Building Geolocated Business Inventories from Digital Trace Data: A Tobacco Retail Use Case

Sara Lafia (PhD)[1], Yoonsang Kim (PhD)[1], Simon Page (BA)[1], Chandler Carter (MA)[1], Shyanika W. Rose (PhD)[2], Sherry Emery (PhD)[1], Ganna Kostygina (PhD)[1] *

[1] NORC at the University of Chicago, 300 E. Randolph Street, Suite 4600, Chicago, IL 60601

[2] Department of Behavioral Science and Center for Health, Engagement, and Transformation, University of Kentucky College of Medicine, Lexington, KY, USA

* Corresponding Author: Kostygina-Anna@norc.org

Abstract

Digital trace data have become central to social science research, yet restricted platform access and data quality limitations complicate efforts to build reliable data collection pipelines. This study examines social media marketing by tobacco retailers, a task hindered by the lack of geolocation information in platform data, which prevents researchers from linking online marketing to local contexts. To overcome this limitation, we develop an innovative approach that identifies tobacco shops/vendors and their physical locations, and systematically links these entities to their online presence.

Using tobacco retailers as a case study, we present a generalizable framework that integrates Google Maps and Yelp listings and uses a hybrid record linkage strategy to construct geolocated, entity-level inventories. Associated social media accounts are identified and validated through manual review, enabling integration with population surveys to support spatial and inferential analyses. More broadly, this framework informs the development of sampling frames and analytic datasets in settings where traditional data sources are incomplete and platform access is rapidly evolving.

Key Words: Digital trace data; Entity resolution; Record linkage; Geospatial data integration; Social media marketing; Survey data integration

1. Introduction

1.1 Digital Trace Data for Public Health Research

The tobacco and nicotine product marketplace has undergone substantial transformation over the past decade, characterized by rapid diversification of non-cigarette tobacco products, including e-cigarettes, nicotine pouches, heated tobacco products, hookah products, and emerging nicotine analog formulations.[1 2] Along with these changes, tobacco marketing has increasingly shifted from traditional advertising channels toward digital and social media environments that facilitate direct consumer engagement, rapid dissemination of promotional content, and targeted communication strategies.[3-8]

Tobacco manufacturers, retailers, and influencers use social media to promote tobacco products, communicate brand identity, engage consumers, as well as cross-promote tobacco with other health-harming products, like alcoholic beverages.[5-7 9 10] These marketing strategies frequently incorporate discounts, giveaways, user-generated content, influencer endorsements, and other engagement mechanisms designed to increase message visibility and dissemination across online communities.[6 11] Retailer-generated social media content represents an increasingly important and understudied component of the digital tobacco marketing ecosystem. Specialty tobacco retailers, vape shops, smoke shops, cigar lounges, and related businesses routinely maintain social media

accounts that promote products, advertise events and discounts, communicate local availability of nicotine and tobacco products, and reinforce community-level tobacco product norms.[5 12] Platforms such as Instagram, Facebook, and TikTok provide retailers with low-cost mechanisms for maintaining a persistent local marketing presence and engaging customers beyond the retail environment.[12]

More broadly, retailer-generated social media activity illustrates both the promise and the challenges of digital trace data for public health research. Online platforms contain large volumes of behavioral and marketing information, but these data are often fragmented across independent systems that were not designed for research purposes. As access to platform data becomes increasingly restricted, researchers must develop new approaches that combine multiple digital sources to construct valid and reproducible measures of social, commercial, and environmental phenomena.

Business listing platforms represent one such source of digital trace data. These systems contain information on business identity, category, location, websites, and other characteristics that can support the development of sampling frames, surveillance systems, and geospatial measures. However, no single source typically contains all information needed for measurement, making integration across multiple data sources necessary.

1.2 The Challenge of Connecting Online Activity to Place

A fundamental challenge in digital trace data research is that geographic context and online activity often exist in separate data systems. Social media platforms contain marketing content and user engagement but rarely provide reliable business location information. Conversely, business listing platforms contain business locations and attributes but generally do not capture online marketing activity itself.

Recent changes in platform structures, content discovery mechanisms, and data access policies have further complicated these challenges.[13-15] Researchers increasingly face restricted access to platform data, including geographic information that could otherwise be used to locate businesses or content.[14] Gaps in monitoring and enforcement of social media marketing restrictions can thus allow unapproved marketing claims and promotion of unauthorized products to reach the target audience.[6 16 17] As a result, identifying businesses, assigning them to geographic locations, and linking them to their online presence has become increasingly difficult.

These challenges are particularly important for studies seeking to characterize localized online marketing environments. Tobacco regulations, product availability, and retail density vary across communities, and retailers increasingly use geographically targeted promotional strategies.[18 19] Local retailers may maintain substantial marketing activity despite limited regulatory visibility, and social media platforms increasingly support geographically targeted promotion.[20-28] Social media amplifies these local marketing effects because individuals disproportionately follow local business accounts and engage with community-specific content.[29] Retailers increasingly capitalize on these engagement patterns, using geotargeted promotional content to reach users in specific markets.

This paper addresses a broader methodological problem that extends beyond tobacco research: **How can researchers systematically link businesses' geographic locations to their social media presence when relevant information exists in separate and incomplete data systems?**

1.3 Tobacco Retail as a Use Case

The present study uses specialty tobacco retail as a case study for addressing this broader methodological challenge. Specialty tobacco retailers and venues routinely maintain dedicated social media accounts and generate product-specific marketing content, making them particularly well suited for studying connections between physical retail environments and online activity.

The study also addresses a longstanding limitation in tobacco control research: the absence of a comprehensive national retailer registry. Existing retailer inventories are commonly constructed

from tobacco retailer licensing records, Synar rosters, or commercial business directories such as Dun & Bradstreet.[30] While these approaches can yield accurate local retailer censuses, they are labor-intensive, reflect retail conditions at a single point in time, and may be unavailable or incomplete in jurisdictions lacking comprehensive retailer licensing systems.

Digital trace data offer an opportunity to overcome these limitations. The framework presented here provides a scalable method for building and refreshing tobacco and nicotine product retailer inventories that can serve as sampling frames for digital surveillance, field audits, and assessments such as the Standardized Tobacco Assessment for Retail Settings (STARS) and Vape Shop STARS (V-STARS).[30-32]

Emerging research has used digital trace data (e.g., Google Maps) to construct geolocated inventories of vape shops in several U.S. states and localities.[33-36] These studies demonstrated that digital business listings can support geographically accurate, resource-efficient identification of vape shops, but were limited to a small number of counties or metropolitan areas, focused exclusively on vape shops rather than the broader tobacco retail landscape, and did not extend beyond store identification to characterize retailer social media marketing presence.

The present study focuses specifically on specialty tobacco and nicotine product retailers, including tobacco shops, vape shops, smoke shops, and related venues in the United States. We focus on specialty retailers because they are more likely to maintain dedicated social media accounts and engage in tobacco-specific digital marketing activity than retailers for whom tobacco products represent only a small component of overall sales, such as convenience stores, gas stations, grocery stores, and pharmacies.

To our knowledge, this study is the first to apply and validate a digital-trace-data approach for constructing a national geolocated inventory of specialty tobacco retailers and venues and the first to systematically link those retailers to social media accounts.

1.4 Study Objectives

This study presents and evaluates a framework for constructing geolocated tobacco retailer inventories from digital trace data. Specifically, we:

1. Construct a national inventory of specialty tobacco retailers and venues using multiple business listing sources;
2. Geolocate retailer locations through address standardization and geocoding;
3. Link retailers and venues to Facebook, Instagram, and TikTok accounts; and
4. Evaluate retailer coverage through comparison with an independent audit dataset.

Using specialty tobacco retailers as a use case, the study demonstrates how business listing data, geospatial information, and social media discovery procedures can be integrated into a unified framework for constructing geolocated business inventories from digital trace data.

2. Conceptual Framework

A recurring challenge in digital trace data research is that information needed to answer a research question is often distributed across multiple disconnected systems. Social media platforms contain information about online marketing activity and user engagement but generally provide limited and increasingly restricted geographic information. Business listing platforms contain business locations, categories, and identifying attributes but provide little information about marketing activity itself. As a result, researchers frequently need to integrate multiple incomplete data sources to construct meaningful measures of real-world phenomena.

The framework developed in this study addresses this challenge by integrating business listings, geographic information, and social media account data into a unified measurement and record linkage system. Using specialty tobacco retailers as a case study, the framework demonstrates how digital trace data can be transformed into geolocated business inventories capable of supporting surveillance and future analytic applications.

The framework consists of five sequential stages as shown in **Figure 1**, which shows the overall workflow used to identify retailers, validate inventory completeness, geocode retailer locations, and link retailers to social media accounts.

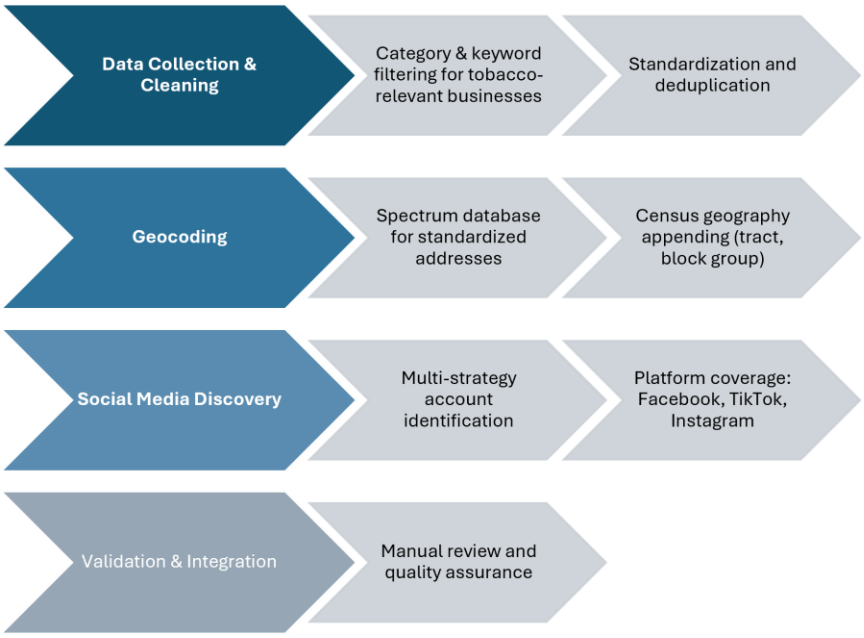


Figure 1. Conceptual framework for constructing geolocated tobacco retailer inventories from digital trace data.

No single source contains all information needed to characterize retailer identity, location, and online activity. Consequently, the framework relies on combining multiple sources that contribute complementary information. Business listings provide information about retailer identity and location; geocoding provides spatial context; and social media account discovery provides access to online marketing activity. The resulting geolocated inventory creates a bridge between physical retail environments and digital marketing systems.

An important feature of this framework is its treatment of inventory construction as a measurement problem rather than a simple data collection exercise. Measurement errors can arise during retailer identification, address standardization, geocoding, entity resolution, and social media account discovery. The framework therefore incorporates quality assurance, manual review, and external validation procedures at multiple stages. In the present study, independent retailer audit data serve as an external benchmark for evaluating inventory completeness, while social media discovery procedures are designed to maximize account recovery across platforms.

Viewed more broadly, this framework is not specific to tobacco retail. The same sequence of business identification, geolocation, entity resolution, record linkage, and validation can be applied whenever geographic context and online activity must be connected across disparate data systems. Accordingly, the tobacco retail application presented here serves as a methodological demonstration of a more general approach to constructing geolocated business inventories from digital trace data.

3. Data Sources

The framework integrates three primary data sources: (1) national business listing data, (2) independently collected retailer audit data used for validation, and (3) retailer social media account information recovered through business metadata and website-based discovery procedures. Together, these sources provide the information necessary to identify retailers, assign geographic locations, validate coverage, and link physical businesses to their online presence. Each source

contributes distinct information that is not fully available from the other sources, making data integration a central component of the framework.

3.1 Business Listing Data

National business listing data were obtained from Google Maps and Yelp through a third-party commercial provider of large-scale business information. Google Maps contributed 90,055 candidate businesses meeting tobacco-related search criteria, while Yelp contributed 35,822 candidate businesses. Both sources provided business names, addresses, category labels, descriptive text, and website URLs when available.

Google Maps and Yelp were selected because they provide broad geographic coverage, are continually updated, and contain business attributes useful for retailer identification and record linkage. In addition to standard business information, Google Maps records included structured web-results fields containing links to external websites and, in some cases, retailer-operated social media accounts. Approximately 53.4% of Google Maps businesses and 53.6% of Yelp businesses included website information that could subsequently be used for social media account discovery.

The use of multiple business listing platforms is important because no single source provides complete coverage of specialty tobacco retailers. Combining sources increases opportunities for retailer identification while providing additional information for validation and quality assurance. More generally, the two sources illustrate a common characteristic of digital trace data systems: different platforms often contain overlapping but incomplete representations of the same underlying population.

3.2 External Validation Data

To evaluate inventory completeness, we used an independently collected tobacco retailer audit dataset from eight rural Appalachian Kentucky counties and one Central Kentucky county that was more urban. The audit included 523 verified tobacco retailers identified through field observation and retailer verification procedures conducted during 2024–2025. Because the present study focuses on specialty tobacco retailers and vape shops, validation analyses were restricted to a subset of 68 audited specialty retailers consisting of 44 tobacco specialty retailers and 24 vape shops.

The Kentucky audit served as an external benchmark independent of the business listing data used for inventory construction. Detailed audit procedures are provided in **Supplementary Methods S1** and reported elsewhere.[30] This benchmark allowed us to evaluate the extent to which retailers identified through digital trace data corresponded to retailers observed in the physical environment. Because the validation dataset was collected independently of the business listing platforms, it provides an opportunity to assess inventory completeness without relying on internal platform characteristics alone.

3.3 Social Media Data

Retailer social media accounts were identified for three platforms: Facebook, Instagram, and TikTok. These platforms were selected because they represent major channels for retailer-generated marketing activity and are commonly used by specialty tobacco retailers to communicate with consumers.

Unlike business listing platforms, social media systems generally contain limited geographic information about businesses. Consequently, social media data entered the framework through a linkage process that connected platform accounts to identified retailer locations using business metadata, website information, and supplemental search procedures. The resulting linked retailer–social media records provide the foundation for subsequent analysis of localized online marketing activity. In the conceptual framework, social media data represent the component of the measurement system that contributes information about online activity, while business listing platforms contribute information about retailer identity and location. The linkage process therefore serves as the mechanism that connects these otherwise separate data systems.

4. Methods

This study developed a multi-stage framework for constructing geolocated inventories of specialty tobacco retailers from digital trace data. Digital business platforms may introduce measurement error through false positives, false negatives, and outdated business records (for example, home address for the business owner listed as the business address).[37] The framework integrates retailer identification, data cleaning, geocoding, social media account discovery, and external validation into a unified workflow shown above in **Figure 1**. Each stage addresses a distinct measurement challenge associated with linking business locations to online marketing activity. The workflow combines entity identification, spatial processing, record linkage, and validation procedures to transform fragmented digital trace data into an analytically useful measurement system.

4.1 Retailer Identification

Identifying specialty tobacco retailers from business listing platforms is fundamentally a measurement problem. Business listing categories are frequently broad, inconsistently assigned, or incomplete, while keyword searches can introduce false positives. Consequently, retailer identification required a combination of automated retrieval procedures and manual quality assurance.

Since the current framework is intended to characterize retailer-generated digital marketing activity rather than point-of-sale product availability (which is more directly assessed by established audit tools such as STARS),[38] category and keyword criteria were designed specifically to identify specialty tobacco and nicotine retailers or venues. Businesses were identified using a combined category- and keyword-based retrieval strategy. Tobacco-related categories included tobacco shops, vape shops, smoke shops, hookah establishments, cigar retailers, head shops, herbal shops, and cannabis dispensaries. Additional candidate businesses were identified through tobacco-related keywords appearing in business names and descriptions. Categories and keywords used in retailer identification are provided in **Supplementary Table S1**.

Because this study focused on specialty tobacco retailers and venues, businesses with primary activities that were unrelated to tobacco retailing were excluded. Convenience stores, gas stations, grocery stores, pharmacies, and other general retailers were not targeted by the search strategy because these businesses are less likely to maintain dedicated tobacco-related social media accounts.

To improve identification accuracy, stratified samples of category-only businesses, keyword-only businesses, and category-plus-keyword businesses were manually reviewed. Findings from these reviews informed iterative refinement of inclusion and exclusion criteria and helped identify common failure modes in automated retailer identification.

Consistent with previous work, categories generally provided higher precision, while keyword searches improved recall by identifying businesses whose assigned categories did not fully reflect their tobacco-related activities. The final retailer inventory therefore relied on a hybrid retrieval strategy that balanced precision and recall. This balance between precision and recall reflects a trade-off common to many digital trace data applications, where maximizing coverage often increases the risk of false positives.

4.2 Geocoding

Following retailer identification and cleaning, business addresses were standardized and geocoded to obtain latitude and longitude coordinates.

Address standardization mitigated inconsistencies arising from formatting differences, abbreviations, incomplete address information, and other common data quality issues encountered in business listing platforms. Addresses were subsequently processed using Spectrum Geocoding software to assign geographic coordinates to each retailer location.

The resulting coordinates transformed business records into spatially referenced observations that could be mapped, aggregated, and linked to additional geographic information. Because geographic context is often absent from social media platforms, geocoding serves as the bridge between physical retailers and future measures of localized online marketing activity. More broadly, geocoding converts retailer records from administrative business listings into observations that can be used in spatial analyses and geographic record-linkage workflows.

4.3 Social Media Account Discovery

After retailer locations were established, the next stage involved identifying retailer-operated social media accounts.

Because businesses rarely share common identifiers across platforms, social media account discovery was treated as a record-linkage problem requiring integration of multiple incomplete information sources. Account discovery relied on three complementary strategies. Rather than relying on a single identifier, the procedure combined business metadata, website information, and platform-specific evidence to establish retailer-to-account linkages.

4.3.1 Structured Metadata

The first strategy leveraged structured business metadata available in Google Maps records. Google Maps frequently included direct website URLs and structured web-results fields containing links to external websites and social media profiles. When available, these links provided high-confidence retailer-to-platform matches.

4.3.2 Website-Based Extraction

For businesses with website information, social media accounts were identified through automated extraction procedures. Websites were scanned for Facebook, Instagram, and TikTok links embedded within:

- navigation menus,
- header and footer links,
- social media icons,
- metadata fields, and
- structured page markup.

Extracted links were standardized and deduplicated prior to account assignment.

4.3.3 Supplemental Discovery Procedures

Businesses lacking website information or structured social media metadata underwent supplemental discovery procedures. Retailer names, available business metadata, and publicly accessible online information were used to identify potential social media accounts.

When multiple candidate accounts were identified for a retailer, heuristic ranking procedures were applied. Candidate accounts were evaluated based on business-name similarity, consistency with retailer website information, and available geographic information associated with the account. The highest-scoring profile was retained as the retailer's primary account. This process can be viewed as an entity-resolution task in which multiple candidate records are evaluated and ranked according to evidence supporting a common underlying business identity.

Additional details regarding social media discovery and account-ranking procedures are provided in **Supplementary Methods S3**.

4.4 Validation

To evaluate inventory completeness, retailer records derived from Google Maps and Yelp were compared against an independently collected audit dataset from Appalachian Kentucky.

The Kentucky retailer audit included 523 verified tobacco retailers identified through field observation and retailer verification procedures. Because the present study focuses exclusively on specialty tobacco retailers and vape shops, validation was restricted to a benchmark sample of 68 specialty retailers (44 tobacco specialty retailers and 24 vape shops).

Retailer records were matched through manual review procedures that relied primarily on street addresses and secondarily on business names. Ambiguous cases were resolved using additional publicly available business information. Detailed matching procedures are provided in **Supplementary Methods S2**.

Retailer recall was calculated as the proportion of audited specialty retailers successfully identified within each business listing source. Because the validation benchmark was collected independently of the digital business listing platforms, recall provides an external assessment of inventory completeness rather than an internal measure derived from platform characteristics alone. This validation process provided an external benchmark for assessing inventory completeness and comparing source performance across Google Maps and Yelp.

5. Results

5.1 Inventory Construction

Application of the retailer identification and cleaning pipeline produced a national inventory of 87,646 businesses from Google Maps and 33,895 businesses from Yelp. Retailers represented a range of specialty tobacco and nicotine product retail categories, including tobacco shops, vape shops, smoke shops, cigar retailers, hookah businesses, head shops, herbal shops, and cannabis dispensaries.

Google Maps yielded approximately three times as many candidate businesses as Yelp, suggesting substantial differences in coverage across business listing platforms. Consistent with subsequent validation results, Google Maps provided substantially broader retailer coverage overall. Yelp identified fewer businesses but contributed supplemental retailer records and business metadata that supported retailer identification and downstream social media account discovery.

These findings illustrate a common challenge in digital trace data applications: no single source provides complete coverage of the target population. Consequently, retailer identification benefited from integrating multiple business listing platforms rather than relying on a single source. More broadly, the results demonstrate that platform selection can substantially influence the size and composition of inventories derived from digital trace data.

5.2 Validation

External validation against the Kentucky retailer audit benchmark demonstrated substantial differences in retailer recall across sources (**Table 1**).

Of the 68 audited specialty tobacco retailers and vape shops included in the validation sample, 62 retailers (91.1%) were identified by at least one business listing source. Google Maps captured 62 of 68 audited retailers (91.1% recall), whereas Yelp identified 13 retailers (19.1% recall). Notably, all retailers identified through Yelp were also identified through Google Maps, indicating that Yelp did not increase retailer recall beyond the businesses already captured by Google Maps within the validation sample.

Google Maps additionally identified two businesses that met audit eligibility criteria but were not included in the Kentucky audit roster. Businesses not recovered through the framework primarily reflected inactive listings, business ownership changes, or retailer name changes occurring between digital data collection and field verification.

Table 1. Recall of specialty tobacco retailers relative to the Kentucky audit benchmark

Source	*Audited Retailers Identified*	*Total Audited Retailers*	*Recall (%)*
Google Maps	62	68	91.1
Yelp	13	68	19.1

These validation results suggest substantial differences in source performance and demonstrate the value of independent benchmarks when constructing inventories from digital trace data. Without external validation, these differences would not have been observable from the business listing data alone.

5.3 Social Media Linkage

Social media account discovery identified substantial variation across platforms and business listing sources (**Table 2**).

Overall, 55,628 Google Maps retailers (63.5%) and 6,299 Yelp retailers (18.6%) were linked to at least one Facebook, Instagram, or TikTok account. Facebook was the most frequently identified platform, followed by Instagram and TikTok.

Table 2. Social media profile recovery by platform and source

Platform	***Google Maps n (%)***	***Yelp n (%)***
Facebook	42,702 (48.7%)	5,625 (16.6%)
Instagram	34,042 (38.8%)	5,083 (15.0%)
TikTok	5,880 (6.7%)	944 (2.8%)
Unique retailers with ≥1 platform	55,628 (63.5%)	6,299 (18.6%)

Note*: Counts for Facebook, Instagram, and TikTok are not mutually exclusive because retailers may maintain accounts on multiple platforms.*

The substantially higher account discovery rates observed for Google Maps were likely attributable to the availability of structured web-results fields and additional business metadata that facilitated retailer-to-platform linkage.

These findings demonstrate that source selection affects not only retailer identification but also downstream linkage success, a consideration that is often overlooked when constructing digital trace data measurement systems. The results also highlight the dependence of record-linkage performance on the availability and quality of platform-specific metadata.

5.4 Spatial Distribution

Retailers were identified in all U.S. states. Geocoding successfully assigned geographic coordinates to 89,905 of 90,055 Google Maps businesses (99.8%) and 35,510 of 35,822 Yelp businesses (99.1%), resulting in minimal geographic data loss.

Retailer concentrations were highest in major metropolitan regions, particularly Southern California, the New York City region, and the Chicago metropolitan area, although substantial numbers of specialty retailers were also identified in nonmetropolitan regions.

Figure 2 depicts the national distribution and concentration of specialty tobacco retailers with identified social media presence.

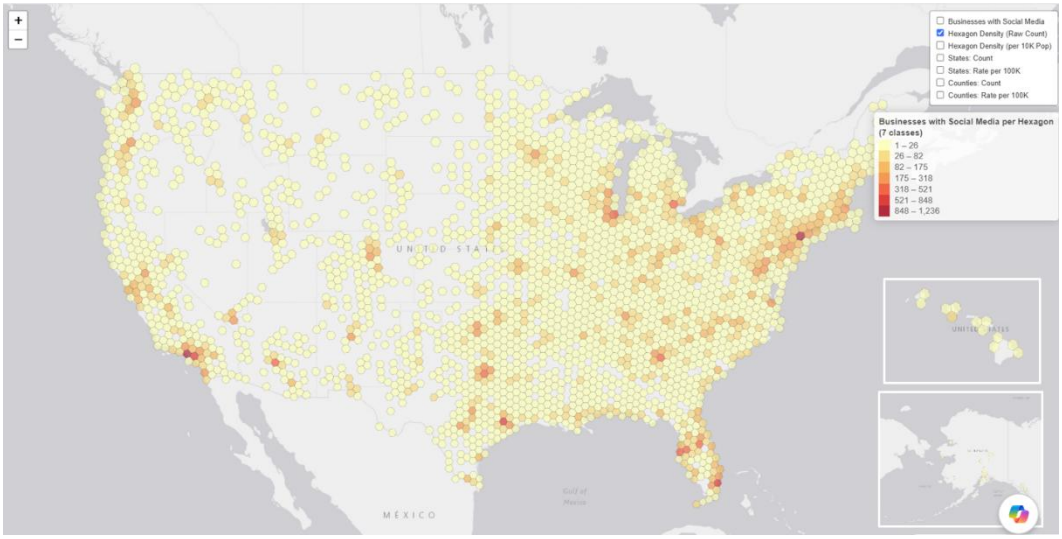


Figure 2. National distribution and concentration of tobacco and nicotine product retailers with identified social media presence

State-level variation in retailer social media density was also substantial. **Figure 3** illustrates the density of specialty tobacco retailers with identified social media accounts per 100,000 population. Oklahoma exhibited the highest retailer social media density (38.5 retailers per 100,000 population; n=1,522), followed by New Mexico (31.0 retailers per 100,000 population; n=651).

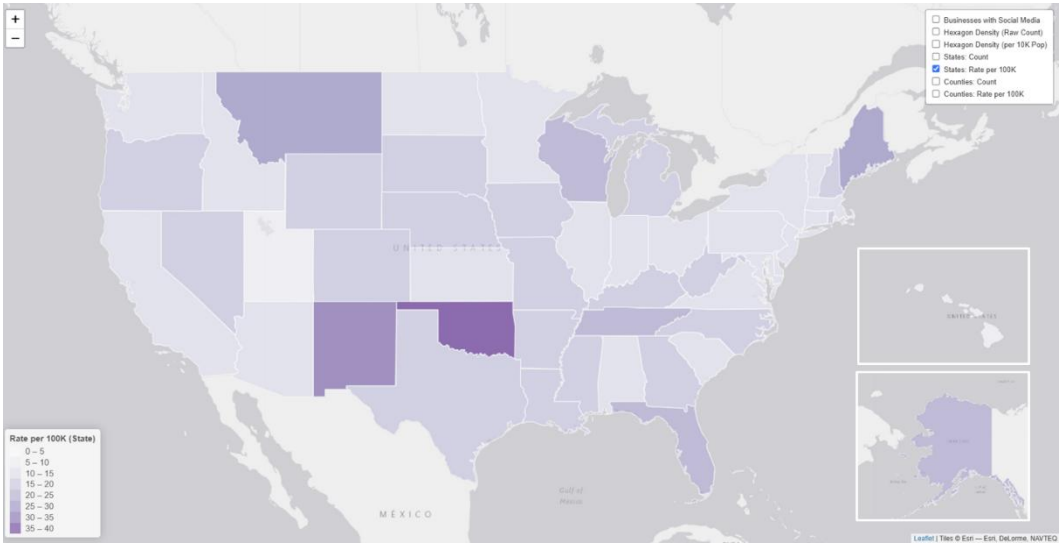


Figure 3. Per capita distribution and concentration of specialty tobacco retailers with social media accounts across U.S. states

Together, these results demonstrate that the framework successfully transforms business listing records into spatially referenced retailer inventories that can support geographic analyses at

multiple scales. The high geocoding success rates further indicate that location information contained within digital business listings can be effectively converted into spatially analyzable data products.

5.5 Coverage Diagnostics

Several barriers to social media account discovery were observed. Approximately half of businesses lacked website information that could support automated account discovery. Additional challenges arose from website-processing limitations, including age-verification systems, JavaScript-rendered content, inaccessible pages, and missing outbound social media links. Platform restrictions and incomplete cross-platform linkage further limited account recovery.

Diagnostic logs maintained throughout the account-discovery workflow were used to classify unresolved cases according to specific failure modes. More than 3,600 age-gated websites were identified among Google Maps businesses. While age-verification systems may reflect responsible retailer practices and regulatory compliance, they also reduced automated access to retailer information and limited opportunities for social media account discovery.

These findings highlight the practical challenges involved in linking physical businesses to online platforms and underscore the importance of documenting uncertainty throughout digital trace data workflows. More broadly, the diagnostics demonstrate that incomplete account recovery reflects both limitations of the underlying data sources and limitations inherent to the record-linkage process itself.

6. Methodological Lessons

The tobacco retail application presented in this study illustrates broader challenges that arise when constructing measurement systems from digital trace data. Although the substantive focus is tobacco retail marketing, the lessons learned extend well beyond this use case and are relevant to researchers integrating business listings, social media platforms, geospatial information, and other digital data sources.

6.1 Coverage Differs Substantially Across Data Sources

One of the most important findings of this study is that inventory completeness depends heavily on the choice of data source. Although Google Maps and Yelp were both designed to provide business information, they differed substantially in retailer coverage and in their ability to support downstream social media account discovery.

Google Maps identified more than three times as many specialty tobacco retailers as Yelp and achieved substantially higher recall when evaluated against the Kentucky audit benchmark. Google Maps also yielded much higher social media account recovery rates. These differences demonstrate that data sources that appear similar at first glance can produce markedly different measurement outcomes.

More generally, digital trace data should not be treated as interchangeable. Source selection can influence inventory completeness, account recovery, geographic representation, and downstream analyses. Validation against external benchmarks is therefore essential whenever possible. Researchers should view platform selection as a methodological decision rather than merely a data acquisition decision, as different sources may produce substantially different representations of the same underlying population.

6.2 No Single Data Source Is Complete

A second lesson is that no single digital trace data source contains all information necessary to construct geolocated marketing indicators.

Business listing platforms provide retailer identity and location but generally do not capture online marketing activity. Social media platforms contain marketing activity but often provide limited

geographic information. Even within business listing platforms, information is incomplete. Some businesses lack websites, some contain inaccurate category assignments, and others may have outdated or missing information.

The framework therefore relies on integrating multiple sources of partial information rather than depending on a single authoritative data source. This observation is consistent with a broader challenge in digital trace data research, where meaningful measurement frequently requires combining multiple incomplete data systems. In this sense, inventory construction can be viewed as a data integration problem in which useful measurement emerges from the combination of complementary sources rather than any individual platform.

6.3 Social Media Linkage Is a Record-Linkage Problem

The process of identifying retailer-operated social media accounts illustrates a broader record-linkage challenge common in digital trace data applications.

Businesses and social media platforms rarely share common identifiers. Retailers may maintain multiple accounts for a single location (e.g., active and legacy accounts), operate both chain-level and location-specific accounts, change business names, maintain inconsistent branding across platforms, or provide incomplete metadata. These factors complicate retailer-to-platform linkage and require the integration of structured metadata, website information, heuristic matching procedures, and manual review.

Importantly, social media account discovery should not be viewed as a binary process in which accounts are either present or absent. Instead, account identification is constrained by platform architecture, metadata quality, website availability, and discoverability. The resulting uncertainty should be acknowledged and documented as part of the measurement process. The challenges encountered here are analogous to entity-resolution problems that arise in many digital trace data applications where records must be linked across systems that lack common identifiers.

6.4 Digital Trace Data Requires Human Review

Although the framework relies heavily on automated data collection and processing, human review remained necessary at multiple stages.

Manual review procedures were used to identify false positives generated through category and keyword searches, refine inclusion criteria, evaluate ambiguous retailer matches, and resolve uncertain social media account linkages. These procedures improved inventory quality and helped identify common failure modes that would not have been apparent through automated processing alone.

This finding highlights an important limitation of fully automated digital trace data workflows. While automation improves scalability, quality assurance often requires human oversight, particularly when entity identification and record linkage are involved. Rather than replacing analyst judgment, automated procedures and manual review operated as complementary components of the measurement process.

6.5 Validation Strengthens Digital Trace Data Measurement

Finally, this study demonstrates the value of external validation. Digital trace data are often used without independent verification, making it difficult to assess coverage, completeness, and measurement error.

Comprehensive field audits, such as the recent Appalachian Kentucky study, and standardized observational protocols such as STARS, remain the reference standard for characterizing product availability, placement, price, and point-of-sale promotion with documented measurement reliability.[30-32] The Appalachian Kentucky study further used ground truthing by field staff to identify additional retailers beyond what was on these lists, however, such approaches are labor intensive and time-limited to the period when data collectors are in the field.[30]

By comparing retailer inventories against an independently collected audit dataset, we were able to estimate retailer recall, identify differences between data sources, and quantify important limitations in inventory construction. Validation therefore served not only as a quality-control procedure but also as a mechanism for understanding the strengths and weaknesses of alternative data sources.

More broadly, validation should be viewed as a core component of digital trace data measurement frameworks, particularly when those frameworks are intended to support surveillance, geospatial analysis, or downstream inference.

Used this way, the framework functions as an enhanced sampling frame: state and local tobacco control programs or research teams could use digitally derived inventories to identify newly opened retailers, flag likely closures between audit waves, and prioritize which stores to visit for STARS or other in-person surveillance, improving the timeliness and completeness of field-based retail marketing surveillance rather than replacing it.[39]

7. Limitations

The framework is subject to several limitations. First, the framework depends on commercial business listing platforms that were not designed for research purposes. Business classifications may be incomplete or inaccurate, retailer information may be outdated, and platform coverage may not fully represent specialty tobacco retailers operating within a geographic area. Additionally, there may be systematic differences in business coverage that are undetermined (i.e., it is unknown if the fewer identified retailers in rural areas in the US is due to no retailers being in those areas or systematic under-referencing of retailers in such areas). Consequently, some businesses may be misclassified or absent from the inventory despite efforts to improve retailer identification through category filtering, keyword searches, and manual review. Like many digital trace data sources, business listings provide an imperfect representation of the underlying population and may introduce coverage and classification error that cannot be fully observed.

Second, validation was conducted using an independently collected retailer audit dataset from a single geographic region. Although the Kentucky audit provided an important external benchmark and demonstrated high retailer recall for Google Maps, coverage may vary across states, metropolitan areas, and business platforms. Additional validation studies in other geographic contexts would strengthen confidence in the generalizability of inventory coverage estimates. Future applications would benefit from validation against multiple independent benchmarks representing diverse geographic and regulatory environments.

Third, retail environments are inherently dynamic. Businesses open, close, relocate, rebrand, and change ownership over time. As a result, any inventory derived from digital trace data should be interpreted as a snapshot of the retail environment corresponding to the period of data collection. Routine updates are therefore necessary to maintain inventory accuracy. This issue is not unique to tobacco retail and represents a broader challenge for measurement systems that rely on continuously evolving digital data sources.

Finally, social media account discovery remains imperfect. Many retailers lack websites, maintain limited online presence, or use platform architectures that complicate automated account identification. Platform-specific functionality, privacy settings, age restrictions, and multiple-account structures may reduce account discoverability. Consequently, failure to identify a social media account should not be interpreted as evidence that a retailer does not maintain one. Instead, the resulting measures should be interpreted as estimates of discoverable retailer social media presence rather than complete measures of platform participation. More generally, retailer-to-platform linkage is an entity-resolution problem in which uncertainty arises from incomplete metadata, inconsistent naming conventions, and the absence of common identifiers across systems.

Despite these limitations, the framework demonstrates that multiple digital trace data sources can be integrated, validated, and geolocated to produce a scalable inventory of specialty tobacco

retailers and their online presence. Such inventories provide a useful foundation for future surveillance, record linkage, and geospatial analyses using digital trace data. The limitations reflect challenges inherent to digital trace data measurement rather than limitations unique to the tobacco retail application presented here.

8. Conclusion

This study demonstrates a scalable framework for constructing geolocated business inventories from digital trace data using specialty tobacco retailers as a methodological case study. By integrating business listing platforms, geospatial processing, social media account discovery, and external validation, the framework addresses a common challenge in digital trace data research: connecting online activity to geographic context when relevant information exists across multiple incomplete data systems. More broadly, the framework illustrates how fragmented digital traces can be transformed into coherent measurement systems through data integration, record linkage, and validation.

The results highlight substantial differences in source performance, with Google Maps providing broader retailer coverage and higher social media account recovery than Yelp. More broadly, the study demonstrates the importance of validation, multi-source integration, and quality assurance when constructing measurement systems from digital trace data. These findings further demonstrate that seemingly similar digital platforms can yield markedly different measurement outcomes, reinforcing the need for careful source evaluation.

Although developed for specialty tobacco retailers, the framework is applicable to a wide range of business, marketing, and public health research settings where geographic locations must be linked to online activity. As access to platform data continues to evolve, methods that integrate business listings, geospatial information, and digital trace data will become increasingly important tools for surveillance, record linkage, and spatial analysis. The same general approach could be applied whenever researchers seek to connect entities, locations, and online activity across otherwise disconnected data systems.

Ultimately, tobacco retail serves as a useful demonstration of a broader methodological principle: digital trace data can be transformed into validated, geolocated measurement systems when multiple incomplete sources are systematically integrated and evaluated. This approach provides a foundation for future work seeking to link online behaviors and activities to physical places and communities. As digital data sources continue to expand, frameworks that emphasize measurement quality, validation, and transparent linkage procedures will become increasingly important for producing reliable and reproducible research.

Funding

Research reported in this publication was supported by the National Institute on Drug Abuse under Award Number R01DA060246. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Declaration of Interests

The authors have no conflicts of interest to declare.

Ethics Approval

Research activities for this project were reviewed and approved by the NORC Institutional Review Board (Protocol ID: 25-07-2305; approval date: July 23, 2025).

Data Availability

Data are proprietary; however, the authors can share aggregated data upon request.

Appendix

Supplementary Table S1. Business Identification Categories and Keywords

Businesses were retained if they matched at least one tobacco-related category or contained one or more tobacco-related keywords in business names, descriptions, or associated metadata fields. Categories were harmonized across Google Maps and Yelp for comparability across platforms.

Business Categories	Tobacco Shop Vape Shop Smoke Shop Hookah Shop Cigar Retailer Head Shop Herbal Shop Cannabis Dispensary Cigar Bar
Search Keywords	tobacco vape nicotine hookah cigar cigarette vapor disposables smoke shop

Supplementary Methods S1. Kentucky Validation Dataset

To evaluate inventory completeness, retailer records identified through Google Maps and Yelp were compared against an independently collected tobacco retailer audit dataset from Appalachian Kentucky. The store audit study has been described in detail elsewhere.[30] The dataset included 523 verified tobacco retailers identified through field observation and retailer verification procedures conducted between June and October 2025 across eight counties. Eligibility for the store audits was that the store needed to sell at least one type of commercial nicotine/tobacco product. Locations like bars or lounges primarily for on-site consumption were excluded. Validation analyses were restricted to specialty tobacco retailers and vape shops to align with the target population represented in the digital inventory. The resulting validation subset included 68 audited specialty retailers, comprising 44 tobacco specialty retailers and 24 vape shops.

Supplementary Methods S2. Validation Matching Protocol

Validation relied on manual reconciliation rather than automated record linkage.

Address Matching

Street addresses served as the primary matching variable. Street name, street number, city, ZIP code, and related location identifiers were reviewed to determine whether audited retailers corresponded to businesses identified within digital business listing datasets. Address matching was especially useful for resolving cases involving business renaming or ownership changes.

Name Matching

Business names were used as a secondary matching criterion. Partial matching was allowed to accommodate business abbreviations, alternate spellings, stylistic naming variations, and branding changes. When exact matches were unavailable, businesses were reviewed manually using additional publicly available business information.

Resolution of Ambiguous Cases

Potential matches that could not be resolved using names and addresses alone were reviewed using publicly available websites, social media profiles, mapping services, and online business directories. Cases involving business closures, relocations, ownership changes, or multiple candidate matches were resolved through consensus review.

Supplementary Methods S3. Social Media Account Discovery Procedures

Social media accounts were identified using a multi-stage discovery procedure combining structured business metadata and website-based extraction. Candidate accounts were collected for Facebook, Instagram, and TikTok.

Structured Metadata Extraction

For Google Maps businesses, social media accounts were identified from:

- Business website URLs
- Structured web-results fields
- Direct social media links provided within business metadata

When a direct social media URL was available, the account was assigned to the retailer without additional matching.

Website-Based Extraction

For businesses with a valid website, website content was scanned for outbound links to Facebook, Instagram, and TikTok profiles. Candidate social media links were extracted from:

- Navigation menus
- Header and footer links
- Social media icons
- Structured metadata (e.g., Open Graph tags)
- JSON-LD markup
- Contact and About pages

Extracted links were standardized to remove tracking parameters and duplicate URLs.

Candidate Account Ranking

When multiple candidate social media accounts were identified for the same retailer and platform, accounts were ranked using a heuristic scoring procedure based on:

- Business name similarity
- Exact matches between retailer and account name received highest priority.
- Partial matches and common abbreviations received lower scores.
- Website consistency

Accounts linking back to the retailer website received additional weight. The highest-scoring account was retained as the retailer's primary account for that platform.

References

1. Borowiecki M, Emery SL, Kostygina G. New recreational nicotine lozenges, tablets, gummies and gum proliferate on the US market. *Tobacco Control* 2022:tc–2022–057673. doi: 10.1136/tc-2022-057673
2. O'Connor R, Schneller LM, Felicione NJ, et al. Evolution of tobacco products: recent history and future directions. *Tobacco Control* 2022;31(2):175–82. doi: 10.1136/tobaccocontrol-2021-056544
3. Czaplicki L, Kostygina G, Kim Y, et al. Characterizing JUUL-related posts on Instagram. *Tobacco Control* 2019:Published Online First: 02 July 2019. doi: 10.1136/tobaccocontrol-2018-054824
4. Kostygina G, Tran H, Kim Y, et al. Heating Up: How Early Twitter Marketing Gave Rise to Organic Word-of-Mouth About Heated Tobacco Products. *Social Media + Society* 2022;8(4):20563051221141851. doi: 10.1177/20563051221141851
5. Kostygina G, Tran H, Schillo B, et al. Industry response to strengthened regulations: amount and themes of flavoured electronic cigarette promotion by product vendors and manufacturers on Instagram. *Tob Control* 2022;31(Suppl 3):s249–s54. doi: 10.1136/tc-2022-057490
6. Kostygina G, Tran H, Carter CC, et al. Promotion of Health-Harming Products on Instagram: Characterizing Strategies Boosting Audience Engagement with Cigar Marketing Messages. *International Journal of Environmental Research and Public Health* 2025;22(8):1285.
7. Kim Y, Kresovich A, Page S, et al. Copromotion across commercial determinants of health: content analysis of smokeless tobacco social media messages referencing alcohol, 2016–2021. *Tobacco Control* 2025:tc–2025–059523. doi: 10.1136/tc-2025-059523
8. Huang J, Duan Z, Kwok J, et al. Vaping versus JUULing: how the extraordinary growth and marketing of JUUL transformed the US retail e-cigarette market. *Tobacco Control* 2019;28(2):146–51. doi: 10.1136/tobaccocontrol-2018-054382
9. Czaplicki L, Kostygina G, Kim Y, et al. Characterising JUUL-related posts on Instagram. *Tobacco Control* 2019:tobaccocontrol–2018–054824. doi: 10.1136/tobaccocontrol-2018-054824
10. Kostygina G, Tran H, Czaplicki L, et al. Developing a theoretical marketing framework to analyse JUUL and compatible e-cigarette product promotion on Instagram. *Tobacco Control* 2022:tobaccocontrol–2021–057120. doi: 10.1136/tobaccocontrol-2021-057120
11. Kresovich A, Norris AH, Carter CC, et al. Deciphering Influence on Social Media: A Comparative Analysis of Influential Account Detection Metrics in the Context of Tobacco Promotion. *Social Media + Society* 2024;10(1):20563051231224268. doi: 10.1177/20563051231224268
12. Brawn Media. Why Facebook Ads Are Crucial for Local Businesses in 2025. 2025 [Available from: https://brawnmediany.com/blog/why-facebook-ads-are-crucial-for-local-businesses-in-2025/.
13. Blakey E. The day data transparency died: How Twitter/X cut off access for social research. *Contexts* 2024;23(2):30–35.
14. Mimizuka K, Brown MA, Yang K-C, et al. Post-post-api age: studying digital platforms in scant data access times. *arXiv preprint arXiv:250509877* 2025
15. Scharfbillig M, Lewandowsky S, Altay S, et al. Fractured reality. 2026
16. Kostygina G, Tran H, Shi Y, et al. 'Sweeter Than a Swisher': amount and themes of little cigar and cigarillo content on Twitter. *Tob Control* 2016;25(Suppl 1):i75–i82. doi: 10.1136/tobaccocontrol-2016-053094 [published Online First: 2016/10/05]
17. Kong G, Lee J, Celentano M, et al. The promotion of premium cigars on social media. *Nicotine and Tobacco Research* 2023;25(Supplement_1):59–64.
18. Cantrell J, Anesetti-Rothermel A, Pearson JL, et al. The impact of the tobacco retail outlet environment on adult cessation and differences by neighborhood poverty. *Addiction* 2015;110(1):152–61.

19. Ribisl KM, D'Angelo H, Feld AL, et al. Disparities in tobacco marketing and product availability at the point of sale: Results of a national study. *Preventive medicine* 2017;105:381–88. doi: 10.1016/j.ypmed.2017.04.010
20. Zhu S-H, Sun JY, Bonnevie E, et al. Four hundred and sixty brands of e-cigarettes and counting: implications for product regulation. *Tobacco control* 2014;23(suppl 3):iii3–iii9.
21. Ali FRM. E-cigarette unit sales, by product and flavor type—United States, 2014–2020. *MMWR Morbidity and mortality weekly report* 2020;69
22. Diaz MC, Silver NA, Bertrand A, et al. Bigger, stronger and cheaper: growth in e-cigarette market driven by disposable devices with more e-liquid, higher nicotine concentration and declining prices. *Tobacco Control* 2025;34(1):65–70.
23. Gammon DG, Loomis BR, Dench DL, et al. Effect of price changes in little cigars and cigarettes on little cigar sales: USA, Q4 2011-Q4 2013. *Tob Control* 2016;25(5):538–44. doi: 10.1136/tobaccocontrol-2015-052343 [published Online First: 20150910]
24. Wang TW, Falvey K, Gammon DG, et al. Sales Trends in Price-Discounted Cigarettes, Large Cigars, Little Cigars, and Cigarillos—United States, 2011–2016. *Nicotine & Tobacco Research* 2017;20(11):1401–06. doi: 10.1093/ntr/ntx249
25. Gammon DG, Rogers T, Coats EM, et al. National and state patterns of concept-flavoured cigar sales, USA, 2012-2016. *Tob Control* 2019;28(4):394–400. doi: 10.1136/tobaccocontrol-2018-054348 [published Online First: 20180801]
26. Gammon DG, Rogers T, Gaber J, et al. Implementation of a comprehensive flavoured tobacco product sales restriction and retail tobacco sales. *Tobacco Control* 2022;31(e2):e104–e10. doi: 10.1136/tobaccocontrol-2021-056494
27. Golden SD, Baggett CD, Kuo TM, et al. Trends in the Number and Type of Tobacco Product Retailers, United States, 2000-2017. *Nicotine Tob Res* 2022;24(1):77–84. doi: 10.1093/ntr/ntab150
28. Apollonio DE, Dennehy CE, Tsourounis C, et al. An observational study of regulatory violations related to online tobacco product marketing and retailer responses to US FDA warning letters. *Tobacco Use Insights* 2024;17:1179173X241300825.
29. Nadeem W, Khani AH, Schultz CD, et al. How social presence drives commitment and loyalty with online brand communities? The role of social commerce trust. *Journal of Retailing and Consumer Services* 2020;55:102136.
30. Feizy S, van de Venne J, Abadi M, et al. Commercial tobacco retail marketing in rural communities in Appalachian Kentucky, 2024. *Preventive Medicine Reports* 2025:103354. doi: https://doi.org/10.1016/j.pmedr.2025.103354
31. Henriksen L, Ribisl KM, Rogers T, et al. Standardized Tobacco Assessment for Retail Settings (STARS): dissemination and implementation research. *Tobacco Control* 2016;25(Suppl 1):i67–i74. doi: 10.1136/tobaccocontrol-2016-053076
32. Kong AY, Eaddy JL, Morrison SL, et al. Using the Vape Shop Standardized Tobacco Assessment for Retail Settings (V-STARS) to Assess Product Availability, Price Promotions, and Messaging in New Hampshire Vape Shop Retailers. *Tob Regul Sci* 2017;3(2):174–82. doi: 10.18001/trs.3.2.5 [published Online First: 20170401]
33. Lee JGL, D'Angelo H, Kuteh JD, et al. Identification of Vape Shops in Two North Carolina Counties: An Approach for States without Retailer Licensing. *International Journal of Environmental Research and Public Health* 2016;13(11):1050.
34. Berg CJ, Schleicher NC, Johnson TO, et al. Vape shop identification, density and place characteristics in six metropolitan areas across the US. *Preventive Medicine Reports* 2020;19:101137. doi: https://doi.org/10.1016/j.pmedr.2020.101137
35. Kim AE, Loomis B, Rhodes B, et al. Identifying e-cigarette vape stores: description of an online search methodology. *Tobacco Control* 2016;25(e1):e19–e23. doi: 10.1136/tobaccocontrol-2015-052270
36. Landini A, Lowenstein C, Pesko MF. Patterns of specialty tobacco retail locations and visitor counts in the USA. *Tobacco Control* 2026:tc–2026–060085. doi: 10.1136/tc-2026-060085
37. D'Angelo H, Fleischhacker S, Rose SW, et al. Field validation of secondary data sources for enumerating retail tobacco outlets in a state without tobacco outlet licensing. *Health & Place* 2014;28:38–44. doi: https://doi.org/10.1016/j.healthplace.2014.03.006

38. Lee JGL, Henriksen L, Myers AE, et al. A systematic review of store audit methods for assessing tobacco marketing and products at the point of sale. *Tobacco Control* 2014;23(2):98–106. doi: 10.1136/tobaccocontrol-2012-050807
39. Hrywna M, Ackerman C, Lee JGL, et al. Development and implementation of a rapid surveillance system for tobacco product retail availability in six U.S. sites. *Preventive Medicine Reports* 2026;66:103343. doi: https://doi.org/10.1016/j.pmedr.2025.103343