Constructing Open Data for the Alzheimer's Disease and Related Dementias Innovation System

Lucy Xiaolu Wang*

September 2026

1 Introduction

Alzheimer's disease and related dementias (ADRD) affect one in ten Americans aged 65 or older and are the most expensive disease in the United States (Chandra, Coile, & Mommaerts, 2023). Chandra et al. (2023) identify innovation in prevention and treatment as a critical area for economic analysis given this enormous social cost. After decades of failed trials, the field has entered its most productive period in a generation. The FDA signaled openness to biomarker-based approval in its 2018 early-Alzheimer's guidance and granted accelerated approval to aducanumab on the amyloid surrogate in June 2021, but the drug was later withdrawn in 2024. Lecanemab and donanemab followed on clinical endpoints in 2023 and 2024, and the first blood-based diagnostic (a plasma p-tau217 ratio test) was cleared in May 2025, promising earlier and cheaper detection. Taken together, these developments make ADRD an important setting for studying how regulations shape the rate and direction of medical innovation.

The same period produced a research-integrity shock that, in our view, deserves systematic investigation rather than passing commentary. Investigative reporting uncovered manipulated images in some of the field's most influential amyloid studies. The revelations culminated in the 2022 *Science* investigation, the 2024 retraction of the field's most-cited amyloid-oligomer paper, and a book-length account of the episode (Piller, 2025). That such a shock could strike the research base of the world's most expensive disease is startling. How the innovation system absorbed it, in whether and how quickly funding, trials, and firm investment reallocated, remains an open empirical question, and answering it requires data linking the research base to the development pipeline.

*xiaoluwang@umass.edu: University of Massachusetts Amherst and Max Planck Institute for Innovation and Competition. I was drawn to this topic after my undergraduate advisor and a distant relative were diagnosed with Alzheimer's disease; as populations age, and with women facing a disproportionately higher risk of the disease, these questions matter to a growing share of families. All errors are mine.

Empirical work on these questions often runs on commercial pipeline databases such as Cortellis, Pharmaprojects, and BioMedTracker. These databases are expensive, license-restricted, and non-redistributable. A mechanism classification cannot be traced to its source, and a published estimate cannot be reproduced without the same subscription. As a result, a policy debate of first-order importance, the ongoing reform of the FDA's accelerated approval pathway, runs on evidence that the research community cannot independently verify.

We propose to present OpenADRD, a fully open data collection that covers the ADRD innovation system from trial registrations to public funding, regulatory decisions, and the underlying research base. We also present the construction and audit protocol behind it.

2 The Data

We construct a novel set of linked panels, built entirely from public sources and redistributable, with each entry linked to its source. The counts below have been independently recomputed from the raw inputs (Section 3).

2.1 Clinical Trials Data

ADRD trial-level panel: 10,757 trial registrations from ClinicalTrials.gov with start dates 2005 through 2025 (registry snapshot August 2026). The panel contains 4,754 registrations indexed to AD or mild cognitive impairment (MCI), after removing dementias caused by other conditions; 1,709 interventional drug trials form the core analysis sample. It adds 6,003 control trials in long-lag neurodegenerative diseases without a comparable surrogate shock (Parkinson's 4,805; ALS 940; Huntington's 258, with overlaps with the AD sample removed). Each trial is coded for target mechanism, primary-endpoint type (biomarker or clinical), designed duration, planned and actual enrollment, phase, and sponsor class. The panel deliberately includes Phase-1 and prevention trials, because the longest-lag segment is where patent-design theory (Budish, Roin, & Williams, 2015) predicts underinvestment and a patients-only scope would discard the observations where the economics matters most.

Drug-to-mechanism dictionary: The binding constraint in trial-registry research is mechanism classification: registry keyword coding leaves 67 percent of AD drug trials unclassified. We hand-compiled a 614-entry drug-name dictionary from the AlzForum therapeutics database and the Cummings et al. annual pipeline reports, covering roughly 580 distinct agents plus tracer and control entries, with alternative names for the same drug matched to one entry. Every entry carries a source URL, and a corrections overlay documents 53 verified fixes with per-row page evidence. The dictionary rebuilds byte-identically from its source files by script, so no hand edit can enter the data silently. Applying it reduces the unclassified share from 67 to 19 percent. The dictionary is not tied to any single registry and is directly reusable in other ADRD projects.

Published pipeline benchmark: Cummings and coauthors publish the field-standard annual census of the Alzheimer's drug pipeline. We digitized their eleven reports (2016–2026) into a consistent machine-readable panel and use it as an external validation target. Our dictionary-coded shares reproduce its qualitative pattern (no post-2021 amyloid surge; a 2016–18 Phase-3 amyloid peak), with level differences explained by trials-versus-agents counting.

Chinese registry panel: A complete extract of the Chinese Clinical Trial Registry (ChiCTR), 127,216 records, with an ADRD subset built under both strict and broad definitions. To the best of our knowledge, no comparable full-registry ChiCTR research extract is publicly available. It supports cross-country designs around the divergent regulatory paths. China's NMPA approved lecanemab in January 2024 on clinical evidence, having never accepted the surrogate basis.

2.2 Funding and Regulatory Data

NIH funding panel: NIH RePORTER ADRD grants (FY2008–2026) mapped to the same mechanism taxonomy as the trials panel, giving a public-side counterpart to the private pipeline, where patent-clock incentives should not operate. The current mechanism coding uses grant titles only, which is deliberately conservative and captures a documented 10 to 14 percent of dollars; a version that also reads the full grant abstracts is in progress.

International regulatory panel: All 37 identifiable national regulatory and reimbursement decisions on the anti-amyloid antibodies across eight jurisdictions, 2021–2026; 36 are verified against primary sources and one relies on a single source. The panel documents that the FDA alone accepted the surrogate basis, while clinical-endpoint approvals of the same molecules later spread internationally. It lets us compare the same drugs across countries that made different choices at different times. The UK case, where the license was granted on the same day reimbursement was refused, separates regulatory standards from market size.

2.3 Research Base Data

This panel is in progress. For each sponsor firm, we measure how closely its scientific knowledge base builds on the specific amyloid research that was later retracted or credibly questioned, using citation links fixed before any public revelation. The list of implicated researchers and papers draws on the public record, including a dated table of events and key people that we built from the book-length account of the episode (Piller, 2025). Combined with the staggered 2021–2024 credibility shocks, these measures let us study how firms responded to news about the reliability of the science they build on. Closeness measures exposure to the affected research, not culpability.

3 Linkages and Reproducibility

The panels are designed to be used together. The dictionary maps drug names to mechanisms, so trials, NIH grants, and the published pipeline censuses can all be grouped the same way. Trials

map to sponsor firms, regulatory decisions map to the drugs they cover, and the research base measures map to the same sponsor firms through citation links built on OpenAlex. A single mechanism and firm vocabulary runs through the whole collection.

Everything is built to be reproduced. Each file is constructed by scripts from raw public inputs, every entry keeps a link to its source, and automated checks confirm that known cases are coded correctly and that totals reconcile before any file changes. We used large language models to facilitate parts of the data construction and auditing, and independent re-runs reproduce all headline counts. The full build code will be released with the data.

4 Value to the Research Community

OpenADRD offers innovation economists a reproducible, citable, zero-cost alternative to commercial pipeline data in one of the most policy-relevant disease areas. Its design supports the standard quasi-experimental toolkit: event studies around the 2018 and 2021 surrogate shocks with within-registry disease controls, endpoint-composition first stages, public-versus-private contrasts, international triple-differences, and shift-share exposure designs around the integrity shocks.

As evidence of fitness for purpose rather than as research findings, the descriptive facts the data establish are already informative. AD drug-trial starts show no jump after the 2018 guidance or the 2021 approval. The amyloid share of disease-modifying starts declines across the policy era, from about 60 percent in the late 2000s to roughly a third in recent years. Amyloid's dominance instead shows up in enrollment, where a few very large trials carry most of the volume, and this concentration pre-dates the policy. The share of Phase 2–3 trials with biomarker primary endpoints rises from 21 to 30 percent after the guidance, while NIH funding composition shows no corresponding break. These facts could not be checked or contested by third parties without the commercial subscriptions they would otherwise require.

5 Analyses in Progress

Several analyses are in progress, and all are scheduled to be presentable by the December meeting. First, we estimate Poisson difference-in-differences models of AD drug-trial starts against the disease controls around the 2018 guidance and the 2021 approval, together with event studies on the amyloid share of the pipeline in counts and in enrollment weights. Second, we test whether designed trial durations changed where the surrogate applies, using ClinicalTrials.gov version histories rather than realized completions, which are confounded by completion selection. This test separates an appropriability channel from a science-momentum alternative that predicts no duration change. Third, we estimate persistence and reallocation after the 2019–20 amyloid failure wave, in both program counts and enrollment weights.

Beyond these aggregate designs, we develop firm-level shift-share designs on the integrity

layer, interacting pre-determined fraud-closeness with the staggered 2021–2024 credibility revelations and pre-2018 amyloid portfolio shares with the 2021 approval. Because the same firms receive both shocks with opposite signs, their relative responses separate belief and appropriability channels that aggregate timing cannot.

6 Dissemination Plan

All inputs are public-source and redistributable, and every component will carry an open license. We aim to release in 2027, ideally when the associated working paper is ready for submission to a peer-reviewed journal. The first release covers the trial and control panels, the mechanism dictionary with its build inputs, the benchmark, funding, and regulatory panels, the full build code and checks, and a codebook, hosted on GitHub with permanently archived releases. Later releases will add the ChiCTR panel and the research-integrity layer. We commit to registering OpenADRD in the i3 data catalog, and we would welcome i3 feedback.

References

Budish, E., Roin, B. N., & Williams, H. (2015). Do firms underinvest in long-term research? Evidence from cancer clinical trials. *American Economic Review*, 105(7), 2044–2085.

Chandra, A., Coile, C., & Mommaerts, C. (2023). What can economics say about Alzheimer's disease? *Journal of Economic Literature*, 61(2), 428–470.

Cummings, J., et al. (2016–2026). Alzheimer's disease drug development pipeline (annual series). *Alzheimer's & Dementia: TRCI*.

Piller, C. (2025). *Doctored: Fraud, arrogance, and tragedy in the quest to cure Alzheimer's*. One Signal Publishers.