Bodies of Knowledge:
Processes of Social Construction and the Epistemology of Empirical Measurement in Computational Social Science

Adam Visokay

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington
2026

Reading Committee:

Sasha Johfre, co-Chair

Tyler McCormick, co-Chair

Sarah Quinn

Program Authorized to Offer Degree:

Department of Sociology

University of Washington

Abstract

Bodies of Knowledge:
Processes of Social Construction and the Epistemology of Empirical Measurement in Computational Social Science

Adam Visokay

Chairs of the Supervisory Committee:

Sasha Johfre[†] and Tyler McCormick[†,§]

[†]Department of Sociology [§]Department of Statistics

This dissertation argues that taking the relationship between conceptualization and measurement seriously as both a theoretical and empirical problem generates contributions that are methodological, substantive, and sociological. Across three empirical studies, I show that the categories through which social scientists observe neighborhoods, bodies, and causes of death are artifacts of measurement systems that carry their own histories and assumptions, and that choosing more conceptually aligned measures can unsettle conclusions that may be taken for granted. Using large language model annotation of Craigslist rental advertisements, I show that administrative neighborhood boundaries overlook the process by which listing agents claim and construct urban space, revealing patterns of spatial representation invisible to coordinate-based approaches. Comparing BMI-, waist circumference-, and total body fat-based obesity prevalence trends of the US adult population from 1999–2018, I find that while BMI-defined obesity rose 13 percentage points, obesity derived from directly observed body fat showed no statistically significant change – and that this aggregate estimate masks a pronounced gendered divergence. Addressing cause-of-death estimation where 40 percent of global deaths are not accompanied by a traditional death certificate, I use a validation dataset to demonstrate that language models applied to verbal autopsy text narratives can match or exceed structured-data classification performance, and introduce `multiPPI++`,

extending prediction-powered inference to multinomial outcomes for valid population-level inference from predicted labels. Together, these studies advance an epistemology of empirical measurement: one in which *how* we observe and categorize the social world is treated not as simply preliminary to quantitative analysis, but as fundamentally constitutive of the knowledge it produces.

Contents

Acknowledgements

This work would not have been possible without the support, care, and inspiration I received from so many along the way. Words cannot do justice, but nevertheless, here are some words. To my parents, Steve and Alison, and my sisters, Lauren and Bethany: thank you for making the home environment I grew up in so supportive and stimulating. Each of you has encouraged me to chase my diverse curiosities from a young age – from Awaiting Shipwrecks, to running, to this most recent academic milestone.

Zoey, Andy and Arlo: thank you for all that you have done to care for me, laugh and play and celebrate with me, and for encouraging me to submit my application to graduate school despite my own reservations at that juncture.

Ben, Summer, Nicole, Fred and Eira: I am so grateful for the time we spent living together in community. Among so many memories, I will especially cherish learning to swim and ride my bike properly (also how to dress). And to Remy: watching your parents prepare and seeing you grow has been so inspiring and humbling – I can't wait for my next visit.

Thank you to Max, Molly, Kristin and Emily, my Sunnyside community, for being so present and invested in one another, and for celebrating the big wins and holding space for the griefs life throws our way. Living together has been a total joy.

Thank you to Nat for helping me reconnect with my goofy, authentic self, for being vulnerable, for the ice cream and the rocks.

To my dear friends in the Soggy Bottom Boys: thank you for your consistent availability to goof off, compete, unwind, or listen to me opine about my niche macroeconomic interests. It is so special to maintain close friendships for decades as we move through different life transitions, even while scattered across the country. Cheers to having so many rocks of the rift.

Thank you to Chris Fox and Adam Smith for believing in me, for devoting your time, energy, and experience to helping me grow as an individual and a teammate, for modeling how to set the bar exceptionally high and execute, and for always reminding me that being a well-rounded, good person comes first and foremost.

I also want to thank the many families who have welcomed me with open arms: The Deals, Shepherds, Batzels, Grinnans, and Hochmans.

Finally, many thanks to my academic community. My terrific mentors – Tyler McCormick, Sasha Johfre, Sarah Quinn, Brandon Stewart, Sara Curran, Audrey Dorélien, Theresa Rocha Beardall, Zack Almquist and Kyle Crowder. Emilio Zagheni, my 2024 PHDS cohort, and the rest of the Max Planck Institute for Demographic Research in Rostock. My rockstar 2023 UW Sociology cohort: Eddie, Sidnee, Luke, Jess, Aidan, Brandon and Kelly. To Lee Coppock, for initially piquing my interest in pursuing graduate school and an academic career.

My many collaborators, especially Denis Peskoff for always being willing to try new things and grow together. The CSDE, CSSS, eScience and RAISE communities at the University of Washington, as well as Jon Froehlich and the Makeability Lab. And to Courtney Allen, Aryaa Rajouria, Ryan DeCarsky, Crystal Yu, Todd Nobles, İhsan Kahveci, Maxine Wright, Chassidy Wen, Lauren Woyczynski, Desiree Salais, Danny Nolan, Mark Igra, and Yehong Deng.

—- CHAPTER 1 ——-

Chapter 1

Introduction

Quantitative social science is fundamentally an interplay between concepts, which may be more or less abstract, and empirical artifacts, which may be more or less measurable. Sometimes, concepts of interest are relatively straightforward and can be measured with considerable precision in ways that align closely with how we understand them. An individual's height, for example, is both a comparatively simple concept – how tall a standing person is at a particular moment – and one that can be measured consistently using instruments and procedures that comport with ordinary understandings of height (e.g., standing barefoot against a wall with a ruler or measuring tape). A person who is 165 centimeters tall is, within a small margin of measurement error, 165 centimeters tall regardless of context.

Yet even seemingly simple concepts become more complicated once measurement gives way to categorization. Translating a continuous quantity like height into categories such as "tall" or "short" necessarily introduces social context into the process of interpretation. Tall for an eight-year-old? Short for a professional basketball player? Tall relative to women in the United States, or men in Northern Europe? The number of centimeters corresponding to "tallness" cannot be determined independently of the social group against which the comparison is made. In everyday life, these contextual assumptions are often implicit enough that categorization feels natural and unproblematic. Individuals, institutions, and societies alike have difficulty resisting the naturalization of socially constructed categories, in part because the world is complex and many classifications appear self-evidently meaningful in

practice. Yet this process of naturalization has profound epistemological consequences. What we "know" about the social world cannot be disentangled from how we choose to interrogate it, and measurement decisions play a central role in shaping the realities that become empirically visible, statistically legible, and institutionally actionable.

Concepts such as neighborhoods, obesity, and causes of death are not naturally given categories waiting to be observed. They are socially constructed, institutionally stabilized, and operationalized through measurement systems that shape the kinds of inferences researchers can make and, ultimately, the forms of knowledge that become socially authoritative. Neighborhoods are contained within rigid and discrete boundaries. Health status is derivative of BMI thresholds. Causes of death fit within a system of standardized international classification of diseases. In each case, measurement is not merely a technical procedure for observing an underlying reality; it is part of the process through which that reality is defined, categorized, and made legible.

This dissertation examines how mismatches between conceptualization and measurement produce divergent forms of social knowledge. Across three empirical domains – urban space, population health, and mortality estimation – I show how, for each concept – neighborhood boundaries, obesity status, and cause of death – our substantive conclusions vary depending upon the measures and classificatory systems used to construct them. Where neighborhood boundaries are understood to begin and end depends on which spatial imaginaries and institutional definitions are privileged. Narratives about the rise of obesity in the United States since 1999 differ substantially depending on whether obesity is operationalized through BMI, waist circumference, or directly measured adiposity. And in regions where death certificates and traditional autopsies are unavailable, understandings of population mortality depend on inferential systems that translate narrative accounts into probabilistic causes of death.

Across these three studies, I draw on a range of methods old and new. Large language model annotation and validation, topic modeling, and spatial regression to recover socially con-

structed spatial claims from unstructured text; non-parametric imputation, survey-weighted estimation with multiple imputation via Rubin's rules, and direct demographic standardization in the tradition of Kitagawa–Oaxaca–Blinder decomposition to compare population health trends across measurement systems; and bag-of-words and large language model classification, ML-based cause-of-death prediction, and statistical calibration via prediction-powered inference to enable valid population inference from narrative data. What unifies these approaches is not a commitment to any single methodological tradition but rather a shared orientation: using the best available tools to quantify the gap between the concepts social scientists care about and the measures through which those concepts are made empirically tractable.

Chapter 2 examines the social construction of urban space by studying the gaps between various "official" neighborhood boundaries and how those boundaries are referenced, understood and contested in practice. Drawing on rental advertisements in Chicago listed on Craigstlist from 2018–2024, I use large language model annotation methods to classify neighborhood identity from the unstructured text in the listing titles and descriptions. This is approach seeks to approximate a more nuanced task than conventional string-matching or entity recognition because it requires distinguishing genuine location *claims* from mere *mentions*. This LLM pipeline, validated against manual annotation and compared to string-matching entity recognition reveals three systematic patterns of spatial representation that would otherwise be overlooked by matching latitude/longitude coordinates to established boundary shapefiles. While I am not the first to conceptualize neighborhoods as social constructs, these findings demonstrate a novel way for social scientists to identify this process of construction at scale using unstructured text.

Chapter 3 asks "How has US adult obesity prevalence changed since 1999?". At first consideration, one might ask themselves, don't we *know* this already? I take a critical measurement perspective to exploring this question, drawing a distinction that is rarely made explicit in public health surveillance: rising BMI, rising adiposity, and rising obesity

may be related, but are fundamentally different in important ways. This orientation to the question brings an important tension to the fore: the WHO defines obesity as excessive fat tissue accumulation that poses an increased risk for adverse health outcomes, but BMI, the predominant screening tool for obesity, measures neither fat nor health outcomes directly. What's more, it is not the case that we do not have measures of fat or adverse health outcomes. This chapter uses data collected in 1999-2018 from the National Health and Nutrition Examination Survey to compare obesity prevalence derived from total body fat percentage measured using dual-energy X-ray absorptiometry (DXA) scans, waist circumference, and BMI. The central finding is stark: while adult obesity prevalence according to BMI rose by 13% since 1999, DXA-based obesity showed no statistically significant change. Decompositions of these aggregate trends by race/ethnicity, age and sex reveal an even more complicated story. Most striking among them that DXA-based obesity actually *fell* for adult women over this time period while it increased at the same rate as BMI-based obesity for men. By focusing on the differences between measures used to construct our understanding of the "obesity epidemic," this chapter presents empirical evidence that unearths and complicates a "settled" narrative about how the US has been getting fatter for decades.

Finally, Chapter 4 contributes to cause-of-death estimation in low-resource settings where most deaths occur outside the formal healthcare system and death certificates are unavailable. Such conditions characterize 40% of deaths worldwide, and having robust and reliable understandings of the various causes are essential to effective policy making. In these settings, verbal autopsies (VA) – structured interviews with the surviving family or caregivers about the circumstances of a death – are the primary tool used to monitor mortality trends. Prior algorithmic approaches to cause-of-death classification have relied predominantly on structured questionnaire data. This chapter adds to this literature by testing the extent to which language models can enable us to make mortality classifications relying only on unstructured text narratives. Then, given these cause of death labels, how can we perform valid statistical inference to understand how population dynamics like aging of policy interventions

affect the cause of death distribution? This chapter introduces *multiPPI++*, an extension of prediction-powered inference (PPI++) to the multinomial classification setting, which gives us leverage on answering this question. Empirical analysis of VA data demonstrates that naive use of predicted causes – without this correction – produces misleading estimates of cause-specific mortality fractions, while multiPPI++ recovers valid confidence intervals even when the prediction model is imperfect or when the relationship between symptoms and causes differs across sites. The chapter illustrates how the categories through which death is made statistically legible are not merely observed but constructed through inferential systems whose assumptions shape what patterns of mortality become visible to researchers and policymakers.

Chapter 2

Social Construction of Urban Space: Using LLMs to Identify Neighborhood Boundaries From Craigslist Ads

Adam Visokay, Ruth Bagley, Ian Kennedy, Chris Hess, Kyle Crowder, Rob Voigt, Denis Peskoff. 2026. In *Proceedings of the Seventh Workshop on NLP+CSS.* Association for Computational Linguistics. San Diego, California.

Abstract. Rental listings offer a window into how urban space is socially constructed through language. We analyze Chicago Craigslist rental advertisements from 2018 to 2024 to examine how listing agents characterize neighborhoods, identifying mismatches between institutional boundaries and neighborhood claims. Through manual and large language model annotation, we classify unstructured listings from Craigslist according to their neighborhood. Further geospatial analysis reveals three distinct patterns: properties with conflicting neighborhood designations due to competing spatial definitions, border properties with valid claims to adjacent neighborhoods, and "reputation laundering" where listings claim association with distant, desirable neighborhoods. Through topic modeling, we identify patterns that correlate with spatial positioning: listings further from neighborhood centers emphasize different amenities than centrally-located units. Natural language processing techniques reveal how definitions of urban spaces are contested in ways that traditional methods overlook.

2.1 Contested Neighborhood Boundaries

Neighborhood location matters for a wide range of individual and collective outcomes (Sampson et al., 2002; Sharkey and Faber, 2014; Minh et al., 2017; Chyn and Katz, 2021). Beyond objective demographic characteristics, the subjective features of a neighborhood—its reputation, status, or stigma—shape resident satisfaction, place attachment, and overall well-being (Tran et al., 2020; Kullberg et al., 2010; Permentier et al., 2011; Otero et al., 2024). Neighborhood reputation also structures the economic value of property, patterns of investment, and the residential mobility that drives neighborhood stratification (Krysan and Crowder, 2017; Evans and Lee, 2020; Kirk, 2024; Korver-Glenn and Mayorga, 2024).

We define **neighborhood identity** not just as a boundary in space, but as the spatial area that corresponds with the patterns of social activity and perceptions of people living in the area. This conceptualization builds on a lineage of research utilizing user-generated content (UGC) to define urban space (Plangprasopchok and Lerman, 2009; Hollenstein and Purves, 2010; Hiippala et al., 2019; Brunila et al., 2023). Because this identity is socially constructed through interaction and language, it is inherently fluid and contested. Identifying a singular "true" neighborhood boundary is not our aim, but rather, mapping the contours of contestation: the systematic slippage between institutional maps and the spatial claims made by social actors.

Neighborhoods are socially constructed through interaction and language (Zelner, 2015; Hohle, 2023; Stuart et al., 2024), yet traditional urban research often relies on rigid administrative boundaries as a proxy for neighborhood identity. In practice, listing agents frequently navigate a tradeoff between geographic fidelity and reputational leverage, substituting symbolic identity for physical proximity when properties sit at the periphery of desirable areas.

[1] *Conflicting conceptions* of the Humboldt Park neighborhood according to the official City of Chicago limits (blue), Zillow's boundary (red) and rental advertisements (points). Orange points depict unit listings claiming to be Humboldt Park while purple points claim elsewhere. The green star denotes the LLM-defined social center of Humboldt park.

[2] Across Chicago, the share of place-based language decreases and generic "boilerplate" language increases in listings for units further from the social center of the neighborhood.

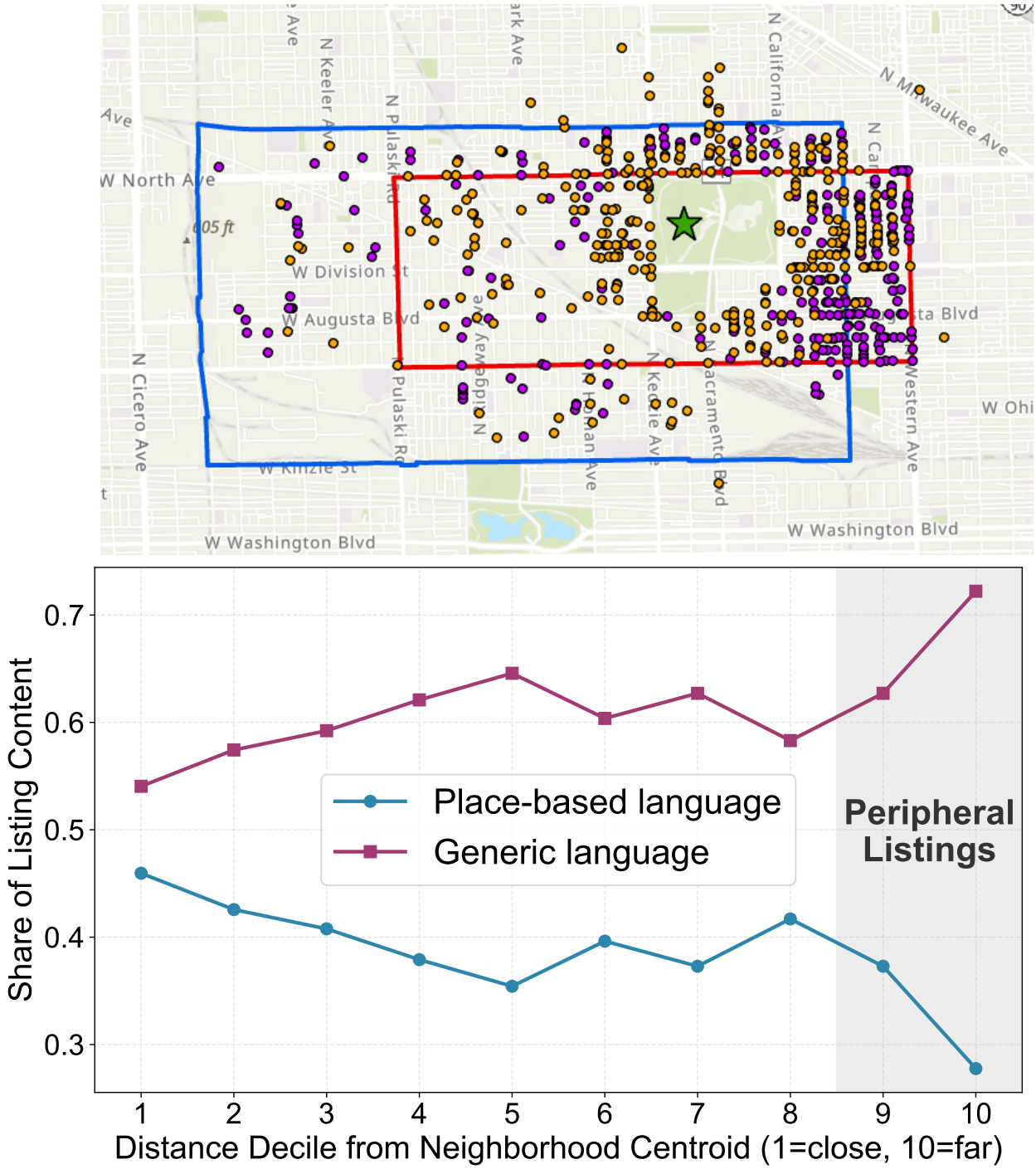


Figure 2.1: Extracting neighborhoods from unstructured rental listings with LLMs (RQ1, Section 2.3) provides insight into the social construction of space (*RQ2*, top) and allows us to study how language changes relative to distance from neighborhood centers (*RQ3*, bottom).

We characterize this behavior not as random noise, but as a systematic distortion of urban space. Still, an important question remains: do spatial claims that depart from institutional

maps always represent strategic "reputation laundering" (Stuart et al., 2024), or might they reflect legitimate disagreements arising from the inherent ambiguity of socially constructed boundaries?

A related challenge for urban sociology has been the difficulty of observing and classifying such distortions at scale. Conventional data is blind to the strategic manipulation of spatial labels, and traditional NLP methods like string-matching struggle to distinguish between casual mentions and strong locational claims. Large Language Models (LLMs) provide a transformative opportunity to recover this latent social variable: claimed neighborhood identity. By evaluating LLMs on Chicago Craigslist rental advertisements (2018–2024), this work provides answers to the following Research Questions:

- **(RQ1) Measurement Viability:** Can zero-shot LLMs accurately identify specific neighborhood claims (vs. mere mentions) in unstructured rental listings compared to more traditional string-matching?
- **(RQ2) Social Location:** Where are neighborhoods actually located according to listing claims, and can we define a meaningful "social center" for each neighborhood?
- **(RQ3) Linguistic Substitution:** How does marketing language vary with spatial location? Specifically, does place-based language change as listings move farther from their claimed neighborhood center?

Section 4.2 describes our spatially-anchored corpus of Chicago listings. Section 2.3 evaluates LLM performance against string-matching baselines **(RQ1)**. Section 2.4 develops a framework for identifying neighborhood "social centers" **(RQ2)**. Section 2.5 and Section 2.5.2 analyze the semantic structure and statistical associations between unit language and spatial positioning **(RQ3)**. Finally, Section 3.4 contextualizes these findings within the broader field of computational social science.

2.2 Craigslist Housing Advertisements

We use data collected from Chicago Craigslist rental advertisements from 2018 to 2024 to identify listings with mismatches between the neighborhood containing the unit and the neighborhood claimed by the listing agent.[1]

2.2.1 Why Craigslist?

These rental listings offer a particularly valuable lens for examining how neighborhoods are socially constructed and contested because Craigslist's platform design creates an unusually unconstrained environment for spatial classification. Unlike many digital platforms that restrict users to predetermined administrative or commonly recognized neighborhood boundaries, Craigslist allows advertisers to freely designate any neighborhood label in their listings with unstructured text. This feature transforms rental advertisements into sites of boundary-making where the socio-spatial imaginary of the city becomes visible.

The neighborhood fields in these listings represent more than mere locational information—they reveal how real estate actors actively participate in constructing, reinforcing, or challenging existing spatial hierarchies. When landlords and property managers assign neighborhood labels to their listings, they engage in acts of spatial categorization that reflect both market strategies and internalized cognitive maps of urban space. These choices may align with officially recognized boundaries, reproducing understandings of place, or deliberately transgress established spatial categories to claim association with perceived higher-status areas.

Preceding computational analysis of Craigslist rental listings explores price distributions (Boeing and Waddell, 2017), neighborhood descriptions (Kennedy et al., 2020; Besbris et al., 2021), housing policy interventions (Boeing et al., 2021), exclusionary language (Stewart

[1]We have been actively collecting data in Chicago since 2018, providing a rich window into the discursive construction of urban space. It includes all available advertisements each day from December 2018 until June 2024 using the webscraper `Helena` (Chasins et al., 2018; Hess and Chasins, 2022). Occasional changes to the architecture of the Craiglist website result in limited periods of data loss, the longest of which was from August 2019 to early October 2019.

et al., 2023), affordability (Hess et al., 2023), and home security (Somashekhar et al., 2024). Holistically, research shows that rental listings on Craigslist align with and appear to reproduce social inequality. Recent work has begun to focus on the importance of neighborhood names and the places those names describe (Schachter et al., 2024), with specific focus on contested naming: when advertisements use neighborhood names that seem to diverge from the name most local residents would use for that space.

2.2.2 Why Chicago?

We focus on Chicago because it stands as a quintessential "city of neighborhoods," where locally recognized community areas hold exceptional cultural, economic, and social significance (Hwang and Sampson, 2014). Chicago is an ideal site for examining the social construction of urban space due to the high salience of its neighborhood boundaries. While the city maintains 77 officially recognized community areas, these rigid, non-overlapping boundaries often fail to represent the fluidity of neighborhood identity in reality. By using Chicago's stable institutional definitions as a point of comparison, we can more effectively identify and quantify how real estate actors use language to challenge or reinforce existing spatial hierarchies. This neighborhood orientation is so deeply embedded in Chicago's social fabric that it shapes how residents understand their place in the city, influences social networks, and structures daily mobility patterns (Kaysen, 2024).

By analyzing patterns in how these spatial designations align with or diverge from official boundaries, we can observe in real time the processes through which neighborhood reputations are reinforced or contested. The negotiated quality of these spatial boundaries becomes particularly visible when examining cases where advertisers claim association with neighborhoods other than those in which units are formally located, according to administrative boundaries. Such instances of spatial repositioning offer a window into the dynamics that shape how urban space is valued, categorized, and ultimately experienced by various stakeholders from listing agents to residents or potential tenants to municipal administrators.

2.3 Detecting Neighborhoods with LLMs

Online rental advertisements are generally unstructured and vary widely between listings. Distinguishing between which neighborhoods are mentioned in a listing from which neighborhood(s) the listing claims to be in is a nuanced task of great importance to social scientists interested in the social construction of urban space. This answer to "which neighborhood does this advertisement claim the unit to be in?" is not always obvious, even to a human annotator. For example,

> ... this fully restored **East Lakeview** property sits on a beautiful tree-lined street located in the heart of the popular **Wrigleyville** neighborhood ...

It is clear based on the language that this advertisement is not merely mentioning these neighborhoods, but staking a strong claim to being located in both. Wrigleyville is a Chicago neighborhood within the larger neighborhood of East Lakeview, so this claim is entirely coherent. However, reconciling such competing claims at scale is a principal challenge inherent to this particular task.

Furthermore, some listings contain mentions of several irrelevant neighborhoods, even dozens like this example:

> ... Disclaimer: Pricing, availability, and specials are subject to change at any time and without notice. HotSpot Rentals services the following neighborhoods: South Loop, Printers Row, Near North, River North, Gold Coast, West Loop, Fulton River District, West Loop Gate, The Loop, Streeterville, Lakeshore East, New East Side, Old Town, Medical District, University Village, Near North, River West, Lincoln Park South, Lakeview, Uptown, Ukrainian Village, Wicker Park, Edgewater, Ravenswood, Bronzeville, Logan Square.

Making the distinction between neighborhood mentions and strong claims that a unit is in a particular neighborhood is a nuance that large language models are particularly well-suited

for compared to existing methods. To make this comparison, first we label the full corpus using a bespoke string-matching approach which serves as the baseline "best practice" which we compare to the Language Model-based labeling. Both sets of labels are evaluated against a subset of 200 manually labeled neighborhood listings. These manual labels were produced by authors of this article.

2.3.1 Manual Labeling

The process for creating our validation set of "gold standard" labels considers three sources of information for each advertisement in the following order. First, if the title field includes a neighborhood name, that becomes the manual label for the neighborhood claim. Then if there is no claim in the title, we consider the body of the listing. This is the largest source of text in each advertisement, and also the most ambiguous with respect to identifying strong claims. When faced with multiple claims – as in the quote above – we take the first claim as the manual label. Finally, if neither the title nor body fields contain a neighborhood claim, we extract a neighborhood claim from the neighborhood field, if it exists. An advertisement only receives the 'unknown' label if there is not a strong claim in any of the three fields. We follow the same prioritization scheme in the string-matching and LM labeling procedures.

2.3.2 Data Pre-Processing

Before labeling we performed standard data pre-processing and de-duplication on the raw text. We removed common boilerplate text that appeared in virtually all Craigslist listings (such as "QR Code Link to This Post"), corrected Unicode translation errors to ensure consistent character rendering, and precisely mapped listings onto Zillow and City of Chicago neighborhood boundaries to confirm geographical validity. We also performed thorough de-duplication of listing entries by title and body text, retaining only the most recent version when multiple entries existed, as Craigslist saves edited listings as separate posts. This preparation distilled a clean dataset of 30,531 unique listings from an initial

corpus of n=128,764 initial observations.

2.3.3 String-Match Labeling Neighborhoods

To determine which Chicago neighborhood a rental advertisement belongs to based on the text in the post we begin with a list of 192 distinct neighborhoods from Zillow, a real-estate marketplace company which provides commercial neighborhood lists for major US cities. Then, we manually construct a comprehensive list of regular expressions that can match the official name and its alternatives (e.g. *wriggleyville, wriglyville, wrigglyville, wrigley ville* for *wrigleyville*). These regex patterns are designed to be flexible with spaces and case-insensitive. Following the same protocol as the manual annotations, we use a function which tries to match neighborhoods in the listing title, body and dedicated neighborhood field. When multiple neighborhoods match, we select the label which appears earliest in the text.

2.3.4 Language Model Labeling Neighborhoods

We prompt GPT-4.1 mini as a high quality relative to cost option.[2] Table 2.1 contains our prompt.

BASE_PROMPT
Extract the Chicago neighborhood from the rental text.
Rules:
- NEVER use the address or zip code to determine the neighborhood
- Choose only explicitly stated neighborhood from possible responses
- If neighborhood is unclear mark it as [unknown]
- Format precision: ‘‘lakeview’’ but not ‘‘x, y, z’’ for ‘‘lakeview residence near x, y, z’’
text: {body}
Possible responses: {zillow_list}
Return only:
label: [neighborhood]

Table 2.1: Prompt format for extracting Chicago neighborhood claims from rental listings.

[2]We compare a sample of regular vs mini vs nano, and other non-OpenAI options. Reasoning models are unnecessary for this extraction task.

We create three separate labeling workflows, one for each title, body and neighborhood field. We input the data independently to avoid leakage. A Zillow neighborhood list is provided to constrain possible responses to items in the list. In the case of unknowns, we prioritize the label from the title, then the body, then the neighborhood field. Only if all three are unknown is the final neighborhood claim labeled 'unknown'.

2.3.5 Label Post-Processing

While we engineer the prompt to provide structured output in the form `label:response`, the outputs still require post-processing. We implement a multi-stage post-processing pipeline to standardize the three LLM labels for each advertisement and assess model performance against manual validation. Through manual review we flag instances of minor spelling discrepancies (e.g. 'lakeview' instead of 'lake view', 'wrigglyville' instead of 'wrigleyville'). For text normalization, we construct a replacement dictionary to correct such instances. We implement the same priority-based claim-selection algorithm as is used for manual annotation and string matching. The post-processing system first examines the listing title label; if no neighborhood is identified, it analyzes the listing body label, and if still unsuccessful, it checks the neighborhood field label; and only then returns 'unknown' classification. This aligns with how potential renters process listing information, prioritizing the most prominent textual elements. When faced with compound neighborhood designations, we prioritize the first neighborhood when multiple were present, just as we do for manual labeling and string-matching. For outputs that contain separators (e.g. 'uptown, ravenswood' or 'avondale/logan square'), we extract the first neighborhood claim before a separator.

2.3.6 Evaluation of Labeling Task

We compare the string-match and LLM labels to the manual labels in the 200-item validation set. Performance metrics are shown in Table 2.2. While accuracy scores are comparable (GPT-4 mini's 85% is marginally better than the string matching's 79%), the

Metric	String Match	GPT-4.1
Accuracy	0.79	**0.85**
Macro Average F1	0.62	**0.70**
Weighted Average F1	0.77	**0.85**

Table 2.2: GPT-4.1 outperforms the string match-based keyword search on the neighborhood claim labeling task. While the performance gain is marginal, the LLM labeling process was cheap, fast and can be scaled up.

disagreements largely reflect the inherent ambiguity of neighborhood classification – a challenge even human annotators face when listings claim multiple neighborhoods. A notable advantage of the LLM approach is its efficiency and scalability. Unlike string matching on keywords, which requires extensive manual configuration of locale specific patterns, the pre-trained LLM can work with a zero-shot prompt, making it adaptable.

2.4 Geospatial Analysis of Neighborhoods

Although Craiglists rental listings do not comprehensively show every available property in the Chicago area, there is still substantial coverage across the city, as seen in Figure 2.2. As a result, we take these rental listings to be a reasonably representative sampling, which we use to identify neighborhoods as they might be conceived of by the people, or at least the realtors, of Chicago.

Neighborhood boundaries are nebulous and hard to define; even the City of Chicago and Zillow, both with access to a great deal of data/information, have developed quite different maps of neighborhood boundaries. Figure 2.1 shows that while the official Zillow and city boundaries of Humboldt Park include most of the posts claiming to be in that neighborhood, none of the borders are the same. In addition, borders between neighborhoods are likely less rigid than an official boundary might suggest; despite being located within the formal boundaries of Humboldt Park, there are a number of listings, mainly around the edges, claiming to be part of other nearby neighborhoods, primarily Ukrainian Village, Wicker Park,

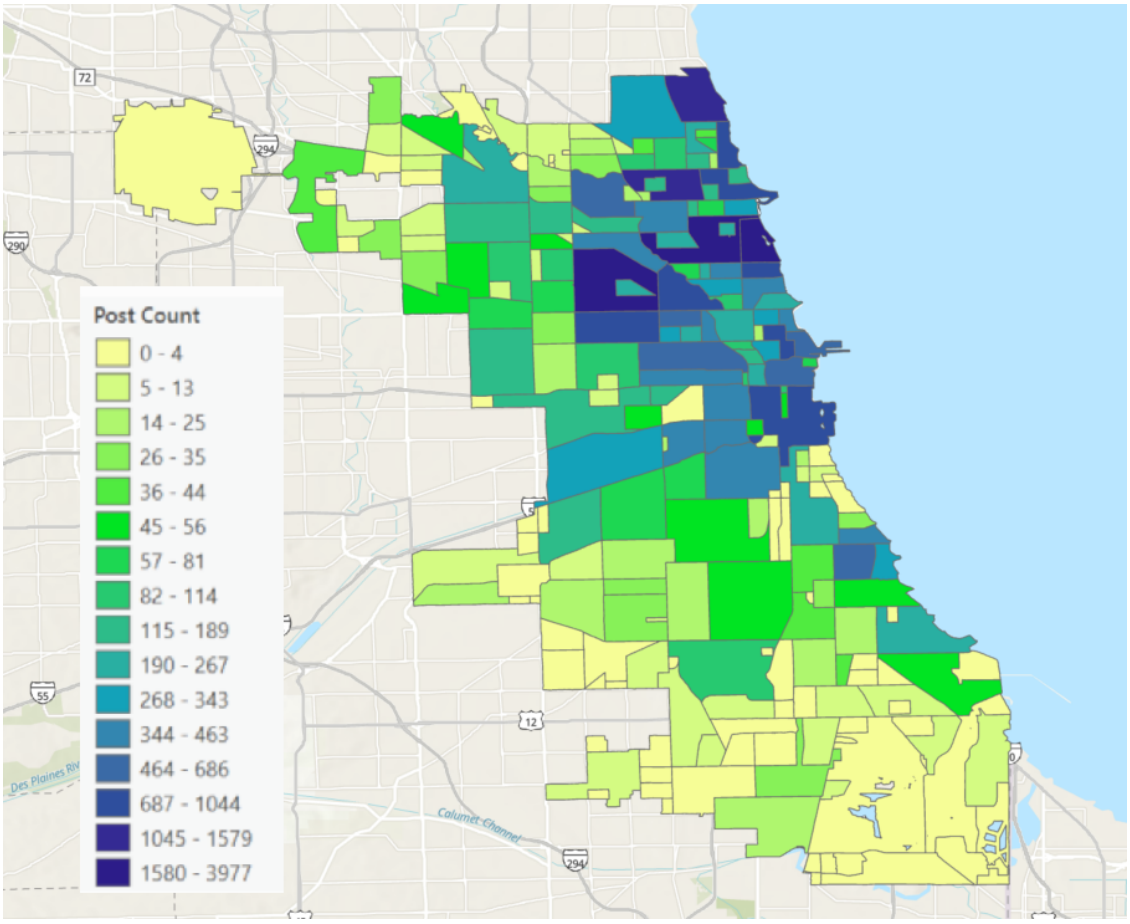


Figure 2.2: The city of Chicago and its constituent neighborhoods, as defined by Zillow, colored to represent the number of posts for properties located in each area

or West Town.

Using the rental listings, we conceptualize the "social center" of the neighborhood as the centroid of all listings that claim to be located within that neighborhood. We use `geopandas` to identify the centroid of all posts claiming to be in the same neighborhood using the latitude and longitude coordinates of all property locations. The map of Humboldt Park suggests this is an effective method, because the social center (represented by the green star) is in fact located in the eponymous park which is considered to be the heart of the neighborhood.

However, not all posts claiming to be in a given neighborhood may have the same strength to their claim; advertisements for apartments at the center of a popular neighborhood like Logan Square likely have different characteristics than posts for units on the fringes of the

neighborhood. The posts on the fringes might even be trying to seem more desirable by claiming to be in a more popular neighborhood, whereas it might be more of an objective description of location for a property actually located on Logan Square. We conceptualize this by identifying how far from the social center a post is, using three metrics for distance: 1) Raw Distance: Euclidean distance between the latitude and longitude coordinates of the post and that of the neighborhood centroid; 2) Relative Distance: raw distance for all posts in the neighborhood projected onto a [0,1] interval using min-max scaling; 3) Z-scored Distance: z-scored distance for posts claiming to be in the same neighborhood.

We use these measures to distinguish a specific subset of the posts: those on the periphery of a neighborhood. We define peripheral posts as those that are in the furthest 20% of posts from the centroid for a given neighborhood (for any neighborhood labels with at least 5 posts).

2.4.1 Comparison of Neighborhood Boundaries

In order to get a more concrete understanding of different conceptions of neighborhoods, we identify two more neighborhoods to explore more in depth: Logan Square and Pilsen.

Logan Square is a well-known neighborhood in Chicago, and also a neighborhood label where a large number (n=2495) of listings claim to be. We can see in Figure 2.4 that the City of Chicago boundaries for Logan Square contain primarily postings claiming to be in the neighborhood (orange points) without many posts claiming to be elsewhere (purple). The Zillow boundaries do encompass Logan Square claims that the Chicago boundaries do not, but the areas excluded by Chicago also have a higher concentration of claims of being in other neighborhoods. In addition, there are a number of listings claiming to be in Logan Square that are outside both formal boundaries–these posts would certainly be part of the 'peripheral' posts we defined earlier. This kind of posting merits further exploration to better understand what is happening in listings that claim to be part of a neighborhood when that might be more likely to be contested.

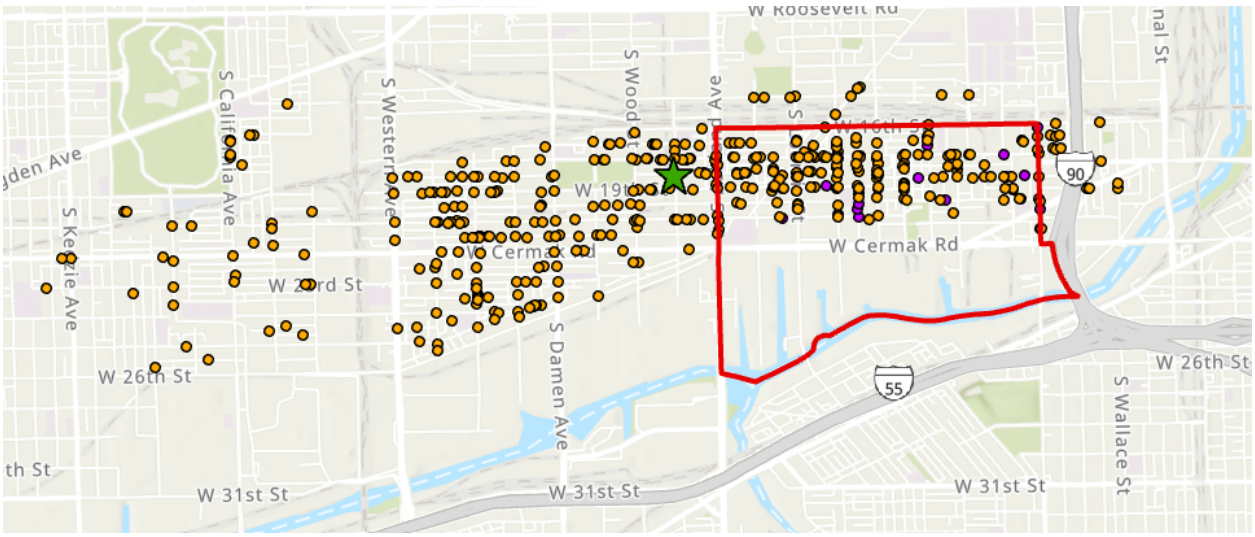


Figure 2.3: Contested boundaries for Pilsen neighborhood according to Zillow's definition (red). Orange points depict listings claiming to be in Pilsen, purple points are listings claiming to be elsewhere). The green star depicts the Pilsen neighborhood centroid. This is an example of what we call *border stretching*, demonstrating how static boundary systems may not capture the neighborhood identities as experienced by the people living in them.

Although the blue City of Chicago boundaries appear to be a better approximation for Logan Square in Figure 2.4, this is not the case for every neighborhood. Pilsen, shown in Figure 2.3, is another well-known neighborhood within Chicago, but it is not included as a distinct neighborhood by the City of Chicago. However, even though Zillow does include Pilsen as an area, it also does not seem to define it in the same way as Craigslist advertisements. Many of the posts claiming to be in Pilsen appear in a clustered way outside of the Zillow boundaries, and even the centroid of the Craigslist-defined neighborhood is outside the Zillow bounds. This could represent a developing neighborhood identity, which may not have been as strong at the time of the map creation, and supports the need for a more dynamic model of neighborhoods.

2.5 Content Analysis of Rental Listing Text

We use the `tomotopy` Python package to identify latent topics in the content posted in Craigslist rental advertisements. `tomotopy` uses Gibbs-sampling and is based on the LDA

approach described in Blei et al. (2003) and Newman et al. (2009). We prepared the text corpus by combining listing titles and body text from the Craigslist dataset. Standard preprocessing was applied using NLTK – lowercasing, removing alphanumeric characters, tokenization, custom stopword filtering, and lemmatization. We remove common real estate jargon that would otherwise dominate the topic distributions without providing meaningful differentiation between the content in different listing types. For the LDA implementation we selected hyperparameters that allow for moderate document sparsity ($\alpha = 0.1$) and greater topic-word concentration ($\eta = 0.01$). We trained for 100 iterations and tried $k = 5, 7$ and 10 topics which returned coherence scores of `0.7344`, `0.7425` and `0.7509`, respectively. After manual review we decided to focus on the $k = 7$ results for our remaining analysis as these had clearer separation than the $k = 5$ topics and are more interpretable than the $k = 10$ results.

2.5.1 Topic Interpretations

We identify seven distinct topics in the Craigslist advertisements, shown in Table 2.3. These topics reveal distinct patterns in how rental listings are framed, which have important implications for understanding the conception of neighborhood reputation and representation in the online rental market.

Topic 1 focuses on furnished short-term rentals, highlighting amenities and comfort with words like "furnished," "month," and "stay," suggesting a market segment catering to temporary residents seeking turnkey living situations. Topic 2 centers on rental requirements and financial considerations, with terms like "fee," "credit," "deposit," and "application," reflecting the administrative and financial aspects of renting. Topics 3 and 4 both relate to property search, with Topic 4 specifically including Spanish-language terms like "alquilar" and "propiedades," indicating efforts to reach Spanish-speaking audiences in Chicago's rental market. Topic 5, the most prevalent across the corpus at approximately 35% of document content, focuses on apartment features and amenities such as "appliance," "stainless," "modern," and "dishwasher,"

Topics	Common Words
1. Furnished Short-Term *3.6% of corpus*	lease, home, month, furnished, mo, amenity, community, view, apartment, offer, blueground, access, neighborhood, enjoy, stay, loop
2. Rental Terms *25.9% of corpus*	fee, lease, included, tenant, credit, street, pay, deposit, application, large, move, heat, per, gas, security, pet, utility
3. Property Search *1.1% of corpus*	apartment, rental, property, place, hill, cheap, grove, find, apt, height, center, agency, search, local, credit
4. Spanish Language *2.1% of corpus*	apartment, rental, property, place, hill, cheap, alquilar, propiedades, buscar, alquileres, height, search, agency, google
5. Amenities *35.0% of corpus*	floor, central, new, appliance, dishwasher, heat, large, space, air, stainless, feature, modern, steel, living, cat, closet
6. Listing Conditions *13.3% of corpus*	subject, unit, change, price, center, property, onsite, hour, availability, special, dog, pricing, studio, fitness, housing, amenity
7. Contact Information *18.9% of corpus*	contact, info, show, click, feature, view, call, id, renovated, included, closet

Table 2.3: Topic interpretations based LDA topic modeling on Chicago Craigslist rental listings.

underscoring the prevalence of interior quality in marketing rental properties. Topic 6 addresses listing conditions and availability, featuring terms related to pricing, special offers, and property policies. Finally, Topic 7 concentrates on contact information and viewing arrangements, with words like "contact," "show," "click," and "schedule," facilitating the connection between potential renters and property managers. The distribution of these topics across advertisements reveals how Chicago's rental market is presented online, with physical features and financial terms dominating the discourse.

2.5.2 Regression Analysis

We estimate the relationship between listing characteristics and proximity to the social center of an associated neighborhood using OLS regression with relative distance to neighborhood center as our primary dependent variable. Our model includes unit characteristics (bedrooms, bathrooms, rent, and square footage) and the topic proportions identified in our LDA analysis. For a full table of regression outputs see Table 2.5 in the Appendix.

This analysis reveals several patterns in spatial representations. Advertisements with more

bedrooms/bathrooms are associated with being further from neighborhood centers, while higher-priced listings are closer to their respective social centers. Most notably, listings with higher proportions of Topic 3 (Property Search) content exhibit increased relative distance from the center of the neighborhood (+0.20, $p<0.001$). Conversely, listings emphasizing apartment amenities (Topic 5) are associated with being closer to the centroid (-0.03, $p<0.01$). These findings indicate that misrepresentation is not random but correlates with specific marketing approaches in rental listings.

2.6 Results

Our analysis of over 30,000 unique Chicago rental listings reveals that neighborhood boundaries are actively renegotiated through strategic linguistic claims. By establishing a "social center" for neighborhoods based on the density of textual claims rather than rigid municipal boundaries, we demonstrate how agents navigate the tradeoff between geographic reality and reputational leverage. The following sections detail our findings in relation to our primary research questions.

2.6.1 Labeling Viability (RQ1)

We find that Large Language Models are highly effective for identifying neighborhood claims within unstructured text, outperforming traditional string-matching techniques. While the accuracy gain is incremental (85% for GPT-4.1 mini vs. 79% for string-matching), the LLM approach can scale beyond Chicago to other urban contexts without requiring reconfiguration or the same level of local real estate knowledge. Our labeling system prioritizes precision by selecting the single strongest neighborhood claim, addressing the inherent ambiguity found in listings that mention multiple areas.

2.6.2 Defining Social Centers (RQ2)

Our geospatial analysis reveals that the "social center" of a neighborhood—defined as the centroid of all property listings claiming that identity—often aligns with local landmarks, such as the eponymous park in Humboldt Park. However, these Craigslist-defined centers frequently diverge from institutional boundaries. We identify significant variation in representation: neighborhoods like Lake View are heavily overrepresented in claims relative to their Zillow-defined footprints, while areas such as Englewood and North Austin are significantly underrepresented, suggesting a lack of strong neighborhood identity or the presence of territorial stigma in the rental market.

2.6.3 Linguistic Substitution (RQ3)

Our analysis characterizes spatial misrepresentation not as random market noise, but as a predictable distortion of urban space. We identify three distinct patterns of spatial claim discrepancies: (1) *Conflicting Conceptions*, where stakeholders disagree on boundaries (e.g. Humboldt Park in Figure 2.1); (2) *Border Stretching*, where listings claim adjacent, plausible neighborhoods (e.g. Pilsen in Figure 2.3); and (3) *Reputation Laundering*, where properties or peripheral areas claim association with distant, desirable neighborhoods.

To identify these patterns systematically, we operationalize "peripheral claims" as those located in the furthest 20% from a neighborhood's social centroid. By comparing these LLM-labeled claims against institutional Zillow boundaries, we find evidence of systematic over and under-representation. Specifically, Lake View is significantly overrepresented–claimed more frequently than geographic distributions would predict–while neighborhoods such as Englewood, North Austin, and Hanson Park are significantly underrepresented. This pattern is consistent with reputation laundering: agents appear to distance properties from stigmatized neighborhood names while strategically claiming higher-status alternatives.

This substitution is statistically observable through a compensatory language pattern. Regression results show that as properties move further from the neighborhood social center,

listing agents substitute symbolic identity for physical proximity. For instance, generic property search language (Topic 3) is positively associated with relative distance from centroid ($+0.20, p < 0.001$), while specific interior amenity language (Topic 5) is negatively associated ($-0.03, p < 0.01$). Further, compared to central listings, the language used in peripheral listings shifts from location specific, non-portable amenities (Topics 1,5) toward more abstract and generic property-search language (Topics 2,3,4,6,7), a pattern illustrated in Figure 2.1.

2.7 Implications Beyond Sociology

The use of LLMs has exploded in recent years, and they can be seen by the general public as a simple, reliable solution to many routine problems. However, it remains an open question how powerful they may be in interdisciplinary research (Ziems et al., 2024). In order to better understand the impact of NLP on a broader scale and help address a specific question in the field of Urban Sociology, we demonstrate the effectiveness of LLMs at a notoriously difficult task: identifying where rental listings claim to be located on a large scale, in order to inform our understanding of processes of social construction of urban space.

However, although the impact of language models on this task may be transformative in the ability to quickly expand the scope of analysis, the quantifiable improvement on simple algorithms designed by experts is perhaps more incremental. In addition, the LLM labeling was certainly not perfectly accurate, suggesting that there is still room for improvement in large language models that might not be captured by tests and benchmarks developed solely within the field of NLP, and that interdisciplinary collaboration could lead to improvements both in NLP methodology and in making research questions and analyses in a broad range of fields more tractable and scalable.

2.8 Limitations

Our analysis of Craigslist rental listings provides valuable insights into neighborhood claims and social construction, but several important limitations should be acknowledged. The process of collecting data from Craigslist presents inherent challenges regarding post uniqueness and identification. Despite our deduplication efforts, the platform's structure makes it difficult to definitively determine which posts represent truly unique listings versus slightly modified versions of the same property.

While our dataset offers substantial coverage across Chicago neighborhoods, it cannot claim to be fully representative of the entire rental market. Craigslist represents just one segment of available rental properties, potentially skewing toward certain price points or property types. Additionally, our data may over represent certain management companies and realtors who post frequently on the platform, as opposed to "mom and pop" owners. These high-volume posters are more likely to use standardized language across multiple listings, which may introduce uniformity in how neighborhoods are described that doesn't reflect broader market patterns.

A fundamental challenge in this research is the absence of an authoritative catalog of Chicago's neighborhoods. As we argue, such a catalog is conceptually impossible. For practical purposes, we relied on Zillow's neighborhood boundaries as our primary reference, but these designations do conflict with local understandings. For example, questions arise about whether "West Loop" constitutes its own neighborhood distinct from "West Loop Gate," or whether "East Rogers Park" should be considered separate from "Rogers Park." These ambiguities reflect the socially constructed nature of neighborhoods themselves. Also, many units along Lake Michigan have a view of the water, and therefore advertise "views of the lake" or "lake views" which can be impossible to distinguish from listings claiming to be a "lake view" without considering the corresponding latitude and longitude coordinates.

Furthermore, assigning a single neighborhood label to each listing proved challenging when advertisements contained multiple, sometimes competing neighborhood claims. Our

hierarchical labeling approach (prioritizing title, then body, then neighborhood field) necessarily simplifies what can be complex spatial positioning strategies employed by listing agents. A rental advertisement might strategically claim association with multiple neighborhoods simultaneously, a nuance our single-label framework cannot fully capture. For instance, a listing located on the border between Logan Square and Humboldt Park might leverage both neighborhood identities depending on the perceived audience and market conditions.

These limitations highlight the inherent complexity of studying socially constructed spatial boundaries through digital traces and suggest opportunities for future research employing more nuanced approaches to neighborhood identification and classification.

Ethics Statement

Deciding how to approach this analysis is a non-trivial decision and following the extensive work in sociology and economics is important. We spoke experts in both fields in scoping and executing this project.

We collect data from Craigslist which, in some cases, contains specific information about individual posters. Craigslist is a public forum whose housing section should not contain much information irrelevant to the housing ads themselves. We do not release these data publicly per the non-commercial use terms of service and ensure no PII appear in any of our writing, results or figures.

Another limitation when working with pretrained foundation models like GPT-4 is a lack of reproducibility, as we do not have access to the training data or the weights. In the interest of reproducibility, we keep annotation costs under $100.

We utilized multiple Generative AI tools (OpenAI's GPT-4/5 and Anthropic's Claude 4.5 Opus/Claude Code) in the production of this manuscript, in the following ways: 1) producing computer code for data cleaning and analysis and 2) iteratively improving the concision and clarity of the writing. We have carefully reviewed all aspects of the manuscript

for accuracy and coherence. All scientific insights, analysis and interpretation of data and scientific conclusions are made solely by the authors.

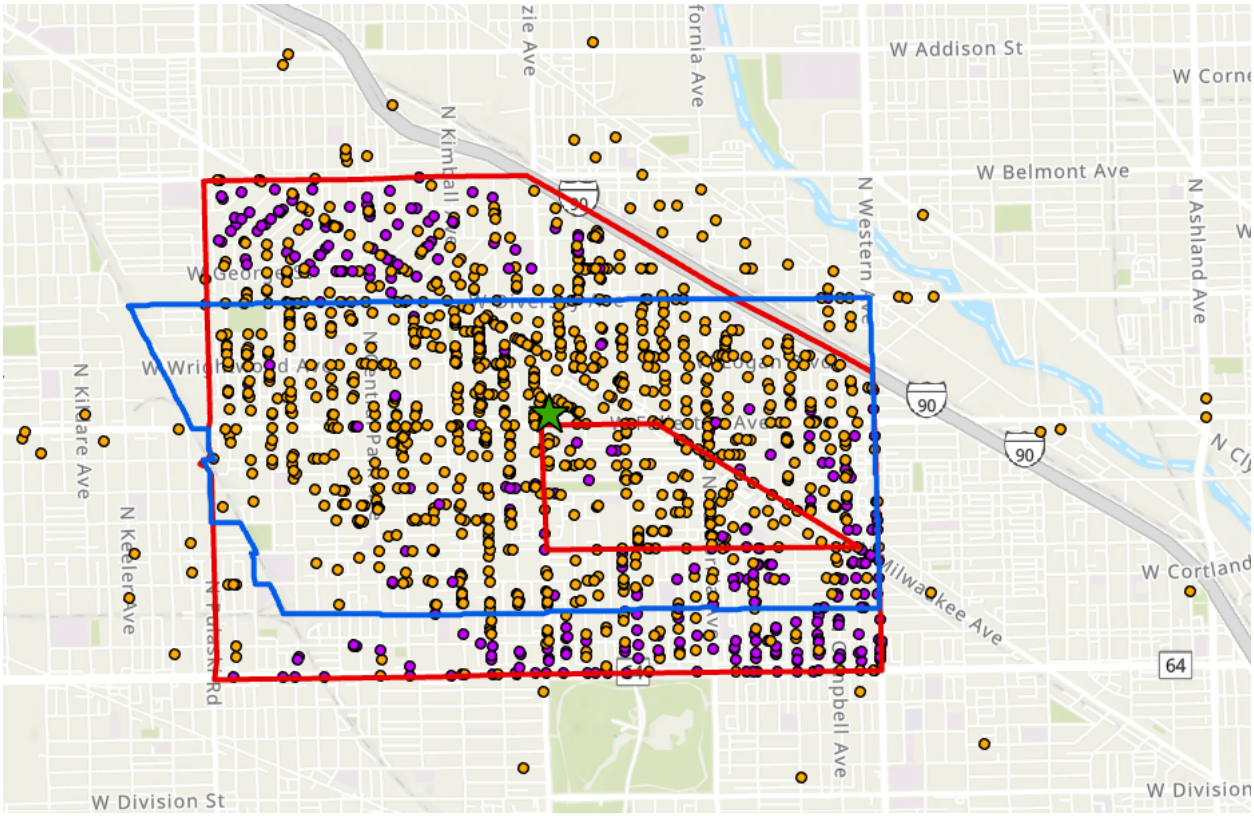


Figure 2.4: Contested boundaries for Logan Square neighborhood according to the official City of Chicago limits (blue), Zillow's definition (red), and claims in Craigslist rental listings (orange points for listings claiming to be in the neighborhood, purple if claiming to be elsewhere); the center of our Craigslist-defined neighborhood is represented by a green star. Logan Square represents a popular neighborhood with many postings and a number of nearby listings claiming to be within the neighborhood.

2.9 Supplementary Tables and Figures

2.9.1 Full Performance Metrics for Neighborhood Claim Validation

2.9.2 Topic Modeling Results

Working with more than 35,000 unique and often verbose advertisements, we fit a topic model on $k = 25$ topics to identify main themes in the rental listings. We do not provide interpretations of the topic, but show the common words in Table 2.6.

	String Match			GPT-4 Mini			
Neighborhood	**Prec.**	**Recall**	**F1**	**Prec.**	**Recall**	**F1**	**Support**
the loop	0.12	0.50	0.20	1.00	1.00	1.00	2
rogers park	1.00	1.00	1.00	0.75	1.00	0.86	3
lake view	0.63	0.86	0.73	0.94	0.73	0.82	22
ranch triangle	–	0.00	0.00	–	0.00	0.00	1
lincoln park	0.90	1.00	0.95	0.94	0.89	0.91	18
fulton river district	–	0.00	0.00	1.00	1.00	1.00	1
south loop	1.00	1.00	1.00	1.00	1.00	1.00	3
park west	–	0.00	0.00	–	0.00	0.00	1
west town	1.00	1.00	1.00	1.00	1.00	1.00	1
logan square	0.88	0.88	0.88	0.89	1.00	0.94	8
lake view east	–	0.00	0.00	0.90	0.90	0.90	10
university village - little italy	–	0.00	0.00	–	0.00	0.00	2
uptown	1.00	0.83	0.91	1.00	0.83	0.91	6
wicker park	1.00	1.00	1.00	1.00	1.00	1.00	4
portage park	1.00	1.00	1.00	1.00	1.00	1.00	2
old town	1.00	0.83	0.91	1.00	0.83	0.91	6
avondale	–	0.00	0.00	–	0.00	0.00	1
west loop gate	–	0.00	0.00	1.00	1.00	1.00	4
streeterville	1.00	1.00	1.00	1.00	1.00	1.00	2
ravenswood	1.00	0.80	0.89	1.00	0.80	0.89	5
wrigleyville	1.00	0.67	0.80	1.00	1.00	1.00	3
river north	0.83	0.71	0.77	0.86	0.86	0.86	7
buena park	1.00	1.00	1.00	0.75	1.00	0.86	3
bucktown	1.00	1.00	1.00	1.00	1.00	1.00	7
kenwood	–	0.00	0.00	1.00	1.00	1.00	1
old irving park	1.00	1.00	1.00	1.00	1.00	1.00	1
edgewater	1.00	1.00	1.00	1.00	1.00	1.00	5
pilsen	1.00	1.00	1.00	1.00	1.00	1.00	7
garfield ridge	1.00	1.00	1.00	1.00	1.00	1.00	1
humboldt park	1.00	1.00	1.00	1.00	1.00	1.00	1
andersonville	1.00	1.00	1.00	1.00	0.67	0.80	3
west rogers park	1.00	1.00	1.00	1.00	1.00	1.00	1
ukrainian village	1.00	1.00	1.00	1.00	1.00	1.00	1
gold coast	0.75	1.00	0.86	0.75	1.00	0.86	3
hyde park	1.00	1.00	1.00	1.00	1.00	1.00	1
roscoe village	1.00	1.00	1.00	0.33	1.00	0.50	2
east garfield park	1.00	1.00	1.00	1.00	1.00	1.00	1
albany park	1.00	1.00	1.00	1.00	1.00	1.00	1
lincoln square	0.33	1.00	0.50	0.00	0.00	0.00	1
south shore	1.00	1.00	1.00	1.00	1.00	1.00	1
hermosa	1.00	1.00	1.00	0.50	1.00	0.67	1
ravenswood manor	–	0.00	0.00	–	0.00	0.00	1
unknown	0.78	0.88	0.82	0.50	0.50	0.50	8
Average Metrics							**Total: 163**
Accuracy		0.79			0.85		
Macro Avg	0.76	0.74	0.62	0.78	0.81	0.70	
Weighted Avg	0.88	0.79	0.77	0.91	0.85	0.85	

Table 2.4: Detailed performance metrics by neighborhood for string matching and GPT-4 Mini classification methods. The Support column indicates the number of test samples for

Variable	Coef.	Std. Err.	t	P>|t|
Intercept	0.10	0.01	10.40	0.00
Bedrooms	0.02	0.00	8.83	0.00
Bathrooms	0.02	0.00	6.25	0.00
Rent	0.00	0.00	-14.56	0.00
Square Footage	0.00	0.00	2.17	0.03
Topic 2 Rental	0.03	0.01	3.40	0.00
Topic 3 Property	0.20	0.01	14.99	0.00
Topic 4 Spanish	0.10	0.01	9.42	0.00
Topic 5 Amenities	-0.03	0.01	-2.98	0.00
Topic 6 Condition	0.02	0.01	1.86	0.06
Topic 7 Contact	0.05	0.01	5.56	0.00

Table 2.5: Regression results demonstrating how relative distance from the center of the neighborhood associates with rental unit characteristics.

Table: Topics for $k = 25$

Topic	Common Words
Topic 1	hyde, property, si, terrace, la, estos, con, mac, elli, street
Topic 2	large, space, floor, living, dining, closet, bedroom, heat, storage, central
Topic 3	contact, call, schedule, photo, tour, please, unit, info, show, actual
Topic 4	space, walk, home, lease, great, living, street, one, loft, large
Topic 5	real, estate, star, text, tour, community, view, lounge, apartment, finish
Topic 6	lease, mo, blueground, month, furnished, amenity, stay, offer, view, access
Topic 7	cat, feature, floor, call, dishwasher, one, fee, lakeview, n, dog
Topic 8	modern, heat, property, central, bus, appliance, call, gas, cat, air
Topic 9	apartment, rental, lake, property, place, hill, alquilar, cheap, new, grove
Topic 10	unit, special, availability, subject, dog, weight, onsite, please, price, various
Topic 11	view, center, lake, amenity, free, onsite, community, michigan, call, window
Topic 12	fee, credit, deposit, tenant, new, security, included, move, pay, pet
Topic 13	hyde, apartment, regent, horas, onsite, market, la, est, e, museum
Topic 14	apartment, rental, lake, property, place, hill, cheap, find, new, grove
Topic 15	studio, property, bjb, internet, complimentary, e, change, picture, subject, fitness
Topic 16	fee, mile, white, home, tile, neighborhood, make, company, new, tenant
Topic 17	housing, water, included, heat, opportunity, landstar, equal, group, act, feature
Topic 18	contact, show, info, price, availability, email, renovated, included, click, web
Topic 19	lake, bus, block, cta, walk, red, minute, away, stop, distance
Topic 20	info, contact, show, click, il, id, feature, friendly, text, n
Topic 21	amenity, view, center, pool, fitness, luxury, outdoor, lounge, garage, window
Topic 22	logan, banker, coldwell, square, blue, real, estate, please, opportunity, equal
Topic 23	hyde, village, height, center, drexel, grand, nuestros, river, maintenance, m
Topic 24	fee, per, cat, lease, dog, application, deposit, one, heat, gas
Topic 25	new, appliance, stainless, steel, floor, central, large, feature, granite, dishwasher

Table 2.6: Topic clusters from LDA topic modeling on Chicago Craigslist rental listings, $k = 25$.

Chapter 3

The "Obesity Epidemic" Is Gendered: Decomposing US Prevalence Trends Since 2000

Abstract. Concern about the "obesity epidemic" in the United States is anchored by a consistent upward trend in Body Mass Index (BMI) over time. Yet it has not been empirically supported that population-level increases in BMI reflect actual increases in excess fat tissue (adiposity). Using direct standardization and machine learning-assisted multiple imputation on National Health and Nutrition Examination Survey (NHANES) data from 1999 to 2018, this study decomposes U.S. obesity prevalence trends across three obesity classification metrics: BMI, waist circumference (WC), and dual-energy X-ray absorptiometry (DXA)—the clinical standard for measuring total body fat. I find that while BMI- and WC-based obesity prevalence rose by 12.8 and 14.3 percentage points, DXA-based obesity showed no statistically significant change over the same period (+0.9 %, $p = 0.7$). Demographic compositional shifts account for less than 3% of the BMI trend, ruling out population aging and racial/ethnic diversification as primary contributors. However, the aggregate stability in DXA masks a profound, structurally gendered divergence: DXA-based obesity rose by nearly 10 percentage points among men, but fell by 7 percentage points among women, even as BMI indicated rapidly rising obesity for both. These findings challenge the foundational narratives linking rising population fatness to adverse health outcomes, demanding a critical reassessment of the metrics used to construct and surveil public health crises.

Keywords: Obesity, Health Disparities, Gender, Measurement, Machine Learning

3.1 Introduction

The "obesity epidemic" emerged as one of the defining public health narratives of the early twenty-first century in the United States. Following the National Institutes of Health (NIH) decision to standardize the classification of obesity at a Body Mass Index (BMI) ≥ 30, the the documented prevalence of obesity among U.S. adults surged from approximately 30 to 43 percent (Ogden et al., 2013; Hales et al., 2020; Emmerich et al., 2024). This trend commanded urgent attention not only because of the rate of increase, but because it coincided with rising rates of type 2 diabetes, metabolic syndrome, cardiovascular disease, and numerous other morbidities (Moore et al., 2017; Roth et al., 2020; Saeedi et al., 2019). For many of these morbidities, the physiological mechanisms linking excess adiposity to adverse outcomes – such as insulin resistance, systemic inflammation, and hypertension – were well-established (Kahn et al., 2006; Poirier et al., 2006). Consequently, the co-occurrence of rising obesity with worsening health outcomes endowed the "obesity epidemic" with immense scientific and social force. Framed and understood as a public health crisis (Saguy, 2013), it reshaped clinical guidelines, restructured public health priorities, and generated a decades-long research apparatus aimed at reversing the observed rise in BMI.

That scientific force rests, however, on a largely unexamined assumption: that BMI tracks actual adiposity well enough at the population level to capture how body fat has changed across demographic groups. In practice, this assumption underwrites the widespread interpretation that the "obesity epidemic" reflects a shared increase in body fat across time, among both men and women, within racial/ethnic groups, etc. Yet, this remains an empirical claim. BMI systematically misclassifies individuals by conflating fat mass with lean mass, and its relationship to body fat composition varies substantially by age, sex, and race/ethnicity (Romero-Corral et al., 2008). While a robust literature documents these limitations at

the individual level and proposes alternative indices (Stanford et al., 2019; Jeong et al., 2023; Wang et al., 2024), whether they produce different prevalence trends over time at the population level has received far less scrutiny. If BMI is a reliable epidemiological signal, population trends derived from it must track trends derived from direct measures of adiposity. If they diverge, the data-driven basis of the obesity epidemic narrative is even more tenuous than previously documented.

In this study, I test that assumption directly using data from the National Health and Nutrition Examination Survey (NHANES). I compare obesity prevalence trends derived from BMI and waist circumference (WC) – the two most widely used anthropometric proxies – against trends derived from dual-energy X-ray absorptiometry (DXA) scans of total body fat, the clinical reference standard for measuring adiposity. Because DXA collection in NHANES has been intermittent since 1999, I multiply impute body fat percentage to recover a consistent series spanning the full study period.

The central finding is stark: BMI and WC followed virtually identical trajectories, rising by 12–14 percentage points over 1999–2018. While these two measures were offset in absolute level, their slopes were statistically indistinguishable, demonstrating a shared trend that was entirely absent in the DXA data. Indeed, DXA-based obesity rose by only 0.9 percentage points over this same period – a change that was not statistically significant ($p = 0.700$). While the two anthropometric measures tracked each other with high fidelity, neither reflected aggregate adiposity as measured by DXA.

A natural concern is that this divergence is a demographic artifact rather than a genuine measurement discrepancy. As the U.S. population aged and became more racially and ethnically diverse, the share of individuals drawn from groups for whom BMI systematically mis-estimates adiposity shifted (Stanford et al., 2019; Perez and Hirschman, 2009) – potentially changing aggregate BMI-based prevalence mechanically, without any change in underlying body fat. To assess this, I decompose the BMI and DXA trends using direct standardization, constructing counterfactual prevalence series that hold demographic composition fixed at

1999 levels while allowing within-group rates to evolve, and vice versa. This analysis finds that demographic compositional change explains less than 3 percent of the cumulative rise in BMI-based obesity over the study period. The divergence between BMI and DXA trends is therefore not an artifact of population aging or diversification; it reflects a genuine difference in what the measures capture.

The implication is that the biological claim embedded in the obesity epidemic narrative – that the U.S. population carried substantially more body fat in 2018 than in 1999, and became less healthy as a result – is considerably weaker than the BMI time series implies. This does not mean cardiometabolic outcomes improved, or that the health conditions rising alongside BMI were imaginary. Those observed outcomes are not contested by the analyses presented here. But if the most direct available measure of adiposity shows no statistically compelling upward trend over the same period, the causal chain connecting population-level fatness to population-level morbidity requires more scrutiny than it has received to date. This has important implications for social policy as interventions specifically targeting body weight may not be the thing we should be prioritizing.

3.2 Background

3.2.1 Defining and Measuring Obesity

It is important to delineate the concept of “obesity" from its many measures. BMI, the most common measure (calculated as weight in kilograms divided by height in meters), is often fallaciously conflated with “obesity,” the concept, which the World Health Organization (WHO) formally defines as “(1) excessive adipose tissue accumulation that (2) presents a risk to health” (World Health Organization, 2023). However conceptually intuitive, some scrutiny reveals that this definition is still under-specified along several axes. Excessive fat relative to which baseline? How much of an increased health risk, and for which health outcomes? BMI ≥ 30 is the dominant metric for measuring obesity even though a ratio of weight to

height measures neither adiposity nor health directly. And we do have different ways to measure adiposity. There is directly measured adiposity with Dual X-ray Absorptiometry (DXA) scans, which is the most reliable measure of fat we have. Depending upon the machine, there is a small amount of irreducible measurement error associated with any individual scan. There is also *imputed* DXA, which tries to capture the same thing conceptually, but is derived from more readily observable correlates. Imputed DXA, like any predicted outcome, contains some amount of prediction error. These measures of adiposity are different conceptually from anthropometric measures like BMI or waist circumference which are targeting different, related constructs, each with their own measurement and prediction error.

Direct or imputed measures of adiposity and anthropometric body measures are also distinct from health. The notion that BMI relates in some systematic way to population health was made when it was repurposed as an epidemiological tool by Keys et al. (1972). The BMI was then further institutionalized in U.S. clinical practice following the NIH's 1998 revision of BMI cutoffs to align with the WHO's global standard. This move was reported at the time by an infamous CNN article that opened "Millions of Americans became "fat" Wednesday – even if they didn't gain a pound." (Cohen and McDermott, 1998).

Despite being a bit tongue in cheek, this characterization was exactly right, because rather than capturing body composition, the BMI is simply a measure of body proportion. It cannot distinguish between fat mass, lean mass, muscle mass, or bone density; a crucial distinction for a definition of obesity that focuses on fat tissue. Furthermore, past work has found that body proportion, as measured by BMI, varies considerably across demographic groups (Romero-Corral et al., 2008; Jeong et al., 2023). Competing measurement systems have been proposed to operationalize the underlying construct as it relates to health more directly. Waist circumference and waist-to-hip ratio target abdominal adiposity specifically, motivated by evidence that central fat deposition is more strongly associated with cardiometabolic risk than peripheral fat (Després, 2006; Fourman et al., 2025). Dual-Energy X-ray Absorptiometry (DXA) scans, the current clinical reference standard for measuring adiposity, distinguishes

regional body fat percentage through differential tissue attenuation of low-dose X-rays, providing a far more direct index of total body fat percentage than any anthropometric proxy can (Laskey, 1996; Wells and Fewtrell, 2006; Shepherd et al., 2017; Borga et al., 2018). DXA-based obesity thresholds are anchored more closely to the physiological construct the WHO definition invokes, though the provenance and accuracy of these cutoffs is itself contested (Ho-Pham et al., 2011; Potter et al., 2025). The coexistence of these measurement systems reflects a foundational ambiguity: obesity as a biological construct – excess fat tissue in individuals – and obesity as a public health phenomenon – a risk-stratifying classification applied to populations – have always been loosely coupled, held together by the practical tractability of BMI rather than its construct validity alone. This tension has recently prompted formal reconsideration of how obesity should be defined clinically.

Because BMI-based measures of obesity can both underestimate and overestimate adiposity and provide inadequate information about health at the individual level, the Lancet released a Diabetes & Endocrinology Commission on Clinical Obesity in 2025 where they proposed a revised framework that integrates BMI with anthropometric measures or direct measures of adiposity and distinguishes *clinical* obesity – characterized by organ dysfunction or physical limitation – from *preclinical* obesity, where excess adiposity is present without current impairment (Rubino et al., 2025). Applying this new definition to the All of Us cohort (2017–2023), **?** find that obesity prevalence increases by 60 percent relative to traditional BMI-based obesity classification. That a redefinition of the same underlying concept produces a 60 percent shift in estimated prevalence is not merely a measurement curiosity: it illustrates that what we collectively “know" about the scope and distribution of this “obesity epidemic" is not a stable empirical fact but a product of the classificatory choices through which the phenomenon is made observable.

3.2.2 Social Disparities, Adiposity, and Health

The scientific case for treating excess adiposity as a health risk is more contingent than public health discourse typically acknowledges. Adipose tissue, particularly visceral fat deposited around the abdominal organs, contributes to insulin resistance, systemic inflammation, and hypertension through well-characterized molecular pathways (Kahn et al., 2006; Poirier et al., 2006). Elevated adiposity has also been associated with increased risk of type 2 diabetes, cardiovascular disease, certain cancers, and all-cause mortality across a wide range of populations (Roth et al., 2020). But not-being-fat is far from a panacea for health. For example, there are individuals classified as obese by BMI but who exhibit healthy cardiometabolic profiles (Wildman et al., 2008); a fact articulated clearly by some fat activists such as through the slogan "health at every size" (HAES) (Rothblum, 1992; Gaesser, 1996; Bacon et al., 2005; Burgard, 2009; Penney and Kirk, 2015). There are also, conversely, individuals within the "normal" BMI range who harbor metabolically harmful levels of visceral fat – a condition sometimes called Metabolically Obese Normal Weight (MONW), Normal Weight Obesity (NWO), or colloquially referred to as "skinny-fat" (Oliveros et al., 2014).

These examples illustrate a deeper problem with treating BMI as a proxy for health risk: the relationship between body weight and adverse outcomes is complicated by factors that can operate independently of adiposity itself. On the biological side, risk depends not on total fat mass but on its distribution and metabolic context – visceral fat carries different implications than subcutaneous fat, and individuals with identical BMIs can have dramatically different cardiometabolic profiles (**?**Oliveros et al., 2014). But the pathway from a high BMI classification to poor health outcomes is not exclusively biological. Being categorized as obese by a clinical or administrative system can itself generate harm. Ernsberger and Haskew (1987); Wooley and Wooley (1981); Polivy and Herman (1985) have shown that the medicalization of fatness, repeated cycles of prescribed dieting, and participation in weight-loss interventions have been linked to disordered eating, depression, and elevated anxiety. Weight stigma in clinical encounters can deter care-seeking and reduce the quality of

treatment received, independently of any physiological effect of excess fat (Puhl and Heuer, 2010; Tomiyama et al., 2018). BMI thus functions less as a clean biological signal than reinforcing socially consequential classifications that activates institutional responses whose health effects are entangled with the underlying adiposity it imperfectly measures. The heterogeneity of outcomes within any BMI category reflects not only the metric's imprecision as a correlate of body fat, but the complexity of the causal pathways, biological and social, through which body size manifests in observable health outcomes.

3.2.3 Obesity as a Population Health Problem

In the late twentieth century we witnessed a dramatic rise in BMI-defined obesity in the United States, from roughly 15 percent in the late 1970's to the levels documented above, prompting the CDC and WHO to frame the trend as a "global epidemic" by the early 2000s (Flegal et al., 1998; Mokdad et al., 2003). This framing commanded unusual cross-disciplinary attention; clinical medicine, public health, nutrition science, demography and sociology converged on rising BMI as a primary driver of preventable morbidity and mortality (Allison et al., 1999; Ferraro and Kelley-Moore, 2003; Scharoun-Lee et al., 2011; Johnston and Lee, 2011). With this emerging consensus came an urgency to reshape clinical guidelines, accelerate pharmaceutical investment in weight-loss interventions, motivate lifestyle and behavior changes, and motivated efforts to redesign the built environment to reduce obesogenic exposures (Wing et al., 2001; Hill et al., 2003; Sallis and Glanz, 2006; Bray and Ryan, 2021).

Within the Demography literature on obesity, a core contribution has been the decomposition of persistent racial and socioeconomic disparities in BMI-defined obesity. Johnston and Lee (2011) decompose the substantial black-white gap in female obesity into contributions from energy intake, energy expenditure, and unobserved factors, finding that overconsumption rather than inactivity is the primary proximal driver – and demonstrating in the process that the choice between BMI and waist-to-height ratio as the adiposity outcome materially affects

estimates of the gap's magnitude. Scharoun-Lee et al. (2011) use latent class analysis to show that intergenerational socioeconomic profiles predict young adult obesity in ways that conventional social mobility measures obscure, with effects that differ substantially by race and gender. Beyond stratification, demographers have also examined how contextual and institutional factors shape obesity risk: Baker et al. (2015) find that children of immigrant mothers are more likely to be have an "obese" BMI than children of U.S.-born mothers – contrary to the immigrant epidemiological paradox – with maternal linguistic isolation and pre-pregnancy weight status as key mediating factors. Barlow (2021) exploits a natural experiment in Nigeria to show that the protective effect of schooling on overweight and obesity documented in high-income countries does not replicate in a low- or middle-income context, underscoring the degree to which associations between social position and BMI-defined obesity are calibrated by contextual factors rather than reflecting universal biological regularities.

A parallel, more critical literature challenged whether the "epidemic" was as empirically grounded as official discourse implied. In response to limitations identified in Mokdad et al. (2003), Flegal et al. (2005) presented updated analyses demonstrating that the association between BMI and all-cause mortality was non-monotonic, with modest overweight associated with lower mortality than "normal" weight in some demographics – findings that generated intense methodological controversy while revealing the fragility of mapping BMI to health risk (Flegal, 2021). Scholarship across critical fat studies, the sociology of health, and feminist theory has interrogated not just the empirical foundations of the obesity epidemic narrative but the social work that narrative performs. This literature frames rising BMI as a moral panic in which statistical associations between weight and health risk are amplified into a broader discourse of bodily deviance, individual failure, and public menace (Campos, 2004; Gard and Wright, 2005). The speed and consensus with which this framing took hold is itself sociologically remarkable: within a decade of the NIH's 1998 revision of BMI cutoffs, obesity had been declared a crisis, reclassified as a disease, and made the target of clinical, pharmaceutical, legislative, and public health intervention at a scale that arguably outran

the evidence base (Kirkland, 2011). Saguy (2013) shows that how fatness is framed – as a medical condition, a moral failing, or a social justice issue – is not determined by scientific evidence alone but by the institutional positions and cultural assumptions of the actors doing the framing, with the medical framing consistently winning out in ways that pathologize bodies and concentrate authority in clinical and public health institutions. That BMI is the instrument through which this pathologization is operationalized raises questions that go beyond measurement validity to the politics of how a metric becomes entrenched, whose bodies bear the costs of its systematic errors, and what institutional interests are served by its persistence.

Weight stigma research documented that anti-fat bias was pervasive in medical settings and everyday social life, and that stigma itself – rather than or in addition to adiposity – was a proximate driver of adverse health outcomes for people classified as obese (Puhl and Heuer, 2010; Tomiyama et al., 2018). Crucially, these critiques did not deny that excess adiposity could impair health; they challenged whether BMI-classified obesity was a coherent biological category, and whether the social costs of the epidemic framing, like amplified stigmatization, were adequately weighed against the purported benefits.

3.2.4 The Obesity Measurement Gap

These debates have unfolded largely without resolving a prior empirical question: whether population-level trends in BMI-defined obesity actually track population-level trends in adiposity-defined obesity. The biological claim underlying the epidemic narrative suggests that the epidemiological signal (rising BMI) and the construct it represents (obesity) follow a similar trajectory. Individual-level evidence that BMI misclassifies adiposity is well established (Romero-Corral et al., 2008; Stanford et al., 2019; Jeong et al., 2023), but measurement error at the individual level does not necessarily aggregate into valid measurements at the population level (Salerno et al., 2025). Moreover, how population trends compare for different measures of obesity is a distinct empirical question, and one whose answer depends on whether

the systematic errors in BMI, as a proxy for adiposity, have changed in magnitude over time and whether those changes are distributed across demographic groups in ways that compound rather than cancel.

The stakes of this question extend directly into the demographic literature described above. Decompositions of racial and socioeconomic disparities in obesity, analyses of contextual and institutional drivers, and studies of cumulative disadvantage over the life course treat BMI-based obesity as a reliable proxy for the underlying *health* construct. If the BMI-DXA obesity gap has widened at different rates across race, sex, or age groups over time, then apparent changes in group disparities may partly reflect shifting measurement error rather than genuine convergence or divergence in body fat, and ultimately, health outcomes.

Recent work has begun to address the gap between DXA- and BMI-based measures of obesity. Aryee et al. (2025) restrict comparisons of BMI and DXA to the 2011-2017 NHANES cycles for adults aged 20-59, finding substantially different prevalence levels across measures, but stopping short of a formal trend analysis including all adults (age 60+) or spanning the full period over which the obesity epidemic narrative was constructed (1999-2017). The intermittent availability of DXA in NHANES also raises methodological challenges: Salerno et al. (2025) show that substituting BMI for directly measured adiposity does not preserve valid obesity inference even when predictive accuracy is high, because prediction error introduces both bias and variance that standard inferential tools fail to propagate. No prior study has decomposed the sources of the BMI-DXA divergence into compositional and within-stratum rate components, or examined how the BMI-DXA gap has itself changed over time and across demographic subgroups. If the gap between anthropometric measures and direct measures of adiposity has widened systematically over the same period that BMI-based obesity prevalence rose most dramatically, the evidentiary basis of the epidemic framing requires reassessment not merely on conceptual grounds, but empirical ones, too. Obesity, as operationalized by BMI, may be better understood as a heterogeneously valid proxy whose aggregate trends can diverge substantially from the underlying biological construct they are

intended to measure – a possibility with direct implications for the research apparatus and policy priorities built upon it.

3.3 Data and Methods

3.3.1 Data

The National Health and Nutrition Examination Survey (NHANES) is a continuous cross-sectional study conducted by the National Center for Health Statistics (NCHS). NHANES employs a stratified, multistage probability design to produce a representative sample of the civilian non-institutionalized U.S. population. Data collection involves an in-home interview (demographics and health behaviors) followed by a physical examination in a Mobile Examination Center (MEC). This analysis uses NHANES data from the 1999–2000 cycle through the 2017–2018 cycle, the last complete pre-pandemic survey cycle. Data collection in the subsequent 2019–2020 cycle was interrupted in March 2020 by the COVID-19 pandemic.

Standard anthropometric measures – including height, weight, and waist circumference – are collected for all MEC participants in every two-year cycle. However, the collection of whole-body Dual-Energy X-ray Absorptiometry (DXA) has been intermittent across the study period. Figure 3.1 shows the data sources used for the subsequent analyses which are described below:

- **1999–2006:** DXA scans were administered to participants aged 8 and older. Because missingness was non-random (correlated with age, race, and high BMI due to equipment weight limits), NCHS provides five multiply imputed datasets which used Sequential Regression Multivariate Imputation (SRMI) to address these biases (NCHS, 2008). These data cover the entire adult age distribution, 20+.

- **2007–2010:** Whole-body DXA scans were not performed during these cycles. I use an

ML-based imputation model trained on variables collected in the demographic survey from the observed cycles to predict body fat percentage for this period (see Section 4.3).

- **2011–2018:** DXA collection resumed with upgraded equipment (Hologic Discovery A) that accommodated higher body weights. However, scans were restricted to adults aged 20 to 59. I augment these data with imputations for adults age 60^{+} using the same ML-based imputation model.

To decompose obesity trends into compositional and rate components, I supplement NHANES with CDC WONDER annual population estimates for 1999–2017 to obtain population distributions by age, sex, and race/ethnicity. Age is recorded or grouped from vital records and aligned with Census population counts. For sex, binary biological classification is derived from administrative records. Race/ethnicity is taken from standardized categories combining Census self-identification and administrative reporting. Additional information about data collection in CDC Wonder is available on the CDC website here. These data provide the marginal population shares necessary for the standardization and counterfactual analyses.

The analysis proceeds in four stages. First, I impute missing BF% values for the full age distribution of adult MEC respondents in 2007-2010 and for adults age 60^{+} in 2011-2018. Second, I apply survey weights to estimate crude obesity prevalence (based separately on different measures) for the US adult population in each survey cycle from 1999–2018. Third, I decompose the gap between BMI- and DXA-based obesity prevalence estimates using direct standardization to U.S. Census population distributions, constructing counterfactual prevalence scenarios that isolate the role of demographic composition. Fourth, I estimate how each strata are differently contributing changes in the gap over time.

3.3.2 Predicting Missing Body Composition Data

DXA-based measures of body fat percentage are not consistently observed across NHANES survey cycles due to both item-level nonresponse and design-induced changes in age eligibility, requiring explicit adjustment to ensure comparability of adiposity estimates over time (see Section 4.2 for details regarding DXA data collection in NHANES). Although missing DXA data among eligible respondents under age 60 in later cycles were judged by NCHS to be approximately missing at random, the exclusion of older adults introduces structural missingness that may bias DXA-based obesity prevalence estimates.

To address this, I impute BF% for respondents lacking DXA measurements using a model trained exclusively on NHANES cycles with full adult age coverage. Specifically, I trained an XGBoost prediction model on the NCHS multiply imputed DXA-based BF% data from the 1999–2004 cycles, which include observations for adults aged 60 and older. The model conditions flexibly on age, sex, race/ethnicity, BMI, waist circumference, and additional anthropometric and health covariates, allowing the relationship between BMI and BF% to vary across the age distribution (see (Schenker et al., 2011) and Appendix Section 3.8.1 for more details about covariates and specifications used for imputation).

To account for heterogeneity in the precision of the NCHS imputed values across demographic subgroups, training observations were assigned weights inversely proportional to the squared standard error of the NCHS multiple imputation estimates within each sex × BMI category cell:

$$w_i \propto \frac{1}{\widehat{\mathrm{SE}}_{g(i)}^{\,2} + c}$$

where $\widehat{\mathrm{SE}}_{g(i)}$ is the published standard error of the NCHS mean body fat estimate for the sex × BMI category cell to which participant i belongs (Schenker et al., 2011), and c is a floor constant that prevents extreme weights in the most precisely measured cells. This scheme concentrates learning on demographic cells where the NCHS imputed values are most reliable,

reducing the influence of observations with high imputation uncertainty. Model performance was assessed via five-fold cross-validation within the 1999–2004 training data, yielding an overall training root mean squared error (RMSE) of 1.828 BF%. Additional details on model specification, tuning, and validation are provided in the Appendix, section 3.8.1.

BF% values from the NCHS multiple imputation (Schenker et al., 2011) are carried forward unchanged for the 1999–2006 cycles in the analytic dataset. The trained model is applied only to respondents for whom no DXA-based BF% is available: (i) all MEC respondents in survey cycles where DXA was not collected (2007 and 2009), and (ii) respondents aged 60 and older in cycles where DXA collection was restricted by design (2011–2018). To propagate imputation uncertainty I follow Centers for Disease Control and Prevention (CDC) (2024) and generate five predicted BF% values per respondent by adding independent draws from a zero-mean Gaussian distribution with standard deviation equal to the model's training residual SD, yielding five completed datasets that are analyzed separately and combined using Rubin's rules.

3.3.3 Estimating Crude Obesity Prevalence

All analyses use NHANES MEC weights, strata, and primary sampling units to account for the complex survey design. For respondents with multiply imputed DXA measures (1999–2006), survey weights are retained unchanged across imputations and combined using Rubin's rules. For survey years without DXA examinations (2007–2010, 2011–2018 age 60^{+}), multiply ML-predicted BF% values are treated as multiple imputations with weights carried unchanged across draws. Unlike official NCHS reports (Ogden et al., 2013; Hales et al., 2020; Emmerich et al., 2024), which age-standardize to the U.S. population in 2000, I produce crude estimates to capture how obesity prevalence evolves alongside the shifting demographic structure of the U.S. population.

Measure-specific crude obesity prevalence in each survey cycle t is estimated as the weighted proportion of binary obesity indicators I:

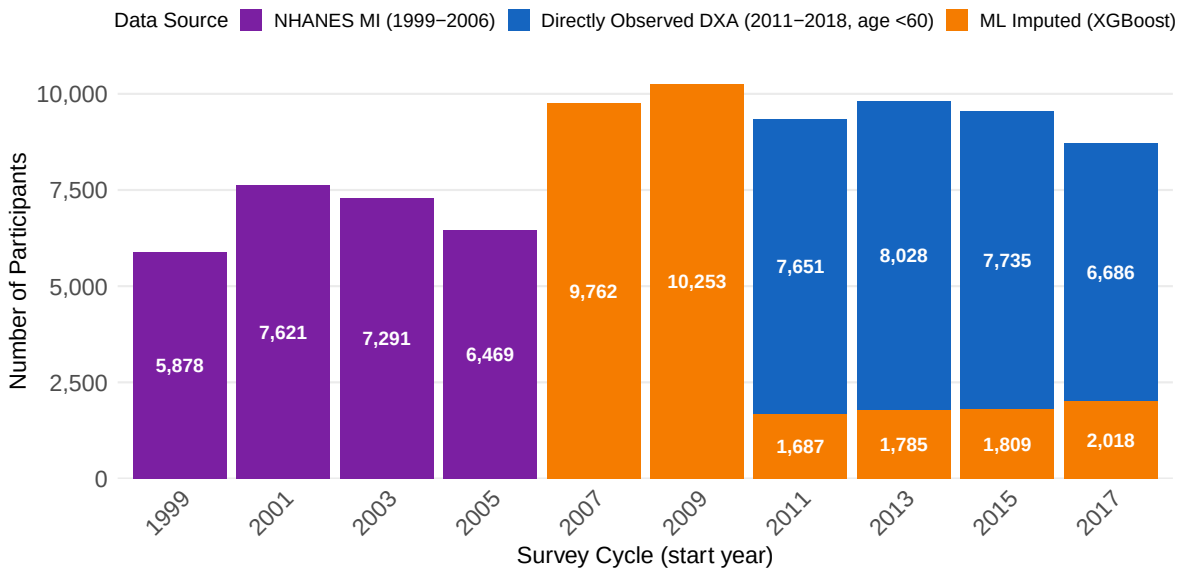


Figure 3.1: This shows the composition of participant data across NHANES survey cycles. Data for 1999–2005 relies on NHANES Multiple Imputation (MI). Directly observed DXA measurements are available for a subset of participants in the 2011–2018 cycles. To address missing data – specifically for the 60^{+} age group in 2011–2018 and all participants in the 2007–2009 cycle – an XGBoost ML model was used for imputation. This model was trained on 1999–2005 data to mitigate selection bias.

$$O_t^m = \sum_i \omega_{it}\, I\left(X_i^m \geq c_{\text{sex}}^m\right) \tag{3.1}$$

for measure $m \in \{\text{BMI}, \text{DXA}, \text{WC}\}$, where ω_{it} denotes survey weights incorporating strata and PSU, and c_{sex}^m denotes sex- and measure-specific obesity cutoffs (BMI ≥ 30; DXA ≥ 30(M), 42(F) BF%; WC ≥ 88(F), 102(M) cm for continuous measure X_i^m) (Potter et al., 2025). Stratum-specific prevalence is estimated by partitioning on age, sex, and race/ethnicity.

3.3.4 Demographic Decomposition of Prevalence Trends

Let $g \in G$ index mutually exclusive demographic strata defined by age group $\in$ {20–39, 40–59, 60^{+}}, sex $\in$ {Female, Male}, and race/ethnicity $\in$ {NH-White, Hispanic, NH-Black, NH-Asian, Other}. Let π_{gt} denote the population share of stratum g at time t (obtained from U.S. Census data), and let O_{gt}^m denote stratum-specific obesity prevalence for obesity

measure $m \in \{\text{BMI, WC, DXA}\}$. Aggregate prevalence can be written as:

$$O_t^m = \sum_{g \in G} \pi_{gt} \, O_{gt}^m \tag{3.2}$$

To isolate the contributions of demographic compositional change and within-stratum rate change to prevalence trends, I apply direct standardization separately for each measure m. Holding population composition fixed at its 1999 distribution while allowing stratum-specific rates to evolve yields:

$$\tilde{O}_t^{m,\text{fix-}\pi} = \sum_{g \in G} \pi_{g,1999} \, O_{gt}^m \tag{3.3}$$

Conversely, holding stratum-specific rates fixed at 1999 levels while allowing composition to change yields:

$$\tilde{O}_t^{m,\text{fix-rate}} = \sum_{g \in G} \pi_{gt} \, O_{g,1999}^m \tag{3.4}$$

The portion of the trend in measure m attributable to compositional shifts is $\Delta_t^{m,\text{comp}} = O_t^m - \tilde{O}_t^{m,\text{fix-}\pi}$, and the portion attributable to changing stratum-specific rates is $\Delta_t^{m,\text{rate}} = O_t^m - \tilde{O}_t^{m,\text{fix-rate}}$. Applying this decomposition to both BMI and DXA separately reveals whether the divergence between the two measures is driven by within-stratum rate changes, compositional shifts, or both.

3.3.5 The BMI–DXA Gap Over Time

I define the aggregate BMI–DXA obesity gap as:

$$\delta_t = O_t^{\text{BMI}} - O_t^{\text{DXA}} = \sum_{g \in G} \pi_{gt} \underbrace{\left(O_{gt}^{\text{BMI}} - O_{gt}^{\text{DXA}}\right)}_{\delta_{gt}} \tag{3.5}$$

where δ_{gt} denotes the stratum-specific measurement discrepancy in survey cycle t. I track δ_t

over the full study period as a descriptive quantity, and examine stratum-specific contributions $C_{gt} = \pi_{gt}\,\delta_{gt}$ to identify populations where high measurement error and significant demographic weight combine to distort aggregate obesity estimates. A negative contribution indicates BMI underestimates obesity relative to DXA for that group; a positive contribution indicates overestimation.

3.4 Results

3.4.1 Obesity Prevalence Trends Across Measures

Figure 3.2 presents crude obesity prevalence estimates across all three measures from 1999 to 2017. BMI-based obesity prevalence increased from 29.9% (95% CI: 27.0–32.8) in 1999–2000 to 42.7% (95% CI: 39.3–46.0) in 2017–2018, a 12.8 percentage point increase over the 19-year period ($p < 0.001$). Waist circumference–based obesity showed the largest absolute increase, rising from 46.2% (95% CI: 41.9–50.6) to 60.6% (95% CI: 57.3–63.8), a 14.3 percentage point rise ($p < 0.001$). DXA-based obesity prevalence showed almost no change, rising from 40.4% (95% CI: 37.2–43.6) to 41.3% (95% CI: 38.2–44.3), a 0.9 percentage point increase that was not a statistically significant change ($p = 0.700$).

Taken at face value, these aggregate trends suggest a divergence between anthropometric measures and direct measures of adiposity. However, this aggregate stability in DXA-based obesity conceals substantial heterogeneity by gender. Figure 3.3 shows that the flat overall DXA trend is the result of offsetting changes among men and women. Among men, DXA-based obesity increases over the study period in a pattern broadly consistent with BMI. Among women, however, DXA-based obesity declines, even as BMI-based obesity rises steadily. The appearance of a stable aggregate trend in adiposity is therefore not indicative of uniform population dynamics, but rather reflects the combination of opposing trajectories across genders.

This divergence is further reflected in the BMI–DXA gap. As shown in Figure 3.4,

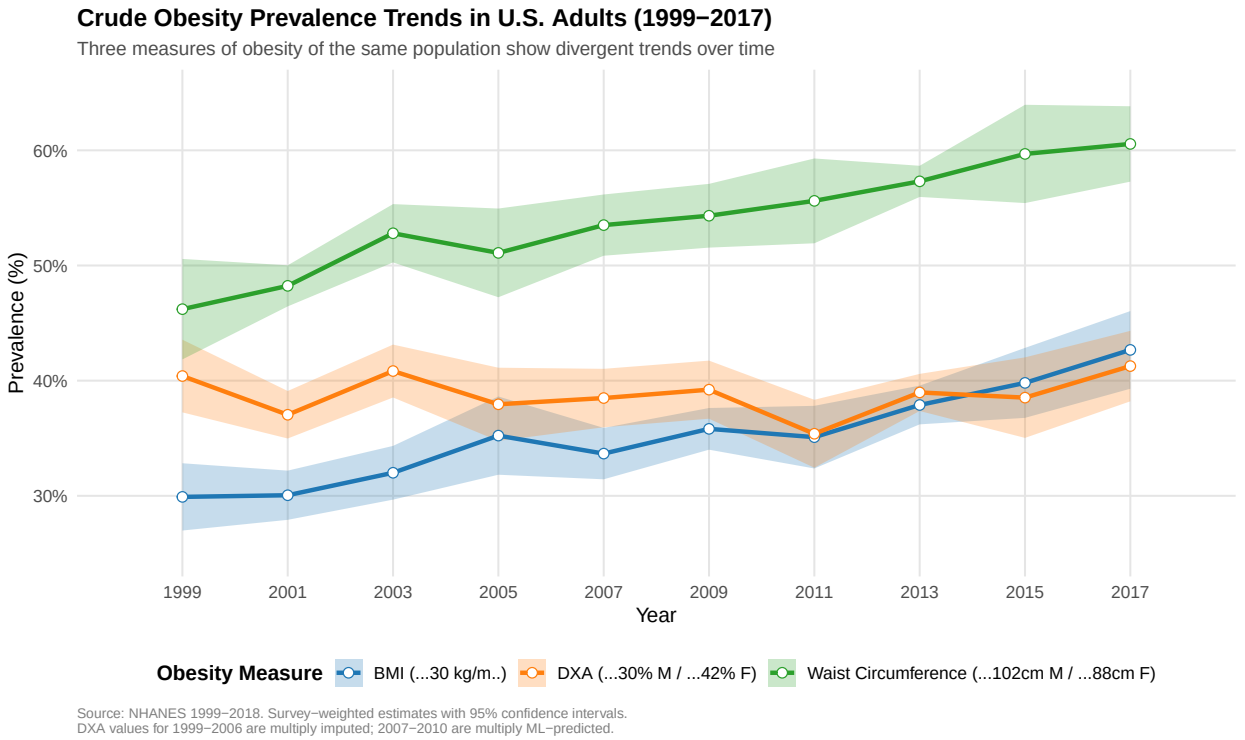


Figure 3.2: Crude obesity prevalence trends by measure, U.S. adults aged ≥20, NHANES 1999–2017. Survey-weighted estimates with 95% confidence intervals. DXA estimates for 1999–2006 use multiply imputed values; 2007–2010 estimates use ML-predicted values.

DXA-based prevalence exceeded BMI-based prevalence in most survey cycles. The BMI-DXA gap was largest at -13.3 percentage points in 1999-2000, narrowed to -7.2 pp by 2005-2006, and continued to shrink through 2017-2018, where it reached -2.1 pp. While this pattern is consistent with prior evidence that BMI underestimates obesity due to its inability to distinguish fat from lean mass Romero-Corral et al. (2008), Figure 3.5 shows that this gap is not uniform across genders. Instead, it reflects distinct and evolving discrepancies between BMI and adiposity for men and women, with BMI increasingly masking the divergence in underlying body composition trends.

3.4.2 Formal Tests of Trend Equality

To assess whether the three measures rose at statistically equivalent rates, I fitted linear probability models of the within-person difference in obesity classification on survey year,

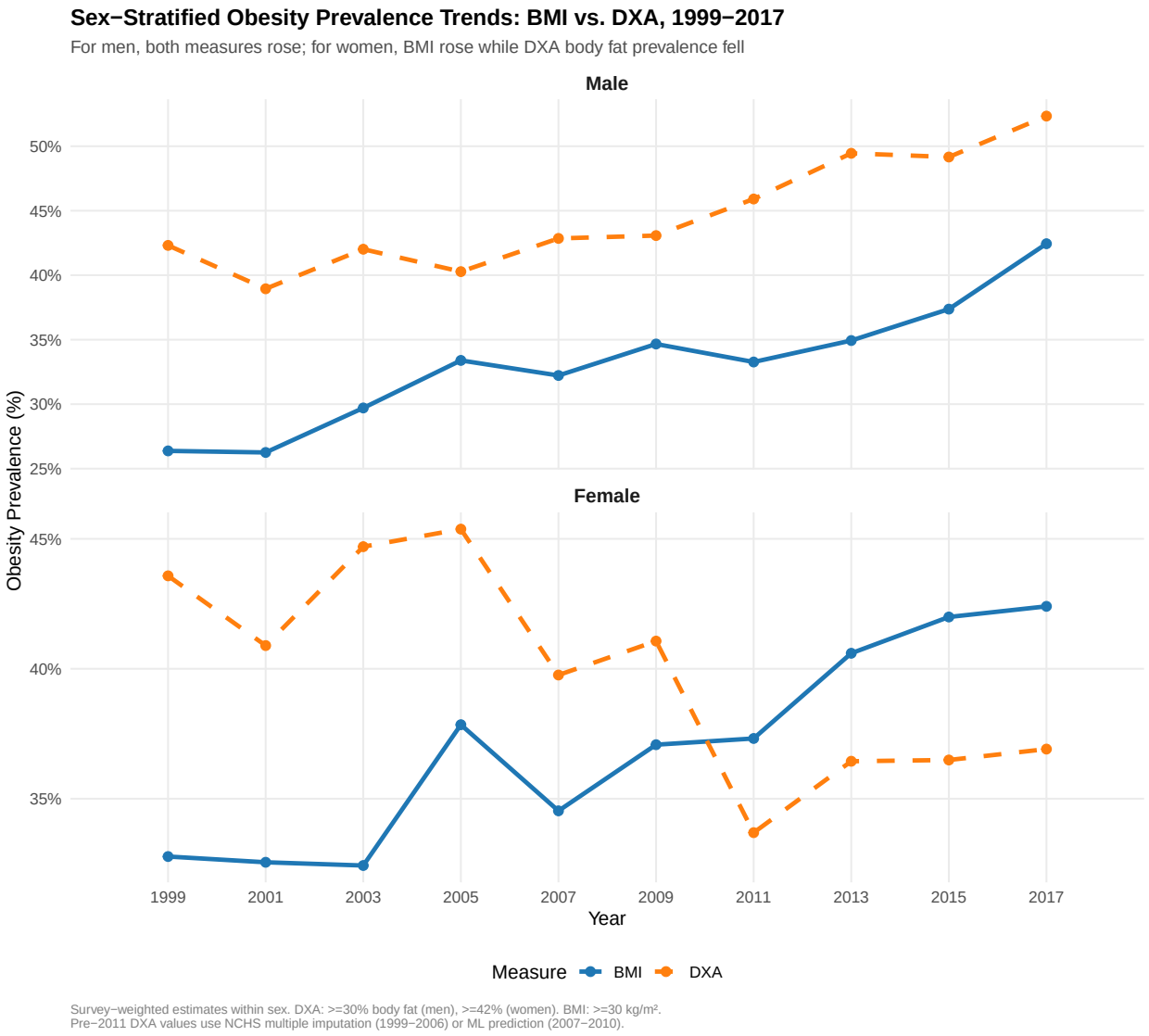


Figure 3.3: Decomposing the BMI- and DXA-obesity trends by gender reveals that the stable aggregate trend is comprised of an increasing trend for men and a decreasing trend for women which roughly cancel.

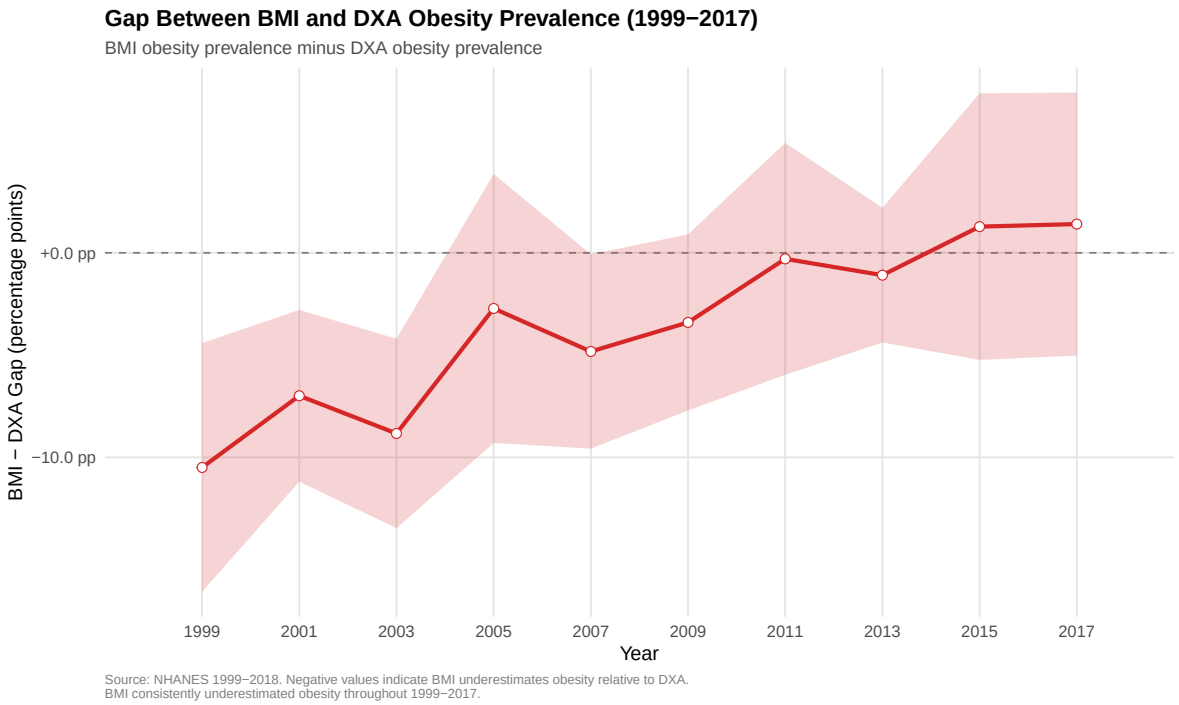


Figure 3.4: Evolution of the BMI–DXA obesity prevalence gap, 1999–2017. Gap calculated as BMI prevalence minus DXA prevalence within demographic strata, then aggregated; negative values indicate BMI underestimates obesity relative to DXA. The gap was largest at –13.3 pp in 1999–2000, narrowed to –3.0 pp in 2015–2016, and stood at –2.1 pp in 2017–2018.

Table 3.1: Crude Obesity Prevalence by Measure, U.S. Adults, NHANES 1999–2018

	BMI $\geq$ 30		DXA BF%[a]		Waist Circ.[b]	
Survey Cycle	Prev. (%)	95% CI	Prev. (%)	95% CI	Prev. (%)	95% CI
1999–2000	29.9	(27.0, 32.8)	40.4	(37.2, 43.6)	46.2	(41.9, 50.6)
2001–2002	30.1	(27.9, 32.2)	37.0	(35.0, 39.1)	48.2	(46.4, 50.0)
2003–2004	32.0	(29.7, 34.3)	40.8	(38.5, 43.1)	52.8	(50.3, 55.3)
2005–2006	35.2	(31.8, 38.6)	38.0	(34.8, 41.1)	51.1	(47.2, 54.9)
2007–2008	33.7	(31.4, 35.9)	38.5	(36.0, 41.0)	53.5	(50.8, 56.2)
2009–2010	35.8	(34.0, 37.6)	39.2	(36.7, 41.7)	54.3	(51.6, 57.1)
2011–2012	35.1	(32.4, 37.8)	35.4	(32.4, 38.3)	55.6	(51.9, 59.3)
2013–2014	37.9	(36.2, 39.6)	39.0	(37.4, 40.6)	57.3	(56.0, 58.7)
2015–2016	39.8	(36.8, 42.8)	38.5	(35.0, 42.0)	59.7	(55.4, 64.0)
2017–2018	42.7	(39.3, 46.0)	41.3	(38.2, 44.3)	60.6	(57.3, 63.8)
Change (pp)	+12.8		+0.9		+14.3	

[a] DXA body fat percentage: $\geq$30% (men), $\geq$42% (women). Values for 1999–2006 use multiply imputed measurements; 2007–2010 use ML-predicted measurements.
[b] Waist circumference: $\geq$102 cm (men), $\geq$88 cm (women).
Survey-weighted estimates with 95% confidence intervals, using Rubin's rules for cycles with multiple imputations.

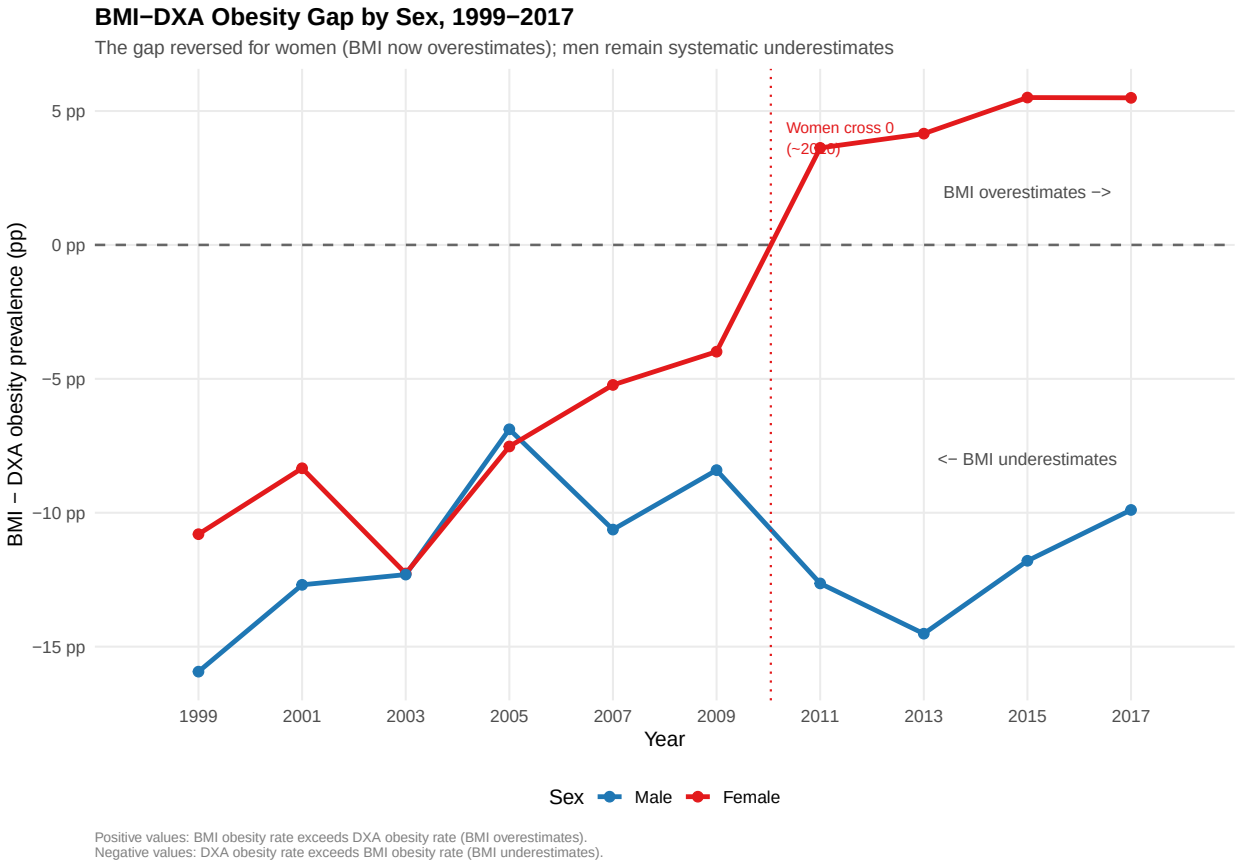


Figure 3.5: Decomposing the BMI- and DXA-obesity prevalence gap by sex shows a divergence between men and women. Convergence is an artifact – BMI consistently underestimates obesity relative to DXA for men over the entire period, but for women, the gap reverses.

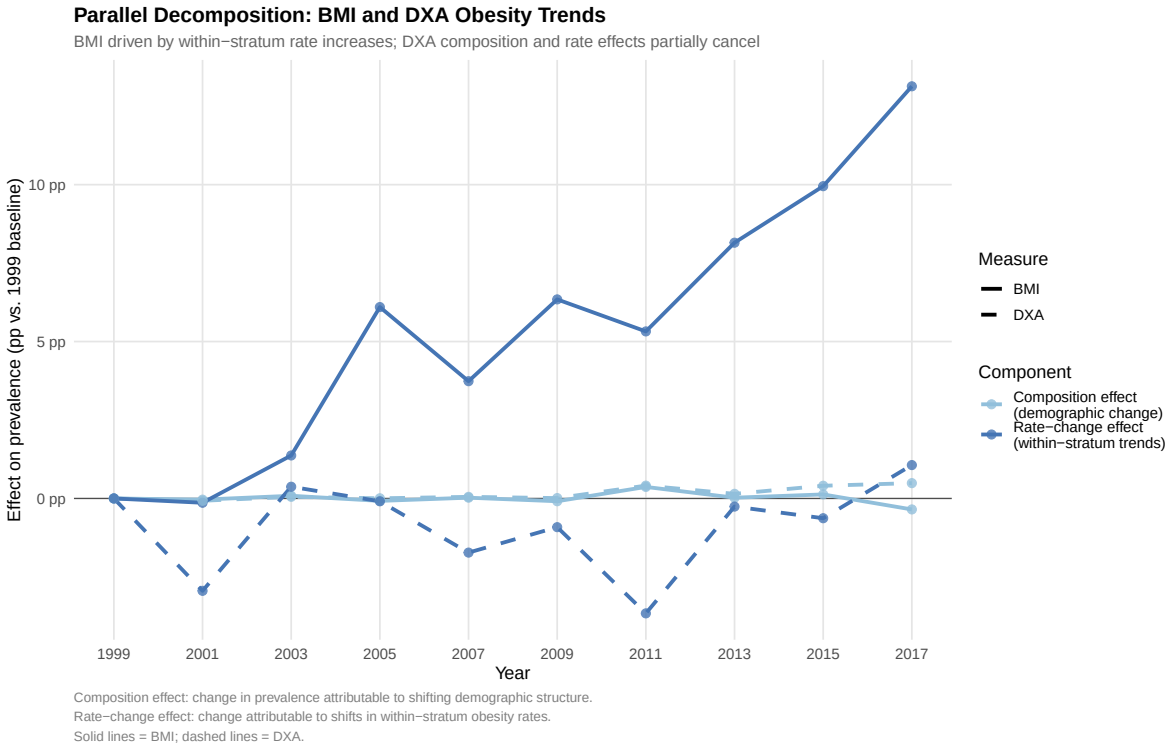


Figure 3.6: Changes to both BMI and DXA trends are driven by rate changes, not composition effects

accounting for complex survey design and multiple imputation. Table 3.2 reports the estimated linear trend for each measure and Table 3.3 reports the pairwise slope equality tests.

BMI-based and WC-based obesity exhibited statistically significant positive trends, while DXA-based obesity showed no significant increase (Table 3.2). BMI-based obesity increased at +0.33% per year (95% CI: 0.25–0.41, $p < 0.001$), WC-based obesity at +0.37% per year (95% CI: 0.28–0.47, $p < 0.001$), and DXA-based obesity at +0.01% per year (95% CI: −0.07 to +0.09, $p = 0.864$). The DXA trend was statistically indistinguishable from zero, in sharp contrast to both anthropometric measures.

Pairwise slope tests confirmed that these differences were statistically significant. The BMI trend exceeded the DXA trend by +0.31% per year (95% CI: 0.25–0.37, $t = 10.57$, $p < 0.001$), and the WC trend exceeded the DXA trend by +0.33% per year (95% CI: 0.26–0.39, $t = 10.10$, $p < 0.001$). By contrast, the BMI and WC trends were statistically indistinguishable from each other (slope difference = −0.05% per year, $t = -1.66$, $p = 0.098$). These results indicate

that anthropometric indicators of obesity (BMI and WC) rose in parallel over 1999–2017, while direct adiposity measurement via DXA showed no significant change.

Table 3.2: Linear Trend in Obesity Prevalence by Measure, NHANES, 1999-2017

Measure	Slope (%/2-yr)	Slope (%/yr)	95% CI (%/yr)	t	p-value
BMI (≥30)	+0.66	+0.33	(0.25, 0.41)	8.05	<0.001
DXA body fat[a]	+0.01	+0.01	(−0.07, 0.09)	0.17	0.864
Waist circ.[b]	+0.74	+0.37	(0.28, 0.47)	7.71	<0.001

Slopes estimated via survey-weighted linear probability model (`svyglm`, Gaussian family) with `year` − 1999 as the continuous predictor. DXA estimates combine 5 imputations via Rubin's rules with Barnard-Rubin degrees-of-freedom correction.
[a] ≥30% body fat (men), ≥42% (women).
[b] ≥102 cm (men), ≥88 cm (women).

Table 3.3: Pairwise Tests of Equal Linear Trend, NHANES 1999–2017

Comparison	Slope Equality			Overall Equality[c]		
	Diff. (%/yr)	95% CI	p-value	F	df	p-value
BMI vs. DXA	+0.31	(0.25, 0.37)	<0.001	108.4	(2, 1356)	<0.001
BMI vs. WC	−0.05	(−0.10, 0.00)	0.098	1518.2	(2, 152)	<0.001
DXA vs. WC	−0.33	(−0.39, −0.26)	<0.001	740.9	(2, 1507)	<0.001

Slope difference = trend in (measure A − measure B) per year. Positive values indicate measure A is rising faster; negative values indicate measure B is rising faster.
[c] Joint Wald F-test of H_0: intercept = 0 *and* slope = 0 (i.e., measures are equal at all time points). The overall test is significant for BMI vs. WC despite the non-significant slope test because BMI and WC have substantially different baseline prevalences (BMI − WC = −18.1 % in 1999, $p < 0.001$).

3.4.3 Stratum-Specific Contributions to the BMI–DXA Gap

Table 3.4 decomposes the aggregate BMI–DXA gap in 2017–2018 into contributions from 30 demographic strata. The total gap in 2017–2018 was −2.1 percentage points under the stratum-weighted decomposition framework. Because some strata exhibit BMI overestimation (positive δ_{gt}) that partially offsets strata with BMI underestimation, individual stratum contributions may exceed the aggregate gap in absolute magnitude. The largest contributor to BMI underestimation was older non-Hispanic White men aged 60^+ ($\pi_{gt} = 14.0\%$, $\delta_{gt} = -26.6\,\%$,

$C_{gt} = -3.73\,\%$). The second largest was older Hispanic men aged 60^{+} ($C_{gt} = -1.29\,\%$), followed by older non-Hispanic Asian men aged 60^{+} ($C_{gt} = -0.57\,\%$) and older non-Hispanic Black men aged 60^{+} ($C_{gt} = -0.45\,\%$).

Among strata exhibiting BMI *overestimation* (positive δ_{gt}), the largest positive contribution came from young non-Hispanic White women aged 20–39 ($\delta_{gt} = +14.3\,\%$, $C_{gt} = +1.13\,\%$), followed by young non-Hispanic Black women ($\delta_{gt} = +22.5\,\%$, $C_{gt} = +0.47\,\%$), middle-aged non-Hispanic White men 40–59 ($C_{gt} = +0.46\,\%$), and middle-aged non-Hispanic White women 40–59 ($C_{gt} = +0.41\,\%$). These positive contributions substantially offset the predominantly negative contributions from older men, so that the aggregate gap is considerably smaller than any single large stratum contribution.

The largest within-stratum discrepancy was observed among older non-Hispanic Asian men aged 60^{+}, where BMI underestimated obesity by 47.0 percentage points relative to DXA ($\delta_{gt} = -47.0\,\%$). In contrast, among young non-Hispanic Black women aged 20–39, BMI overestimated obesity by 22.5 percentage points relative to DXA ($\delta_{gt} = +22.5\,\%$). This 70 percentage point range in measurement error across strata underscores the inadequacy of applying a uniform BMI cutoff across heterogeneous populations.

3.4.4 Demographic Decomposition of BMI and DXA Trends

Table 3.5 presents the decomposition of BMI-based obesity prevalence changes into demographic composition effects and rate effects. Of the 12.8 percentage point increase in BMI-based obesity from 1999 to 2017, only -0.4 % (-2.7%) is attributable to changes in population composition (aging and racial/ethnic diversification). The remaining 13.2 % (102.7%) reflects changes in stratum-specific obesity rates.

The counterfactual prevalence series, which holds 1999 demographic composition constant while allowing stratum-specific rates to evolve, yields a 2017 BMI-based obesity prevalence of 42.8% compared to the observed 42.4%. This near-equivalence demonstrates that demographic compositional change had negligible impact on aggregate BMI-based obesity trends over

Table 3.4: Top 10 Demographic Strata Contributing to Aggregate BMI–DXA Gap, 2017–2018

Stratum	Pop. Share (%)	δ_{gt} (%)[a]	C_{gt} (%)[b]	Sign
60^{+} | Male | NH White	14.0	−26.6	−3.73	−
60^{+} | Male | Hispanic	4.0	−32.4	−1.29	−
20–39 | Female | NH White	7.9	+14.3	+1.13	+
60^{+} | Male | NH Asian	1.2	−47.0	−0.57	−
20–39 | Female | NH Black	2.1	+22.5	+0.47	+
40–59 | Male | NH White	8.2	+5.6	+0.46	+
60^{+} | Male | NH Black	2.8	−16.3	−0.45	−
40–59 | Female | NH White	8.3	+5.0	+0.41	+
20–39 | Female | Hispanic	2.9	+12.2	+0.35	+
20–39 | Male | Hispanic	3.1	+9.6	+0.30	+
All other strata (20)	42.9	–	+0.83	+
Total (30 strata)	100.0	–	−2.09	

[a] $\delta_{gt} = O_{gt}^{\text{BMI}} - O_{gt}^{\text{DXA}}$ (stratum-specific BMI–DXA gap). Negative values indicate BMI underestimates obesity relative to DXA in that stratum.
[b] $C_{gt} = \pi_{gt} \times \delta_{gt}$ (contribution to aggregate gap). Individual contributions sum to the aggregate gap of −2.09 %. Because positive contributions (BMI overestimates) partially offset negative contributions (BMI underestimates), some stratum contributions exceed the aggregate gap in absolute magnitude. NH = Non-Hispanic.
Strata ranked by absolute contribution magnitude.

this period. The composition effect varied modestly across cycles, ranging from −0.09 % in 2005–2006 and 2009–2010 to +0.37 % in 2011–2012, never accounting for more than 6.5% of the cumulative change in any cycle.

Table 3.5: Decomposition of BMI-Based Obesity Prevalence Changes, 1999–2017

	BMI Prevalence (%)		Change from 1999 (%)		
Survey Cycle	Observed	CF (Fixed Pop.)[a]	Total	Composition[b]	% Composition
1999–2000	29.6	29.6	0.0	0.0	–
2001–2002	29.5	29.5	−0.2	−0.0	21.6
2003–2004	31.1	31.0	+1.5	+0.1	5.8
2005–2006	35.7	35.7	+6.0	−0.1	−1.3
2007–2008	33.4	33.4	+3.8	+0.0	0.6
2009–2010	35.9	36.0	+6.3	−0.1	−1.4
2011–2012	35.3	35.0	+5.7	+0.4	6.5
2013–2014	37.8	37.8	+8.2	+0.0	0.3
2015–2016	39.7	39.6	+10.1	+0.1	1.3
2017–2018	42.4	42.8	+12.8	−0.4	−2.7

[a] Counterfactual (CF) prevalence holding population structure fixed at 1999 distribution.
[b] Composition effect = Observed − Counterfactual. Negative values indicate demographic shifts slightly reduced BMI-based prevalence relative to within-group rate changes alone.
Population structure defined by 30 demographic strata: 3 age groups (20–39, 40–59, 60^{+}) × 2 sexes × 5 race/ethnicity categories. Pre-2011 cycles used 24 strata due to smaller sample sizes.

3.5 Discussion

3.5.1 Limitations

Several limitations delineate the scope of the findings described above. First, the recovery of a consistent DXA time series relies on an imputation strategy that assumes the multivariate relationship between anthropometrics (BMI, waist circumference), demographic covariates, and true body fat percentage remained stationary between the 1999–2004 training period and the 2007–2018 estimation period. If the underlying mapping of lean-to-fat mass shifted fundamentally over these two decades, the imputed estimates for the 60^{+} cohort in later cycles could contain structural bias. However, given the model's flexibility and the sheer

magnitude of the observed BMI-DXA gap across fully observed strata, it is highly improbable that a gradual drift in anthropometric mapping could account for the near-total divergence in the aggregate trends.

Second, while this paper demonstrates a profound decoupling of aggregate adiposity from BMI, it stops short of causal epidemiological modeling regarding the concurrent rise in cardiometabolic morbidities or other health outcomes. As noted, rates of type 2 diabetes and cardiovascular disease did increase over this period. By demonstrating that the prevalence in obesity according to directly measured adiposity is not rising in the aggregate and actually declining among women, this analysis does not contest the observation of worsening health outcomes, but rather problematizes the causal vector. The scope of this paper does not adjudicate alternative mechanisms; however, the findings strongly suggest that future population health research must investigate whether observed morbidity and mortality trends are driven by upstream environmental determinants (e.g., diet quality, built environment, social stigma) independent of adiposity, or by the biomechanical and metabolic consequences of weight-bearing load (which BMI captures) rather than fat mass.

Finally, identifying the massive structural heterogeneities in measurement error across race, age, and sex highlights BMI's limitations as a biological proxy, but a full accounting of how this specific metric became entrenched as an administrative and clinical infrastructure falls outside the scope of this demographic analysis. The 70-percentage-point swing in measurement error between older Asian men and young Black women underscores that standard BMI cutoffs operate as a structurally racialized and gendered classification system. Tracing how a proxy with such varied validity continues to dictate public health surveillance and resource allocation remains an important task for the sociology of science and critical algorithm studies.

3.5.2 Conclusion

This paper makes four interrelated contributions that correspond to the empirical and conceptual threads discussed throughout. First, it provides direct population-level evidence on the construct validity of BMI as a measure of adiposity, extending individual-level critiques to a national population trend analysis. Second, it complicates the demographic literature on racial, gender and age disparities in obesity by showing that the measurement error BMI introduces is not uniform across groups, raising the possibility that apparent changes in group gaps partly reflect a shifting measurement artifact. Third, it subjects the "obesity epidemic" narrative to empirical scrutiny using the biological construct that narrative presupposes, finding that the increasing trend in obesity is at most conflicted and at least considerably more ambiguous than BMI-based surveillance implies. Fourth, it closes the specific gap identified in recent work by providing a full 1999–2017 trend analysis, including for adults aged 60+, with formal decomposition of the sources of BMI–DXA divergence across demographic strata.

How we measure obesity determines what we observe. BMI-based surveillance told one story since 1999: a sustained rise in obesity prevalence that became the empirical backbone of a widespread moral panic (Campos, 2004; Gard and Wright, 2005; Saguy, 2013). Measuring total body fat directly with DXA scans – the closest available proxy to the biological construct that panic presupposes – tells a different story. By the stricter standard of direct adiposity measurement, obesity prevalence was not statistically distinguishable between 1999 and 2017.

The finding that BMI and DXA diverge so substantially at the population level extends the individual-level construct validity literature in an important direction. Prior work established that BMI misclassifies individual-level adiposity at non-trivial rates, with error that varies systematically across demographic groups Romero-Corral et al. (2008); Stanford et al. (2019); Jeong et al. (2023). The concern raised here is distinct: individual measurement error need not aggregate into valid population-level inference (Salerno et al., 2025), and the present results confirm that it does not. The subgroup decomposition reveals that "obesity", as operationalized by BMI thresholds, is not a single construct applied consistently across the

population but rather an average over meaningful variation, laundered by aggregate prevalence figures. The same cutoff that fails to detect adiposity in a 65-year-old man overestimates it in a 25-year-old woman, and the structural errors are growing wider over time in both directions.

Taken together, these findings carry direct implications for the demographic literature that uses BMI-defined obesity to decompose racial and socioeconomic disparities. Work in this tradition treats BMI as a reliable proxy for excess adiposity: the underlying health construct of interest. Because BMI–DXA obesity gap has widened at different rates across race, sex and age groups over time, the apparent changes in group disparities in existing research may be confounded by measurement error rather than capturing genuine convergence or divergence in population adiposity. Johnston and Lee (2011) noted that the choice between BMI and waist-to-hip ratio materially affected estimates of the Black–White obesity gap's magnitude. The present results suggest the problem is even more pervasive and structurally heterogeneous than any single comparison of two anthropometric measures implies.

The direct-standardization decomposition sharpens the picture on the epidemic claim considerably. Nearly the entire 12.8 percentage-point rise in BMI-based obesity between 1999 and 2017 is attributable to within-stratum rate changes – increasing BMI obesity within every age, sex, and race/ethnicity cell simultaneously – while shifts in demographic composition contributed very little. The rise is therefore "real" in the sense of being uncontaminated by demographic change; the question is what, within each group, it is actually measuring. Within the same demographic cells that saw rising BMI prevalence, DXA-based body fat obesity did not rise in parallel – in most female strata it fell – while waist circumference rose nearly in lockstep with BMI. This three-way divergence (BMI $\uparrow$, WC $\uparrow$, DXA $\rightarrow$) is consistent with a redistribution of body fat from peripheral subcutaneous to central visceral deposits, a mechanism that would leave total body fat percentage unchanged while increasing abdominal adiposity and, if anything, worsening cardiometabolic risk – a finding with implications for which constructs future surveillance should prioritize. This pattern has gone undetected

in BMI-based international surveillance: a recent cross-national analysis finds that obesity growth has plateaued in high-income countries (NCD Risk Factor Collaboration (NCD-RisC), 2026), but the present results suggest that even a "plateau" may overstate adiposity trends when the outcome is measured directly.

Even the aggregate stability of DXA-based obesity prevalence conceals meaningful underlying change. The decomposition of DXA-based prevalence shows that a flat aggregate trajectory is a statistical artifact produced by offsetting forces: population aging places upward pressure on adiposity while within-stratum DXA rates have declined across much of the population, most notably among women. The "obesity epidemic," when understood as a trend in excess body fat rather than body mass index, is thus not a uniform population-wide phenomenon but a gender-differentiated process – one that may have directed attention and resources toward groups whose adiposity was in fact declining.

The temporal narrative is particularly vulnerable to this critique. The finding that aggregate DXA-based obesity prevalence was statistically flat from 1999 to 2017 – the precise period in which BMI-based surveillance anchored claims of an accelerating American epidemic – illustrates that the trend we measure depends substantially on the instrument we choose to measure it with. A construct that means different things in different bodies, that tracks body fat only loosely and inconsistently across demographic groups, and whose trend diverges from direct adiposity measurement over the very period in which it anchored a public health epidemic, is a fragile foundation for the policy, clinical, and social apparatus built upon it.

These results do not imply that population health has improved, nor that excess adiposity is irrelevant to health outcomes. Rather, they suggest that measurement deserves to be treated as an empirical question rather than a background assumption. Where possible, researchers should triangulate across multiple conceptually aligned indicators – distinguishing total adiposity, fat distribution, and body mass – and evaluate how alternative operationalizations alter substantive conclusions. Trends should be examined within demographic strata rather than inferred from aggregate patterns that may conceal offsetting dynamics. For surveillance

and policy, BMI may retain utility as a simple screening tool or index of weight-bearing burden, but it should not be assumed to provide a stable measure of adiposity across populations or over time. The broader implication is not to replace one metric with another, but to ask more precisely what is being measured and why – and to recognize that the answer may differ by group, by era, and by the question being posed.

3.6 Acknowledgements

I thank Amy Wang for excellent research assistance. I am grateful for the helpful assistance and comments received from Sasha Johfre, Tyler McCormick, Emilio Zagheni, Sarah Quinn, and participants of the reading groups at UW Sociology. I gratefully acknowledge the resources provided by the International Max Planck Research School for Population, Health and Data Science (IMPRS-PHDS). Partial support for this research came from a Shanahan Endowment Fellowship and a Eunice Kennedy Shriver National Institute of Child Health and Human Development training grant, T32 HD101442, to the Center for Studies in Demography & Ecology at the University of Washington. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Additional support for this research came from support to CSDE from the College of Arts & Sciences, the UW Provost, eSciences Institute, the Evans School of Public Policy & Governance, College of Built Environment, School of Public Health, the Foster School of Business, and the School of Social Work.

3.7 Reproducibility and Data Availability

All data used for this study are publicly available for download from the CDC (https://wwwn.cdc.gov/nchs/nhanes/default.aspx) and (https://wonder.cdc.gov/bridged-race-v2020.html).

Notes

1. For detailed description of the machine learning imputation, please see Appendix 3.8.1.

2. Code to reproduce my analysis is available at github.com/avisokay/obesity_trends.

3.8 Chapter Appendix

3.8.1 Machine Learning Model Performance

Model specification. DXA body fat percentage was predicted using a gradient-boosted tree model (XGBoost) trained on the NCHS multiply imputed 1999–2004 NHANES data (Schenker et al., 2011). The model was specified with the following hyperparameters, held fixed across all imputation methods to allow clean comparisons: learning rate $\eta = 0.05$, maximum tree depth of 6, subsampling fraction of 0.8, column subsampling fraction of 0.8, and minimum child weight of 5. The number of boosting rounds was selected by five-fold cross-validation with early stopping (patience = 30 rounds) up to a maximum of 1,000 rounds; cross-validation selected the full 1,000 rounds. Training observations were assigned heterogeneous weights

$$w_i \propto \frac{1}{\widehat{\mathrm{SE}}_{g(i)}^{\,2} + c}$$

derived from published NCHS standard errors stratified by sex and BMI category (Schenker et al., 2011) with $c = 0.01$; Table 3.6 reports the resulting SE and weight for each cell. Weights span a 7.5-fold range, from 3.44 for under-weight males (BMI < 20) to 25.71 for normal-range males and females (BMI 25–< 30), reflecting substantially higher NCHS imputation uncertainty at the tails of the BMI distribution.

Training performance and imputation uncertainty. The weighted model achieved a training RMSE of 1.83 percentage points with negligible mean bias (< 0.001 %), and a residual standard deviation of 1.83 % that was used as the noise scale for generating the five imputation draws. Between-imputation variability for predicted respondents had a mean SD of 1.71 % (median 1.67 %; 95th percentile 2.81 %), consistent with appropriate propagation of prediction uncertainty through the five completed datasets and Rubin's rules combination.

Table 3.6: NCHS-Derived Training Weights by Sex and BMI Category

Sex	BMI Category	NCHS SE (%)	Training Weight
Male	< 20	0.53	3.44
Male	20–<25	0.25	13.79
Male	25–<30	0.17	25.71
Male	30–<35	0.18	23.58
Male	35–<40	0.30	10.00
Male	40+	0.50	3.85
Female	< 20	0.52	3.57
Female	20–<25	0.26	12.89
Female	25–<30	0.17	25.71
Female	30–<35	0.19	21.69
Female	35–<40	0.28	11.31
Female	40+	0.31	9.43

Note: SE is the published standard error of the NCHS multiply imputed mean body fat percentage for each sex × BMI category cell (Schenker et al., 2011). Training weight $w_i \propto 1/(\mathrm{SE}^2 + 0.01)$; weights span a 7.5-fold range.

Alternative imputation strategies. Three additional imputation strategies were evaluated prior to adopting the NCHS-weighted approach. First, I trained an XGBoost model without weights on the full 1999–2006 NCHS multiple imputation data with pooled Gaussian residual noise. This model achieved an overall out-of-sample test RMSE of 3.01 percentage points (2.86 % on the imputed-data test partition). A *calibrated bootstrap* variant used 50 XGBoost replicates trained on bootstrap samples of the 1999–2006 data, with imputation noise drawn from stratum-specific (sex × age × race/ethnicity) residual distributions anchored to 2011–2018 observed DXA. A third variant explored ridge regression, which performed substantially worse (test RMSE ≈11.9 %) and was not used further.

To assess sensitivity, predictions from the weighted model (Method 4) were compared against those from the calibrated bootstrap model (Method 3) for the overlapping set of respondents requiring imputation. The two methods correlated at $r = 0.948$ with a mean difference of 0.26 percentage points (Method 4 minus Method 3), and produced near-identical distributions of imputed body fat percentage (Method 3: $33.1 \pm 7.3\%$; Method 4: $33.3 \pm 7.0\%$). All substantive prevalence estimates and trend findings in this paper are unchanged when

Method 3 is substituted for Method 4.

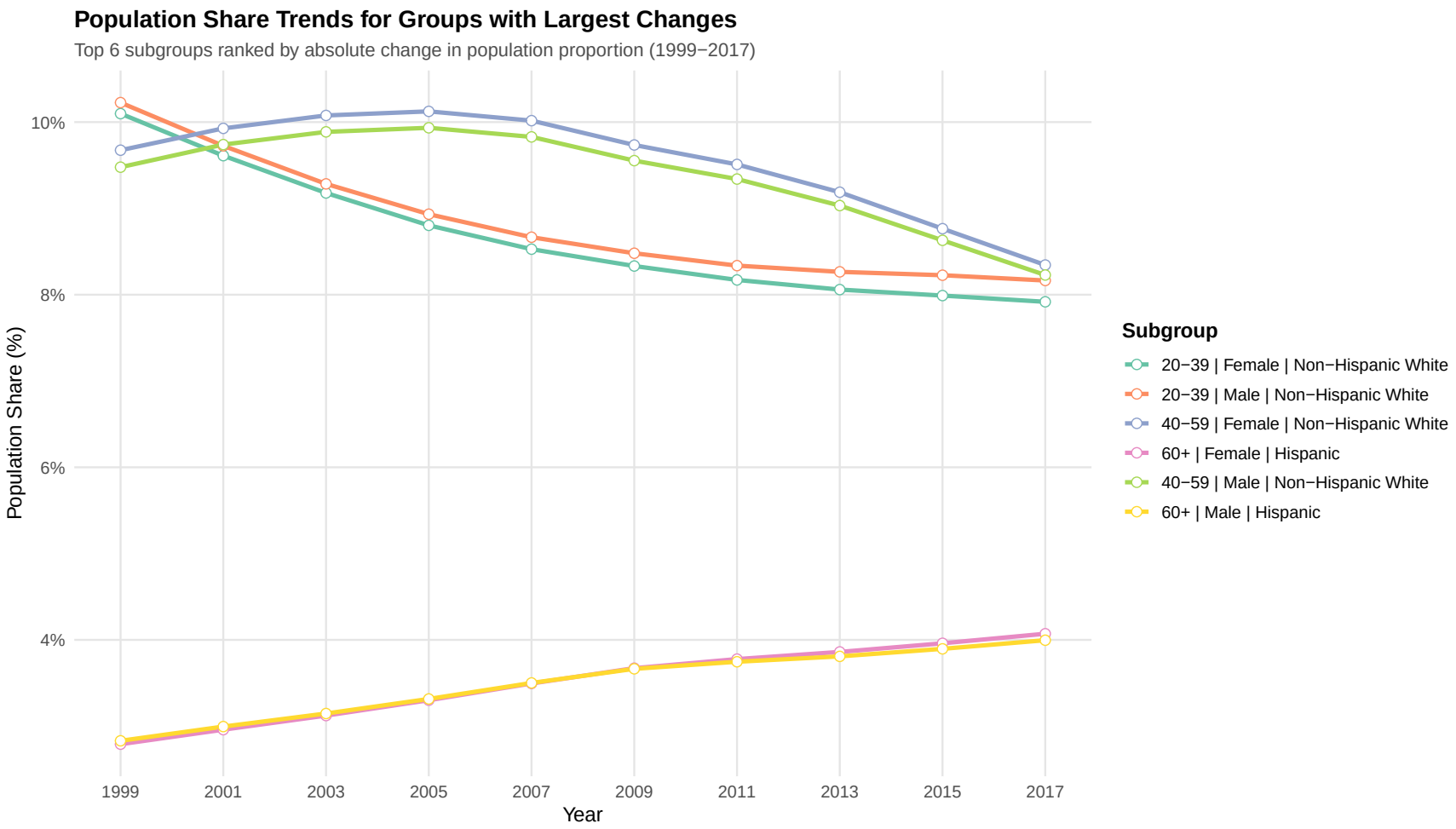
3.8.2 ML Robustness Checks

Trend Attenuation. The primary concern with regression-to-the-mean in an out-of-sample prediction is that the model compresses time trends toward the training-era average, artificially flattening any “true" rise during the out of sample period. To test for attenuation, I exploit the 2011-2017 overlap window where adults 20-59 years of age have both observed DXA values and a model prediction for the same person. I compare the two slopes using weighted least-squares linear trends in DXA obesity prevalence fit separately for the observed and predicted series over 2011-2017, weighting by the inverse squared standard error of each prevalence estimate. Attenuation would manifest in a predicted slope flatter than the observed slope: a miscalibrated model compressing a “true" rise towards zero. Instead, I find the observed slope for these years was 0.22pp/yr (flat, p=0.86) while the predicted slope (based on the model trained on 1999-2004 data) was 1.96pp/yr (rising, p=0.034). The headline aggregate finding DXA trend is therefore if anything biased *upward*. In other words, the “true" DXA would look even flatter if the prediction model was unbiased.

Observed-only replication. The paper's central claim – that DXA-based obesity prevalence was statistically flat over 1999-2017 while BMI rose 12.8 percentage points – rests partly on imputed body fat estimates. DXA scans were not collected in 2007-2010 (any age) or for adults aged 60 and older in 2011-2017; the 1999-2004 values derive from NCHS multiple imputation rather than direct measurement. For a test of robustness I exclude all imputed values, restricting to adults aged 20-59 in cycles 2011-2017 using only directly observed DXA measurements (n = 10,864). The BMI obesity slope is 0.87 pp/yr (95% CI: 0.16 to 1.58, $p = 0.034$) while the DXA obesity slope is 0.22 pp/yr (95% CI: -1.74 to 2.18, $p = 0.68$). That is to say, the BMI-DXA divergence is present without any model predictions, any multiple imputation, or any data from outside the directly measured window. Over the four observed

cycles, BMI obesity rose approximately 4.83 pp while DXA obesity showed no statistically significant increase.

3.8.3 Figures


Population Share Trends for Groups with Largest Changes
Top 6 subgroups ranked by absolute change in population proportion (1999−2017)
Population Share (%)
10%
8%
6%
4%
1999
2001
2003
2005
2007
2009
2011
2013
2015
2017
Year
Subgroup
20−39 | Female | Non−Hispanic White
20−39 | Male | Non−Hispanic White
40−59 | Female | Non−Hispanic White
60+ | Female | Hispanic
40−59 | Male | Non−Hispanic White
60+ | Male | Hispanic
Source: NHANES 1999−2022 and U.S. Census population estimates.

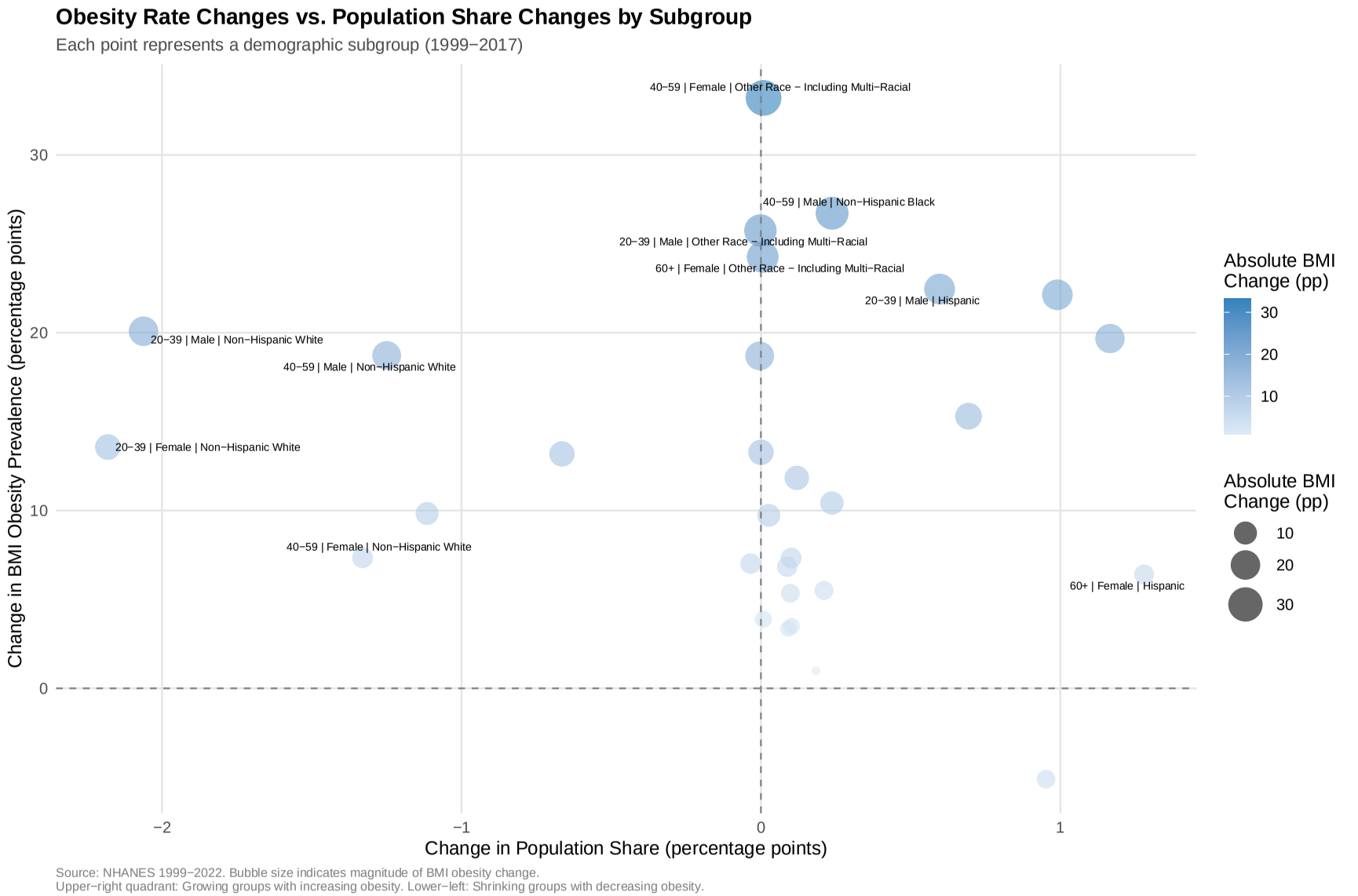
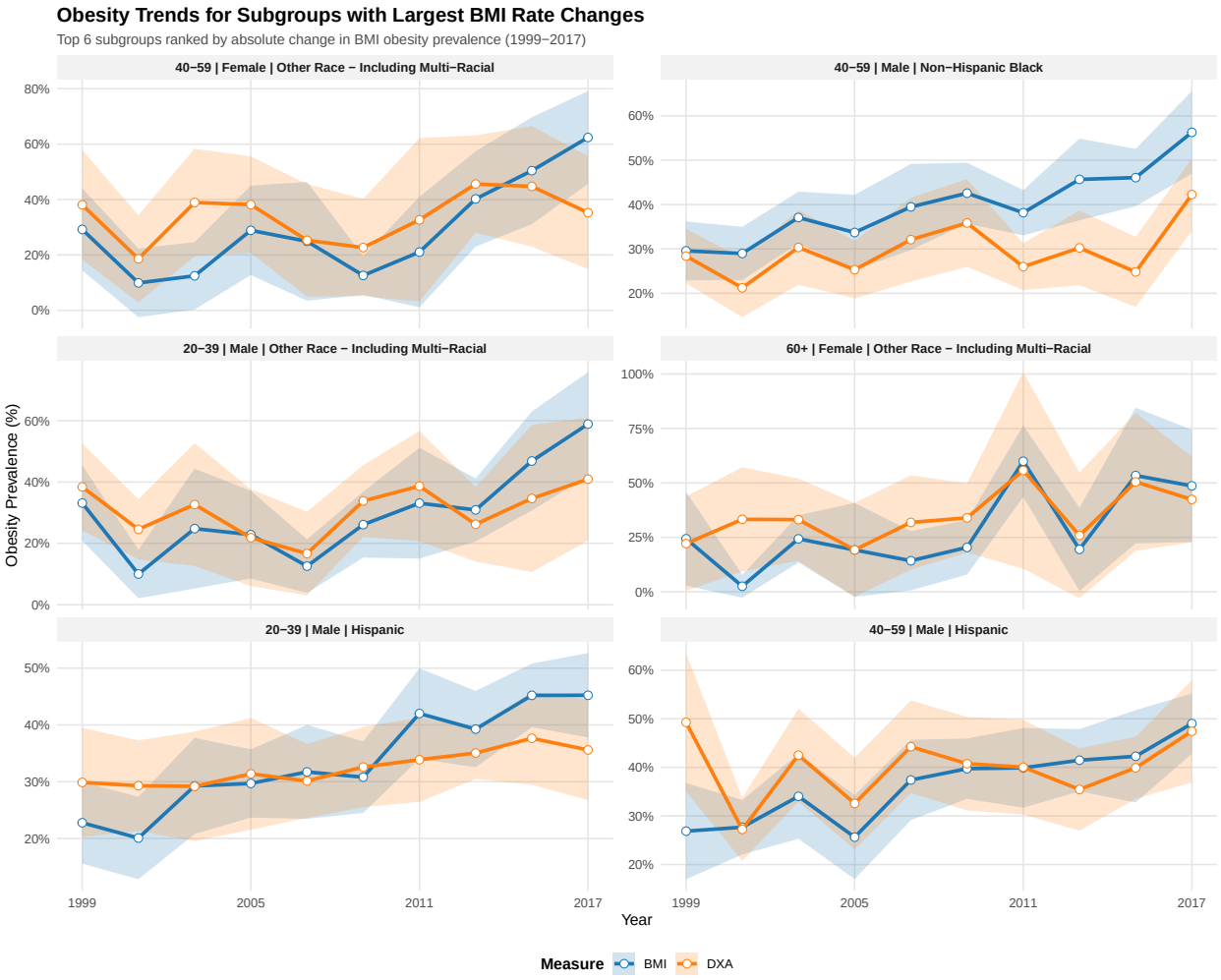
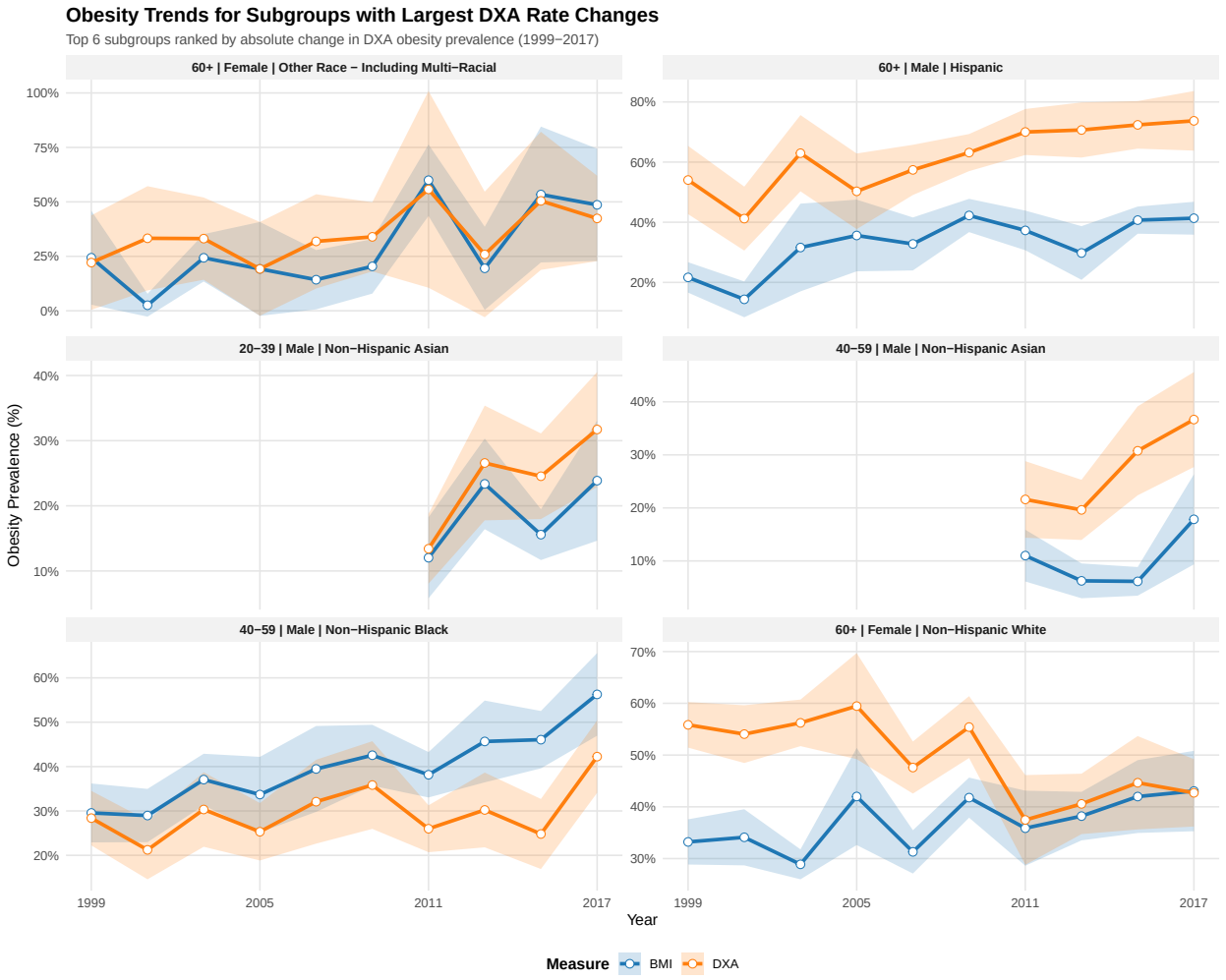
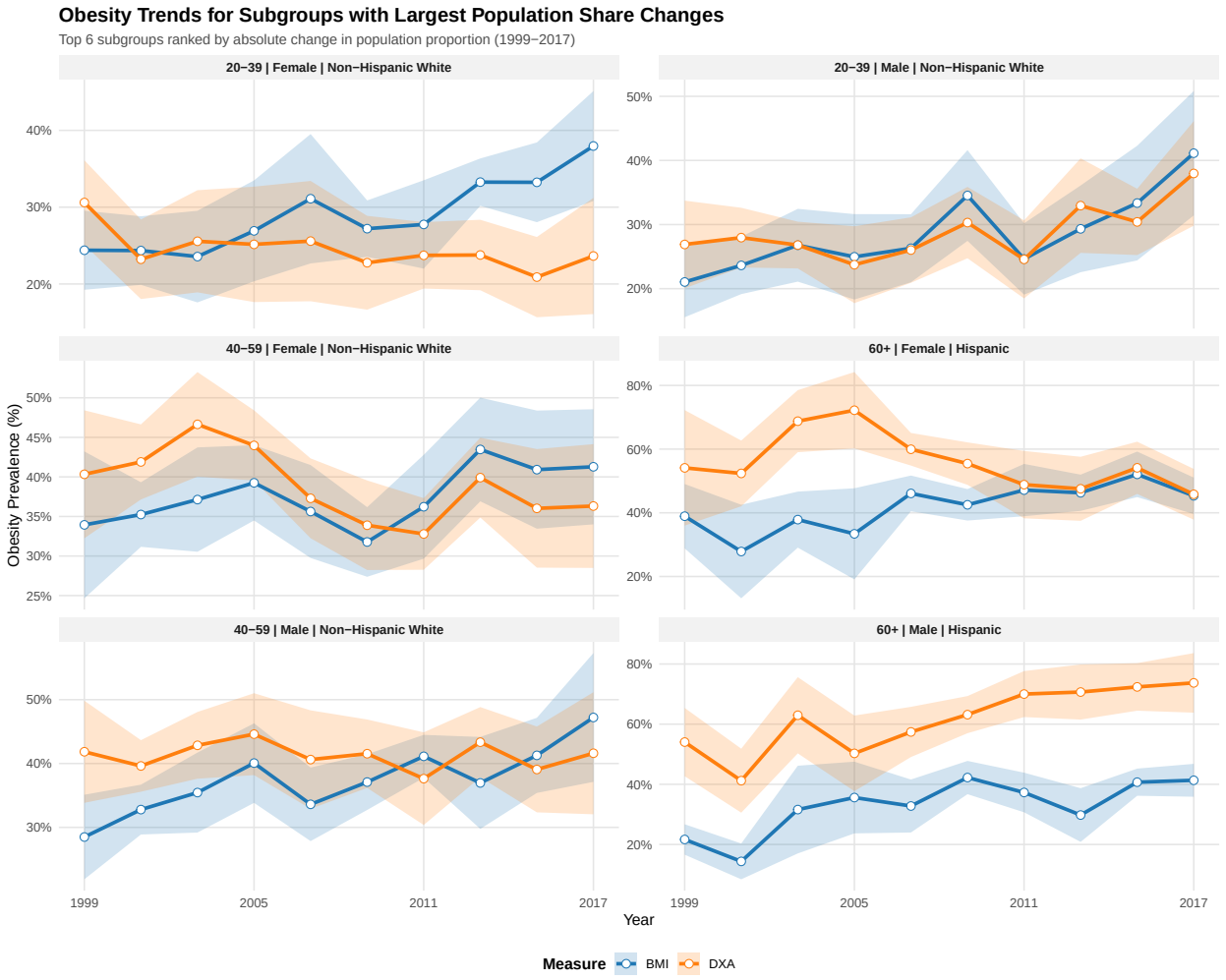


Figure 3.7: Change in BMI-based obesity prevalence versus change in population share by demographic subgroup, 1999–2017. Each bubble represents one of 30 demographic strata defined by age group, sex, and race/ethnicity; bubble size indicates the absolute magnitude of the BMI obesity rate change. Nearly all strata experienced increases in BMI-based obesity prevalence regardless of whether their population share grew or shrank, concentrating points in the upper half of the plot. The absence of a positive correlation between the two axes—growing groups are not systematically more likely to have rising obesity rates—illustrates why demographic compositional change explains little of the aggregate BMI trend: the rise is broad-based within strata rather than driven by the changing mix of groups. Source: NHANES 1999–2018 and U.S. Census population estimates.


Obesity Trends for Subgroups with Largest BMI Rate Changes
Top 6 subgroups ranked by absolute change in BMI obesity prevalence (1999−2017)
40−59 | Female | Other Race − Including Multi−Racial
40−59 | Male | Non−Hispanic Black
20−39 | Male | Other Race − Including Multi−Racial
60+ | Female | Other Race − Including Multi−Racial
20−39 | Male | Hispanic
40−59 | Male | Hispanic
Obesity Prevalence (%)
Year
1999
2005
2011
2017
Measure
BMI
DXA
Source: NHANES 1999−2018. Survey−weighted estimates.
DXA values for 1999−2006 are multiply imputed; 2007−2010 are multiply ML−predicted.


Obesity Trends for Subgroups with Largest DXA Rate Changes
Top 6 subgroups ranked by absolute change in DXA obesity prevalence (1999–2017)
60+ | Female | Other Race – Including Multi–Racial
60+ | Male | Hispanic
20–39 | Male | Non–Hispanic Asian
40–59 | Male | Non–Hispanic Asian
40–59 | Male | Non–Hispanic Black
60+ | Female | Non–Hispanic White
Obesity Prevalence (%)
Year
1999
2005
2011
2017
Measure
BMI
DXA
Source: NHANES 1999–2018. Survey–weighted estimates.
DXA values for 1999–2006 are multiply imputed; 2007–2010 are multiply ML–predicted.


Obesity Trends for Subgroups with Largest Population Share Changes
Top 6 subgroups ranked by absolute change in population proportion (1999–2017)
20–39 | Female | Non–Hispanic White
20–39 | Male | Non–Hispanic White
40–59 | Female | Non–Hispanic White
60+ | Female | Hispanic
40–59 | Male | Non–Hispanic White
60+ | Male | Hispanic
Obesity Prevalence (%)
Year
1999
2005
2011
2017
Measure
BMI
DXA
Source: NHANES 1999–2018. Survey–weighted estimates.
DXA values for 1999–2006 are multiply imputed; 2007–2010 are multiply ML–predicted.

Chapter 4

From Narratives to Numbers: Valid Inference Using Language Model Predictions From Verbal Autopsy Narratives

Shuxian Fan*, Adam Visokay*, Kentaro Hoffman, Stephen Salerno, Li Liu, Jeffrey T. Leek, Tyler H. McCormick. 2024. In *Proceedings of the First Conference on Language Modeling.* COLM 2024, Philadelphia, Pennsylvania.

Abstract. In settings where most deaths occur outside the healthcare system, verbal autopsies (VAs) are a common tool to monitor trends in causes of death (COD). VAs are interviews with a surviving caregiver or relative that are used to predict the decedent's COD. Turning VAs into actionable insights for researchers and policymakers requires two steps (i) predicting likely COD using the VA interview and (ii) performing inference with predicted CODs (e.g. modeling the breakdown of causes by demographic factors using a sample of deaths). In this paper, we develop a method for valid inference using outcomes (in our case COD) predicted from free-form text using state-of-the-art NLP techniques. This method, which we call `multiPPI++`, extends recent work in "prediction-powered inference" to multinomial classification. We leverage a suite of NLP techniques for COD prediction and, through empirical analysis of VA data, demonstrate the effectiveness of our approach in handling transportability issues. `multiPPI++` recovers ground truth estimates, regardless of which NLP model produced predictions and regardless of whether they were produced by a more accurate predictor like GPT-4-32k or a less accurate predictor like KNN. Our findings demonstrate the practical importance of inference correction for public health decision-making

and suggests that if inference tasks are the end goal, having a small amount of contextually relevant, high quality labeled data is essential regardless of the NLP algorithm.

4.1 Introduction

Verbal Autopsies (VAs) are a pivotal tool in public health research, particularly in contexts where access to formal medical records is limited (Soleman et al., 2006; Thomas et al., 2018). VA serves as a method to ascertain causes of death (COD) by collecting information from relatives or witnesses in circumstances necessitated by resource constraints, remote locations, or inadequate healthcare infrastructure (Baqui et al., 2006; Byass et al., 2012; Wahab et al., 2017; Blanco et al., 2021).VA interviews include a structured yes/no questionnaire and an open text narrative where the interviewees provide additional information about the illness or death in their own words.

After the interview, a COD is assigned either by clinician review or, more commonly, by using statistical algorithms. Finally, since financial and logistical constraints prevent VAs from being performed on all deaths, researchers and policymakers use statistical tools to summarize patterns in COD (e.g. the breakdown of infectious disease deaths by age, race, sex, etc).

We propose a method for valid inference using the open text VA narratives, summarized in Figure 4.1. To do this, we address two pressing challenges. First, we use text narratives exclusively to classify COD. Structured VA interviews are long (typically around an hour) and emotionally taxing for respondents (Surek-Clark, 2020). Narratives provide an opportunity for respondents to describe the circumstances around the death of someone they knew well in their own words, without needing to translate clinical jargon or answer irrelevant questions. Using narratives alone could also dramatically reduce the length of the interview, both minimizing impact on the respondent and allowing enumerators to collect more VAs. Recent work (Danso et al., 2013; Blanco et al., 2021; Manaka et al., 2022; Cejudo et al., 2023) has

begun exploring the potential of natural language processing (NLP) for COD classification in VA. Our work furthers this discussion by leveraging state-of-the-art NLP tools. Also, unlike prior NLP work which focuses on *predictive performance* (Naradowsky et al., 2012; Liu et al., 2019; Ye et al., 2021; Peskoff and Stewart, 2023), our goal is to utilize these predictions in *statistical inference* to understand patterns in COD.

Second, we propose a method for inference that corrects for misclassification in CODs predicted by NLP. Training and evaluating COD prediction algorithms is extremely challenging (Murray et al., 2011; Fottrell and Byass, 2010; Murray et al., 2014; Clark et al., 2018; Li et al., 2020), in part because building training sets with reliably labeled CODs is taxing. Clinicians could review VAs and assign causes, but this diverts critical resources from patient care and, since most deaths occur outside healthcare facilities, a physical autopsy is impossible. Hence, it is appealing to create a larger training set by combining the limited number of labeled VAs in one domain with other labeled VAs from different domains (e.g. locations). However, the diversity of cultural practices, linguistic nuances, and variations in interview techniques (not to mention potential biological factors) means that an algorithm trained on labeled VAs in one context may perform poorly in another. Our inferential methods must account for additional uncertainty that arises when most COD labels are predicted, not known. We also need to adjust for transportability so that VAs labeled in other contexts can be incorporated effectively into the training set. To do this, we extend prediction-powered inference (`PPI++`; Angelopoulos et al., 2023a,b) to settings of multinomial classification, showcasing its adaptability to a broader range of prediction problems. We propose an estimator, called `multiPPI++`, used to draw valid statistical inference using predictions from any arbitrary black-box machine learning model given access to a small subset of ground truth labels for multinomial data.

The Workflow for Valid Inference Using `multiPPI++` for VA Narratives.

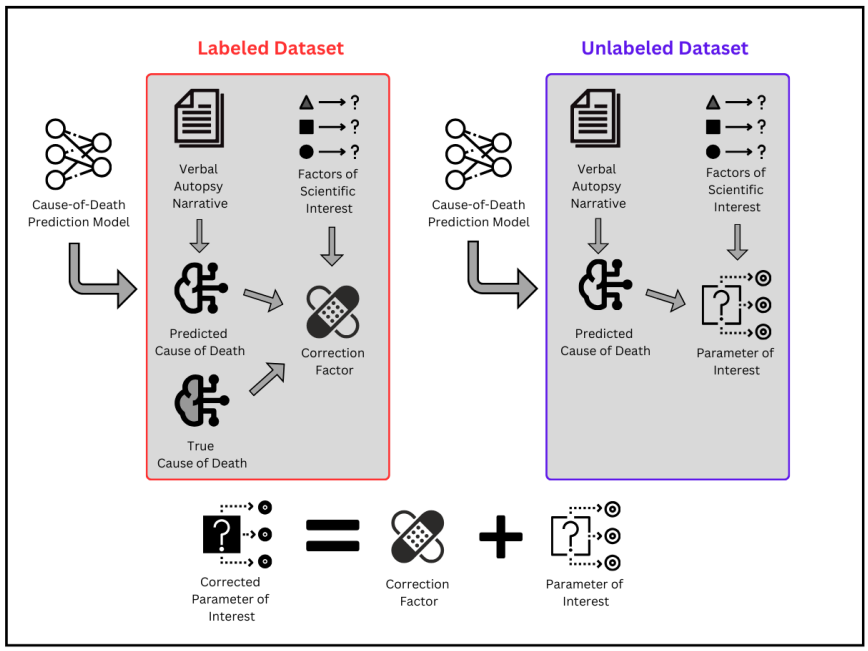


Figure 4.1: Overview of `multiPPI++` correction. Ground truth labels and predicted labels are used separately to perform the same inference task in Domain A. We use the difference between these estimates as a correction factor in Domain B where ground truth labels are not available.

4.2 Population Health Metrics Research Consortium Narratives

The Population Health Metrics Research Consortium (PHMRC) dataset was collected in 2005 to provide a high quality, standardized dataset that includes ground truth COD labels from traditional autopsies, structured questionnaire responses, and free-text open narrative responses (Murray et al., 2011). It is one of a very small number of VA datasets with clinically validated causes. These data were curated from six sites across four countries: (*Andhra Pradesh and Uttar Pradesh, India; Bohol, Philippines; Dar es Salaam and Pemba Island, Tanzania; and Mexico City, Mexico*). An example PHMRC narrative follows: *"the deceased had been burnt and had lost mental balance and died within 1.5 hours of the accident"* (Ground Truth from Traditional Autopsy: **Fires**).

We focus our analysis on adult deaths, excluding those under 12 years of age, in the de-identified narratives released in Flaxman et al. (2018). CODs are categorized into five broader groups: non-communicable diseases, communicable diseases, external causes, maternal causes, and AIDS or tuberculosis (AIDs-TB), based on International Classification of Diseases Tenth Revision (ICD-10) codes (McCormick et al., 2016). See Appendix 4.6.1 for the full mapping from ICD-10 codes to COD labels.

Under these five broad classes, there still remains considerable heterogeneity in the COD distribution between sites. Figure 4.2 illustrates this phenomenon. While "non-communicable" diseases (e.g., various cancers, COPD, stroke) are the most common COD across all six sites, the second most common cause is site-specific. Namely, two sites identify communicable diseases (e.g., malaria, pneumonia) as the second most common COD, while external causes (e.g., suicide, homicide, traffic accidents) are reported in three sites and AIDs-TB in one site. In addition, the difference in rates between the first and second most common CODs varies considerably, with, for example, a 5:4 ratio in Pemba versus a 6:1 ratio in Mexico.

These observations suggest that while information from one site may provide some insights into another, caution is warranted when extrapolating NLP model performance from one location to all others. In practice, this lack of transportability between sites poses a significant public health challenge and is well-documented in the VA literature on predicting COD (e.g., see McCormick et al., 2016). However, without a means of addressing this issue for downstream statistical inference, it may be necessary to establish COD data collection systems at each new site of interest, an endeavor that would likely prove prohibitively expensive. We address this explicitly in Section 4.3.2.

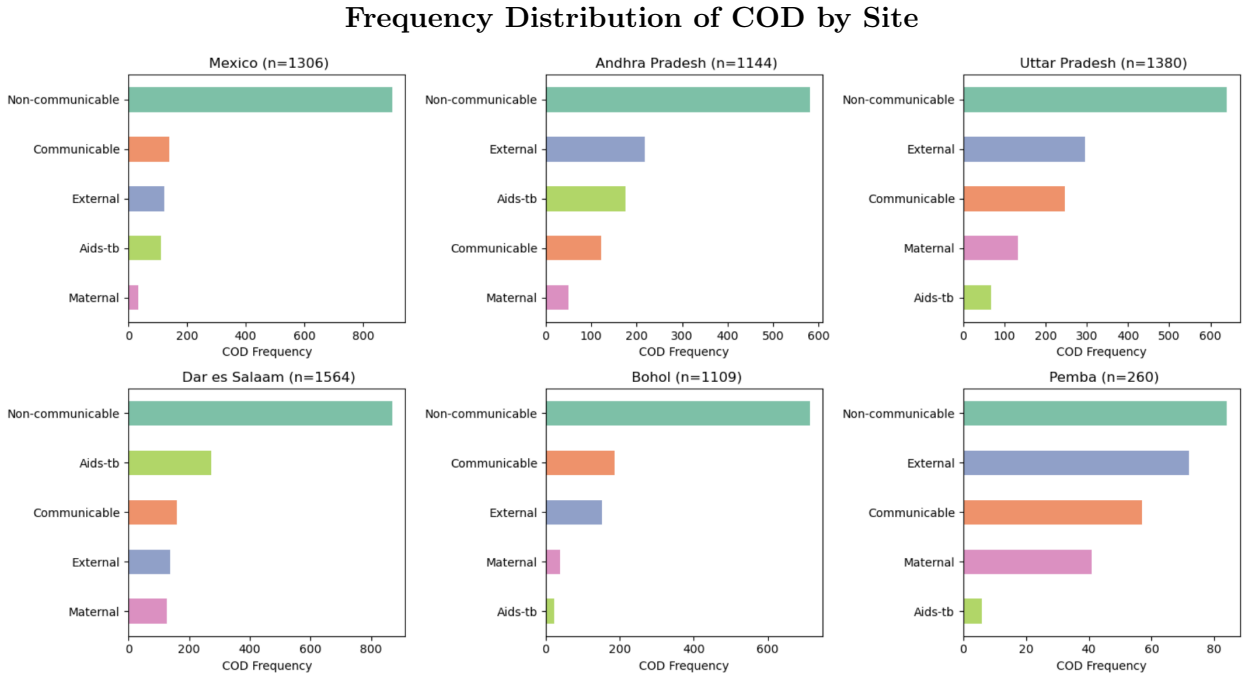


Figure 4.2: While non-communicable disease is the most common COD in each site, relative prevalence of each COD varies considerably.

4.3 Proposed Analytic Workflow

4.3.1 NLP for VA Narratives

The first aim of this work is to assess the predictive performance of contemporary NLP methods for classifying COD from free-text narratives, rather than structured interviews. To predict COD from the free-text narratives, we analyze a range of NLP techniques, including various methods based on a bag-of-words (BoW) representation and more sophisticated pre-trained models such as Bidirectional Encoder Representations from Transformers (BERT; Devlin et al., 2018) and large language models (LLMs) such as the recently popular Generative Pre-trained Transformer (GPT model; Lai et al., 2023).

To generate BoW representations of VA narratives, we employed `scikit-learn`'s `CountVectorizer`, `LabelEncoder`, and `TfidfVectorizer` (Kramer, 2016; Pedregosa et al., 2011), as well as `nltk`'s `word_tokenize` (Millstein, 2020) for data pre-processing. After processing, the initial token count of 556, 713 was reduced to 218, 804. We then applied three supervised classifiers to

predict COD using the BoW representations and ground truth cause of death labels: Support Vector Machine (SVM), K Nearest Neighbors (KNN), and Naive Bayes (NB). SVM was implemented with a third-degree linear kernel and $C = 1.0$, while KNN used cosine distance and 9 nearest neighbors, chosen by incrementing from $K = 3$ until accuracy saturation. The BERT$_{BASE}$ encoder was configured with five transformer layers, to include two input layers of size 256, a main layer of size 768, an intermediate layer of size 512, and a fine-tuning output layer of size 5. As the VA narratives in our data were de-identified, we were able to further assess the predictive performance of OpenAI's GPT models (GPT-3.5, GPT-3.5-turbo, GPT-4, and GPT-4-32K Biswas, 2023). After exploratory study, we selected GPT-4-32K for its larger context window. We used a zero-shot prompt (see Figure 4.3) with a temperature set to zero for all tasks. For all of the methods, we split the data into an 80% training set and 20% testing set, where we held out the true COD labels in the testing set to assess each method's predictive accuracy and F1-score. For BERT$_{BASE}$, training was conducted over two epochs.

We faced a series of challenges in engineering the LM prompt, namely, coercing the LM to consistently return outputs constrained to our exact list of output classes [communicable, non-communicable, external, maternal, aids-tb]. Earlier iterations of the prompt yielded responses such as "the patient died due to external causes." While perhaps semantically correct, validating outputs such as this at scale is intractable and would require additional steps (regex, fuzzy string matching, etc.). Instead, we found that including our desired output classes between `<option>` html tags and explicitly asking the LM to return only a choice from the list of options in the prompt improved results considerably. Future work may include few shot prompting with select domain-specific examples or retrieval augmented generation from a larger database of narratives.

GPT-4 Zero-Shot Prompt Used for COD Prediction

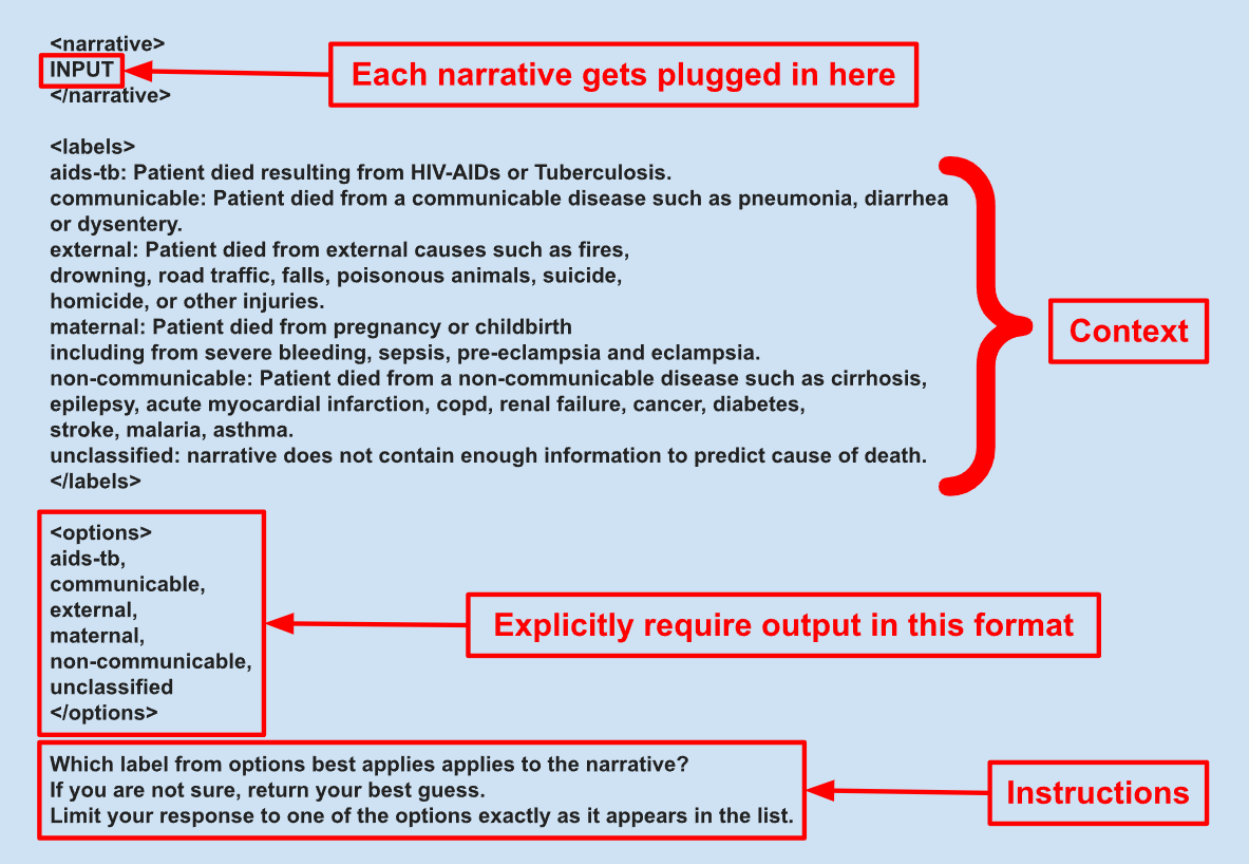


Figure 4.3: The zero-shot prompt explicitly tags the VA narrative, provides minimal context for each COD label, lists an explicit set of COD options to coerce a constrained output, and provides direct instructions pointing to the `<narrative><label><option>` tags.

4.3.2 Valid Statistical Inference with `multiPPI++` Correction

The second aim of this work addresses the challenge of conducting valid downstream inference on covariates associated with predicted COD labels. As long as the NLP model is not always 100% accurate, uncritical inference using NLP predictions will always contain bias, which we must correct for before trusting the results. Note that while the PHMRC dataset is complete with ground truth COD labels from traditional, physical autopsies, this is often not practical due to resource constraints. Particularly in vulnerable areas without healthcare infrastructure, COD is abstracted solely from VA without access to means of verification. Recent works have shown that reifying algorithmically-derived outcomes (here,

COD) for downstream inference can lead to biased coefficient estimates and anti-conservative inference (Wang et al., 2020; Angelopoulos et al., 2023a,b; Miao et al., 2023; Egami et al., 2023; Ogburn et al., 2021; Hoffman et al., 2024).

To address this, we employ a recent technique known as prediction-powered inference (`PPI`; Angelopoulos et al., 2023a) and its extension, `PPI++` (Angelopoulos et al., 2023b) and derive its appropriate form in the context of multiclass logistic regression. These methods operate by using a small subset of ground truth COD labels to *rectify* the coefficient estimates and standard errors from a regression based on predicted COD labels. To that end, assume that our data are comprised of n *labeled* (l) observations, $\mathcal{L} = \{(Y_{l,i}, \hat{Y}^f_{l,i}, X_{l,i}, Z_{l,i}); i = 1, \ldots, n\}$, and N *unlabeled* (u) observations, $\mathcal{U} = \{(\hat{Y}^f_{u,i}, X_{u,i}, Z_{u,i}); i = n+1, \ldots, n+N\}$, where Y denotes the true cause of death, $\hat{Y}^f$ denotes the cause of death predicted from an NLP function, $f(Z)$, X denotes the covariates of interest, and Z denotes the VA text narratives. In statistical inference, one can frame estimating statistical parameters, θ, as minimizing an objective function of the form:

$$\theta^{\ Ground\ Truth} = \underset{\theta \in \mathbb{R}^d}{\arg\min}\, \mathbb{E}[l_\theta(X, Y)], \tag{4.1}$$

where $l_\theta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is a convex loss function in $\theta \in \mathbb{R}^d$. In our case, the loss function $l_\theta(X, Y)$ is defined as the negative log conditional likelihood of the multinomial logistic regression. In practice, replacing the observed COD, Y, with $\hat{Y}^f$ yields what we refer to as a *Naive* estimate:

$$\theta^{Naive} = \underset{\theta \in \mathbb{R}^d}{\arg\min}\, \mathbb{E}[l_\theta(X, \hat{Y}^f)]. \tag{4.2}$$

The Naive estimate relies heavily on the assumption that the NLP model's predictions closely resemble the true CODs.

`PPI++`, and our extension, `multiPPI++`, seeks to utilize all available information from the

unlabeled and labeled data, by minimizing the following objective function:

$$\theta_{\lambda}^{PPI++} = \underset{\theta \in \mathbb{R}^d}{\arg\min} \left\{ \mathbb{E}[l_\theta(X_l, Y_l)] + \lambda \left(\mathbb{E}[l_\theta(X_u, \hat{Y}_u^f)] - \mathbb{E}[l_\theta(X_l, \hat{Y}_l^f)] \right) \right\}, \tag{4.3}$$

where $\lambda \in [0, 1]$ is a tuning parameter which controls how much weight is placed on using the real labels versus the predicted labels. Note that this estimator combines the ground truth estimator from the labeled subset, which is statistically valid but may lack sufficient sample size, with the Naive estimator, which provides additional information and data from the unlabeled subset.

We extend the application of the `PPI++` approach from its logistic regression example to the multinomial case. Angelopoulos et al. (2023b) mention this is possible, but do not provide details of the derivation, and their model has overparameterization issues. We instead derive a different parameterization that avoids identifiability issues, and we complete the derivation of `PPI++` correction for multiclass logistic regression. We refer to this variant as "`multiPPI++`", where, for the multinomial classification problem with K classes and outcomes $Y_i \in \{0, \ldots, K-1\}$, our loss function takes the form

$$l_\theta(X, Y) = -\frac{1}{n} \sum_{i=1}^{n} X_i^T \theta_{Y_i} + \log \left(\sum_{k=1}^{K-1} e^{X_i^T \theta_k} \right).$$

We present the `multiPPI++` adjustment procedures in Algorithm 1. See Appendix 4.6.2 and 4.6.3 for complete derivations.

4.4 Inference with COD Predicted from VA Narratives

4.4.1 Experimental Setup

To study the performance of the proposed methods, we designed an experiment in which we synthetically removed ground truth COD labels from the PHMRC dataset and replaced them with predicted COD from the NLP methods detailed above. Specifically, for each of

the the six sites in the PHMRC dataset, we trained the NLP methods on the other five sites and used the resulting models to predict COD in the sixth site (see Figure 4.4).

We compared the predicted outcomes to the withheld, true outcomes for the sixth site in terms of predicted accuracy and F1-score. We then carried forward these predictions to a downstream inferential model in which we regressed site-specific COD on the age of the decedent, to illustrate the performance of the `multiPPI++` estimator. For each site, we retained 20% of the labeled outcomes and replaced the remaining 80% with the NLP predictions, to mirror the *labeled* and *unlabeled* subsets of the data necessary for the `multiPPI++` approach. We compared the results of the proposed estimator to two additional estimators: (i) the 'ground truth' estimator, which utilized the full set of true COD labels, and (ii) the 'Naive' estimator, which used only the predicted COD, treating them as if they were observed. Note that, in many practical settings, the ground truth estimator is not feasible, but it is included here as a point of reference. Further,

due to the variability in data between sites and the fact that the NLP model was not trained on this particular site, we anticipate that the regression coefficients estimated using the NLP-predicted COD will differ from those estimated using the true COD.

We exemplify one representative set of experiments in which the Uttar Pradesh site is excluded from the NLP model training and used as the validation site to assess the NLP models' predictive performance and the downstream inferential results.

4.4.2 NLP Prediction Results

First, using the "leave one site out" procedure described in Figure 4.4, we trained each NLP model to predict COD from narratives on five sites, with the sixth site reserved as a validation set to use for assessing the methods. As evidenced by the COD distribution displayed in Figure 2, we are dealing with considerable class imbalance within and between sites. We address this in a sensitivity analysis (see Figures 29-54 in A.6) where we train each model on a stratified sample with an 80/20 split for each class to ensure an equal proportion of

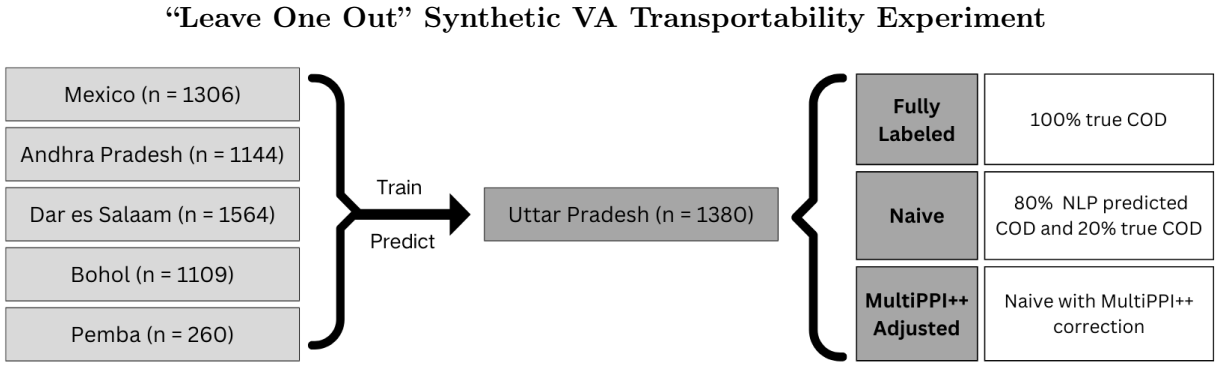


Figure 4.4: The classic and BERT models are trained with VA narratives from five other sites and used to predict COD from Uttar Pradesh narratives. The zero shot GPT-4 model predicts COD without any site specific narrative fine-tuning. Ground truth, Naive and `multiPPI++` corrected inference is performed using these Uttar Pradesh predictions.

training examples for each class. The results from this analysis are not substantively different from our main results where we do not account for the class imbalance in our modeling. Alternative resampling methods such as SMOTE, ADASYN or Tomek Links (Chawla et al., 2002; He et al., 2008; Tomek, 1976) could be applied to correct for the imbalanced distribution of COD. However, even in the absence of such considerations, we show that multiPPI++ still obtains valid estimates in downstream inference using predictions from imperfect NLP modeling. This extensibility of our approach is one of the advantages of the multiPPI++.

We focus our subsequent analysis one representative site, Uttar Pradesh, but share results from the other sites in the Appendix. These results for Uttar Pradesh are summarized in Table 4.1 and Figure 4.5, which contain the four models with highest accuracy. All models scored between 0.60-0.75 accurate, with GPT-4 yielding the highest accuracy (0.75). F1-scores were more uniform, with all but KNN getting F1-scores between 0.71-0.74. KNN scored 0.66. These results seem to point at GPT-4 either being slightly better or nearly equivalent to the other NLP methods we tested. GPT-4's predictions matched the performance of the best examples from the VA literature, despite using only using only the VA text narratives and not the structured questionnaire (James et al., 2011; Murray et al., 2014; Serina et al., 2015).

A unique feature of LLMs compared to other prediction algorithms is that they have the capacity, and in fact the tendency, to output class labels that do not exist in the ground

Metrics	BoW w/ SVM	BoW w/ KNN	BoW w/ NB	BERT	GPT-4
Accuracy	0.68	0.63	0.60	0.55	0.75
F1-Score	0.74	0.66	0.72	0.71	0.73

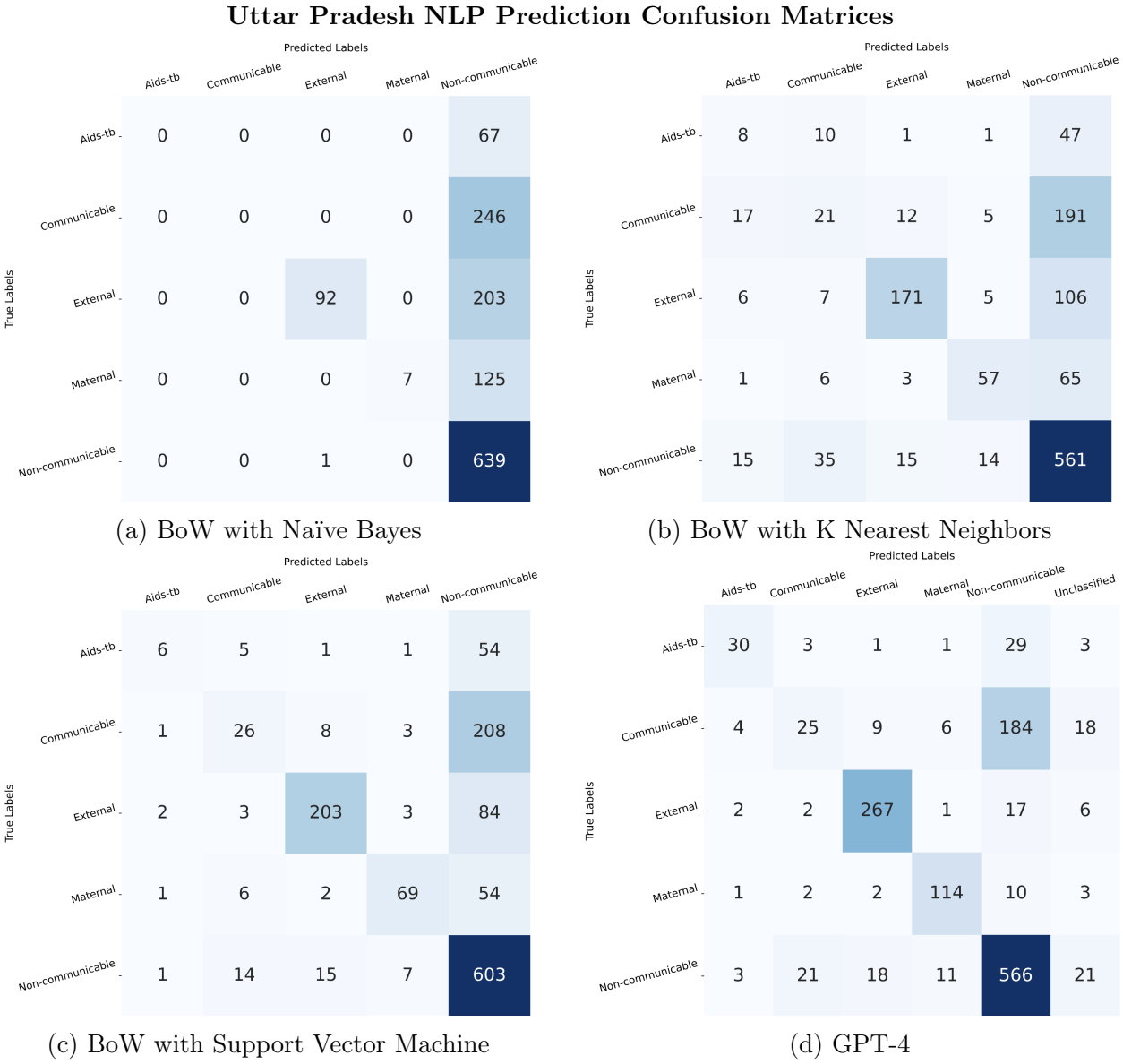
Table 4.1: GPT-4 is far more accurate than competing NLP models and is among the best in terms of F1-score. Using only text narratives, GPT-4 performs on par with the best predictions in the VA literature, which are produced with much richer data in the form of structured questionnaires.

labels. In our analysis, GPT-4 predicted 1503 out of 6763 COD labels as "unclassified". Upon manual review, these 1503 narratives contained no useful information for COD inference. The modal "unclassified" narrative was "the interviewee thanked the interviewer" (n=178). While the accuracies appear similar, GPT-4 correctly identified a flaw in the data cleaning in the PHMRC dataset, which suggest that the accuracies among the other methods may be overinflated. For each of the combination of six sites and five NLP model, we generated predictions and used this to generate regression coefficients.

For the sake of brevity, we present the four models with the best performance for the Uttar Pradesh site, NB, KNN, SVM, and GPT-4. Figure 4.6 displays point estimates and 95% confidence intervals for the ground truth, Naive and `multiPPI++` corrected inference.

4.4.3 Inferential Model Results

In the case of NB and SVM, which had some of the lower F1-scores, there is a large discrepancy between the ground truth and the predicted values, demonstrating that replacing 80% of the causes of death with predicted causes of death yields significantly different regression coefficients for the age of the decedent. `multiPPI++` correction rectifies estimates back to baseline, demonstrating its ability to account for (i) model inaccuracy and (ii) transportability bias. Interestingly, we observe that the Naive regression coefficients from KNN and GPT-4 are not as distant from the ground truth as the other two methods, even though KNN had the worst F1-score. Figure 4.7 shows the association between F1-score and



(a) BoW with Naïve Bayes

(b) BoW with K Nearest Neighbors

(c) BoW with Support Vector Machine

(d) GPT-4

Figure 4.5: For VA narratives from Uttar Pradesh, most of the COD misclassifications are assigned non-communicable COD label. Naive Bayes mostly predicts non-communicable and achieves 0.60 accuracy in-part because non-communicable is overwhelmingly the most common ground truth COD.

`multiPPI++` λ, which also fail to show any significant association. This implies that F1-score alone may not be a good predictor of the quality of downstream statistical inference.

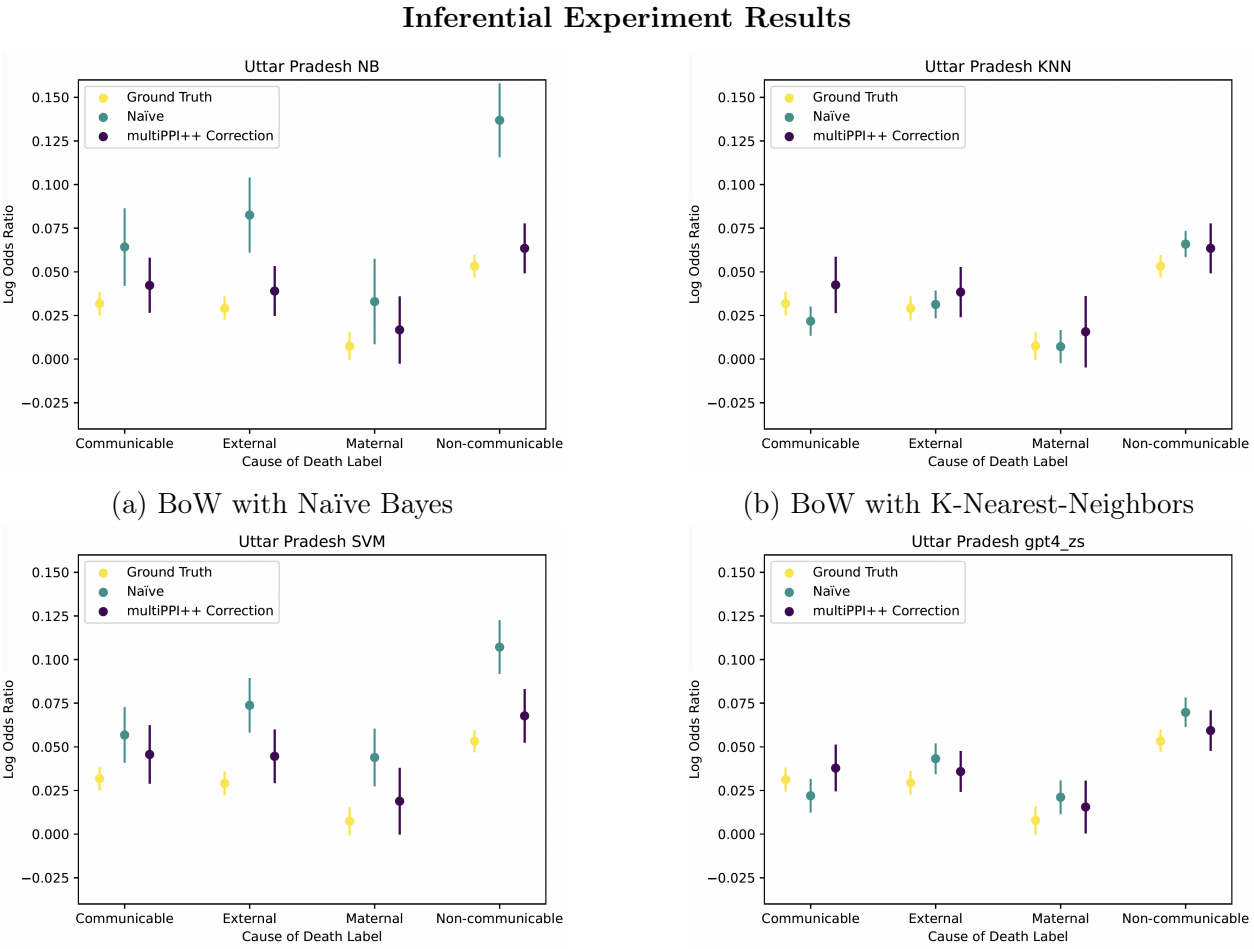
As for the widths of the confidence intervals, we observe that ground truth naturally had the thinnest, as it had the largest sample size and used real data points. On the other hand, the Naive estimate and `multiPPI++` correction show similar confidence interval widths. Since the

widths of the confidence interval for `multiPPI++` is the sum of the widths of the Naive interval plus the width of the correction term, this seems to indicate that performing the `multiPPI++` correction does not seem to add a significant amount of uncertainty. This demonstrates that it can correct for transportability bias without much additional information cost. This is further evidenced by the seeming lack of association between the tuning parameter λ, which weights the use of the predicted outcomes in the downstream inferential model and the accuracy of the predictions. As λ is a function of both the outcome and the associated features, we conclude that better predicted outcomes carry no additional useful information for parameter estimation.

Figure 6 demonstrates that multiPPI++ improves the accuracy of statistical estimates using a post-facto correction term estimated using relatively small amounts of labeled data. As this is applied after model predictions are made, multiPPI++ does not improve or degrade prediction accuracy. This procedure provides a correction for the biased statistical inference parameters and the deflated uncertainty intervals around them, given NLP produced estimates of likely COD from VA narrative predictions. This is a nuanced yet important distinction as these downstream statistical inference parameters are what describe the relationship between covariates, such as age, and cause of death distributions. This same intuition extends to other domains where the end goal is to use inference to guide decision making where NLP predictions are used in place of ground truth labels.

4.5 Discussion

This paper uses a variety of NLP models to predict COD using VAs and, subsequently, develops a statistical method for valid inference using these predicted CODs. This research not only contributes to the advancement of predictive modeling in public health, but also underscores the significance of parameter interpretation in guiding decisions within diverse and dynamic health scenarios.



(a) BoW with Naïve Bayes

(b) BoW with K-Nearest-Neighbors

(c) BoW with Support Vector Machine

(d) GPT-4-32k

Figure 4.6: Ground truth, Naive estimation, and `multiPPI++` correction of estimation of log odds ratios for multinomial regression for age in Uttar Pradesh. `multiPPI++` correction both recovers point estimates and reflects the increased uncertainty corresponding to the use of predicted COD rather than known COD with wider confidence intervals.

Based on our research, there are several points that we believe are worthy of further investigation. First, we noticed that better predictive accuracy in the COD classifications did not necessarily translate to improved parameter estimation when used as a generative model for downstream statistical inference, at least not under the accuracy ranges examined in this paper. In situations where the user is ultimately interested in accurate parameter estimation, as is often the case in public health, care must be taken to differentiate between the most accurate model for prediction versus the most accurate model for parameter estimation, as it is possible that very cheap models such as KNN can be as useful as more expensive,

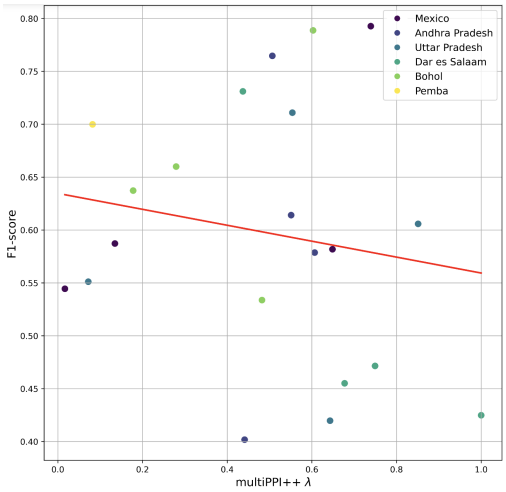


Figure 4.7: There appears no strong global relationship between `multiPPI++` λ and F1-score, though associations do emerge when subsetting by site, hinting at a Yule-Simpson effect.

newer models like GPT-4. For example, the cost to train and predict COD with BoW representations and KNN, NB and SVM was trivial, whereas performing the same task with GPT-4 cost $3000USD.

Second, the extreme class imbalance in COD drives many of our prediction results. While it is not possible in practice to directly balance across CODs, there are broader design questions that could substantially improve inference. How, for example, should a researcher spend their very limited budget to label relevant VAs? And are there other ways of categorizing deaths that provide richer information than the five categories we use? Third, there is evidence that NLP techniques such as large language models perform worse in non-English languages Navigli et al. (2023); Lai et al. (2023); Hirschberg and Manning (2015), which is a major consideration since VAs are common in non-English settings. This could introduce bias to potential VA inference in other languages (Surek-Clark, 2020). This is a limitation of our simulation study, since it does not account for differential translation quality across languages. Future work might instead leverage NLP models specifically pre-trained for multiple languages or medical text (Peskoff et al., 2021; Singhal et al., 2023).

Acknowledgements

We thank the University of Washington eScience institute for granting us Microsoft Azure computing credits used to produce all of our NLP predictions. We also thank Aidan Andronicos, Robert Fatland, Musashi Hinck, Eddie Hock, Jesse Peirce, Denis Peskoff, Nels Schimek, Sander Schulhoff, and Brandon Stewart for their help.

4.6 Chapter Appendix

4.6.1 ICD-10 COD Classification

Mapping 34 PHMRC All-Cause Mortality Labels to Five Broad COD Labels		
ICD-10 Code	**All-Cause Mortality Label**	**Broad COD Label**
K74	cirrhosis	non-communicable
G40	epilepsy	non-communicable
J18	pneumonia	communicable
J44	copd	non-communicable
I21	acute myocardial infarction	non-communicable
X00-X09	fires	external
N17-N19	renal failure	non-communicable
C34	lung cancer	non-communicable
O00-O99	maternal	maternal
W65-W74	drowning	external
I00-I99	other cardiovascular diseases	non-communicable
B20-B24	aids	aids-tb
Varies	other non-communicable diseases	non-communicable
W00-W19	falls	external
V01-V99	road traffic	external
E10-E14	diabetes	non-communicable
A00-B99	other infectious diseases	communicable
A15-A19	tb	aids-tb
X60-X84	suicide	external
Varies	other injuries	external
C53	cervical cancer	non-communicable
I60-I69	stroke	non-communicable
B50-B54	malaria	communicable
J45	asthma	non-communicable
C18-C20	[illegible]	[illegible]

4.6.2 PPI and PPI++: overview

We formalize our problem setting while reviewing the prediction-powered inference PPI and PPI++ approach (Angelopoulos et al., 2023a,b). The goal of PPI is to estimate a d-dimensional parameter $\boldsymbol{\theta}$ with n labeled data points $(X_{li}, Y_{li}) \overset{\text{iid}}{\sim} \mathbb{P}, i \in \{1, \ldots, n\}$, and N unlabeled data points $X_{ui} \overset{\text{iid}}{\sim} \mathbb{P}_X, i \in \{1, \ldots, N\}$. Additionally, there is a black-box model f that uses the features to predict the outcomes based on labeled and unlabeled data, denoted as $\hat{Y}_l^f = (\hat{Y}_{l1}^f, \ldots, \hat{Y}_{ln}^f)$ and $\hat{Y}_u^f = (\hat{Y}_{u1}^f, \ldots, \hat{X}_{uN}^f)$. The prediction rule f is assumed to be independent of the observed data. While PPI performs inference on a correction term to incorporate these black-box predictions; PPI++ improves the methodology to be more computationally efficient and develops a modification that adapts to the accuracy of the prediction rule f by introducing an additional parameter λ that interpolates between classical and prediction-powered inference.

The starting point of PPI and PPI++ approach is to define the population losses:

$$L(\theta) = \mathbb{E}[l_\theta(X, Y)], \text{ and}$$

$$L^f(\theta) = \mathbb{E}[l_\theta(X, \hat{Y}^f)].$$

It is recognized that

$$L^{f_l}(\theta) = \mathbb{E}[l_\theta(X_l, \hat{Y}_l^f)] = \mathbb{E}[l_\theta(X_u, \hat{Y}_u^f)] = L^{f_u}(\theta).$$

Therefore the "rectified" loss is defined as:

$$L^{PPI}(\theta) = L_n(\theta) + L_N^{f_u}(\theta) - L_n^{f_l}(\theta)$$

where

$$L_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} l_\theta(X_{li}, Y_{li}),$$

$$L_n^{f_l}(\theta) = \frac{1}{n}\sum_{i=1}^{n} l_\theta(X_{li}, \hat{Y}_{li}^f)), \text{ and}$$

$$L_N^{f_u}(\theta) = \frac{1}{N}\sum_{i=1}^{N} l_\theta(X_{ui}, \hat{Y}_{ui}^f).$$

The rectified loss leads to the prediction-powered point estimate:

$$\hat{\theta}^{PPI} = \arg\min_{\theta\in\mathbb{R}^d} L^{PPI}(\theta).$$

`PPI` constructs the confidence set $C_\alpha^{PPI} = \{\theta \in \mathbb{R}^d \text{ s.t. } \mathbf{0} \in C_\alpha^\nabla(\theta)\}$ for true parameter θ^* by forming $(1-\alpha)$-confidence sets, $C_\alpha^\nabla(\theta)$, for the gradient of the loss $\nabla L^{PPI}(\theta)$. This approach works well for a fixed θ but suffers from computational issue when forming the confidence set for every $\theta \in \mathbb{R}^d$. `PPI++` derives the asymptotic normality about $\hat{\theta}$. This result allows the construction of $(1-\alpha)$-confidence intervals $C_\alpha^{PPI} = \{\hat{\theta}_j^{PPI} \pm z_{1-\alpha/2} \times \hat{\sigma}_j/\sqrt{n}\}$, where $\hat{\sigma}_j^2$ is a consistent estimate of the asymptotic variance of $\hat{\theta}_j^{PPI}$ and z_q is the q-quantile of standard normal distribution. Similar confidence intervals centered around $\hat{\theta}^{PPI}$ can be constructed when the inference is performed on the whole θ^* vector rather than a single coordinate θ_j^*, which results in far more efficient algorithm.

Moreover, `PPI++` generalizes `PPI` to allow the inference to adapt to the accuracy of the supplied predictions, yielding estimations that are never worse than the classical inference. This approach is called "power tuning" by incorporating a tuning parameter $\lambda \in [0, 1]$ in the rectified loss, leading to the following estimator:

$$\hat{\theta}_\lambda^{\texttt{PPI++}} = \arg\min_{\theta} L_\lambda^{\texttt{PPI++}}(\theta),$$

where

$$L_\lambda^{\texttt{PPI++}}(\theta) = L_n(\theta) + \lambda(L_N^{f_u}(\theta) - L_n^{f_l}(\theta)).$$

`PPI++` recovers the original `PPI` approach when $\lambda = 1$ and reduces to classical inference when $\lambda = 0$. One can adaptively choose a data-dependent tuning parameter $\hat{\lambda}$ to maximize

the estimation and inferential efficiency.

Algorithm 1: `multiPPI++`: Prediction-powered inference for Multinomial Classification

Input: labeled K-category COD data $\{(X_{li}, Y_{li})\}_{i=1}^{n}$, unlabeled data $\{X_{ui}\}_{i=1}^{N}$, NLP model f, significance level $\alpha \in (0,1)$, coefficient index $j \in [d(K-1)]$

1. Optimally select tuning parameter $\hat{\lambda}$ `// set tuning parameter`
2. $\hat{\theta}^{\mathtt{m}}_{\hat{\lambda}} = \underset{\theta \in \mathbb{R}^{d(K-1)}}{\arg\min}\ L^{\mathtt{m}}_{\hat{\lambda}}(\theta)$ `// multiPPI++ estimator`
3. $\hat{H} = \frac{1}{N+n}(\sum_{i=1}^{n} \psi''(X_{li}^T \hat{\theta}^{\mathtt{m}}_{\hat{\lambda}}) X_{li} X_{li}^T + \sum_{i=1}^{N} \psi''(X_{ui}^T \hat{\theta}^{\mathtt{m}}_{\hat{\lambda}}) X_{ui} X_{ui}^T)$, where $\psi(\theta, x) = \log\left(\sum_{k=1}^{K-1} e^{x^T \theta_k}\right), \theta_k \in \mathbb{R}^d$ `// empirical Hessian`
4. $\hat{\Sigma} = \hat{H}^{-1}(\frac{n}{N}\hat{V}_f + \hat{V}_\Delta)\hat{H}^{-1}$, where

   $\hat{V}_f = \hat{\lambda}^2 \widehat{\mathrm{Cov}}_{N+n}((\psi'(X_{li}^T \hat{\theta}^{\mathtt{m}}_{\hat{\lambda}}) - \hat{Y}_{li}^f))X_{li})$ and
   $\hat{V}_\Delta = \widehat{\mathrm{Cov}}_n((1-\hat{\lambda})(\psi'(X_{li}^T \hat{\theta}^{\mathtt{m}}_{\hat{\lambda}}) + (\hat{\lambda}\hat{Y}_{li}^f - Y_{li})X_{li})$ `// covariance estimator`

Output:
Prediction-powered point estimates $\hat{\theta}^{\mathtt{m}}_{\hat{\lambda}}$ and confidence interval
$\mathcal{C}^{\mathtt{m}}_{\alpha} = \left(\hat{\theta}^{\mathtt{m}}_{\hat{\lambda},j} \pm z_{1-\alpha/2}\sqrt{\hat{\Sigma}_{jj}/n}\right)$ for coordinate j

4.6.3 `multiPPI++`

In K-class multinomial classification problem with outcomes $Y_i \in \{0, \ldots, K-1\}$ and covariates $X_i \in \mathbb{R}^d$, the parameter of interests θ is a $d(K-1)$-dimensional vector whose k-th block $\theta_k \in \mathbb{R}^d$ of components represent parameters for class $k \in \{1, \ldots, K-1\}$ excluding the reference class to avoid overparameterization. The loss function in classical inference takes the form

$$l_\theta(X, Y) = -\frac{1}{n}\sum_{i=1}^{n} X_i^T \theta_{Y_i} + \log\left(\sum_{k=1}^{K-1} e^{X_i^T \theta_k}\right),$$

Defining

$$\psi(\theta, x) = \log\left(\sum_{k=1}^{K-1} e^{x^T \theta_k}\right),$$

and the `multiPPI++` loss is given by

$$\begin{aligned} L_\lambda^{\mathtt{m}}(\theta) &= L_n(\theta) + \lambda(L_N^{f_u}(\theta) - L_n^{f_l}(\theta)) \\ &= -\frac{1}{n}\sum_{i=1}^{n}(X_{li}\theta_{Y_{li}} - \psi(\theta, X_{li})) \\ &- \frac{\lambda}{N}\sum_{i=1}^{N}(X_{ui}^T\theta_{\hat{Y}_{ui}^f} - \psi(\theta, X_{ui})) + \frac{\lambda}{n}\sum_{i=1}^{n}(X_{li}^T\theta_{\hat{Y}_{li}^f} - \psi(\theta, X_{li})) \end{aligned}$$

Angelopoulos et al. (2023b) shows that the parameter λ can be optimally tuned to minimize the asymptotic variance of the prediction-powered estimate, leading to better estimation and inference than both the classical and `PPI` strategies. Particularly, in finite samples, we can obtain the plug-in estimate

$$\hat{\lambda} = \frac{1}{2(1+\frac{n}{N})} \times \frac{\mathrm{Tr}\left(\widehat{H}_{\hat{\theta}_{PPI}}^{-1}\left(\widehat{\mathrm{Cov}}_n(\nabla l_{\hat{\theta}_{PPI}}, \nabla l_{\hat{\theta}_{PPI}}^f) + \widehat{\mathrm{Cov}}_n(\nabla l_{\hat{\theta}_{PPI}}^f, \nabla l_{\hat{\theta}_{PPI}})\right)\widehat{H}_{\hat{\theta}_{PPI}}^{-1}\right)}{\mathrm{Tr}\left(\widehat{H}_{\hat{\theta}_{PPI}}^{-1}\widehat{\mathrm{Cov}}_{N+n}(\nabla l_{\hat{\theta}_{PPI}}^f)\widehat{H}_{\hat{\theta}_{PPI}}^{-1}\right)}$$

where $\hat{\theta}_{PPI} = \hat{\theta}_\lambda^{\mathtt{PPI++}}$ is obtained by taking a fixed $\lambda \in [0, 1]$.

We summarize the `multiPPI++` adjusted inference procedures for multinomial logistic regression coefficients on predicted CODs by an NLP model f against covariate X, under significance level $\alpha \in (0, 1)$ in Algorithm 1.

4.6.4 Sensitivity Analysis

In a sensitivity analysis, we varied our data splitting strategy, which revealed nuanced insights regarding the behavior of the classifier. Notably, when maintaining the split proportions (80/20) within each COD, the `PPI++` classifier exhibited minimal utilization of the labeled data, as indicated by the relatively small values of λ. Conversely, when employing an 80/20 split on the entire dataset, this lead to a significant information loss for the minority classes, resulting in larger λ values for certain splits. Since an (80/20) on the whole dataset more closely resembles what one would see in reality on a prospective study while an (80/20) split by COD is resembles a retrospective study, this illustrates the importance of purposefully choosing an appropriate data splitting strategy to match the desired analysis.

4.7 Ethics Statement

As this paper covers the study of the lives and deaths of real human beings, the ethical aspect of this investigation was considered of paramount importance. First, to ensure that any potential personally identifiable information was not mishandled, multiple layers of data security were employed, the primary means being the use of a secure Microsoft Azure cluster for the running of local models. Data concerns are always paramount. We are fortunate to be working with de-identified data, however, when working with more sensitive data there are a number of open-source LLMs which can be run locally (Llama-3, meditron, etc) to ensure safe and ethical research practices. Additionally, many institutions, public and private, buy access to local OpenAI servers. Our use of these methods is illustrative, aiming to lay a foundational understanding of how to perform valid inference with NLP predictions. When communicating with outside organizations such as OpenAI's GPT models, we ensured that all data set were from the publicly available PHMRC data, which has been previously deidentified. Finally, in accordance with the importance of studying a sensitive topic such as VAs deserve, we have done our utmost to ensure that our study's objectives and tools utilized, to the best of our abilities, gives a fair, and respectful analysis of subject matter in question so that its results may be used for the furtherment of public health goals.

4.8 Reproducibility Statement

To ensure the reproducibility of our results, we have published our code repository with our entire pipeline - NLP training, predicting, inference and visualizations - to Github. Please direct questions about data or modeling via email to `avisokay` or `fansx` at `uw.edu`. All data is available in the data folder in the Github repository.

Chapter 5

Conclusion

This dissertation began with a simple observation: we have a tendency to understand the complex social world through categories. While it can be at times difficult to recognize, such categories are not neutral containers for natural, pre-existing “truths", but artifacts of social construction through measurement systems that each carry their own assumptions, historical legacies, and limitations. Through three empirical chapters focusing on urban space, US population obesity, and global mortality estimation, I have shown that the substantive conclusions we draw from any empirical analyses crucially depend upon the alignment of our conceptual constructs of interest and the measurement systems we use to make those concepts legible to inquiry. This concluding chapter summarizes the contributions of each substantive chapter, reflects on their shared limitations, and identifies some directions for future work.

5.1 Contributions

5.1.1 Comptational Methods for Urban Sociology

Chapter 2 contributes to a long tradition in urban sociology that treats neighborhoods as socially produced rather than administratively fixed entities. Operationalizing this contestation empirically has proved challenging: census tracts, zip codes, or administrative neighborhood boundaries offer scale and consistency at the cost of treating boundaries as

fixed, while ethnographic approaches offer depth of the particular at the expense of scale. Through a novel application of language model annotation to a corpus of hundreds of thousands of Craigslist rental advertisements, this chapter demonstrates a feasible way to identify unit-level neighborhood identity. These neighborhood claim labels suggest nuanced contours of social space that better capture how listing agents conceptualize urban space. The LLM annotation pipeline I present, validated against manual annotation and benchmarked against conventional entity recognition with string-matching, offers a generalizeable tool for computational social scientists interested in recovering latent spatial claims from unstructured text at scale.

5.1.2 Rethinking the US “Obesity Epidemic"

Chapter 3 contributes to the demographic and epidemiological literatures on population health as it relates to fatness by subjecting a foundational claim – that US obesity has been consistently on the rise for decades – to an empirical validity test using the same data used to produce the CDC's official estimates of obesity prevalence. The primary contributions are both methodological and substantive. First, through non-parametric imputation with principled uncertainty propagation I am the first to estimate a DXA-based obesity trend that is both representative of US adults, including those aged 60+, and spans the entire modern NHANES era since 1999. Second, I show that the two most widely used anthropometric measures of obesity (BMI and WC) and the clinical reference standard for adiposity (DXA) have been on diverging trajectories during the study period.

The finding that BMI-based obesity rose 13 percentage points while DXA-based obesity stayed at the 1999 level for the subsequent two decades challenges the evidence-based foundation on which a substantial research apparatus (from clinical guidelines to pharmaceutical investments to public health interventions) has been built. The subsequent demographic decomposition analyses reveal the BMI rise is not an artifact of population compositional change but is broad-based within demographic cells. The same is not true, however, for

DXA-based obesity, where prevalence actually fell during this period for several groups, most notably for female respondents. These sex-stratified findings carry direct implications for studies of gender and health, and for decomposition analyses of racial disparities in obesity that rely on BMI as a valid proxy. More broadly, the chapter advances demography's methodological toolkit by demonstrating how survey-weighted multiple imputation, complex sample design adjustments, and direct standardization can be combined to recover valid population-level prevalence estimates from intermittently collected biomarker data which may be more conceptually aligned with the construct of interest, like directly measured adiposity as a screening tool for obesity.

5.1.3 Valid inference with mortality estimates

Chapter 4 contributes to global public health and to the rapidly growing applied statistical literature on inference with predicted data. In global health, verbal autopsy remains the primary tool for monitoring cause-specific mortality in low and middle income countries. Prior work has focused on the structured questionnaire portion of the VA, and has treated cause-of-death labels as an end in itself without adequately addressing the downstream inferential problem of relying on outcomes of interested that have been predicted rather than directly observed. This chapter fills two important gaps. First, this is among the first studies to demonstrate the efficacy of relying solely on the unstructured text portion from VA's for cause-specific prediction. The second is extending the PPI++ estimator to the multinomial setting (multiPPI++), facilitating calibrated statistical inference with predicted data. By requiring only a small set of verified labels, this method offers promise to applied researchers operating with constrained budgets for primary-data collection. This also has the potential to reduce the burden on VA respondents, by moving from a tedious and repetitive structured questionnaire to a more adaptive and less structured process.

5.2 Limitations and Future Directions

5.2.1 Neighborhoods

The neighborhood analysis presented here is necessarily partial in that it fundamentally captures a single perspective on the social construction of urban space. Craigslist rental advertisement listings as a data source reflect how listing agents (e.g. landlords and/or property managers) strategically represent properties to prospective tenants. This is an important and consequential perspective to understand, since listing agents mediate access to housing and their spatial understandings can shape the information environment in which renters search. However, renters likely occupy their own spatial imaginaries, shaped by personal experience, social networks, etc. Long-term residents of a city or established community organizations construct and defend neighborhood identities through different channels than Craigslist advertisements, whether it be through local press, social media, civic associations, or art. These are alternative potential data sources which would better capture alternative perspectives, and perspectives derived from these data may align or conflict with the perspectives of listing agents.

Beyond expanding the spatial and temporal scope of this research, a natural extension would contrast the listing-agent perspective recovered here with renter or community-stakeholder-generated spatial claims. This could be derived from reviews on platforms like Yelp, Google Maps, posts in neighborhood specific Facebook groups, Reddit communities, or responses to survey instruments that ask respondents to draw their current conceptions of neighborhood boundaries on a map. Triangulating across these data sources would allow researchers to map not just the gap between institutional and market representations of neighborhood identity as I do in Chapter 2, but a more robust ecology of competing spatial claims. The LLM annotation pipeline developed here is readily adaptable to these alternative data sources, and doing so could move the analysis of one type of actor's construction of urban space in Chicago to a more complete sociological understanding of how neighborhoods

are made, sustained, and transformed.

5.2.2 Obesity Trends

The empirical findings presented in Chapter 3 are specific to the United States and to the 1999–2018 period covered by NHANES with intermittent DXA collection. Whether the divergence between BMI and DXA trends observed here is a feature of the US context specifically or a more general property of how anthropometric proxies perform as population surveillance instruments remains an open question. The persistence of BMI as the dominant instrument for obesity surveillance in the face of mounting and extensive evidence its limitations also raises questions that go beyond what any empirical analysis can answer. BMI's entrenchment reflects not only tractability, but the institutional inertia of a patchwork of classification systems that have been embedded a myriad of social systems (e.g. clinical guidelines, insurance reimbursement qualification, public health reporting). Understanding why BMI resists displacement is a question for the sociology of science and technology studies.

5.2.3 Verbal Autopsy

Chapter 4 demonstrates the value of free-form text narratives as an important source of signal for cause-of-death classification from VA. It also reveals an important limitation of the current data infrastructure: the narratives available in our most comprehensive validation datasets were not collected with NLP-based analysis in mind. Interview protocols have historically optimized for structured questionnaire responses, with text narratives receiving less thorough consideration, yielding often times brief or formulaic narratives that underutilized the potential richness of caregiver accounts. A direct implication of this work is that VA data collection instruments should be redesigned to prioritize the elicitation of detailed and more comprehensive text narratives, even if this comes at the expense of the structured tabular data.

A related limitation is that current methods – both text based and structured, and those

which combine them to jointly estimate mortality – perform considerably better on common causes of death than on rare ones. Conditions that account for a small fraction of deaths in any given definition are underrepresented in training data, and both prediction accuracy and the statistical power of downstream inference degrade sharply. These rare causes are also often of particular import (e.g. disease outbreaks) but data limitations make it particularly difficult to design and implement timely interventions. Addressing this limitation will require coordinated investment in developing integrated methods which can operate with pooled existing data as well as expanding both the volume and context diversity of labeled VA data to improve rare cause coverage. The multiPPI++ framework developed here is designed to be robust to systematically biased prediction, but it cannot substitute for the underlying sparsity of data, and is dependent upon high quality validation data.

The three studies described in this dissertation share a common commitment to taking measurement seriously as a substantive, theoretical and empirical problem beyond simply a technical preliminary to quantitative social science. Each shows that our understanding of social phenomena hinges critically upon not just *what* we observe, but *how* those observations are made. With this dissertation I urge a renewed focus upon the epistemology of empirical measurement. Looking differently, whether with better instruments, more direct measures, or methods capable of recovering latent constructs from emerging unstructured data sources, can unsettle conclusions that have come to seem obvious. That unsettling is not an end in itself but rather a starting point. The analyses presented here creates both an obligation to be more precise about what we claim to know and an opening for more engagement with the gap between the complex messy social worlds we inhabit and the vast multiplicity of its possible representations.

Bibliography

D. B. Allison, K. R. Fontaine, J. E. Manson, J. Stevens, and T. B. VanItallie. Annual deaths attributable to obesity in the united states. *JAMA*, 282(16):1530–1538, Oct 1999. doi: 10.1001/jama.282.16.1530.

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023a. doi: 10.1126/science.adi6000.

Anastasios N. Angelopoulos, John C. Duchi, and Tijana Zrnic. PPI++: Efficient prediction-powered inference, 2023b. URL https://arxiv.org/abs/2311.01453.

E. K. Aryee, S. Zhang, E. Selvin, and M. Fang. Prevalence of obesity with and without confirmation of excess adiposity among us adults. *JAMA*, 333(19):1726–1728, May 2025. doi: 10.1001/jama.2025.2704.

Lindo Bacon, Judith S. Stern, Marta D. Van Loan, and Nancy L. Keim. Size acceptance and intuitive eating improve health for obese, female chronic dieters. *Journal of the American Dietetic Association*, 105(6):929–936, 2005.

Elizabeth H. Baker, Michael S. Rendall, and Margaret M. Weden. Epidemiological paradox or immigrant vulnerability? obesity among young children of immigrants. *Demography*, 52 (4):1295–1320, Jun 2015. doi: 10.1007/s13524-015-0404-3.

A. H. Baqui, G. L. Darmstadt, E. K. Williams, V. Kumar, T. U. Kiran, D. Panwar, V. K. Srivastava, R. Ahuja, R. E. Black, and M. Santoshama. Rates, timing and causes of neonatal deaths in rural India: Implications for neonatal health programmes. *Bulletin of the World Health Organization*, 84(9):706–713, 2006. doi: 10.2471/BLT.05.026443.

Pepita Barlow. The effect of schooling on women's overweight and obesity: A natural experiment in nigeria. *Demography*, 58(2):685–710, Apr 2021. doi: 10.1215/00703370-8990202.

Max Besbris, Ariela Schachter, and John Kuk. The unequal availability of rental housing information across neighborhoods. *Demography*, 58(4):1197–1221, 2021.

Som S. Biswas. Role of ChatGPT in public health. *Annals of Biomedical Engineering*, 51(5): 868–869, 2023. doi: 10.1007/s10439-023-03172-7.

Alberto Blanco, Alicia Perez, Arantza Casillas, and Daniel Cobos. Extracting cause of death from verbal autopsy with deep learning interpretable methods. *IEEE Journal of Biomedical and Health Informatics*, 25(4):1315–1325, 2021. doi: 10.1109/JBHI.2020.3005769.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Geoff Boeing and Paul Waddell. New insights into rental housing markets across the United States: Web scraping and analyzing Craigslist rental listings. *Journal of Planning Education and Research*, 37(4):457–476, 2017.

Geoff Boeing, Max Besbris, Ariela Schachter, and John Kuk. Housing search in the age of big data: Smarter cities or the same old blind spots? *Housing Policy Debate*, 31(1):112–126, 2021.

Magnus Borga, Jaro West, Jimmy D. Bell, et al. Advanced body composition assessment: From body mass index to body composition profiling. *Journal of Investigative Medicine*, 66(5):1–9, 2018. doi: 10.1136/jim-2018-000722.

George A. Bray and Donna H. Ryan. Evidence-based weight loss interventions: Individualized treatment options to maximize patient outcomes. *Diabetes, Obesity and Metabolism*, 23 (S1):50–62, Feb 2021. doi: 10.1111/dom.14200.

Mikael Brunila, Jack LaViolette, Sky CH-Wang, Priyanka Verma, Clara Féré, and Grant McKenzie. Toward a critical toponymy framework for named entity recognition: A case study of Airbnb in New York City. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4676–4695, Singapore, 2023. Association for Computational Linguistics.

Deb Burgard. What is "Health at Every Size"? In Esther Rothblum and Sondra Solovay, editors, *The Fat Studies Reader*, pages 42–53. New York University Press, New York, 2009. doi: 10.18574/nyu/9780814777435.003.0010.

Peter Byass, Daniel Chandramohan, Samuel J. Clark, Lucia D'Ambruoso, Edward Fottrell, Wendy J. Graham, Abraham J. Herbst, Abraham Hodgson, Sennen Hounton, Kathleen Kahn, Anand Krishnan, Jordana Leitao, Frank Odhiambo, Osman A. Sankoh, and Stephen M. Tollman. Strengthening standardised interpretation of verbal autopsy data: the new InterVA-4 tool. *Global Health Action*, 5:1–8, 2012. doi: 10.3402/GHA.V5I0.19281.

Paul Campos. *The Obesity Myth: Why America's Obsession with Weight is Hazardous to Your Health*. Gotham Books, 2004.

Ander Cejudo, Arantza Casillas, Alicia Pérez, Maite Oronoz, and Daniel Cobos. Cause of death estimation from verbal autopsies: Is the open response redundant or synergistic? *Artificial Intelligence in Medicine*, 143:102622, 2023.

Centers for Disease Control and Prevention (CDC). National Health and Nutrition Examination Survey (NHANES): Dual-Energy X-ray Absorptiometry (DXA) Procedures Manual, 2024. URL https://cdc.gov. Accessed: April 21, 2026.

Sarah E. Chasins, Maria Mueller, and Rastislav Bodik. Rousillon: Scraping distributed hierarchical web data. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, pages 963–975, 2018.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. doi: 10.1613/jair.953.

Eric Chyn and Lawrence F. Katz. Neighborhoods matter: Assessing the evidence for place effects. *Journal of Economic Perspectives*, 35(4):197–222, 2021.

Samuel J. Clark, Zehang Li, and Tyler H. McCormick. Quantifying the contributions of training data and algorithm logic to the performance of automated cause-assignment algorithms for verbal autopsy, 2018. arXiv preprint arXiv:1803.07141.

Elizabeth Cohen and Anne McDermott. Who's fat? New definition adopted. *CNN*, June 1998. URL http://www.cnn.com/HEALTH/9806/17/weight.guidelines/. Accessed: April 15, 2026.

Samuel Danso, Eric Atwell, and Owen Johnson. Linguistic and statistically derived features for cause of death prediction from verbal autopsy text. In *Language Processing and Knowledge in the Web: 25th International Conference, GSCL 2013*, pages 47–60. Springer, 2013.

Jean-Pierre Després. Abdominal obesity: the most prevalent cause of the metabolic syndrome and related cardiometabolic risk. *European Heart Journal Supplements*, 8:B4–B12, May 2006. doi: 10.1093/eurheartj/sul002. URL https://doi.org/10.1093/eurheartj/sul002.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Naoki Egami, Musashi Jacobs-Harukawa, Brandon M. Stewart, and Hanying Wei. Using large language model annotations for valid downstream statistical inference in social science: Design-based semi-supervised learning, 2023. arXiv preprint arXiv:2306.04746.

Samuel D. Emmerich, Cheryl D. Fryar, Bryan Stierman, and Cynthia L. Ogden. Obesity and severe obesity prevalence in adults: United States, August 2021–August 2023. NCHS Data Brief 508, National Center for Health Statistics, September 2024. URL https://www.cdc.gov/nchs/products/databriefs/db508.htm.

Paul Ernsberger and Paul Haskew. Rethinking obesity: An alternative view of its health implications. *The Journal of Obesity and Weight Regulation*, 6(2):57–137, 1987.

Megan Evans and Barrett A. Lee. Neighborhood reputations as symbolic and stratifying mechanisms in the urban hierarchy. *Sociology Compass*, 14(10):1–15, 2020.

Kenneth F. Ferraro and Jessica A. Kelley-Moore. Cumulative disadvantage and health: Long-term consequences of obesity? *American Sociological Review*, 68(5):707–729, 2003. doi: 10.2307/1519759.

Abraham D. Flaxman, Lisa Harman, Jonathan Joseph, Jonathan Brown, and Christopher J.L. Murray. A de-identified database of 11,979 verbal autopsy open-ended responses. *Gates Open Research*, 2:18, 2018. doi: 10.12688/gatesopenres.12812.1.

K. M. Flegal, M. D. Carroll, R. J. Kuczmarski, and C. L. Johnson. Overweight and obesity in the united states: prevalence and trends, 1960-1994. *International Journal of Obesity and Related Metabolic Disorders*, 22(1):39–47, Jan 1998. doi: 10.1038/sj.ijo.0800541.

K. M. Flegal, B. I. Graubard, D. F. Williamson, and M. H. Gail. Excess deaths associated with underweight, overweight, and obesity. *JAMA*, 293(15):1861–1867, Apr 2005. doi: 10.1001/jama.293.15.1861.

Katherine M. Flegal. The obesity wars and the education of a researcher: A personal account. *Progress in Cardiovascular Diseases*, 67:75–79, 2021. doi: 10.1016/j.pcad.2021.06.009.

E. Fottrell and P. Byass. Verbal autopsy: Methods in transition. *Epidemiologic Reviews*, 32 (1):38–55, 2010. doi: 10.1093/epirev/mxq003.

Lindsay T. Fourman, Aya Awwad, Alba Gutiérrez-Sacristán, Camille A. Dash, Julia E. Johnson, Allison K. Thistle, Nikhita Chahal, Sara L. Stockman, Mabel Toribio, Chika Anekwe, Arijeet K. Gattu, and Steven K. Grinspoon. Implications of a New Obesity Definition Among the All of Us Cohort. *JAMA Network Open*, 8(10):e2537619, October 2025. doi: 10.1001/jamanetworkopen.2025.37619. URL https://doi.org.

Glenn A. Gaesser. *Big Fat Lies: The Truth About Your Weight and Your Health*. Fawcett Columbine, New York, 1996.

Michael Gard and Jan Wright. *The obesity epidemic: Science, morality and ideology*. Routledge/Taylor & Francis Group, 2005.

Craig M. Hales, Margaret D. Carroll, Cheryl D. Fryar, and Cynthia L. Ogden. Prevalence of obesity and severe obesity among adults: United States, 2017–2018. NCHS Data Brief 360, National Center for Health Statistics, February 2020. URL https://www.cdc.gov/nchs/products/databriefs/db360.htm.

Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328, 2008. doi: 10.1109/IJCNN.2008.4633969.

Chris Hess and Sarah E. Chasins. Informing housing policy through web automation: Lessons for designing programming tools for domain experts. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–9, 2022.

Chris Hess, Rebecca J. Walter, Ian Kennedy, Arthur Acolin, Alex Ramiller, and Kyle Crowder. Segmented information, segregated outcomes: Housing affordability and neighborhood representation on a voucher-focused online housing platform and three mainstream alternatives. *Housing Policy Debate*, 33(6):1511–1535, 2023.

Tuomo Hiippala, Anna Hausmann, Henrikki Tenkanen, and Tuuli Toivonen. Exploring the linguistic landscape of geotagged social media content in urban environments. *Digital Scholarship in the Humanities*, 34(2):290–309, 2019.

James O. Hill, Holly R. Wyatt, George W. Reed, and John C. Peters. Obesity and the environment: where do we go from here? *Science*, 299(5608):853–855, Feb 2003. doi: 10.1126/science.1079857.

Julia Hirschberg and Christopher D. Manning. Advances in natural language processing. *Science*, 349(6245):261–266, 2015. doi: 10.1126/science.aaa8685.

Lan T. Ho-Pham, Lesley V. Campbell, and Tuan V. Nguyen. More on body fat cutoff points. *Mayo Clinic Proceedings*, 86(6):584, 2011. doi: 10.4065/mcp.2011.0097. Letter.

Kentaro Hoffman, Stephen Salerno, Awan Afiaz, Jeffrey T. Leek, and Tyler H. McCormick. Do we really even need data?, 2024. arXiv preprint arXiv:2401.08702.

Randolph Hohle. Rusty gardens: Stigma and the making of a new place reputation in Buffalo, New York. *American Journal of Cultural Sociology*, 11(2):193–219, 2023.

Livia Hollenstein and Ross Purves. Exploring place through user-generated content: Using Flickr to describe city cores. *Journal of Spatial Information Science*, 1:1–18, 2010.

Jackelyn Hwang and Robert J. Sampson. Divergent pathways of gentrification: Racial inequality and the social order of renewal in Chicago neighborhoods. *American Sociological Review*, 79(4):726–751, 2014.

Spencer L. James, Abraham D. Flaxman, Christopher J. Murray, and Population Health Metrics Research Consortium. Performance of the tariff method: Validation of a simple additive algorithm for analysis of verbal autopsies. *Population Health Metrics*, 9(1):31, 2011.

Su-Min Jeong, Dong Hoon Lee, Leandro F. M. Rezende, and Edward L. Giovannucci. Different correlation of body mass index with body fatness and obesity-related biomarker according to age, sex and race-ethnicity. *Scientific Reports*, 13:3472, 2023. doi: 10.1038/s41598-023-30527-w.

David W. Johnston and Wang-Sheng Lee. Explaining the female black-white obesity gap: A decomposition analysis of proximal causes. *Demography*, 48(4):1429–1450, Sep 2011. doi: 10.1007/s13524-011-0064-x.

Steven E Kahn, Rebecca L Hull, and Kristina M Utzschneider. Mechanisms linking obesity to insulin resistance and type 2 diabetes. *Nature*, 444(7121):840–846, 2006.

Ronda Kaysen. Apartment rent renewal rates are rising. *The New York Times*, 2024.

Ian Kennedy, Chris Hess, Amandalynne Paullada, and Sarah Chasins. Racialized discourse in Seattle rental ad texts. *Social Forces*, 99(4):1432–1456, 2020.

Ancel Keys, Flaminio Fidanza, Martti J. Karvonen, Noboru Kimura, and Henry L. Taylor. Indices of relative weight and obesity. *Journal of Chronic Diseases*, 25(6):329–343, July 1972. ISSN 0021-9681. doi: 10.1016/0021-9681(72)90027-6. URL https://www.sciencedirect.com/science/article/pii/0021968172900276.

Richard Kirk. Legitimising displacement: Academic discourse, territorial stigmatisation and gentrification. *Urban Studies*, 61(13):2492–2512, 2024.

Anna Kirkland. The environmental account of obesity: a case for feminist skepticism. *Signs: Journal of Women in Culture and Society*, 36(2):463–486, 2011. doi: 10.1086/655916.

Elizabeth Korver-Glenn and Sarah Mayorga. *A Good Reputation: How Residents Fight for an American Barrio*. University of Chicago Press, 2024.

Oliver Kramer. Scikit-learn. In *Machine Learning for Evolution Strategies*, pages 45–53. 2016.

Maria Krysan and Kyle Crowder. *Cycle of Segregation: Social Processes and Residential Stratification*. Russell Sage Foundation, New York, 2017.

Agneta Kullberg, Toomas Timpka, Tommy Svensson, Nadine Karlsson, and Kent Lindqvist. Does the perceived neighborhood reputation contribute to neighborhood differences in social trust and residential wellbeing? *Journal of Community Psychology*, 38(5):591–606, 2010.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning, 2023. arXiv preprint arXiv:2304.05613.

M. A. Laskey. Dual-energy x-ray absorptiometry and body composition. *Nutrition*, 12(1): 45–51, Jan 1996. doi: 10.1016/0899-9007(95)00017-8.

Zehang Richard Li, Tyler H. McCormick, and Samuel J. Clark. Using Bayesian latent Gaussian graphical models to infer symptom associations in verbal autopsies. *Bayesian Analysis*, 15(3):781, 2020.

Ran Liu, Joseph L. Greenstein, Sridevi V. Sarma, and Raimond L. Winslow. Natural language processing of clinical notes for improved early prediction of septic shock in the ICU. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6103–6108, 2019. doi: 10.1109/EMBC.2019.8857819.

Thokozile Manaka, Terence van Zyl, and Deepak Kar. Improving cause-of-death classification from verbal autopsy reports. In *Southern African Conference for Artificial Intelligence Research*, pages 46–59. Springer, 2022.

Tyler H. McCormick, Zehang Richard Li, Clara Calvert, Amelia C. Crampin, Kathleen Kahn, and Samuel J. Clark. Probabilistic cause-of-death assignment using verbal autopsies. *Journal of the American Statistical Association*, 111(515):1036–1049, 2016. doi: 10.1080/01621459.2016.1152191.

Jiacheng Miao, Xinran Miao, Yixuan Wu, Jiwei Zhao, and Qiongshi Lu. Assumption-lean and data-adaptive post-prediction inference, 2023. arXiv preprint arXiv:2311.14220.

Frank Millstein. *Natural Language Processing with Python: Natural Language Processing Using NLTK*. Frank Millstein, 2020.

Anita Minh, Nazeem Muhajarine, Magdalena Janus, Marni Brownell, and Martin Guhn. A review of neighborhood effects and early child development: How, where, and for whom, do neighborhoods matter? *Health & Place*, 46:155–174, 2017.

A. H. Mokdad, E. S. Ford, B. A. Bowman, W. H. Dietz, F. Vinicor, V. S. Bales, and J. S. Marks. Prevalence of obesity, diabetes, and obesity-related health risk factors, 2001. *JAMA*, 289(1):76–79, Jan 2003. doi: 10.1001/jama.289.1.76.

Justin X Moore, Naman Chaudhary, and Tomi Akinyemiju. Metabolic syndrome prevalence by race/ethnicity and sex in the united states, national health and nutrition examination survey, 1988–2012. *Preventing Chronic Disease*, 14:E24, 2017. doi: 10.5888/pcd14.160287.

Christopher J.L. Murray, Alan D. Lopez, Kenji Shibuya, and Rafael Lozano. Verbal autopsy: Advancing science, facilitating application. *Population Health Metrics*, 9(1), 2011. doi: 10.1186/1478-7954-9-18.

Christopher J.L. Murray, Rafael Lozano, Abraham D. Flaxman, Peter Serina, et al. Using verbal autopsy to measure causes of death: The comparative performance of existing methods. *BMC Medicine*, 12(1), 2014. doi: 10.1186/1741-7015-12-5.

Jason Naradowsky, Sebastian Riedel, and David Smith. Improving NLP through marginalization of hidden syntactic structure. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 810–820, Jeju Island, Korea, 2012. Association for Computational Linguistics.

Roberto Navigli, Simone Conia, and Björn Ross. Biases in large language models: Origins, inventory, and discussion. *Journal of Data and Information Quality*, 15(2):1–21, 2023. doi: 10.1145/3597307.

NCD Risk Factor Collaboration (NCD-RisC). Obesity rise plateaus in developed nations and accelerates in developing nations. *Nature*, 653:510–518, 2026. doi: 10.1038/s41586-026-10383-0. URL https://doi.org.

NCHS. Technical documentation for the 1999–2004 dual energy x-ray absorptiometry (dxa) multiple imputation data files, 2008. URL https://wwwn.cdc.gov/nchs/data/nhanes/public/1999/documents/dxa_techdoc.pdf. Accessed: 2025-12-12.

David Newman, Arthur Asuncion, Padhraic Smyth, and Max Welling. Distributed algorithms for topic models. *Journal of Machine Learning Research*, 10:1801–1828, 2009.

Elizabeth L. Ogburn, Kara E. Rudolph, Rachel Morello-Frosch, Amber Khan, and Joan A. Casey. A warning about using predicted values from regression models for epidemiologic inquiry. *American Journal of Epidemiology*, 190(6):1142–1147, 2021.

Cynthia L. Ogden, Margaret D. Carroll, Brian K. Kit, and Katherine M. Flegal. Prevalence of obesity among adults: United States, 2011–2012. NCHS Data Brief 131, National Center for Health Statistics, October 2013. URL https://www.cdc.gov/nchs/products/databriefs/db131.htm.

E. Oliveros, V. K. Somers, O. Sochor, K. Goel, and F. Lopez-Jimenez. The concept of normal weight obesity. *Progress in Cardiovascular Diseases*, 56(4):426–433, Jan 2014. doi: 10.1016/j.pcad.2013.10.003.

Gabriel Otero, Quentin Ramond, María Luisa Méndez, Rafael Carranza, Felipe Link, and Javier Ruiz-Tagle. The damages of stigma, the benefits of prestige: Examining the consequences of perceived residential reputations on neighbourhood attachment. *Urban Studies*, 61(3):462–494, 2024.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Tarra L. Penney and Sara F. L. Kirk. The Health at Every Size paradigm and obesity: Missing empirical evidence may help push the reframing obesity debate forward. *American Journal of Public Health*, 105(5):e38–e42, May 2015. doi: 10.2105/AJPH.2015.302552.

Anthony Daniel Perez and Charles Hirschman. The changing racial and ethnic composition of the US population: Emerging american identities. *Population and Development Review*, 35(1):1–51, mar 2009. doi: 10.1111/j.1728-4457.2009.00260.x.

M. Permentier, G. Bolt, and M. Van Ham. Determinants of neighbourhood satisfaction and perception of neighbourhood reputation. *Urban Studies*, 48(5):977–996, 2011.

Denis Peskoff and Brandon Stewart. Credible without credit: Domain experts assess generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 427–438, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.37.

Denis Peskoff, Viktor Hangya, Jordan Boyd-Graber, and Alexander Fraser. Adapting entities across languages and cultures. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3725–3750, Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.315.

Anon Plangprasopchok and Kristina Lerman. Constructing folksonomies from user-specified relations on Flickr. In *Proceedings of the 18th International Conference on World Wide Web*, pages 781–790, 2009.

Paul Poirier, Thomas D Giles, George A Bray, Yuling Hong, Judith S Stern, F Xavier Pi-Sunyer, and Robert H Eckel. Obesity and cardiovascular disease: pathophysiology, evaluation, and effect of weight loss: an update of the 1997 american heart association scientific statement on obesity and heart disease from the obesity committee of the council on nutrition, physical activity, and metabolism. *Circulation*, 113(6):898–918, 2006.

Janet Polivy and C. Peter Herman. Dieting and binging: A causal analysis. *American Psychologist*, 40(2):193–201, 1985.

Adam W. Potter, Geoffrey C. Chin, David P. Looney, and Karl E. Friedl. Defining overweight and obesity by percent body fat instead of body mass index. *The Journal of Clinical Endocrinology & Metabolism*, 110(4):e1103–e1107, 2025. doi: 10.1210/clinem/dgae341.

Rebecca M. Puhl and Chelsea A. Heuer. Obesity stigma: important considerations for public health. *American Journal of Public Health*, 100(6):1019–1028, Jun 2010. doi: 10.2105/AJPH.2009.159491.

Abel Romero-Corral, Virend K. Somers, Justo Sierra-Johnson, Randal J. Thomas, Kent R. Bailey, Maria L. Collazo-Clavell, Thomas G. Allison, Josef Korinek, John A. Batsis, and Francisco Lopez-Jimenez. Accuracy of body mass index to diagnose obesity in the US adult population. *International Journal of Obesity*, 32(6):959–966, 2008. doi: 10.1038/ijo.2008.11.

Gregory A Roth, George A Mensah, Catherine O Johnson, Giovanni Addolorato, Enrico Ammirati, Larry M Baddour, Noel C Barengo, et al. Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the gbd 2019 study. *Journal of the American College of Cardiology*, 76(25):2982–3021, 2020.

Esther D. Rothblum. The stigma of "women's weight": Social and economic realities. *Feminism & Psychology*, 2(1):61–73, 1992.

Francesco Rubino, David E. Cummings, et al. Definition and diagnostic criteria of clinical obesity. *The Lancet Diabetes & Endocrinology*, 13(3):221–262, March 2025. doi: 10.1016/S2213-8587(24)00316-4.

Pouya Saeedi, Inga Petersohn, Paraskevi Salpea, Belma Malanda, Suvi Karuranga, Nigel Unwin, Stephen Colagiuri, et al. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the international diabetes federation diabetes atlas, 9th edition. *Diabetes Research and Clinical Practice*, 157:107843, 2019.

Abigail Cope Saguy. *What's wrong with fat?* Oxford University Press, New York, 2013. ISBN 978-0-19-985708-1.

Stephen Salerno, Kentaro Hoffman, Awan Afiaz, Anna Neufeld, Tyler H. McCormick, and Jeffrey T. Leek. Do we really even need data? a modern look at drawing inference with predicted data, 2025. URL https://arxiv.org/abs/2512.05456.

James F. Sallis and Karen Glanz. The role of built environments in physical activity, eating, and obesity in childhood. *The Future of Children*, 16(1):89–108, 2006. doi: 10.1353/foc.2006.0009.

Robert J. Sampson, Jeffrey D. Morenoff, and Thomas Gannon-Rowley. Assessing "neighborhood effects": Social processes and new directions in research. *Annual Review of Sociology*, 28(1):443–478, 2002.

Ariela Schachter, John Kuk, Max Besbris, and Lydia Ho. What's in a name? place misrepresentation and neighbourhood stigma in the online rental market. *Urban Studies*, 61(16):3050–3068, 2024.

Melissa Scharoun-Lee, Penny Gordon-Larsen, Linda S. Adair, Barry M. Popkin, Jay S. Kaufman, and Chirayath M. Suchindran. Intergenerational profiles of socioeconomic (dis)advantage and obesity during the transition to adulthood. *Demography*, 48(2):625–651, Apr 2011. doi: 10.1007/s13524-011-0024-5.

N. Schenker, L. G. Borrud, V. L. Burt, L. R. Curtin, K. M. Flegal, J. Hughes, C. L. Johnson, A. C. Looker, and L. Mirel. Multiple imputation of missing dual-energy x-ray absorptiometry data in the national health and nutrition examination survey. *Statistics in Medicine*, 30(3):260–276, Feb 2011. doi: 10.1002/sim.4080.

Peter Serina, Ian Riley, Andrea Stewart, Spencer L. James, Abraham D. Flaxman, Rafael Lozano, et al. Improving performance of the tariff method for assigning causes of death to verbal autopsies. *BMC Medicine*, 13(1), 2015. doi: 10.1186/s12916-015-0527-9.

Patrick Sharkey and Jacob W. Faber. Where, when, why, and for whom do residential contexts matter? moving away from the dichotomous understanding of neighborhood effects. *Annual Review of Sociology*, 40(1):559–579, 2014.

John A. Shepherd, Bennett K. Ng, Markus J. Sommer, and Steven B. Heymsfield. Body composition by dxa. *Bone*, 104:101–105, 2017. doi: 10.1016/j.bone.2017.06.010.

Karan Singhal, Shekoofeh Azizi, Tao Tu, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023. doi: 10.1038/s41586-023-06291-2.

Nadia Soleman, Daniel Chandramohan, and Kenji Shibuya. Verbal autopsy: Current practices and challenges. *Bulletin of the World Health Organization*, 84(3):239–245, 2006.

Mahesh Somashekhar, Chris Hess, Ian Kennedy, and Kyle Crowder. How do real estate actors advertise in mixed-income neighborhoods? the importance of home security. *Socius*, 10:23780231241260253, 2024.

Fatima Cody Stanford, Minyi Lee, and Chin Hur. Race, ethnicity, sex, and obesity: Is it time to personalize the scale? *Mayo Clinic Proceedings*, 94(2):362–363, 2019. doi: 10.1016/j.mayocp.2018.10.014. PMID: 30711132, PMCID: PMC6818706.

Remy Stewart, Chris Hess, Ian Kennedy, and Kyle Crowder. Move-in fees as a residential sorting mechanism within online rental markets. *Cityscape*, 25(1):239, 2023.

Forrest Stuart, Charles R. Collins, Bocar Wade, Rebecca D. Gleit, and Caylin Louis Moore. Where do neighbourhood reputations come from? analysing Chicago community areas using a systematic neighbourhood reputation score, 1985–2020. *Urban Studies*, page 00420980241297088, 2024.

Clarissa Surek-Clark. Verbal autopsy interview standardization study: A report from the field. In *The Anthropological Demography of Health*. Oxford University Press, 2020. doi: 10.1093/oso/9780198862437.003.0011.

Lisa-Marie Thomas, Lucia D'Ambruoso, and Dina Balabanova. Verbal autopsy in health policy and systems: A literature review. *BMJ Global Health*, 3(2):e000639, 2018.

Ivan Tomek. Two modifications of CNN. In *IEEE Transactions on Systems, Man, and Cybernetics*, 1976. doi: 10.1109/TSMC.1976.4309452.

A. Janet Tomiyama, Deborah Carr, Ellen M. Granberg, Brenda Major, Eric Robinson, Angelina R. Sutin, and Alexandra Brewis. How and why weight stigma drives the obesity 'epidemic' and harms health. *BMC Medicine*, 16(1):123, Aug 2018. doi: 10.1186/s12916-018-1116-5.

Emma Tran, Kim Blankenship, Shannon Whittaker, Alana Rosenberg, Penelope Schlesinger, Trace Kershaw, and Danya Keene. My neighborhood has a good reputation: Associations between spatial stigma and health. *Health & Place*, 64:102392, 2020.

Abdul Wahab, Ifta Choiriyyah, and Siswanto Agus Wilopo. Determining the cause of death: Mortality surveillance using verbal autopsy in Indonesia. *American Journal of Tropical Medicine and Hygiene*, 97(5):1461–1468, 2017. doi: 10.4269/AJTMH.16-0815.

Siruo Wang, Tyler H. McCormick, and Jeffrey T. Leek. Methods for correcting inference based on outcomes predicted by machine learning. *Proceedings of the National Academy of Sciences*, 117(48):30266–30275, 2020.

Sujing Wang, Jie Shen, Woon-Puay Koh, Jian-Min Yuan, Xiang Gao, Yinshun Peng, Yaqing Xu, Shuxiao Shi, Yue Huang, Ying Dong, and Victor W Zhong. Comparison of racial/ethnic-specific bmi cutoffs for categorizing obesity severity: a multicountry prospective cohort study. *Obesity*, 32(10):1958–1966, 2024. doi: 10.1002/oby.24129. PMID: 39223976, PMCID: PMC11421961.

J. C. K. Wells and M. S. Fewtrell. Measuring body composition. *Archives of Disease in Childhood*, 91:612–617, 2006. doi: 10.1136/adc.2005.085522.

Rachel P. Wildman, Paul Muntner, Kristi Reynolds, Aileen P. McGinn, Swapnil Rajpathak, Judith Wylie-Rosett, and Mary R. Sowers. The obese without cardiometabolic risk factor clustering and the normal weight with cardiometabolic risk factor clustering: prevalence and correlates of 2 phenotypes among the us population (nhanes 1999-2004). *Archives of Internal Medicine*, 168(15):1617–1624, Aug 2008. doi: 10.1001/archinte.168.15.1617.

Rena R. Wing, Michael G. Goldstein, Kelly J. Acton, Leann L. Birch, John M. Jakicic, James F. Sallis, Delia Smith-West, Robert W. Jeffery, and Richard S. Surwit. Behavioral science research in diabetes: Lifestyle changes related to obesity, eating behavior, and physical activity. *Diabetes Care*, 24(1):117–123, Jan 2001. doi: 10.2337/diacare.24.1.117. URL https://doi.org/10.2337/diacare.24.1.117.

Susan C. Wooley and O. Wayne Wooley. Should obesity be treated at all? In L. A. Cioffi, W. P. T. James, and T. B. Van Itallie, editors, *The Body Weight Regulatory System: Normal and Disturbed Mechanisms*, pages 333–337. Raven Press, New York, 1981.

World Health Organization. Obesity, 2023. URL https://www.who.int/health-topics/obesity#tab=tab_1. Accessed on March 29, 2025.

Zihuiwen Ye, Pengfei Liu, Jinlan Fu, and Graham Neubig. Towards more fine-grained and reliable NLP performance prediction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3703–3714. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.324.

Sarah Zelner. The perpetuation of neighborhood reputation: An interactionist approach. *Symbolic Interaction*, 38(4):575–593, 2015.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291, 2024.