Title

Validation of secondary data sources for enumerating marijuana dispensaries in a state commercializing marijuana

Authors

Cao, Yiwen
Carrillo, Angelina S
Jankowska, Marta M
et al.

Publication Date

2020-10-01

DOI

10.1016/j.drugalcdep.2020.108183

Peer reviewed

Validation of Secondary Data Sources for Enumerating Marijuana Dispensaries in a State Commercializing Marijuana

Yiwen Cao[1], Angelina S. Carrillo[1], Marta M. Jankowska[2], Yuyan Shi[1]*

[1] Department of Family Medicine and Public Health, University of California San Diego, CA, USA

[2] Qualcomm Institute/Calit2, University of California San Diego, CA, USA

***Corresponding author**:

Yuyan Shi

9500 Gilman Drive, MC0628, La Jolla, CA 92093-0628, USA

Phone number: 1(858)534-4273

Email address: yus001@ucsd.edu

Abstract

Objectives: To assess 1) the validity of online crowdsourcing platforms in enumerating licensed brick-and-mortar marijuana dispensaries and 2) the validity of state licensing directory and online crowdsourcing platforms in enumerating active brick-and-mortar marijuana dispensaries in California.

Methods: We obtained business lists from California Bureau of Cannabis Control (BCC) licensing directory and three online crowdsourcing platforms (Weedmaps, Leafly, and Yelp) in May 2019. Calls were made to verify street address, operation status, dispensary category (recreational-only, medical-only, recreational & medical), and presence of storefronts in May-July 2019. Validity measures, including sensitivity, specificity, positive predictive value, and negative predictive value, were calculated when applicable.

***Results*:** In identifying licensed dispensaries in BCC, Leafly had the highest sensitivity (.66) and Yelp had the highest specificity (.87). The dispensary category posted on online crowdsourcing platforms in over 25% licensed dispensaries and the dispensary category claimed in call verification in over 10% licensed dispensaries disagreed with the approved category in BCC. There were 2,121 businesses combined from BCC and online crowdsourcing platforms, among which 826 were verified to be active brick-and-mortar dispensaries. Weedmaps had the highest sensitivity (.80) and Yelp had the highest negative predictive value (.74) in identifying verified dispensaries. Weedmaps overall had the highest sensitivity in all three dispensary categories. Weedmaps had the highest sensitivity in more populated counties whereas BCC had the highest sensitivity in less populated counties.

Conclusions: Each secondary data source has strengths and limitations. The findings inform surveillance and research regarding how to best strategize data use when resources are limited.

Key Words

Marijuana dispensaries; brick-and-mortar outlets; marijuana commercialization; crowdsourcing; sensitivity and specificity

Abbreviations

BCC – California Bureau of Cannabis Control

PPV – Positive Predictive Value

NPV – Negative Predictive Value

1. Introduction

Following recreational marijuana legalization and commercialization in the US, marijuana dispensaries have served as a major venue for marijuana retail sales in neighborhoods. Nonetheless, research on the impacts of marijuana dispensaries on public health remains limited (Berg et al., 2018). Availability, accessibility, and point-of-sale marketing of retail outlets have been associated with attitudes, perceptions, and health behaviors in tobacco and alcohol literature (Anderson et al., 2009; Campbell et al., 2009; Finan et al., 2019; Henriksen et al., 2008; Lovato et al., 2011; McCarthy et al., 2009; Paynter and Edwards, 2009; Reitzel et al., 2011; Smith and Foxcroft, 2009). Marijuana dispensaries may impact marijuana-related outcomes in a similar manner. They may increase availability and accessibility of marijuana (Paschall and Grube, 2020), promote greater awareness and consumption through marketing activities (D'Amico et al., 2018; Fiala et al., 2018), increase product appeal such as through increased quality and potency (Orens et al., 2018), diversify product variation such as vaping devices and edibles (Tormohlen et al., 2019), reduce prices through mass production and introduction of competition (Hall and Lynskey, 2016), and shape social norms favorable of marijuana use (Berg et al., 2018; Lipperman-Kreda and Grube, 2018).

A major challenge in understanding the availability and retail environments of marijuana dispensaries is identifying a complete and accurate list of marijuana dispensaries in neighborhoods. In a state operating a statewide licensing system, one can obtain the official licensing directories from government databases. Nonetheless, most of these directories are updated infrequently. More importantly, they do not reflect the operation status of dispensaries in reality or capture unlicensed dispensaries that are common in areas with weak law enforcement.

Business directories provided by commercial providers (e.g., InfoUSA, Dun & Bradstreet) are commonly used to identify tobacco, alcohol, and food retail outlets when state licensing directories are unavailable or unsatisfactory (Carlos et al., 2017; D'Angelo et al., 2014; Gustafson et al., 2012; Lake et al., 2010; Liese et al., 2010; Powell et al., 2011; Seliske et al., 2012). Unfortunately, these commercial databases had not systematically gathered information on marijuana dispensaries by the time of this study. One can also conduct a field census with direct search and observation to enumerate a certain type of business in a geographic area. It is considered to be the best practice in outlet identification and often used to validate the business lists obtained from commercial databases (D'Angelo et al., 2014; Gustafson et al., 2012; Liese et al., 2010; Powell et al., 2011; Seliske et al., 2012). The limitation of field census is obvious: the required efforts and resources increase exponentially as the geographic area of interest expands. Due to practical and budget concerns, most tobacco, alcohol, and food outlet studies that adopted this method searched retail outlets in smaller regions such as a county (D'Angelo et al., 2014; Gustafson et al., 2012; Liese et al., 2010). State-level field censuses, especially in a large state like California, are nearly nonexistent.

In light of the challenges of using conventional approaches to identify marijuana dispensaries, existing studies have primarily relied upon a single or a few online crowdsourcing platforms, such as Weedmaps, Leafly, and Yelp, to obtain dispensary information voluntarily submitted by dispensary owners and marijuana users (Freisthler and Gruenewald, 2014; Freisthler et al., 2016; Mair et al., 2015; Shi et al., 2018; Shi et al., 2016; Shi, 2016; Shih et al., 2019). Because these platforms serve as online communities to promote dispensaries, products, and share experiences, they are perceived to be more up-to-date and comprehensive than official licensing directories. Particularly, these platforms provide data on both licensed and unlicensed

dispensaries. Despite the increasingly common use of online crowdsourcing platforms in marijuana research, the validity of this approach has not been comprehensively assessed at statewide level. To date, only two studies have conducted validation in a single county (both in Los Angeles County), one before recreational marijuana commercialization (Pedersen et al., 2018) and one after the commercialization (Pedersen et al., 2020) in California.

In this study, we examined the validity of using secondary data sources, including the state licensing directory and commonly used online crowdsourcing platforms, in enumerating brick-and-mortar marijuana dispensaries across the entire state of California. California is the most populous state with the longest history of medical marijuana legalization (since 1996) in the US. In November 2016 California legalized recreational marijuana and in January 2018 California initiated retail sale of recreational marijuana in dispensaries. California now has the largest legal marijuana market in the world, with sales rising from $2.5 billion in 2018 to $3.1 billion in 2019 (Mcgreevy, 2019). Although California allows delivery services, in this study, we concentrated only on brick-and-mortar marijuana dispensaries because delivery-only providers do not have storefronts to showcase and promote products. In addition, the wide geographic coverage of delivery services (usually the entire city or county) contributes little variation in marijuana availability at neighborhood level.

We offered a protocol for identifying dispensaries that can be replicated in other large geographic regions with marijuana retail sales. We aimed to answer two research questions. *The first question* was to what extent online crowdsourcing platforms are valid in enumerating licensed brick-and-mortar dispensaries. The motivation was that many dispensaries in California operated without a license. (Pedersen et al., 2020) Even for licensed dispensaries, how they operate in practice may not agree with what was approved in the license. Findings from the first

question will provide quantifiable evidence on the level of agreement between state licensing directory and online crowdsourcing platforms, add surveillance data point on the operation of unlicensed dispensaries, and inform policymakers regarding the validity of using online crowdsourcing platforms as alternatives when state licensing directory is not publicly accessible or licensing information is inadequate (e.g., no street address). *The second question* was to what extent state licensing directory and online crowdsourcing platforms are valid in enumerating the universe of active brick-and-mortar dispensaries. The motivation was that a single data source may not capture all active dispensaries in California and the information in a data source may not agree with how dispensaries operate in practice. Findings from the second question will provide quantifiable evidence on the strengths and weaknesses of each data source, inform surveillance and research regarding how to best strategize data use when resources are limited, and demonstrate the need for combining multiple data sources and verifying information to obtain the universe of dispensaries in a large geographic area. Because recreational-only, medical-only, and recreational & medical dispensaries co-existed in California, we also assessed validity measures by dispensary category. Dispensaries may tend to promote themselves on online crowdsourcing platforms in larger counties with keen competition, we hence further assessed validity measures by county population size.

2. Methods

2.1 Data Sources

In May 2019, we obtained marijuana business lists from multiple secondary data sources: 1) the state official licensing directory was obtained from the California Bureau of Cannabis Control (BCC) online license search portal, and 2) business directories were obtained from three

commonly used online crowdsourcing platforms, including Weedmaps, Leafly, and Yelp. Weedmaps and Leafly specialize in marijuana business listings, whereas Yelp provides general business listings encompassing various types of industries. Key words "marijuana", "weed", "cannabis", and "dispensary" that were commonly used in Yelp to describe marijuana-related businesses were used to search records on Yelp. All four secondary data sources contained information on business name, street address, phone number, and delivery services, but dispensary category (recreational only, medical only, recreational & medical) was only available on BCC, Weedmaps, and Leafly.

2.2 Online Data Cleaning

Because business listings included both delivery services and brick-and-mortar dispensaries, we first removed businesses if the online information indicated that they only provided delivery services without storefronts. We then removed duplicated records by hand if two or more dispensaries within a single data source had the same business name and street address. We further combined records from all four data sources and removed duplicated records across data sources. The cleaned, combined database included 2,121 unique businesses (Figure 1).

2.3 Call Verification

From May to July 2019, eight trained research associates aged 21 or older called the 2,121 unique businesses to verify their street address, operation status, category of business, and presence of storefronts (Figure 1). Each call took fewer than 5 minutes on average. As commonly done in compliance check inspections of tobacco product retailers, (FDA, 2020) the research associates did not reveal the research purpose of the calls. Instead, they identified themselves as interested customers who were considering a visit in near future. To determine

dispensary category, researchers asked if a doctor's recommendation or a patient registration card was required to enter the dispensary and make purchase. An affirmative response indicated the dispensary category to be medical only. If the response was negative yet customers with a doctor's recommendation or a patient registration card were eligible for reduced tax rates, the dispensary was categorized as recreational & medical. The remaining dispensaries were considered to be recreational only. Up to five calls were made to each business in different business hours and/or on different business days to determine operation status. If a dispensary could not be reached after five call attempts, researchers checked its recent online activities on Weedmaps, Leafly, Yelp, and Google Map Reviews. If the dispensary had any online activity within the past month (e.g., posted customer reviews, posted promotional offers), it would be considered active[1]. After removing inactive businesses, businesses not selling marijuana, and businesses without storefronts during the verification procedure, the 2,121 unique records were reduced to 826 businesses (Figure 1). These 826 dispensaries constituted the call-verified, combined database of active brick-and-mortar dispensaries in California.

2.4 Statistical Analysis

Validity statistics, including sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were computed for each of the four secondary data sources when applicable. Definitions and calculations were described in Technical Note S1.

To compute validity statistics, a gold standard must be defined that can identify the "true positive" and the "true negative". Field census is typically considered the gold standard in retail outlet research. However, it is infeasible in this study due to budget and time constraints for a

[1] Only two dispensaries were verified to be active based on their recent online activities. All the remaining 824 dispensaries were verified to be active by calls. We referred verified dispensaries as "call-verified" throughout the remaining of the manuscript with the understanding that two dispensaries were verified based on online activities.

statewide census. Two gold standards were adopted alternatively to answer the two research questions. To answer the first question regarding the validity of online crowdsourcing platforms in enumerating licensed brick-and-mortar marijuana dispensaries, the first gold standard was whether a record was listed in the BCC state licensing directory (Yes="true positive", No="true negative"). To answer the second question regarding the validity of state licensing directory and online crowdsourcing platforms in enumerating active brick-and-mortar marijuana dispensaries, the second gold standard was whether a record was included in the call-verified, combined database of active dispensaries (Yes="true positive", No="true negative").

We must also define a test that can identify the "positive test" and the "negative test" in validity statistics calculations. Two tests were conducted. The first test was whether a record was present in a given data source after online data cleaning (Yes="positive test", No="negative test"). We used this test to examine the validity of using a single data source with simple online data cleaning for dispensary identification, an approach requiring moderate resources. The second test was whether a record passed call verification; in other words, whether the record was verified to be an active brick-and-mortar dispensary (Yes="positive test", No="negative test"). We used this test to examine the validity of using a single data source with simple online data cleaning plus call verification for dispensary identification, an approach requiring much more resources.

To illustrate these validity statistics in the context of this study, we provide an example below (equations and explanations in Technical Note S1). In this example, the data source of interest is Weedmaps, the gold standard is whether a record on Weedmaps was present in the BCC state licensing directory, and the test is whether a record was present on Weedmaps after online data cleaning. *Sensitivity* measures the probability of a record present on Weedmaps

conditional on the record being included in the BCC directory, calculated as the number of records that were present on both Weedmaps and the BCC directory divided by the number of records present on the BCC directory. *Specificity* measures the probability of a record absent on Weedmaps conditional on the record being excluded from the BCC directory, calculated as the number of records that were neither present on Weedmaps nor present on the BCC directory divided by the number of records excluded from the BCC directory. *PPV* measures the probability of a record included in the BCC directory conditional on the record being present on Weedmaps, calculated as the number of records that were present on both Weedmaps and the BCC directory divided by the number of records present on Weedmaps. *NPV* measures the probability of a record excluded from the BCC directory conditional on the record being absent on Weedmaps, calculated as the number of records that were neither present on Weedmaps nor present on the BCC directory divided by the number of records being absent on Weedmaps. You will notice that specificity and NPV cannot be calculated in this example, because we were not able to identify a "true negative", a record that was excluded from Weedmaps and also absent in the BCC directory. In fact, not all validity statistics were applicable to a combination of a gold standard and a test with the current study design (details in Technical Note S1).

Following tobacco outlet research (D'Angelo et al., 2014), we considered validity statistics 0-0.2 to be poor, 0.21-0.4 to be fair, 0.41-0.6 to be moderate, 0.61-0.8 to be good, and 0.81-1.0 to be very good. R Version 3.5.3 (package "epiR") was used to calculate 95% confidence intervals for all the validity statistics. We computed overall statistics as well as the statistics by dispensary category (recreational only, medical only, recreational & medical) and county population size (over or fewer than one million population). Locations of call-verified active brick-and-mortar dispensaries in California were mapped with ArcGIS Version 10.5.

3. Results

3.1 Online Data Cleaning and Call Verification Results

A total of 2,121 business records were combined from BCC and the three online crowdsourcing platforms after online data cleaning. BCC, Weedmaps, Leafly, and Yelp had 630, 811, 535, and 1,468 records included in the combined database, respectively. The overlaps across the data sources were presented in Figure S1. Only 240 records were present in all four data sources.

Following call verification, the 2,121 records were reduced to 826, which were confirmed to be active brick-and-mortar dispensaries. Among the 1,295 records removed during call verification, 56.0% were closed, 4.2% were not open yet, 38.0% were not selling marijuana, and 1.8% had no storefronts (Figure 1). BCC, Weedmaps, Leafly, and Yelp had 486, 659, 459, and 471 records included in these 826 verified dispensaries, respectively. The overlaps across the data sources were presented in Figure S2. The 826 records included 77 recreational-only, 65 medical-only, and 684 recreational & medical dispensaries. The dispensary category was based on self-reporting by dispensary staff in call verification.

3.2 Validity Statistics

Table 1 (details in Table S1) reports validity statistics using the BCC licensing directory as the gold standard. When the test was whether being present on each online crowdsourcing platform after online data cleaning, Leafly had good sensitivity (.70) and Weedmaps and Yelp had moderate sensitivity (.59 and .53, respectively). It indicated that 70% of the BCC licensing directory could be found on Leafly. Leafly also had very good PPV (.83), yet Yelp's PPV was only fair (.23). It indicated that 83% of Leafly records were included in the BCC licensing

directory. When the test was whether passing call verification, Leafly still had the highest sensitivity (good: .66) and PPV (very good: .90), and Yelp had the highest specificity (very good: .87) and NPV (good: .76). It indicated that, call-verified Leafly records performed the best for identifying truly licensed dispensaries and call-verified Yelp records performed the best for identifying truly unlicensed dispensaries in this scenario.

Table 2 (details in Table S2) reports validity statistics using the call-verified, combined database as the gold standard. When the test was whether being present in each data source after online data cleaning, Weedmaps had the highest sensitivity (good: .80) and BCC, Leafly, and Yelp all had moderate level of sensitivity ranging from .56 to .59. It indicated that 80% of the call-verified, combined database of active dispensaries could be found on Weedmaps. Leafly and Weedmaps had very good PPV (.86 and .81, respectively), and Yelp's PPV was only fair (.32). It indicated that 86% of Leafly records were included in the call-verified, combined database of active dispensaries. When the test was whether passing call verification, sensitivity statistics remained the same as when the test was whether being present in each data source. This was because call-verified businesses in each data source were a subset of the businesses included in each data source before call verification, such that the numerators and denominators for sensitivity calculation remained the same. Yelp had the highest NPV (good: .74) and Leafly had the lowest NPV (poor: .17). It indicated that call-verified Yelp records performed the best for identifying truly not active brick-and-mortar dispensaries.

3.3 Validity Statistics by Dispensary Category

Table 3 reports the agreement between BCC, online crowdsourcing platforms, and call verification in terms of the category of the 630 licensed dispensaries. Approximately 25% of the licensed dispensaries on Weedmaps and 29% of the licensed dispensaries on Leafly posted their

category that disagreed with what was approved in the BCC license. Approximately 12% of the call-verified, licensed dispensaries stated their category in call verification that disagreed with what was approved in the BCC license. Most of the businesses that stated an unapproved category on online crowdsourcing platforms and/or in call verification claimed themselves to be recreational & medical when they were only licensed for recreational-only or medical-only.

Table S3 quantifies category-specific validity statistics when the gold standard was whether being present in the BCC licensing directory. Leafly had the highest sensitivity in recreational-only and recreational & medical categories and Weedmaps had the highest sensitivity in medical-only category, regardless of the definition of a test. Table S4 quantifies category-specific validity statistics when the gold standard was whether being present in the call-verified, combined database. When the test was whether being present in each data source after online data cleaning, Weedmaps had the highest sensitivity in identifying recreational-only and medical-only dispensaries, yet BCC had the highest sensitivity in identifying recreational & medical dispensaries. When the test was whether passing call verification, Weedmaps overall had the highest sensitivity in all three categories.

3.4 Validity Statistics by County Population Size

In 2019, California had 16 counties with a population size above one million and 42 counties with a population size below one million. Table S5 reports validity statistics by county population size when the gold standard was whether being present in the BCC licensing directory. Leafly had the highest sensitivity regardless of test definition and county population size. Table S6 reports validity statistics by county population size when the gold standard was whether being present in the call-verified, combined database. Regardless of test definition,

Weedmaps had the highest sensitivity in more populated counties and BCC had the highest sensitivity in less populated counties.

3.5 Mapping of Call-verified, Active Dispensaries

Call-verified, active brick-and-mortar dispensaries were mapped in Figure S3 by dispensary category and county population size. Los Angeles County had the largest number of dispensaries, followed by Riverside County and San Diego County.

4. Discussion

This study is the first to assess the validity of secondary data sources in identifying brick-and-mortar marijuana dispensaries across a large state. We reported the validity of online crowdsourcing platforms in enumerating licensed dispensaries and the validity of state licensing directory and online crowdsourcing platforms in enumerating active dispensaries.

Regarding the validity of using online crowdsourcing platforms in identifying the BCC licensing directory, all three online crowdsourcing platforms were able to include over 50% records in the BCC directory, with Leafly containing the largest number of licensed dispensaries (70%). These findings suggested that the online crowdsourcing platforms could serve as a reasonable proxy for the licensing directory. It evidences the validity for many existing and future studies to utilize online crowdsourcing platforms for dispensary identification, especially if a licensing system is not open to the public or is updated infrequently. It should be noted, however, that the dispensary category registered in the BCC directory may be mismatched with the "de facto" category in which dispensaries operated. Over 25% licensed dispensaries on online crowdsourcing platforms posted their category that disagreed with the BCC license and over 10% call-verified, licensed dispensaries stated their category in call verification that

disagreed with the BCC license. Particularly, most of such dispensaries claimed themselves to be recreational & medical while they were only licensed for recreational only or medical only. Such disagreement might be intentionally used as a means of attracting customers or be reflective of how dispensaries operate in practice.

Regarding the validity of using the state licensing directory in identifying active brick-and-mortar dispensaries, over 20% licensed dispensaries did not pass call verification. This indicated that business licenses may not accurately represent businesses' operation status in reality. For instance, a business may have been closed before its license is expired and a business may not be open yet even though its license has been approved. In the final 826 call-verified dispensaries, 58.8% (486) were included in the BCC licensing directory. This indicated that the BCC directory failed to capture unlicensed dispensaries, which accounted for over 40% of the total active dispensaries in California. Solely relying on a state licensing directory would overestimate active, licensed dispensaries whereby overlook active, unlicensed dispensaries.

Regarding the validity of using online crowdsourcing platforms in identifying active brick-and-mortar dispensaries, Weedmaps had a nearly very good sensitivity; it contributed 80% of the records in the final call-verified, combined database. It had the highest sensitivity in identifying recreational-only and medical-only dispensaries. It was also the most sensitive database in identifying dispensaries in more populated counties, which were mostly urban areas. The high concentration of dispensaries and intense competition in urban areas may motivate more businesses to promote themselves on this highly visible and popular platform (Pedersen et al., 2018). Leafly had the lowest sensitivity in identifying active dispensaries. It also had the lowest sensitivity in identifying all three dispensary categories. It is likely because the costs of advertising on Leafly were substantially higher than other online crowdsourcing platforms

specialized in marijuana (Marijuanaseo, 2020). Only 32% of the businesses listed on Yelp were verified to be active brick-and-mortar dispensaries. This is not surprising because Yelp, which provides a general business listing service not specifically designed for marijuana industry, had more records irrelevant to marijuana dispensary than Weedmaps and Leafly.

Taken together, no single secondary data source could provide a reasonably complete and accurate list of active brick-and-mortar dispensaries in a large state like California. We recommend surveillance and research to consider their unique strengths and weaknesses when a single data source is used to minimize required resources. When resources are available, we recommend the integration of multiple secondary data sources, preferably including a licensing directory and multiple online crowdsourcing platforms, as well as verification through phone calls such as what has been done in this study or through even better approaches such as a field census. The verification could considerably improve the accuracy of the data compiled from secondary data sources.

Our findings were overall consistent with the two smaller-scale studies conducted in California, both in Los Angeles County. One was conducted in 2016-2017, before recreational marijuana dispensaries were allowed to open (Pedersen et al., 2018). This study obtained medical marijuana dispensary information from five online crowdsourcing platforms. Weedmaps was suggested to be the most accurate and up-to-date platform, contributing to 95% of the final records. Call verification was conducted in 10% of the dispensaries and found to generally align with the information posted on online crowdsourcing platforms. The other study was conducted in 2018-2019, after recreational marijuana dispensaries were allowed to open (Pedersen et al., 2020). It extracted data from Weedmaps and Yelp and verified dispensary information through site visits. About 80% dispensaries that were determined to be active through online data

cleaning were confirmed to be active in site visits, and licensed dispensaries accounted for roughly 40% of the active dispensaries. Neither study reported validity statistics for each specific data source. Our study expanded on the prior research by covering a much larger geographic region, computing detailed validity statistics for each data source by dispensary category and county population size, and by using two gold standards and two tests to demonstrate validities in different scenarios and for different purposes.

This study has limitations. First, due to the lack of feasibility of conducting a field census in such a large geographic region, phone calls were made to verify information obtained from secondary data sources. While this approach was cost effective, businesses not listed in these secondary data sources were excluded from the analysis, potentially the smaller, unlicensed dispensaries that did not intend to promote themselves on online crowdsourcing platforms because of cost and law enforcement concerns. Future research using field census approach is warranted to assess to what extent unlicensed dispensaries were underrepresented in our study. We could also have misclassified dispensaries as inactive if they provided incorrect contacts or could not be reached after multiple call attempts. Search terms in Yelp may not successfully capture all marijuana-related businesses. As a result of these caveats, our call-verified, combined database would be an underestimation instead of the true "universe" of the active dispensaries in California. Second, validity measures were not all applicable in some scenarios where "true negative" or "false positive" could not be identified with the current study design. Third, regulations on online crowdsourcing platforms have been rapidly evolving. Before our data collection, Weedmaps served as the major platform to advertise and promote dispensaries including the unlicensed ones in California. Right after our data collection, California regulators required Weedmaps to remove unlicensed businesses from its website. By January 2020,

Weedmaps had removed over 2,000 businesses (Branfalt, 2020). Weedmaps may no longer be a good data resource for identifying unlicensed dispensaries, particularly in California, even though it had satisfactory validity statistics in our study. Future studies should consider alternative crowdsourcing platforms that post unlicensed dispensary information. Fourth, we evaluated the three most commonly used online crowdsourcing platforms. The findings may not be applicable to other platforms such as Wheresweed. The findings were not applicable to commercial providers of business listings, either, such as InfoUSA and Dun & Bradstreet that recently incorporated marijuana businesses into their databases. Finally, findings may not be generalizable to the identification of delivery-only services or dispensaries in other states.

Notwithstanding the limitations, the findings of this study provide empirical evidence regarding the validity of using secondary data sources to identify brick-and-mortar marijuana dispensaries in a large geographic region. The data collection and verification protocol and validity statistics could be used by local governments and communities to best strategize regular surveillance on the availability and accessibility of marijuana dispensaries and their compliance to laws. Future research could also use these findings to replicate dispensary identification in other states where marijuana has been commercialized. We hope a comprehensive and accurate enumeration of marijuana dispensaries could facilitate future research evaluating marijuana dispensaries and their impacts on public health.

5. Conclusion

Each secondary data source has its strengths and limitations in identifying brick-and-mortar marijuana dispensaries. Surveillance and research are encouraged to utilize these findings to best strategize data use when resources are limited. When resources are available, we

recommend the use of both a licensing directory and online crowdsourcing platforms with call verification to enumerate a comprehensive and reasonably accurate list of active brick-and-mortar marijuana dispensaries in large geographic regions.

References

Anderson, P., de Bruijn, A., Angus, K., Gordon, R., Hastings, G., 2009. Impact of alcohol advertising and media exposure on adolescent alcohol use: a systematic review of longitudinal studies. https://pubmed.ncbi.nlm.nih.gov/19144976/. Alcohol Alcohol 44(3), 229-243.

Berg, C.J., Henriksen, L., Cavazos-Rehg, P.A., Haardoerfer, R., Freisthler, B., 2018. The emerging marijuana retail environment: Key lessons learned from tobacco and alcohol retail research. https://pubmed.ncbi.nlm.nih.gov/29421347/. Addict Behav 81, 26-31.

Branfalt, T., 2020. Weedmaps removes 2,700 unlicensed dispensaries. Available at https://www.ganjapreneur.com/weedmaps-removes-2700-unlicensed-dispensaries/. Accessed on 04/13/2020. , Ganjapreneur.com.

Campbell, C.A., Hahn, R.A., Elder, R., Brewer, R., Chattopadhyay, S., Fielding, J., Naimi, T.S., Toomey, T., Lawrence, B., Middleton, J.C., Task Force on Community Preventive, S., 2009. The effectiveness of limiting alcohol outlet density as a means of reducing excessive alcohol consumption and alcohol-related harms. https://pubmed.ncbi.nlm.nih.gov/19944925/. Am J Prev Med 37(6), 556-569.

Carlos, H.A., Gabrielli, J., Sargent, J.D., 2017. Validation of commercial business lists as a proxy for licensed alcohol outlets. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5438553/. Bmc Public Health 17.

D'Amico, E.J., Rodriguez, A., Tucker, J.S., Pedersen, E.R., Shih, R.A., 2018. Planting the seed for marijuana use: Changes in exposure to medical marijuana advertising and subsequent adolescent marijuana use, cognitions, and consequences over seven years. https://pubmed.ncbi.nlm.nih.gov/29779761/. Drug Alcohol Depend 188, 385-391.

D'Angelo, H., Fleischhacker, S., Rose, S.W., Ribisl, K.M., 2014. Field validation of secondary data sources for enumerating retail tobacco outlets in a state without tobacco outlet licensing. https://pubmed.ncbi.nlm.nih.gov/24742811/. Health Place 28, 38-44.

FDA, 2020. CTP compliance & enforcement. Available at https://www.fda.gov/tobacco-products/compliance-enforcement-training/ctp-compliance-enforcement. Accessed on July 6, 2020.

Fiala, S.C., Dilley, J.A., Firth, C.L., Maher, J.E., 2018. Exposure to marijuana marketing after legalization of retail sales: Oregonians' experiences, 2015-2016. https://pubmed.ncbi.nlm.nih.gov/29161062/. Am J Public Health 108(1), 120-127.

Finan, L.J., Lipperman-Kreda, S., Abadi, M., Grube, J.W., Kaner, E., Balassone, A., Gaidus, A., 2019. Tobacco outlet density and adolescents' cigarette smoking: a meta-analysis. https://tobaccocontrol.bmj.com/content/28/1/27. Tob Control 28(1), 27-33.

Freisthler, B., Gruenewald, P.J., 2014. Examining the relationship between the physical availability of medical marijuana and marijuana use across fifty California cities. https://pubmed.ncbi.nlm.nih.gov/25156224/. Drug Alcohol Depend 143, 244-250.

Freisthler, B., Ponicki, W.R., Gaidus, A., Gruenewald, P.J., 2016. A micro-temporal geospatial analysis of medical marijuana dispensaries and crime in Long Beach, California. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4861677/. Addiction 111(6), 1027-1035.

Gustafson, A.A., Lewis, S., Wilson, C., Jilcott-Pitts, S., 2012. Validation of food store environment secondary data source and the role of neighborhood deprivation in Appalachia, Kentucky. https://pubmed.ncbi.nlm.nih.gov/22914100/. Bmc Public Health 12, 688.

Hall, W., Lynskey, M., 2016. Evaluating the public health impacts of legalizing recreational cannabis use in the United States. https://pubmed.ncbi.nlm.nih.gov/27082374/. Addiction 111(10), 1764-1773.

Henriksen, L., Feighery, E.C., Schleicher, N.C., Cowling, D.W., Kline, R.S., Fortmann, S.P., 2008. Is adolescent smoking related to the density and proximity of tobacco outlets and retail cigarette advertising near schools? https://pubmed.ncbi.nlm.nih.gov/18544462. Prev Med 47(2), 210-214.

Lake, A.A., Burgoine, T., Greenhalgh, F., Stamp, E., Tyrrell, R., 2010. The foodscape: classification and field validation of secondary data sources. https://pubmed.ncbi.nlm.nih.gov/20207577/. Health Place 16(4), 666-673.

Liese, A.D., Colabianchi, N., Lamichhane, A.P., Barnes, T.L., Hibbert, J.D., Porter, D.E., Nichols, M.D., Lawson, A.B., 2010. Validation of 3 food outlet databases: completeness and geospatial accuracy in rural and urban food environments. https://pubmed.ncbi.nlm.nih.gov/20961970/. Am J Epidemiol 172(11), 1324-1333.

Lipperman-Kreda, S., Grube, J.W., 2018. Impacts of marijuana commercialization on adolescents' marijuana beliefs, use, and co-use with other substances. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6347575/. J Adolesc Health 63(1), 5-6.

Lovato, C., Watts, A., Stead, L.F., 2011. Impact of tobacco advertising and promotion on increasing adolescent smoking behaviours. https://pubmed.ncbi.nlm.nih.gov/14583977. Cochrane Database Syst Rev(10), CD003439.

Mair, C., Freisthler, B., Ponicki, W.R., Gaidus, A., 2015. The impacts of marijuana dispensary density and neighborhood ecology on marijuana abuse and dependence. https://pubmed.ncbi.nlm.nih.gov/26154479/. Drug Alcohol Depend 154, 111-116.

Marijuanaseo, 2020. Weedmaps Vs. Leafly: Which is better for your dispensary? Available at https://www.marijuanaseo.com/weedmaps-vs-leafly/. Accessed on 04/14/2020.

McCarthy, W.J., Mistry, R., Lu, Y., Patel, M., Zheng, H., Dietsch, B., 2009. Density of tobacco retailers near schools: Effects on tobacco use among students. https://tobaccocontrol.bmj.com/content/25/1/75. Am J Public Health 99(11), 2006-2013.

Mcgreevy, P., 2019. California now has the biggest legal marijuana market in the world. Its black market is even bigger. Available at https://www.latimes.com/california/story/2019-08-14/californias-biggest-legal-marijuana-market. Accessed on 04/09/2020., Los Angeles Times.

Orens, A., Light, M., Lewandowski, B., Rowberry, J., Saloga, C., 2018. Market size and demand for marijuana in Colorado 2017 market update. Available at https://www.colorado.gov/pacific/sites/default/files/MED%20Demand%20and%20Market%20%20Study%20%20082018.pdf. Accessed on 04/08/2020. Marijuana Policy Group.

Paschall, M.J., Grube, J.W., 2020. Recreational marijuana availability in Oregon and use among adolescents. https://pubmed.ncbi.nlm.nih.gov/31959327/. Am J Prev Med 58(2), E63-E69.

Paynter, J., Edwards, R., 2009. The impact of tobacco promotion at the point of sale: A systematic review. https://pubmed.ncbi.nlm.nih.gov/19246438. Nicotine Tob Res 11(1), 25-35.

Pedersen, E.R., Firth, C., Parker, J., Shih, R.A., Davenport, S., Rodriguez, A., Dunbar, M.S., Kraus, L., Tucker, J.S., D'Amico, E.J., 2020. Locating medical and recreational cannabis outlets for research purposes: Online methods and observational study. https://pubmed.ncbi.nlm.nih.gov/32130141/. J Med Internet Res 22(2), e16853.

Pedersen, E.R., Zander-Cotugno, M., Shih, R.A., Tucker, J.S., Dunbar, M.S., D'Amico, E.J., 2018. Online methods for locating medical marijuana dispensaries: Practical considerations for

future research. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6625809/. Cannabis 1(2), 22-35.

Powell, L.M., Han, E., Zenk, S.N., Khan, T., Quinn, C.M., Gibbs, K.P., Pugach, O., Barker, D.C., Resnick, E.A., Myllyluoma, J., Chaloupka, F.J., 2011. Field validation of secondary commercial data sources on the retail food outlet environment in the U.S. https://pubmed.ncbi.nlm.nih.gov/21741875/. Health Place 17(5), 1122-1131.

Reitzel, L.R., Cromley, E.K., Li, Y., Cao, Y., Dela Mater, R., Mazas, C.A., Cofta-Woerpel, L., Cinciripini, P.M., Wetter, D.W., 2011. The effect of tobacco outlet density and proximity on smoking cessation. https://pubmed.ncbi.nlm.nih.gov/21164089/. Am J Public Health 101(2), 315-320.

Seliske, L., Pickett, W., Bates, R., Janssen, I., 2012. Field validation of food service listings: a comparison of commercial and online geographic information system databases. https://pubmed.ncbi.nlm.nih.gov/23066385/. Int J Environ Res Public Health 9(8), 2601-2607.

Shi, Y., Cummins, S.E., Zhu, S.H., 2018. Medical marijuana availability, price, and product variety, and adolescents' marijuana use. https://pubmed.ncbi.nlm.nih.gov/30060862/. J Adolesc Health 63(1), 88-93.

Shi, Y., Meseck, K., Jankowska, M.M., 2016. Availability of Medical and Recreational Marijuana Stores and Neighborhood Characteristics in Colorado. https://pubmed.ncbi.nlm.nih.gov/27213075/. J Addict 2016, 7193740.

Shi, Y.Y., 2016. The availability of medical marijuana dispensary and adolescent marijuana use. https://pubmed.ncbi.nlm.nih.gov/27471020/. Prev Med 91, 1-7.

Shih, R.A., Rodriguez, A., Parast, L., Pedersen, E.R., Tucker, J.S., Troxel, W.M., Kraus, L., Davis, J.P., D'Amico, E.J., 2019. Associations between young adult marijuana outcomes and

availability of medical marijuana dispensaries and storefront signage. https://pubmed.ncbi.nlm.nih.gov/31183908/. Addiction 114(12), 2162-2170.

Smith, L.A., Foxcroft, D.R., 2009. The effect of alcohol advertising, marketing and portrayal on drinking behaviour in young people: systematic review of prospective cohort studies. https://pubmed.ncbi.nlm.nih.gov/19200352/. Bmc Public Health 9, 51.

Tormohlen, K.N., Brooks-Russell, A., Ma, M., Schneider, K.E., Levinson, A.H., Johnson, R.M., 2019. Modes of marijuana consumption among Colorado high school students before and after the initiation of retail marijuana sales for adults. https://pubmed.ncbi.nlm.nih.gov/30807274/. J Stud Alcohol Drugs 80(1), 46-55.

Tables and Figures

Figure 1. Online Data Cleaning and Call Verification Procedures

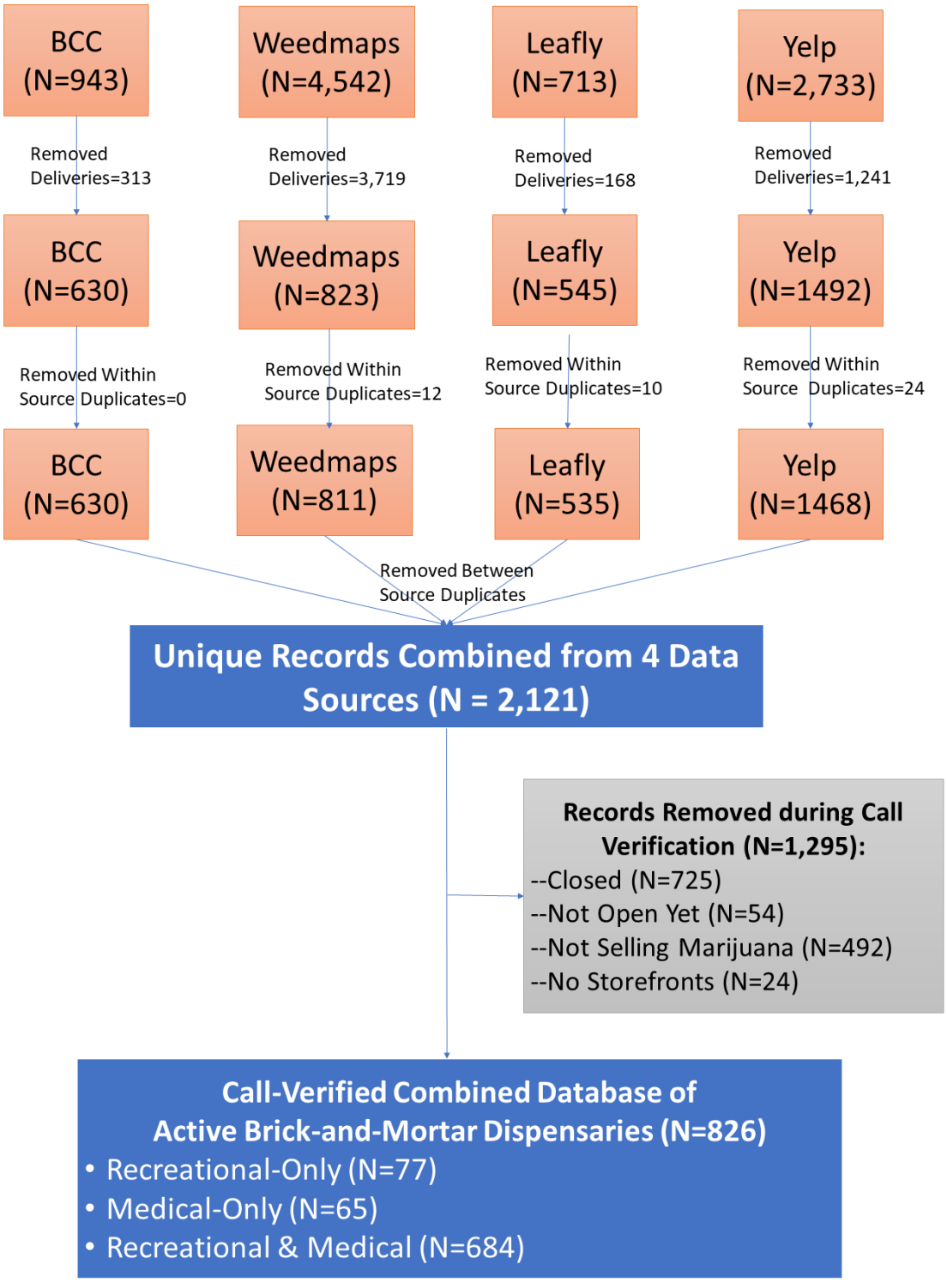
BCC (N=943)
Weedmaps (N=4,542)
Leafly (N=713)
Yelp (N=2,733)
Removed Deliveries=313
Removed Deliveries=3,719
Removed Deliveries=168
Removed Deliveries=1,241
BCC (N=630)
Weedmaps (N=823)
Leafly (N=545)
Yelp (N=1492)
Removed Within Source Duplicates=0
Removed Within Source Duplicates=12
Removed Within Source Duplicates=10
Removed Within Source Duplicates=24
BCC (N=630)
Weedmaps (N=811)
Leafly (N=535)
Yelp (N=1468)
Removed Between Source Duplicates
Unique Records Combined from 4 Data Sources (N = 2,121)
Records Removed during Call Verification (N=1,295):
--Closed (N=725)
--Not Open Yet (N=54)
--Not Selling Marijuana (N=492)
--No Storefronts (N=24)
Call-Verified Combined Database of Active Brick-and-Mortar Dispensaries (N=826)
• Recreational-Only (N=77)
• Medical-Only (N=65)
• Recreational & Medical (N=684)

Table 1. Validity of Using the 3 Online Crowdsourcing Platforms to Identify the BCC Licensing Directory

	Sensitivity	Specificity	Positive Predictive Value (PPV)	Negative Predictive Value (NPV)
Gold Standard = Whether being Present in the BCC Licensing Directory; Test = Whether being Present on Each Crowdsourcing Platform				
Weedmaps	.59 (.55, .62)	NA	.45 (.42, .49)	NA
Leafly	.70 (.67, .74)	NA	.83 (.80, .86)	NA
Yelp	.53 (.49, .57)	NA	.23 (.21, .25)	NA
Gold Standard = Whether being Present in the BCC Licensing Directory; Test = Whether Passing Call Verification				
Weedmaps	.58 (.54, .62)	.34 (.29, .38)	.56 (.52, .59)	.36 (.31, .41)
Leafly	.66 (.62, .69)	.49 (.39, .60)	.90 (.87, .93)	.17 (.13, .22)
Yelp	.52 (.48, .56)	.87 (.85, .89)	.69 (.65, .73)	.76 (.74, .79)

Notes**:** 95% Confidence Intervals are reported in parentheses. BCC: California Bureau of Cannabis Control. Specificity and PPV were not calculated when the test was whether being present on each online crowdsourcing platform because we were not able to identify "true negative" (business records that were excluded from each crowdsourcing platform and absent in the BCC directory).

Table 2. Validity of Using the BCC Licensing Directory and the 3 Online Crowdsourcing Platforms to Identify Active Brick-and-Mortar Dispensaries

	Sensitivity	Specificity	Positive Predictive Value (PPV)	Negative Predictive Value (NPV)
Gold Standard = Whether being Present in the Call-Verified, Combined Database; Test = Whether being Present in Each Data Source				
BCC	.59 (.55, .62)	NA	.77 (.74, .80)	NA
Weedmaps	.80 (.77, .82)	NA	.81 (.78, .84)	NA
Leafly	.56 (.52, .59)	NA	.86 (.83, .89)	NA
Yelp	.57 (.54, .60)	NA	.32 (.30, .35)	NA
Gold Standard = Whether being Present in the Call-Verified, Combined Database; Test = Whether Passing Call Verification				
BCC	.59 (.55, .62)	NA	NA	.30 (.26, .34)
Weedmaps	.80 (.77, .82)	NA	NA	.48 (.42, .53)
Leafly	.56 (.52, .59)	NA	NA	.17 (.14, .21)
Yelp	.57 (.54, .60)	NA	NA	.74 (.71, .76)

Notes**:** 95% Confidence Intervals are reported in parentheses. BCC: California Bureau of Cannabis Control. Specificity and NPV were not calculated when the test was whether being present in each data source because we were not able to identify "true negative" (business records that were excluded from each data source and absent in call-verified, combined database). Specificity and PPV were not calculated when the test was whether passing call verification because no records could be categorized as "false positive" (business records that passed call verification but were excluded from the call-verified, combined database).

Table 3. Agreement between Dispensary Category Licensed by BCC, Dispensary Category Posted on Online Crowdsourcing Platforms, and Dispensary Category Stated in Call Verification in the 630 BCC Licensed Dispensaries

	BCC Licensing Directory # (%)				
	Recreational Only	**Medical Only**	**Recreational & Medical**	***Missing (no category information)***	***Total***
Weedmaps					
Recreational Only	10 (1.54)	0 (.00)	57 (9.05)	0 (0)	67 (10.63)
Medical Only	2 (.32)	29 (4.60)	26 (4.13)	2 (.32)	59 (9.37)
Recreational & Medical	3 (.48)	2 (.32)	231 (36.67)	4 (.63)	240 (38.10)
Missing (not on Weedmaps)	18 (2.86)	40 (6.35)	200 (31.75)	6 (.95)	264 (41.90)
Total	33 (5.24)	71 (11.27)	514 (81.59)	12 (1.90)	630
Leafly					
Recreational Only	10 (1.54)	0 (0)	78 (12.38)	0 (0)	88 (13.97)
Medical Only	1 (.16)	27 (4.29)	30 (4.76)	2 (.32)	60 (9.52)
Recreational & Medical	6 (.95)	4 (.63)	249 (39.52)	4 (.63)	263 (41.75)
Missing (not on Leafly)	16 (2.54)	40 (6.35)	157 (24.92)	6 (.95)	219 (34.76)
Total	33 (5.24)	71 (11.27)	514 (81.59)	12 (1.90)	630
Call-verified, Combined Database					
Recreational Only	15 (2.38)	1 (.16)	34 (5.40)	0 (0)	50 (7.94)
Medical Only	0 (0)	29 (4.60)	4 (.63)	2 (.32)	35 (5.56)
Recreational & Medical	7 (1.11)	9 (1.43)	380 (60.32)	4 (.63)	400 (63.49)
Missing (not on call-verified database)	11 (1.75)	32 (5.08)	96 (15.24)	6 (.95)	145 (23.02)
Total	33 (5.24)	71 (11.27)	514 (81.59)	12 (1.90)	630

Notes: 95% Confidence Intervals are reported in parentheses. BCC: California Bureau of Cannabis Control. Yelp provided no standardized information on dispensary category.

Supplementary Materials

Technical Note S1. Validity Measures

S1.1. General definitions

	Gold Standard - *Yes*	Gold Standard - *No*	**Total**
Positive Test	True Positive a	False Positive b	Total Positive Tests a + b
Negative Test	False Negative c	True Negative d	Total Negative Tests c + d
Total	Total Records Satisfying Gold Standard a + c	Total Records Failing Gold Standard b + d	Total Records a + b + c + d

- Sensitivity

$= \frac{a}{a+c}$

$= \frac{true\ positive}{true\ positive + false\ negative}$

= Probability of positive test conditional on satisfying gold standard

- Specificity

$= \frac{d}{b+d}$

$= \frac{true\ negative}{true\ negative + false\ positive}$

= Probability of negative test conditional on failing gold standard

- Positive Predictive Value (PPV)

$= \frac{a}{a+b}$

$= \frac{true\ positive}{true\ positive + false\ positive}$

= Probability of satisfying gold standard conditional on positive test

- Negative Predictive Value (PPV)

$= \frac{d}{c+d}$

$= \frac{true\ negative}{false\ negative + true\ negative}$

= Probability of failing gold standard conditional on negative test

S1.2. Definitions when the gold standard is whether being present in the BCC licensing directory and the test is whether being present on each online crowdsourcing platform after online data cleaning

- Sensitivity: probability of a record present on the online crowdsourcing platform conditional on the record being included in the BCC directory, calculated as the number of records that were present on the online crowdsourcing platform AND the BCC directory divided by the number of records present on the BCC directory.
- Specificity: probability of a record absent on the online crowdsourcing platform conditional on the record being excluded from the BCC directory, calculated as the

number of records that were neither present on the online crowdsourcing platform nor present on the BCC directory divided by the number of records excluded from the BCC directory. (*not applicable because true negative cannot be identified*)

- PPV: probability of a record included in the BCC directory conditional on the record being present on the online crowdsourcing platform, calculated as the number of records that were present on the online crowdsourcing platform AND the BCC directory divided by the number of records being present on the online crowdsourcing platform.
- NPV: probability of a record excluded from the BCC directory conditional on the record being absent on the online crowdsourcing platform, calculated as the number of records that were neither present on the online crowdsourcing platform nor present on the BCC directory divided by the number of records being absent on the online crowdsourcing platform. (*not applicable because true negative cannot be identified*)

S1.3. Definitions when the gold standard is whether being present in the BCC licensing directory and the test is whether passing call verification

- Sensitivity: probability of an online crowdsourcing record passing call verification conditional on the record being included in the BCC directory, calculated as the number of records on the online crowdsourcing platform that passed call verification AND were present on the BCC directory divided by the number of records present on the BCC directory.
- Specificity: probability of an online crowdsourcing record failing call verification conditional on the record being excluded from the BCC directory, calculated as the number of records on the online crowdsourcing platform that failed call verification AND were absent in the BCC directory divided by the number of records excluded from the BCC directory.
- PPV: probability of an online crowdsourcing record included in the BCC directory conditional on the record passing call verification, calculated as the number of records on the online crowdsourcing platform that passed call verification AND were present on the BCC directory divided by the number of records on the online crowdsourcing platform that passed call verification.
- NPV: probability of an online crowdsourcing record excluded from the BCC directory conditional on the record failing call verification, calculated as the number of records on the online crowdsourcing platform that failed call verification AND were absent in the BCC directory divided by the number of records on the online crowdsourcing platform that failed call verification.

S1.4. Definitions when the gold standard is whether being present in the call-verified, combined database and the test is whether being present in each data source after online data cleaning

- Sensitivity: probability of a record present in the data source conditional on the record being included in the call-verified, combined database, calculated as the number of records that were present in the data source AND the call-verified, combined database divided by the number of records included in the call-verified, combined database.
- Specificity: probability of a record absent in the data source conditional on the record being excluded from the call-verified, combined database, calculated as the number of

records that were neither present in the data source nor present in the call-verified, combined database divided by the number of records excluded from the call-verified, combined database. (*not applicable because true negative cannot be identified*)

- PPV: probability of a record included in the call-verified, combined database conditional on the record being present in the data source, calculated as the number of records that were present in the data source AND the call-verified, combined database divided by the number of records being present in the data source.
- NPV: probability of a record excluded from the call-verified, combined database conditional on the record being absent in the data source, calculated as the number of records that were neither present in the data source nor present in the call-verified, combined database divided by the number of records excluded from the data source. (*not applicable because true negative cannot be identified*)

S1.5. Definitions when the gold standard is whether being present in the call-verified, combined database and the test is whether passing call verification

- Sensitivity: probability of a record passing call verification conditional on the record being included in the call-verified, combined database, calculated as the number of records in the data source that passed call verification AND were present in the call-verified, combined database divided by the number of records included in the call-verified, combined database.
- Specificity: probability of a record failing call verification conditional on the record being excluded from the call-verified, combined database, calculated as the number of records in the data source that failed call verification AND were absent in the call-verified, combined database by the number of records excluded from the call-verified, combined database. (*not applicable because false positive cannot be identified*)
- PPV: probability of a record included in the call-verified, combined database conditional on the record passing call verification, calculated as the number of records in the data source that passed call verification AND were present in the call-verified, combined database divided by the number of records in the data source that passed call verification. (*not applicable because false positive cannot be identified*)
- NPV: probability of a record excluded from the call-verified, combined database conditional on the record failing call verification, calculated as the number of records in the data source that failed call verification AND were absent in the call-verified, combined database divided by the number of records in the data source that failed call verification.

BCC	Weedmaps	Leafly	Yelp	N
●	●	●	●	240
●	●	●		94
●	●		●	26
●		●	●	51
	●	●	●	26
●	●			9
●		●		59
●			●	19
	●	●		8
	●		●	100
		●	●	14
●				132
	●			308
		●		43
			●	992
630	811	535	1,468	2,121

Figure S1. Overlaps across BCC and the 3 Online Crowdsourcing Platforms after Online Data Cleaning

BCC	Weedmaps	Leafly	Yelp	N
●	●	●	●	238
●	●	●		93
●	●		●	26
●		●	●	49
	●	●	●	25
●	●			9
●		●		33
●			●	13
	●	●		7
	●		●	79
		●	●	8
●				25
	●			182
		●		6
			●	33
486	659	459	471	826

Figure S2. Overlaps across BCC and the 3 Online Crowdsourcing Platforms after Call Verification

Table S1. Detailed Data for Table 1 Validity Statistics

	Satisfying Gold Standard	Failing Gold Standard
Gold Standard = Whether being Present in the BCC Licensing Directory; Test = Whether being Present on Each Crowdsourcing Platform		
Weedmaps Positive Test	369	442
Weedmaps Negative Test	261	NA
Leafly Positive Test	444	91
Leafly Negative Test	186	NA
Yelp Positive Test	336	1132
Yelp Negative Test	294	NA
Gold Standard = Whether being Present in the BCC Licensing Directory; Test = Whether Passing Call Verification		
Weedmaps Positive Test	366	293
Weedmaps Negative Test	264	149
Leafly Positive Test	413	46
Leafly Negative Test	217	45
Yelp Positive Test	326	145
Yelp Negative Test	304	987

Notes: BCC: California Bureau of Cannabis Control.

Table S2. Detailed Data for Table 2 Validity Statistics

	Satisfying Gold Standard	Failing Gold Standard
Gold Standard = Whether being Present in the Call-Verified, Combined Database; Test = Whether being Present in Each Data Source		
BCC Positive Test	486	144
BCC Negative Test	340	NA
Weedmaps Positive Test	659	152
Weedmaps Negative Test	167	NA
Leafly Positive Test	459	76
Leafly Negative Test	367	NA
Yelp Positive Test	471	997
Yelp Negative Test	355	NA
Gold Standard = Whether being Present in the Call-Verified, Combined Database; Test = Whether Passing Call Verification		
BCC Positive Test	486	NA
BCC Negative Test	340	144
Weedmaps Positive Test	659	NA
Weedmaps Negative Test	167	152
Leafly Positive Test	459	NA
Leafly Negative Test	367	76
Yelp Positive Test	471	NA
Yelp Negative Test	355	997

Notes: BCC: California Bureau of Cannabis Control.

Table S3. Validity of Using the 3 Online Crowdsourcing Platforms to Identify the BCC Licensing Directory, by Dispensary Category

	Sensitivity	Specificity	Positive Predictive Value (PPV)	Negative Predictive Value (NPV)
Recreational Only				
Gold Standard = Whether being Present in the BCC Licensing Directory (Category Specific); Test = Whether being Present on Each Crowdsourcing Platform (Category Specific)				
Weedmaps	.30 (.16, .49)	.93 (.91, .94)	.12 (.06, .21)	.98 (.97, .99)
Leafly	.30 (.16, .49)	.87 (.84, .90)	.10 (.05, .18)	.96 (.95, .98)
Gold Standard = Whether being Present in the BCC Licensing Directory (Category Specific); Test = Whether Passing Call Verification (Category Specific)				
Weedmaps	.30 (.16, .49)	.96 (.95, .97)	.20 (.10, .34)	.98 (.97, .99)
Leafly	.39 (.23, .58)	.96 (.94, .97)	.30 (.17, .46)	.97 (.95, .98)
Yelp	.30 (.16, .49)	.98 (.98, .99)	.25 (.13, .41)	.99 (.98, .99)
Medical Only				
Gold Standard = Whether being Present in the BCC Licensing Directory (Category Specific); Test = Whether being Present on Each Crowdsourcing Platform (Category Specific)				
Weedmaps	.41 (0.29, 0.53)	.61 (.58, .64)	.07 (.05, .10)	.94 (.91, .95)
Leafly	.38 (0.27, 0.50)	.93 (.91, .95)	.36 (.26, .48)	.93 (.91, .95)
Gold Standard = Whether being Present in the BCC Licensing Directory (Category Specific); Test = Whether Passing Call Verification (Category Specific)				
Weedmaps	.32 (.22, .45)	.97 (.96, .98)	.46 (.32, .61)	.95 (.94, .97)
Leafly	.32 (.22, .45)	.99 (.98, 1.00)	.77 (.58, .90)	.93 (.91, .95)
Yelp	.20 (.11, .31)	.99 (.99, 1.00)	.48 (.29, .67)	.97 (.96, .97)
Recreational & Medical				
Gold Standard = Whether being Present in the BCC Licensing Directory (Category Specific); Test = Whether being Present on Each Crowdsourcing Platform (Category Specific)				
Weedmaps	.45 (.41, .49)	.92 (.89, .94)	.83 (.78, .87)	.64 (.61, .68)
Leafly	.48 (.44, .53)	.63 (.56, .70)	.77 (.72, .81)	.33 (.28, .38)
Gold Standard = Whether being Present in the BCC Licensing Directory (Category Specific); Test = Whether Passing Call Verification (Category Specific)				
Weedmaps	.58 (.53, .62)	.53 (.48, .57)	.53 (.49, .57)	.57 (.53, .62)
Leafly	.65 (.60, .69)	.74 (.68, .80)	.86 (.82, .90)	.46 (.41, .51)
Yelp	.52 (.48, .57)	.89 (.87, .91)	.67 (.62, .71)	.82 (.80, .84)

Notes**:** 95% Confidence Intervals are reported in parentheses. BCC: California Bureau of Cannabis Control. Records on Yelp were not evaluated when the test was whether being present on each online crowdsourcing platform because Yelp provided no standardized information on dispensary category. Records on Yelp were evaluated when the test was whether passing call verification because we obtained dispensary category information during call verification.

Table S4. Validity of Using the BCC Licensing Directory and the 3 Online Crowdsourcing Platforms to Identify Active Brick-and-Mortar Dispensaries, by Dispensary Category

	Sensitivity	Specificity	Positive Predictive Value (PPV)	Negative Predictive Value (NPV)
Recreational Only				
Gold Standard = Whether being Present in the Call-Verified, Combined Database (Category Specific); Test = Whether being Present in Each Data Source (Category Specific)				
BCC	.19 (.11, .30)	.98 (.97, .99)	.45 (.28, .64)	.93 (.92, .95)
Weedmaps	.26 (.17, .37)	.93 (.91, .95)	.24 (.15, .35)	.94 (.92, .95)
Leafly	.18 (.10, .29)	.90 (.88, .92)	.14 (.08, .23)	.92 (.90, .94)
Gold Standard = Whether being Present in the Call-Verified, Combined Database (Category Specific); Test = Whether Passing Call Verification (Category Specific)				
BCC	.65 (.53, .75)	NA	NA	.97 (.96, .98)
Weedmaps	.64 (.52, .74)	NA	NA	.97 (.96, .98)
Leafly	.56 (.44, .67)	NA	NA	.96 (.95, .97)
Yelp	.52 (.40, .63)	NA	NA	.98 (.97, .99)
Medical Only				
Gold Standard = Whether being Present in the Call-Verified, Combined Database (Category Specific); Test = Whether being Present in Each Data Source (Category Specific)				
BCC	.45 (.32, .57)	.95 (.94, .97)	.41 (.29, .53)	.96 (.94, .97)
Weedmaps	.75 (.63, .85)	.60 (.56, .63)	.12 (.09, .15)	.97 (.95, .98)
Leafly	.35 (.24, .48)	.94 (.92, .95)	.31 (.21, .43)	.95 (.93, .96)
Gold Standard = Whether being Present in the Call-Verified, Combined Database (Category Specific); Test = Whether Passing Call Verification (Category Specific)				
BCC	.54 (.41, .66)	NA	NA	.97 (.95, .98)
Weedmaps	.77 (.65, .86)	NA	NA	.98 (.97, .99)
Leafly	.46 (.34, .59)	NA	NA	.96 (.94, .97)
Yelp	.45 (.32, .57)	NA	NA	.98 (.97, .99)
Recreational & Medical				
Gold Standard = Whether being Present in the Call-Verified, Combined Database (Category Specific); Test = Whether being Present in Each Data Source (Category Specific)				
BCC	.56 (.52, .59)	.53 (.47, .59)	.74 (.70, .78)	.33 (.29, .38)
Weedmaps	.39 (.35, .43)	.96 (.93, .98)	.96 (.93, .98)	.40 (.37, .44)
Leafly	.38 (.34, .41)	.69 (.62, .75)	.79 (.74, .83)	.26 (.22, .30)
Gold Standard = Whether being Present in the Call-Verified, Combined Database (Category Specific); Test = Whether Passing Call Verification (Category Specific)				
BCC	.59 (.55, .62)	NA	NA	.50 (.46, .54)

Weedmaps	.82 (.79, .85)	NA	NA	.70 (.66, .75)
Leafly	.56 (.53, .60)	NA	NA	.42 (.38, .47)
Yelp	.59 (.55, .62)	NA	NA	.80 (.78, .82)

Notes**:** 95% Confidence Intervals are reported in parentheses. BCC: California Bureau of Cannabis Control. Records on Yelp were not evaluated when the test was whether being present in each data source because Yelp provided no standardized information on dispensary category. Records on Yelp were evaluated when the test was whether passing call verification because we obtained dispensary category information during call verification.

Table S5. Validity of Using the 3 Online Crowdsourcing Platforms to Identify the BCC Licensing Directory, by County Population Size

	Sensitivity	Specificity	Positive Predictive Value (PPV)	Negative Predictive Value (NPV)
Counties with Over 1 Million Population				
Gold Standard = Whether being Present in the BCC Licensing Directory; Test = Whether being Present on Each Crowdsourcing Platform				
Weedmaps	.59 (.54, .64)	NA	.39 (.35, .43)	NA
Leafly	.69 (.65, .74)	NA	.81 (.76, .85)	NA
Yelp	.60 (.55, .64)	NA	.20 (.18, .23)	NA
Gold Standard = Whether being Present in the BCC Licensing Directory; Test = Whether Passing Call Verification				
Weedmaps	.59 (.54, .63)	.34 (.29, .38)	.49 (.44, .53)	.43 (.38, .49)
Leafly	.65 (.61, .70)	.47 (.36, .59)	.88 (.84, .91)	.19 (.13, .25)
Yelp	.58 (.54, .63)	.87 (.84, .89)	.65 (.60, .70)	.83 (.81, .85)
Counties with Fewer than 1 Million Population				
Gold Standard = Whether being Present in the BCC Licensing Directory; Test = Whether being Present on Each Crowdsourcing Platform				
Weedmaps	.58 (.50, .65)	NA	.79 (.71, .86)	NA
Leafly	.73 (.66, .80)	NA	.89 (.83, .93)	NA
Yelp	.38 (.30, .45)	NA	.43 (.35, .51)	NA
Gold Standard = Whether being Present in the BCC Licensing Directory; Test = Whether Passing Call Verification				
Weedmaps	.57 (.49, .64)	.36 (.19, .56)	.85 (.78, .91)	.11 (.05, .19)
Leafly	.66 (.58, .73)	.59 (.33, .82)	.95 (.89, .98)	.14 (.07, .24)
Yelp	.36 (.29, .43)	.93 (.86, .98)	.92 (.83, .97)	.42 (.35, .49)

Notes: 95% Confidence Intervals are reported in parentheses. BCC: California Bureau of Cannabis Control.

Table S6. Validity of Using the BCC Licensing Directory and the 3 Online Crowdsourcing Platforms to Identify Active Brick-and-Mortar Dispensaries, by County Population Size

	Sensitivity	Specificity	Positive Predictive Value (PPV)	Negative Predictive Value (NPV)
Counties with Over 1 Million Population				
Gold Standard = Whether being Present in the Call-Verified, Combined Database; Test = Whether being Present in Each Data Source				
BCC	.52 (.48, .55)	NA	.76 (.72, .80)	NA
Weedmaps	.82 (.79, .85)	NA	.79 (.76, .82)	NA
Leafly	.50 (.47, .54)	NA	.86 (.83, .90)	NA
Yelp	.61 (.57, .65)	NA	.31 (.28, .33)	NA
Gold Standard = Whether being Present in the Call-Verified, Combined Database; Test = Whether Passing Call Verification				
BCC	.52 (.48, .55)	NA	NA	.25 (.21, .30)
Weedmaps	.82 (.79, .85)	NA	NA	.54 (.48, .60)
Leafly	.50 (.47, .54)	NA	NA	.14 (.10, .18)
Yelp	.61 (.57, .65)	NA	NA	.78 (.75, .80)
Counties with Fewer than 1 Million Population				
Gold Standard = Whether being Present in the Call-Verified, Combined Database; Test = Whether being Present in Each Data Source				
BCC	.87 (.81, .92)	NA	.80 (.74, .86)	NA
Weedmaps	.72 (.64, .78)	NA	.91 (.85, .95)	NA
Leafly	.75 (.68, .82)	NA	.84 (.77, .90)	NA
Yelp	.42 (.35, .50)	NA	.45 (.37, .53)	NA
Gold Standard = Whether being Present in the Call-Verified, Combined Database; Test = Whether Passing Call Verification				
BCC	.87 (.81, .92)	NA	NA	.62 (.48, .74)
Weedmaps	.72 (.64, .78)	NA	NA	.20 (.11, .32)
Leafly	.75 (.68, .82)	NA	NA	.36 (.25, .49)
Yelp	.42 (.35, .50)	NA	NA	.47 (.40, .55)

Notes: 95% Confidence Intervals are reported in parentheses. BCC: California Bureau of Cannabis Control.

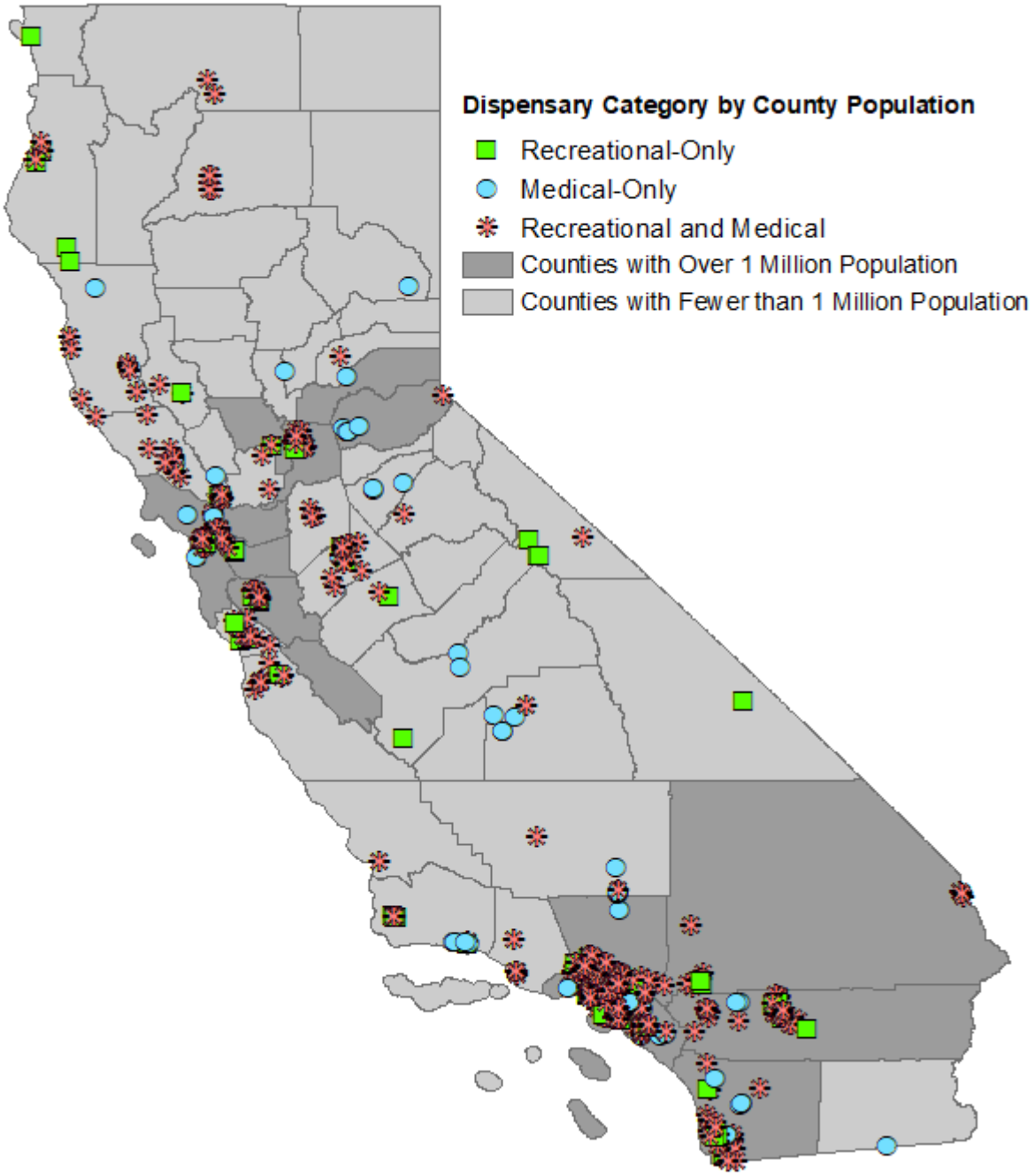

Figure S3. Mapping Call-Verified, Active Brick-and-Mortar Dispensaries in California in 2019, by Dispensary Category and County Population Size