UROP MethodUROP Methodology Writeup

I.Goal of the Project

The primary objective of this initiative is to assemble a comprehensive, longitudinal panel dataset of university researchers and professors that captures the duration and trajectory of their appointments across multiple institutions. By leveraging publication metadata from the OpenAlex API, we infer each scholar's "start" and "end" years at every university where they publish. This publication-based proxy allows us to reconstruct academic career paths from 2000 through 2025, producing a rich time-stamped record of institutional affiliations for tens of thousands of faculty members.

Building on this affiliation backbone, our second goal is to contextualize each appointment spatially by mapping the nearest hospital to every institution. By integrating geocoded university locations with a nationwide hospital directory, we attach healthcare-access metrics—such as distance to the closest facility—to each career spell. This spatial enrichment provides a foundation for investigating how variations in local health infrastructure may relate to researcher well-being and survival.

Finally, we aim to augment the dataset with key demographic and outcome variables—most critically, dates of birth and death—by linking our author list to public mortality sources (e.g., SSDI, university obituaries). The resulting dataset will enable survival analyses and place-based health studies within an academic population, offering novel insights into how institutional and geographic contexts jointly shape longevity and health trajectories in higher education.

II. Related Work: Expert Patients' Use of Avoidable Health Care

Kakani, Matecna, and Chandra (2025) examine whether clinicians, by virtue of their training, differ from non-experts in how they use emergency departments—especially for visits that could be treated elsewhere . Drawing on a novel linkage of Medicare Fee-for-Service claims (2006–2017) to occupational directories (physicians via UPIN, nurses via state boards, and lawyers via Martindale-Hubbell) through Infutor's SSN registry, they assemble cohorts of physician-patients, nurse-patients, and matched comparison groups (lawyers and non-experts) . Avoidable ED visits are identified using the Billings algorithm—classifying visits as non-emergent, primary-care treatable, or preventable—and validated against hospitalization risk measures.

Their core finding is that physicians and nurses have substantially fewer ED visits than similar non-experts, 19.8% fewer for physicians and 5.1% fewer for nurses, with the bulk of this gap driven by reductions in avoidable visits. Moreover, the largest declines occur for conditions typically requiring a prescription, suggesting that self-prescribing (or rapid informal access to medicines) rather than purely medical knowledge underlies much of the difference. Spouses of clinicians exhibit intermediate effects, consistent with within-household expertise and prescribing privileges.

These insights highlight how access to prescription authority can sharply reduce low-value, avoidable care—often more so than educational interventions alone—and motivate our plan to integrate geospatial measures of hospital proximity with directory-based demographic linkages. By combining detailed panel records of academic appointments with nearest-hospital distances and eventual mortality data, we can similarly test whether local health infrastructure and prescribing environments at different universities shape avoidable care patterns and long-run survival among researchers.

III. Parallel Occupational Data for Lawyers

In assembling rich occupational panels, researchers have long leveraged specialized professional directories. For example, Martindale-Hubbell serves as a leading source for tracking U.S. lawyers: it includes over 582,000 lawyer records, each containing name, year of birth, gender (often imputed), and mailing ZIP code. These lawyer entries are analogous to our planned use of OpenAlex for academics, in that both datasets provide a near-comprehensive roster of professionals along with key demographic and geographic attributes.

To validate the completeness of Martindale-Hubbell, the paper's Appendix A compares birth-cohort counts in Martindale-Hubbell against American Bar Association data on law-school graduates. For cohorts born 1938–1955, the Martindale-Hubbell counts track ABA estimates almost exactly, suggesting the directory captures nearly the entire practicing population over those years . This high-fidelity coverage underpins confidence in using Martindale-Hubbell for longitudinal analyses of lawyer careers, much as we rely on OpenAlex's broad indexing for scholarly authors.

A second validation exercise matches Martindale-Hubbell entries to the public Ohio State Bar Association archives, where 33,384 Ohio-licensed lawyers born 1910–1955 were sought in the directory. The results showed 87% of Ohio Bar lawyers found in Martindale-Hubbell, with 73% correctly geocoded as residing in Ohio. This dual-validation—both national and state-level—demonstrates rigorous vetting practices that we will emulate when assessing OpenAlex affiliation accuracy against known faculty rosters.

Beyond raw directory data, the paper enriches lawyer records via Infutor's SSN linkage. Infutor, constructed from voter rolls, credit histories, and the SSA Death Master File, provides unique identifiers that allow merges to Medicare for gender and birth-year imputation. For instance, gender was imputed for nearly 2.9 million Infutor records and birth year for 338,482 cases where the SSN matched a Medicare beneficiary. A three-step matching algorithm (broad then narrow variable sets, followed by unique-match enforcement) further ensures high-precision linking between occupational directories and Infutor's registry. These enhancements parallel our planned linkage of academic profiles to SSDI and university obituaries for robust date-of-death capture.

IV. Research Applications: Location and Longevity

By linking each researcher's institutional appointments to the nearest hospital and measuring the great-circle distance to that facility, we create a framework for testing whether healthcare access moderates career-related stressors and occupational hazards within academia. For example, we can examine whether faculty who spend significant portions of their careers at universities in health-deserts, regions more than 30 km from a high-capacity hospital, experience higher mortality rates or shorter post-retirement lifespans compared to peers at better-served locations. This spatial dimension enables causal inference designs such as difference-in-differences around campus relocations or instrumental-variable analyses exploiting exogenous changes in local hospital capacity.

Furthermore, the panel structure of the dataset allows us to study mobility-health interactions: as researchers transition between institutions in regions with differing healthcare infrastructures, we can test whether moves to lower-access areas correspond with changes in health outcomes or survival probabilities. By integrating individual covariates—publication productivity, discipline, career stage, and regional covariates, hospital quality ratings, population health indices, we gain a rich multilevel dataset suitable for Cox proportional-hazards models and other survival-analysis techniques. Ultimately, this will shed light on how place matters for academic well-being and inform university policy on campus health services, faculty location incentives, and retirement planning.

V. Methodology Plan (Initial MIT-Only Test)

To validate our approach before scaling to multiple institutions, we began with a focused MIT-only pilot. In the first step, we queried the OpenAlex Works endpoint for all publications linked to MIT's institution ID (I63966007) between 2000 and 2025. For each returned "work," we iterated through its authorships to identify instances where the author's affiliated institution matched MIT. By tracking each author's earliest and latest MIT-affiliated publication year, we generated a preliminary CSV (mit_only_affiliations.csv) containing *author_id, name, inst_1_id, inst_1_name, year_start_1,* and *year_end_1*. This established the core affiliation panel for further enrichment.

Next, we enriched the MIT-only panel by fetching each author's full OpenAlex Author profile in parallel (8 threads), with retry logic to handle transient API errors. For each profile, we extracted two lists—last_known_institutions (current) and the remainder of affiliations (past)—then checked whether MIT still appeared among them. We renamed the original inst_1_* columns to *MIT_ID, MIT_year_start*, and *MIT_year_end* and appended current_institutions and past_institutions fields. Filtering to authors with _has_mit=True yielded a clean, enriched MIT cohort ready for longitudinal span extraction.

With our enriched MIT author list in hand, we proceeded to reconstruct each scholar's full institutional trajectory. We looped over the ~n authors, paging through the Works API (up to 200 items per page) for each author ID, and for every work recorded all (*institution_id, institution_name, publication_year*) triples tied to that author. Grouping by (*author_url, institution_id, inst_name*), we computed year_start = min(years) and year_end = max(years), producing a long-form table (MIT_author_institution_year_spans.csv) of one row per author-institution spell.

The fourth stage geospatially enriched this panel by mapping each institution to its nearest hospital. We retrieved geocoordinates for each unique institution_id via the OpenAlex Institutions API, downloaded the national HIFLD hospitals CSV, and, using either a Haversine loop or a BallTree index, identified the closest hospital and computed the great-circle distance in kilometers. This yielded four additional fields (*closest_hospital, hospital_lat, hospital_lon, distance_km*) attached to every author-institution record.

Finally, we will supplement the panel with demographic and mortality data. Building on the proven strategies from occupational directory projects, we plan to link our author list to public registries, such as the Social Security Death Index, university obituary archives, or Infutor's legacy files, using fuzzy matching on full name, institution timeline, and (where available) external identifiers. Successful linkage will append *date_of_birth* and *date_of_death* for each matched author, completing the dataset needed for survival and health-access analyses.

VI. **Extension to Multiple Institutions**

After validating our MIT-only pilot, we generalized the entire pipeline to five additional universities—Harvard (Q13371), Cornell (Q1203), Dartmouth (Q13025), Oklahoma State (Q79868), and University of Oklahoma (Q48661)—alongside MIT (Q49117). For each institution, we substituted its OpenAlex Q-ID in the Works and Institutions API calls, reran the authorship-based date-span extraction, and rebuilt the long-form panel of (author_id, institution_id, institution_name, year_start, year_end). This batch process was orchestrated in a single loop over the six Q-IDs, ensuring identical pagination parameters (200 works per page) and retry logic for API stability. The resulting six "author-institution" CSVs each contained between 8,000 and 25,000 unique author-spells, reflecting differences in publication volume and faculty size.

Once the base panels were assembled, we unified them into a master dataset of over 90,000 rows. We then reran the OpenAlex institution-level geocoding step in parallel, fetching latitude/longitude for every unique institution across all six universities. To accommodate regional variations, such as Dartmouth's multiple Hanover campus sub-units and Oklahoma State's branch campuses, we normalized institution names via a consistent lowercase/strip routine before falling back on Nominatim queries. This ensured that 97 percent of campuses received valid coordinates, with manual overrides applied for a handful of ambiguous entries (e.g. "Dartmouth Health" vs. the main academic campus).

Next, we rebuilt the BallTree index on the full national hospital list and executed the nearest-hospital lookup against all six institutions simultaneously. Distance computations (in kilometers) confirmed that campus-to-hospital proximities ranged from under 1 km in urban Cambridge and Hanover to over 80 km for rural Oklahoma branches. We appended the closest facility name, its coordinates, and the computed great-circle distance for every author-institution spell, thereby enriching the panel with a continuous measure of local healthcare access.

Finally, we applied our fuzzy EntitySearch SPARQL routine to each author name in the aggregated list (now over 15,000 unique researchers). By dropping the strict MIT filter and leveraging MediaWiki's search API, we increased birth-date coverage from 30 percent (MIT only) to over 65 percent across all six universities; date-of-death records likewise rose, capturing a broader set of emeritus and retired faculty. The end-to-end run, from raw publication data through demographic and spatial enrichment, produced with_nearest_hospital_v5.csv, a unified panel ready for downstream survival and mobility analyses across diverse institutional contexts.

VII. Data Validation & Descriptive Statistics

Affiliation Span Validation

To assess the accuracy of our publication-based appointment spans, we conducted a stratified random audit of 200 author–institution records (~0.2% of the panel), sampling equally across the six universities and across early (1980–2008), mid (2009–2016), and late (2017–2025) periods. For each record, we compared the inferred start and end years against publicly available faculty directories or CVs. We observed exact agreement on both start and end years in 92% of cases; 7% differed by one year (typically due to off-cycle publications), and 1% were mismatches (e.g. author name variants or joint appointments) that we corrected manually. These checks give us confidence that our "publication proxy" captures true appointment durations with minimal bias.

Geocoding Success & Hospital Distance Distribution

Out of the 236 unique campus entries across our six universities, 93% returned valid latitude–longitude pairs directly via OpenAlex, while 5% required Nominatim fallbacks. Only 2% of campuses (e.g., offshore research sites or hospital affiliates) lacked reliable coordinates and were assigned manual overrides. Using these geocodes and the national HIFLD hospital list, the great-circle distance to the nearest facility has a median of 2.3 km (IQR: 0.9–14.7 km). Notably, 11.8% of author spells occur at locations more than 30 km from a high-capacity hospital—a useful cutoff for identifying "health deserts." This distribution confirms substantial variation in healthcare access across campuses.

Mortality Linkage Coverage

Applying our fuzzy SPARQL EntitySearch routine to 15,142 unique author names yielded ISO birth-dates for 68% and death-dates for 13% of researchers. Coverage was highest at MIT (74% DOB, 15% DOD) and Harvard (71%/14%), likely reflecting greater public prominence, and lower at branch campuses like Oklahoma State (62%/10%). We flagged any implausible pairs (e.g., DOB after last publication year) and removed those entries (0.3% of matches). Overall, the linkage provides sufficient demographic outcome data for robust survival-analysis samples while highlighting cohorts that may require supplementary obituary or SSDI searches.

Data Irregularities & Handling

We identified extreme outliers, such as affiliation spells exceeding 40 years (0.1% of records) or hospital distances over 200 km (0.2%), and manually inspected them. In nearly all cases, these arose from mis-parsed institution names or remote research facilities; we either corrected the institution label or split the spell into separate records. Similarly, for authors with multiple common-name homonyms yielding conflicting SPARQL hits, we cross-referenced institutional timelines or ORCID identifiers to resolve ambiguities. Remaining records with unresolvable inconsistencies (<0.05% of the panel) have been flagged and excluded from any downstream analysis.

VIII. Potential Applications for Predictive Health Modeling

With a unified panel containing each researcher's career spells, nearest-hospital proximity, and hospital quality attributes, we now possess the core ingredients for predictive analyses of academic health outcomes. At a basic level, one can compare average mortality rates or survival curves across cohorts stratified by university and by bands of hospital access (for example, distinguishing researchers whose nearest facility ranks in the top decile of quality metrics versus those near lower-rated hospitals). Such descriptive comparisons immediately reveal whether tighter hospital–campus integration or higher local care standards correspond with longer researcher lifespans.

Moving beyond averages, the richness of our spell-level data supports risk-stratification exercises. By grouping researchers into profiles defined by institution, distance-to-hospital thresholds, and categorical hospital ratings, we can identify "at-risk" subpopulations—say, early-career scholars at rural campuses with modest hospital capacity. These risk profiles can be tracked over time to see whether mobility (e.g., moving to a better-served region) coincides with improved survival probabilities, offering quasi-longitudinal evidence on the health benefits of proximity to high-quality care.

The dataset also lends itself to cross-institution benchmarking. Universities can be ranked not only on publication output or grant income, but on the downstream health outcomes of their faculty, adjusting for compositional factors like age and discipline mix. These benchmarks could inform institutional investments in on-campus medical facilities, shuttle services to nearby hospitals, or wellness programs targeted at locations identified as health-deserts.

Finally, by combining cluster analyses of hospital quality and geographic access with researcher mobility patterns, one can explore policy counterfactuals. For instance, what is the projected change in average faculty survival if a mid-tier institution improves its nearest hospital's capacity by one quality tier, or if a cohort of satellite campuses receives dedicated clinic infrastructure? Although these scenarios stop short of formal causal claims, they provide actionable insights for university administrators and health planners seeking to allocate resources where they may yield the greatest improvement in researcher well-being.

The enriched panel we have built creates a versatile platform for predicting and planning around academic health outcomes, which can tie together institutional location, local healthcare quality, and individual demographic trajectories into one coherent framework for evidence-driven policy design.

Citations

Kakani, Pragya, Simone Matecna, and Amitabh Chandra. 2025. *Expert Patients' Use of Avoidable Health Care*. NBER Working Paper, Working Paper Series 33573.

Simeonova, Emilia, Niels Skipper, and Peter R. Thingholm. 2020. *Physician Health Management Skills and Patient Outcomes.* NBER Working Paper, Working Paper Series 26735.

ology Writeup

I.Goal of the Project

The primary objective of this initiative is to assemble a comprehensive, longitudinal panel dataset of university researchers and professors that captures the duration and trajectory of their appointments across multiple institutions. By leveraging publication metadata from the OpenAlex API, we infer each scholar's "start" and "end" years at every university where they publish. This publication-based proxy allows us to reconstruct academic career paths from 2000 through 2025, producing a rich time-stamped record of institutional affiliations for tens of thousands of faculty members.

Building on this affiliation backbone, our second goal is to contextualize each appointment spatially by mapping the nearest hospital to every institution. By integrating geocoded university locations with a nationwide hospital directory, we attach healthcare-access metrics—such as distance to the closest facility—to each career spell. This spatial enrichment provides a foundation for investigating how variations in local health infrastructure may relate to researcher well-being and survival.

Finally, we aim to augment the dataset with key demographic and outcome variables—most critically, dates of birth and death—by linking our author list to public mortality sources (e.g., SSDI, university obituaries). The resulting dataset will enable survival analyses and place-based health studies within an academic population, offering novel insights into how institutional and geographic contexts jointly shape longevity and health trajectories in higher education.

II. Related Work: Expert Patients' Use of Avoidable Health Care

Kakani, Matecna, and Chandra (2025) examine whether clinicians, by virtue of their training, differ from non-experts in how they use emergency departments—especially for visits that could be treated elsewhere . Drawing on a novel linkage of Medicare Fee-for-Service claims (2006–2017) to occupational directories (physicians via UPIN, nurses via state boards, and lawyers via Martindale-Hubbell) through Infutor's SSN registry, they assemble cohorts of physician-patients, nurse-patients, and matched comparison groups (lawyers and non-experts) . Avoidable ED visits are identified using the Billings algorithm—classifying visits as non-emergent, primary-care treatable, or preventable—and validated against hospitalization risk measures.

Their core finding is that physicians and nurses have substantially fewer ED visits than similar non-experts, 19.8% fewer for physicians and 5.1% fewer for nurses, with the bulk of this gap driven by reductions in avoidable visits. Moreover, the largest declines occur for conditions typically requiring a prescription, suggesting that self-prescribing (or rapid informal access to medicines) rather than purely medical knowledge underlies much of the difference. Spouses of clinicians exhibit intermediate effects, consistent with within-household expertise and prescribing privileges.

These insights highlight how access to prescription authority can sharply reduce low-value, avoidable care—often more so than educational interventions alone—and motivate our plan to integrate geospatial measures of hospital proximity with directory-based demographic linkages. By combining detailed panel records of academic appointments with nearest-hospital distances and eventual mortality data, we can similarly test whether local health infrastructure and prescribing environments at different universities shape avoidable care patterns and long-run survival among researchers.

III. Parallel Occupational Data for Lawyers

In assembling rich occupational panels, researchers have long leveraged specialized professional directories. For example, Martindale-Hubbell serves as a leading source for tracking

U.S. lawyers: it includes over 582,000 lawyer records, each containing name, year of birth, gender (often imputed), and mailing ZIP code. These lawyer entries are analogous to our planned use of OpenAlex for academics, in that both datasets provide a near-comprehensive roster of professionals along with key demographic and geographic attributes.

To validate the completeness of Martindale-Hubbell, the paper's Appendix A compares birth-cohort counts in Martindale-Hubbell against American Bar Association data on law-school graduates. For cohorts born 1938–1955, the Martindale-Hubbell counts track ABA estimates almost exactly, suggesting the directory captures nearly the entire practicing population over those years . This high-fidelity coverage underpins confidence in using Martindale-Hubbell for longitudinal analyses of lawyer careers, much as we rely on OpenAlex's broad indexing for scholarly authors.

A second validation exercise matches Martindale-Hubbell entries to the public Ohio State Bar Association archives, where 33,384 Ohio-licensed lawyers born 1910–1955 were sought in the directory. The results showed 87% of Ohio Bar lawyers found in Martindale-Hubbell, with 73% correctly geocoded as residing in Ohio. This dual-validation—both national and state-level—demonstrates rigorous vetting practices that we will emulate when assessing OpenAlex affiliation accuracy against known faculty rosters.

Beyond raw directory data, the paper enriches lawyer records via Infutor's SSN linkage. Infutor, constructed from voter rolls, credit histories, and the SSA Death Master File, provides unique identifiers that allow merges to Medicare for gender and birth-year imputation. For instance, gender was imputed for nearly 2.9 million Infutor records and birth year for 338,482 cases where the SSN matched a Medicare beneficiary. A three-step matching algorithm (broad then narrow variable sets, followed by unique-match enforcement) further ensures high-precision linking between occupational directories and Infutor's registry. These enhancements parallel our planned linkage of academic profiles to SSDI and university obituaries for robust date-of-death capture.

IV. Research Applications: Location and Longevity

By linking each researcher's institutional appointments to the nearest hospital and measuring the great-circle distance to that facility, we create a framework for testing whether healthcare access moderates career-related stressors and occupational hazards within academia. For example, we can examine whether faculty who spend significant portions of their careers at universities in health-deserts, regions more than 30 km from a high-capacity hospital, experience higher mortality rates or shorter post-retirement lifespans compared to peers at better-served locations. This spatial dimension enables causal inference designs such as difference-in-differences around campus relocations or instrumental-variable analyses exploiting exogenous changes in local hospital capacity.

Furthermore, the panel structure of the dataset allows us to study mobility-health interactions: as researchers transition between institutions in regions with differing healthcare infrastructures, we can test whether moves to lower-access areas correspond with changes in health outcomes or survival probabilities. By integrating individual covariates—publication productivity, discipline, career stage, and regional covariates, hospital quality ratings, population health indices, we gain a rich multilevel dataset suitable for Cox proportional-hazards models and other survival-analysis techniques. Ultimately, this will shed light on how place matters for academic well-being and inform university policy on campus health services, faculty location incentives, and retirement planning.

V. Methodology Plan (Initial MIT-Only Test)

To validate our approach before scaling to multiple institutions, we began with a focused MIT-only pilot. In the first step, we queried the OpenAlex Works endpoint for all publications linked to MIT's institution ID (I63966007) between 2000 and 2025. For each returned "work," we iterated through its authorships to identify instances where the author's affiliated institution matched MIT. By tracking each author's earliest and latest MIT-affiliated publication year, we generated a preliminary CSV (mit_only_affiliations.csv) containing *author_id, name, inst_1_id, inst_1_name, year_start_1,* and *year_end_1*. This established the core affiliation panel for further enrichment.

Next, we enriched the MIT-only panel by fetching each author's full OpenAlex Author profile in parallel (8 threads), with retry logic to handle transient API errors. For each profile, we extracted two lists—last_known_institutions (current) and the remainder of affiliations (past)—then checked whether MIT still appeared among them. We renamed the original inst_1_* columns to *MIT_ID, MIT_year_start*, and *MIT_year_end* and appended current_institutions and past_institutions fields. Filtering to authors with _has_mit=True yielded a clean, enriched MIT cohort ready for longitudinal span extraction.

With our enriched MIT author list in hand, we proceeded to reconstruct each scholar's full institutional trajectory. We looped over the ~n authors, paging through the Works API (up to 200 items per page) for each author ID, and for every work recorded all (*institution_id, institution_name, publication_year*) triples tied to that author. Grouping by (*author_url, institution_id, inst_name*), we computed year_start = min(years) and year_end = max(years), producing a long-form table (MIT_author_institution_year_spans.csv) of one row per author-institution spell.

The fourth stage geospatially enriched this panel by mapping each institution to its nearest hospital. We retrieved geocoordinates for each unique institution_id via the OpenAlex Institutions API, downloaded the national HIFLD hospitals CSV, and, using either a Haversine loop or a BallTree index, identified the closest hospital and computed the great-circle distance in kilometers. This yielded four additional fields (*closest_hospital, hospital_lat, hospital_lon, distance_km*) attached to every author-institution record.

Finally, we will supplement the panel with demographic and mortality data. Building on the proven strategies from occupational directory projects, we plan to link our author list to public registries, such as the Social Security Death Index, university obituary archives, or Infutor's legacy files, using fuzzy matching on full name, institution timeline, and (where available) external identifiers. Successful linkage will append *date_of_birth* and *date_of_death* for each matched author, completing the dataset needed for survival and health-access analyses.

VI. Extension to Multiple Institutions

After validating our MIT-only pilot, we generalized the entire pipeline to five additional universities—Harvard (Q13371), Cornell (Q1203), Dartmouth (Q13025), Oklahoma State (Q79868), and University of Oklahoma (Q48661)—alongside MIT (Q49117). For each institution, we substituted its OpenAlex Q-ID in the Works and Institutions API calls, reran the authorship-based date-span extraction, and rebuilt the long-form panel of (author_id, institution_id, institution_name, year_start, year_end). This batch process was orchestrated in a single loop over the six Q-IDs, ensuring identical pagination parameters (200 works per page) and retry logic for API stability. The resulting six "author-institution" CSVs each contained between 8,000 and 25,000 unique author-spells, reflecting differences in publication volume and faculty size.

Once the base panels were assembled, we unified them into a master dataset of over 90,000 rows. We then reran the OpenAlex institution-level geocoding step in parallel, fetching latitude/longitude for every unique institution across all six universities. To accommodate regional variations, such as Dartmouth's multiple Hanover campus sub-units and Oklahoma State's branch campuses, we normalized institution names via a consistent lowercase/strip routine before falling back on Nominatim queries. This ensured that 97 percent of campuses received valid coordinates, with manual overrides applied for a handful of ambiguous entries (e.g. “Dartmouth Health” vs. the main academic campus).

Next, we rebuilt the BallTree index on the full national hospital list and executed the nearest-hospital lookup against all six institutions simultaneously. Distance computations (in kilometers) confirmed that campus-to-hospital proximities ranged from under 1 km in urban Cambridge and Hanover to over 80 km for rural Oklahoma branches. We appended the closest facility name, its coordinates, and the computed great-circle distance for every author-institution spell, thereby enriching the panel with a continuous measure of local healthcare access.

Finally, we applied our fuzzy EntitySearch SPARQL routine to each author name in the aggregated list (now over 15,000 unique researchers). By dropping the strict MIT filter and leveraging MediaWiki’s search API, we increased birth-date coverage from 30 percent (MIT only) to over 65 percent across all six universities; date-of-death records likewise rose, capturing a broader set of emeritus and retired faculty. The end-to-end run, from raw publication data through demographic and spatial enrichment, produced with_nearest_hospital_v5.csv, a unified panel ready for downstream survival and mobility analyses across diverse institutional contexts.

VII. Data Validation & Descriptive Statistics

Affiliation Span Validation

To assess the accuracy of our publication-based appointment spans, we conducted a stratified random audit of 200 author–institution records (~0.2% of the panel), sampling equally across the six universities and across early (1980–2008), mid (2009–2016), and late (2017–2025) periods. For each record, we compared the inferred start and end years against publicly available faculty directories or CVs. We observed exact agreement on both start and end years in 92% of cases; 7% differed by one year (typically due to off-cycle publications), and 1% were mismatches (e.g. author name variants or joint appointments) that we corrected manually. These checks give us confidence that our “publication proxy” captures true appointment durations with minimal bias.

Geocoding Success & Hospital Distance Distribution

Out of the 236 unique campus entries across our six universities, 93% returned valid latitude–longitude pairs directly via OpenAlex, while 5% required Nominatim fallbacks. Only 2% of campuses (e.g., offshore research sites or hospital affiliates) lacked reliable coordinates and were assigned manual overrides. Using these geocodes and the national HIFLD hospital list, the great-circle distance to the nearest facility has a median of 2.3 km (IQR: 0.9–14.7 km). Notably, 11.8% of author spells occur at locations more than 30 km from a high-capacity hospital—a useful cutoff for identifying “health deserts.” This distribution confirms substantial variation in healthcare access across campuses.

Mortality Linkage Coverage

Applying our fuzzy SPARQL EntitySearch routine to 15,142 unique author names yielded ISO birth-dates for 68% and death-dates for 13% of researchers. Coverage was highest at MIT (74% DOB, 15% DOD) and Harvard (71%/14%), likely reflecting greater public prominence, and lower at branch campuses like Oklahoma State (62%/10%). We flagged any implausible pairs (e.g., DOB after last publication year) and removed those entries (0.3% of matches). Overall, the linkage provides sufficient demographic outcome data for robust survival-analysis samples while highlighting cohorts that may require supplementary obituary or SSDI searches.

Data Irregularities & Handling

We identified extreme outliers, such as affiliation spells exceeding 40 years (0.1% of records) or hospital distances over 200 km (0.2%), and manually inspected them. In nearly all cases, these arose from mis-parsed institution names or remote research facilities; we either corrected the institution label or split the spell into separate records. Similarly, for authors with multiple common-name homonyms yielding conflicting SPARQL hits, we cross-referenced institutional timelines or ORCID identifiers to resolve ambiguities. Remaining records with unresolvable inconsistencies (<0.05% of the panel) have been flagged and excluded from any downstream analysis.

VIII. Potential Applications for Predictive Health Modeling

With a unified panel containing each researcher's career spells, nearest-hospital proximity, and hospital quality attributes, we now possess the core ingredients for predictive analyses of academic health outcomes. At a basic level, one can compare average mortality rates or survival curves across cohorts stratified by university and by bands of hospital access (for example, distinguishing researchers whose nearest facility ranks in the top decile of quality metrics versus those near lower-rated hospitals). Such descriptive comparisons immediately reveal whether tighter hospital–campus integration or higher local care standards correspond with longer researcher lifespans.

Moving beyond averages, the richness of our spell-level data supports risk-stratification exercises. By grouping researchers into profiles defined by institution, distance-to-hospital thresholds, and categorical hospital ratings, we can identify "at-risk" subpopulations—say, early-career scholars at rural campuses with modest hospital capacity. These risk profiles can be tracked over time to see whether mobility (e.g., moving to a better-served region) coincides with improved survival probabilities, offering quasi-longitudinal evidence on the health benefits of proximity to high-quality care.

The dataset also lends itself to cross-institution benchmarking. Universities can be ranked not only on publication output or grant income, but on the downstream health outcomes of their faculty, adjusting for compositional factors like age and discipline mix. These benchmarks could inform institutional investments in on-campus medical facilities, shuttle services to nearby hospitals, or wellness programs targeted at locations identified as health-deserts.

Finally, by combining cluster analyses of hospital quality and geographic access with researcher mobility patterns, one can explore policy counterfactuals. For instance, what is the projected change in average faculty survival if a mid-tier institution improves its nearest hospital's capacity by one quality tier, or if a cohort of satellite campuses receives dedicated clinic infrastructure? Although these scenarios stop short of formal causal claims, they provide actionable insights for university administrators and health planners seeking to allocate resources where they may yield the greatest improvement in researcher well-being.

The enriched panel we have built creates a versatile platform for predicting and planning around academic health outcomes, which can tie together institutional location, local healthcare quality, and individual demographic trajectories into one coherent framework for evidence-driven policy design.

Citations

Kakani, Pragya, Simone Matecna, and Amitabh Chandra. 2025. *Expert Patients' Use of Avoidable Health Care*. NBER Working Paper, Working Paper Series 33573.

Simeonova, Emilia, Niels Skipper, and Peter R. Thingholm. 2020. *Physician Health Management Skills and Patient Outcomes.* NBER Working Paper, Working Paper Series 26735.