0%
ADRD	7.9%	10.0%
Death	2.3%	3.8%
Mean Age	64.7	66.1
Sex Female	58.2%	54.0%
HTN Dx	58.1%	69.5%
Stroke Dx	0.8%	0.9%
COPD Dx	1.5%	1.5%
Obesity Dx	19.5%	19.9%
CVD Dx	18.7%	22.4%
Cancer Dx	14.6%	14.6%

Site 4 (CPRD) is an EHR primary care database in the United Kingdom. The database includes EHR data from across the country and contains records for over 10,000,000 patients.

Table 5: Summary table for site 4, stratified by treatment arm. HTN is hypertension and CVD is cardiovascular disease.

Feature	Met	Sulf
Total	141,136	28,786
Trt	83.1%	16.9%
ADRD	17.8%	16.1%
Death	31.2%	54.3%
Mean Age	65.3	69.6
Sex Female	42.4%	43.0%
HTN Dx	51.6%	44.2%
Stroke Dx	25%	25.2%
COPD Dx	5.4%	5.7%
Cancer Dx	10.5%	15.7%

Site 5 (Clalit) is an EHR database in Israel, containing three decades of data for over 5,000,000 patients in a closed healthcare system. In preparation for a real-world federated trial across all five sites, we have built the framework for and executed a single-site TTE at Clalit. The estimated HR of initiating metformin vs. sulfonylureas on ADRD was .880, with 95% CI .860 - .900.

Table 6: Summary table for site 5, stratified by treatment arm. HTN is hypertension.

Feature	Met	Sulf
Total	468,227	48,376
Trt	88.4%	11.6%
ADRD	14.6%	23.6%
Death	33.1%	66%
Age 50-59	32.6%	28.7%
Age 60-69	35.1%	30.7%
Age 70-79	23.3%	26.8%
Age 80-89	8.0%	11.9%
Age 90+	1.0%	2.0%
Sex Female	51.1%	50.9%
HTN Dx	57.3%	46.2%
BMI $\geq$ 30	42.7%	31.8%