Give a Second Thought to the Secondary City: New applications of the USPS City State file

Derick Brown[1], Joe McMichael[2]

[1,2]RTI International, 3040 E Cornwallis Rd, Durham, NC, 27709

Abstract

As address-based sampling (ABS) has become the most prevalent method for implementing mail and in-person surveys of the U.S. household population, techniques to improve the implementation of such surveys are still being uncovered. The purpose of this paper is to discuss an underutilized data file and methods we have developed with it to improve the quality of ABS frames and field work. The United States Postal Service leases a data product called the City State file. Among other data elements, this file contains a crosswalk of every ZIP code in the U.S. to the city names associated with each ZIP code. There is a "preferred last line" city for each ZIP code, and then additionally a series of "secondary cities" associated with each ZIP code. Some ZIP codes have no secondary cities, others have upwards of thirty. Our research suggests that the secondary city is an often-underutilized tool that can improve the implementation of ABS methods in the following ways. (1) Improve geocoding accuracy and precision of addresses on an ABS frame, (2) match field enumerated addresses to a known ABS frame, and (3) assist with location of addresses on the ground during field work. Through our experience working with the secondary cities, these innovations can be readily applied to any study using ABS to improve the quality of the study through a more accurate frame and more efficient field work.

Key Points

- **City State Product—** USPS file containing a comprehensive list of city, county, and Post Office names associated with each ZIP Code. Used for a variety of mail processing functions.
- **Preferred Last Line—** Every ZIP code in the United States has a single preferred last line (PLL) city for mailing purposes.
- **Secondary City—** Some ZIP codes have secondary cities which are accepted as a city name on a mailing.

The City State Product can improve…

1. **Geocoding Accuracy—** Secondary cities help improve geocoding precision in geographically clustered areas.
2. **Address matching—**PLL cities improve the match rate of non-standardized addresses.
3. **Field Locating—** Secondary cities can be provided to field staff to assist with locating addresses on the ground.

Visit us at **abs.rti.org** for interactive visualizations, white papers, and other ABS resources.

1. Secondary City Overview

What is a secondary city? This can be most clearly illustrated through a simple real-life example. One of our authors sought to look up his home town on a copy of the vendor provided Computerized Delivery Sequence File (CDS) to see how many addresses were listed in the town. The town is Eldersburg; a small suburban town in central Maryland. Thus, he subset down the CDS to only include addresses in Maryland and ran a frequency

of all city names. To his shock and dismay, he could not find an address in Maryland with the city name "Eldersburg". Alphabetically, the city names in Maryland went from "Edgewood" directly to "Elk Mills", skipping over where Eldersburg should be entirely and looking as though Eldersburg does not exist at all in the CDS. Rather than panicking that his hometown is entirely unrepresented on the CDS, or even worse, fearing that it had fallen off a cliff, an investigation was in order. The next step was looking to see if Eldersburg's ZIP code, 21784, appeared on the CDS. And it did! 17,012 addresses were contained in ZIP code 21784. However, we definitively know none of these 17,012 addresses have "Eldersburg" as a city name. That led our author to see what city name(s) corresponded to addresses ZIP code 21784 on the CDS. There was only one name for all addresses, and it is "Sykesville".

Further investigation led us to the City State Product, provided by the United States Postal Service (USPS). Through this we were able to determine that Eldersburg is a secondary city for ZIP code 21784, and Sykesville is the preferred last line (PLL) city for the ZIP code. That is, the PLL city is the preferred last line on an addressed item in the mail. If the secondary city is written as the city of the destination address for a mailed item, it will still arrive at its destination, but the PLL city is preferred by the USPS as the city name. As it turns out, the vendor provided CDS is standardized such that every address has the PLL city listed for every address' ZIP code. Thus, for ZIP code 21784, and all ZIP codes within the CDS, there is only one city name for all addresses in the same ZIP code.

Figure 1: Secondary Cities Linked to PLL City

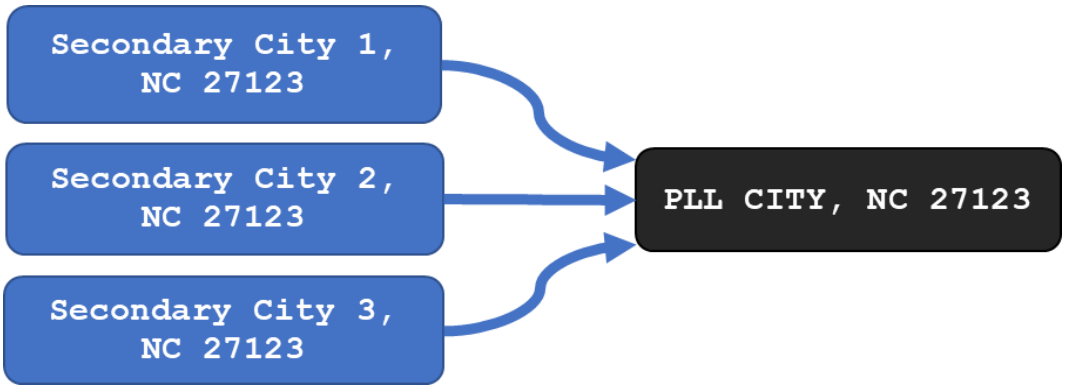


Figure 1 depicts this relationship between secondary city and PLL city for the fictional ZIP code 27123 in North Carolina. In this example there are three secondary cities for the ZIP code, all of which are linked to the PLL city. Note that the arrows move in one direction, from the secondary cities to the PLL city, indicating that the secondary cities get standardized to the PLL city, and not in the opposite direction.

This relationship between ZIP code, PLL city, and secondary city will be explored within this paper. We will then proceed to apply secondary city information to ABS surveys, in hopes that it can improve quality. Specifically, we demonstrate methods that can (1) improve geocoding of sample frames, (2) assist in matching non-standardized addresses, and (3) provide a resource to survey field staff to aid in locating address on the ground.

1.1 City State Product

The City State Product is available for purchase through the USPS Address Information System (United States Postal Service, 2018a). The purpose of the City State Product is to assist with the standardization and accuracy of mailing. The USPS describes the application of the City State Product as such: "City State Product is for mailers who need assistance with the following processing functions: ZIP Code validation, ZIP Code Assignment, finance number assignment, county code/name identification, street name matching, ZIP Code and carrier route changes" (United States Postal Service, 2018b). The data file is updated monthly by the USPS. There are multiple record types within the City State Product. In our research of secondary cities, we utilized the Detail Record portion of the City State Product.

The City State Product data file contains many elements: ZIP code classification, post office facility information, mail delivery information, preferred last line (PLL) city, secondary city, unique identification keys for the PLL and secondary city, and ZIP code finance number (United States Postal Service, 2018c). The data file is structured such that there is one record for each secondary city for each ZIP code in the United States, plus an additional record for the PLL city. Every ZIP code has a single PLL city. Thus, if a ZIP code has two secondary cities that correspond to it, then in the data file there will be three records for the ZIP code; one for each of the two secondary cities, and a third corresponding to the PLL city. For the research contained in this paper we used three data elements found in the Detail Record portion of City State Product, specifically ZIP code, PLL City, and Secondary City.

Note that there is no linkage from address to secondary city available. That is, there is no data that links which secondary city (if any) corresponds to a specific address. Without on the ground knowledge, there would be no way to make the determination that the author's childhood home was in the geographic portion of ZIP code 21784 that corresponds to the secondary city Eldersburg, rather than the PLL city of Sykesville. The only linkage available is the linkage contained in the City State Product at the ZIP code level, as described above.

This paper is limited to the secondary and PLL cities contained within the Detail Record portion of City State Product. There are numerous other ways to apply the City State Product to ABS surveys. To learn how we used the City State Product to identify Only Way to Get Mail (OWGM) PO Boxes, see McMichael and Brown, 2018.

1.2 Exploring secondary cities

We explored the density of secondary cities by ZIP code to obtain a sense of what secondary cities look like within the City State Product data file. That is, how many secondary cities correspond to each ZIP code? Figure 1 provides a sense of the density of secondary cities within ZIP codes. The February 2018 City State Product and February 2018 residential Computerized Delivery Sequence files were used for all subsequent analyses in section 1.2.

Table 1: Secondary cities per ZIP code

	ZIP Codes		Total CDS Addresses	
Number of Secondary Cities	**#**	**%**	**#**	**%**
0	23,690	58.0	67.6 million	46.7
1	9,551	23.4	33.2 million	22.9
2	3,606	8.8	17.4 million	12.0
3	1,721	4.2	9.8 million	6.7
4-6	1,708	4.2	11.9 million	8.2
7-9	408	1.0	3.2 million	2.2
10+	188	0.5	1.7 million	1.2
Total	**40,872**	**100.0**	**144.9 million**	**100.0**

Note: Percentages may not sum to 100 due to rounding.

There are 40,872 United States ZIP codes represented in the February 2018 City State Product. Table 1 displays that the majority of ZIP codes, 58.0% (n=23,690), do not have any secondary cities. That is, within the City State Product data file, there is only a single record for those ZIP codes, with the PLL city. Another 23.4% (n=9,551) of ZIP codes only have one secondary city, which indicates that within the data file there will be two records corresponding to each ZIP code; one for the PLL and one for the single secondary city. As can be seen in table 1, the distribution of secondary cities per ZIP code is strongly skewed right. The number of addresses within each ZIP code categorization follow a similar skewed right distribution, however there is a greater proportion of addresses in the right of the distribution relative to the number of ZIP codes. In fact, most addresses in the residential CDS exist in ZIP codes with at least one secondary city 53.3% (n=77.2 million), even though the majority of ZIP codes do not have a secondary city. For this paper we are only interested in the ZIP codes where there is at least one secondary city. Thus, table 2 is similar to table 1, however it is subset to only the ZIP codes which have one or more secondary cities (n=17,182).

Table 2: Secondary cities per ZIP code amongst ZIP codes with a secondary city

	ZIP Codes		Total CDS Addresses	
Number of Secondary Cities	**#**	**%**	**#**	**%**
1	9,551	55.6	33.2 million	43.0
2	3,606	21.0	17.4 million	22.6
3	1,721	10.0	9.8 million	12.7
4-6	1,708	9.9	11.9 million	15.4
7-9	408	2.4	3.2 million	4.1
10+	188	1.1	1.7 million	2.3
Total	**17,182**	**100.0**	**77.2 million**	**100.0**

Note: Percentages may not sum to 100 due to rounding.

As can be seen in table 2, the majority of ZIP codes with a secondary city only have a single associated secondary city, 55.6% (n=9,551), but the majority of addresses occur in ZIP codes with more than one secondary city, 57.0% (n=44.0 million). Similar to table 1, the number of addresses per ZIP code categorization follows a skewed right distribution with a greater proportion of addresses in the right of the distribution relative to the number of ZIP codes. There is a mean of 4.0 secondary cities per ZIP code which is reflected by the very skewed right distribution of secondary cities per ZIP code. However, the mean number of secondary cities per address is 2.6. This indicates that that very few addresses exist in the ZIP codes with many secondary cities within the "10+" category. Yet, the most interesting ZIP codes are those with high secondary city counts within the category. The most secondary cities within a single ZIP code is 31, which occurs in ZIP code 41465, which is in the heart of the Appalachian Mountains in Kentucky. Only 5,253 addresses exist within the ZIP code. The PLL name for the ZIP code is Salyersville, but obviously the region is known by many other names[1].

2. ABS applications

A whole paper could be written about the rabbit trail of exploring secondary cities within the City State Product data file, but that would not provide much practicality. While some exploration is necessary to provide context and background, at this point we will focus on what we can actually do with secondary cities in address-based survey applications. There are three different hypotheses explored within this paper: 1) Can secondary cities be used to improve the quality of geocoding on an ABS frame? 2) Can secondary cities be used to assist with matching non-standardized addresses to a standardized source? 3) Can

[1] All 31 secondary cities for ZIP code 41465: Bethanna, Burning Fork, Carver, Cisco, Conley, Cutuno, Cyrus, Duco, Edna, Elsie, Ever, Flat Fork, Foraker, Fredville, Fritz, Gapville, Gifford, Hager, Harper, Hendricks, Ivyton, Lickburg, Logville, Maggard, Marshallville, Mashfork, Salyersville, Seitz, Stella, Sublett, Swampton.

secondary cities be used to assist with field staff locating addresses on the ground? The remainder of this paper will provide some evidence that, to varying degrees, the answer to all three questions is "yes".

2.1 Geocoding

We geocoded the July 2017 vendor provided residential CDS with the Street Map Premium (SMP) TomTom (Version 2017 Release 1) geocoder to obtain a latitude and longitude for every address. When geocoding an address, there are multiple levels of precision that are returned by a geocoding software. The most precise level is a "Point Address", and then there are six other levels which are less precise to varying degrees. Those additional levels are listed as follows from most precise to least precise: street address, postal extension, street intersection, street name, postal, and locality (Esri, 2015). A more precise level is desirable because a more precise level indicates a more reliable and accurate latitude and longitude which should then lead to a more accurate census geography derived from the latitude and longitude. Thus, the more addresses which are geocoded as point addresses, the better. The July 2017 CDS file included 142.7 million residential addresses, 68.9% (n=98.3 million) of these geocoded to a point address, and the remaining 31.1% (n=44.4 million) of the addresses geocoded to something less precise than a point address.

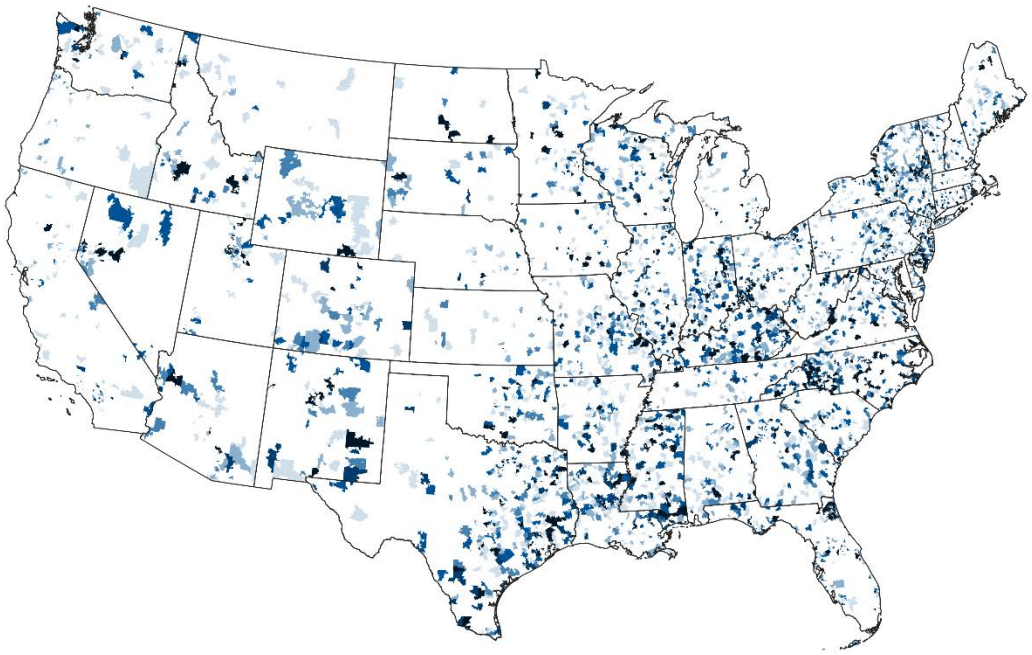
In attempts to increase the number of addresses that geocode to a point address and improve the overall geocoding precision, we used the secondary city associated with each address' ZIP code from the July 2017 City State Product and then processed that data through the SMP TomTom geocoder. Specifically, for every address with a ZIP code that had at least one secondary city we created an additional record(s) for the address with each secondary city associated with the ZIP code. For an address in the aforementioned ZIP code 41465 with 31 secondary cities, there would then be 31 records created, each with a different secondary city as the city name in each record. Of the 44.4 million address which geocoded to less than a point address, 21.7 million addresses have at least one secondary city associated with the ZIP code. Once a record was created for each secondary city associated with each address' ZIP code, there were 57.4 million records created from the 21.7 million addresses, an average of 2.65 records per address.

Of the 21.7 million addresses attempted to be geocoded with a secondary city, .15% resulted in a more precise geocode, 93.83% precision was unchanged, and 6.02% precision decreased. Thus for 31,766 addresses, precision improved relative to the initial level. For the purposes of this classification, if an address had multiple secondary cities associated with its ZIP code and thus multiple records geocoded, we considered the secondary city record(s) with the most precise geocode level. Amongst the 31,766 addresses with improvement, 28.2% (n=8,953) geocoded to a point address in the secondary city geocoding. The remaining 22,827 addresses with improvement were classified as follows: street address 70.8% (n=22,500), and postal 1.0% (n=313). Within the national residential CDS, the portion of addresses with improved geocoding is very small, .022%. At first glance this suggests that the gain from secondary city is minimal, and perhaps not worth the extra effort. However, a closer examination of the results suggests there is some utility for them.

Figure 2 displays a density map of ZIP code tabulation areas (ZCTAs). For the purposes of this paper, a ZCTA is the simplest way to visually represent the geographic boundaries

of a ZIP code. For more information on ZCTAs, please refer to the United States Census documentation (United States Census Bureau). Within figure 2, the darker the shading of the ZCTA, the more addresses there are with improved geocoding precision as a result of the secondary city geocoding.

Figure 2: Map of addresses with improved geocoding by ZIP code



As is readily apparent from viewing the map in figure 2, there is a lot of white area. This suggests that there are many ZIP codes in the nation for which no addresses had improved geocoding from the secondary city geocoding, and the data bears this out. Only 11.8% (n=4,840) of ZIP codes nationally have any addresses with improved geocoding. This indicates that there is a geographic clustering of these addresses with improved geocoding and the addresses with improvements are not evenly dispersed across the country. The minimal amount of dark blue areas in the map suggests that amongst ZIP codes with improved geocoding, only a small portion have many addresses with improved precision. This indicates an additional element of clustering among the ZIP codes with addresses that have improved geocoding. The data supports this notion as well. Among the ZIP codes with improved geocoding, a median of 3 addresses had improved geocoding precision, however, a maximum of 426 addresses in a single ZIP code were improved. In the 95th percentile, 27 addresses per ZIP code were improved, and 80 addresses per ZIP code were improved in the 99th percentile. This clustering is further illustrated by the fact that 55% of addresses with improved geocodes occur in only 10% of ZIP codes with improved geocoding. **In other words, a large amount of the addresses with improved geocode precision occur in a small number of geographical areas.**

While the improvements in precision from secondary city geocoding may have a minimal amount of relevance when looking at the nation in aggregate, the geographic clustering suggests a real utility for this method. If an ABS survey is conducted in one of the states or counties where the geocoding precision improves for a large number of addresses, and geocoding is required for a sampling effort, then the secondary city geocoding method is worth consideration and should add to the accuracy of the sampling frame.

2.2 Matching addresses

A common occurrence in survey research is the collection of non-standardized addresses. Whereas the CDS file is standardized with the respect to the street address[2] and the city name, addresses that are submitted by the public without standardization can have numerous oddities that would not be seen in standardized addresses. For example, the street suffix "avenue" may be spelled out in its entirety as "avenue" in a non-standardized address when in a standardized address it would be abbreviated as "AVE". While there are services that can rectify issues of standardization associated with the street address, these do not necessarily account for issues associated with secondary cities. For example, when referencing the author's hometown, he would never had written "Sykesville", the PLL city. Instead, "Eldersburg" would be the city referred to. Thus, whenever attempting to match non-standardized addresses, issues of non-standardization of the city name, in addition to the street address itself, should be considered.

We encountered this issue when attempting to match 10,445 field enumerated (FE) addresses from a national listing to the CDS file. The FE street addresses were standardized to rectify any irregularities. The standardized FE addresses underwent further manual cleaning and subsequently matched to the CDS at a rate of 43.5%, resulting in 4,537 addresses matching. After the initial matching, the city name of the unmatched addresses was converted to the PLL of the addresses' ZIP code. This resulted in an additional 54 FE addresses matching to the CDS, which was 1.2% of all addresses which matched. This illustrates how city names of addresses collected from the field will pose an obstacle when attempting to match to a standardized address unless the city name provided is taken into consideration. Without considering the city name provided and how it may not be the PLL city, then addresses could go unmatched which otherwise should be matching to the standardized source.

2.3 Field work

Revisiting the example of the author's hometown secondary city of Eldersburg, and the PLL city of Sykesville for ZIP code 21784 will help illustrate another application of the secondary city to survey research. If a field interviewer were to look for the childhood address of our author, all indications on the ground would be that the interviewer was in Eldersburg, not Sykesville. Yet if the address was sampled for a survey it would be sampled how it appears on the CDS, as "123 Fictitious St, Sykesville, MD 21784". We presume there are occurrences where a secondary city can compromise the ability of field interviewers to locate an address on the ground. Thus, for in-person ABS surveys, we are

[2] For the purposes of this paper, a "street address" consists of seven elements: the street number, street pre-direction, street name, street suffix, street post-direction, secondary unit designation, and secondary unit number. For example, refer to the following fictitious address with the seven elements in the same order that they are previously listed: 123 W MAIN ST E APT 1.

starting to include a list of all secondary cities associated with each sampled address for field staff to utilize. While we suspect it will be difficult to empirically detect any benefits of this approach given that address location rates are already very high, we anticipate receiving anecdotal evidence that the secondary city assisted the field staff on the ground.

Here is a recent example where providing the secondary city of the ZIP code has been of assistance for a field interviewer. The sampled address was in Minneapolis, MN 55433, where Minneapolis is the PLL for ZIP code 55433. However, when searching for the address via a map application, the address was said to be in "Coon Rapids". Fortunately, through the materials provided, the field interviewer knew that Coon Rapids is indeed a secondary city for ZIP code 55433, and thus could proceed to the address identified as being in Coon Rapids with confidence that it was the sampled address. While these occurrences may be uncommon, we deem it a worthwhile effort with zero downside to include the secondary cities for field staff's reference.

3. Conclusion

As indicated by the "Exploring secondary cities" section 1.2, there is a lot of information that can be gleaned through the secondary cities on the USPS City State Product data file. This paper includes a relatively shallow exploration of secondary cities as a means to provide context for the applications to ABS research that follow. Each secondary city within each ZIP code has its own story to tell, and the history related to the designation of some secondary cities is fascinating, albeit irrelevant to our research. As a part of this endeavor we researched the relationship between urbanicity and the characteristics of the secondary cities on the City State Product data file. While interesting, the research is not necessarily applicable to purpose of this paper, but we suspect it is just one of many research ventures that could be pursued with the City State product and secondary cities.

Our hope is that we have presented compelling evidence that there are numerous applications of secondary cities to ABS surveys. Specifically, secondary cities can be used to improve geocode precision in geographically clustered areas. Despite the small number of addresses with improved geocode precision nationally, the clustering of the addresses with improved geocodes could prove a boon for any surveys conducted in those areas. Secondary cities can also be leveraged when trying to match non-standardized addresses to a standardized source. If the city name of the non-standardized address is not given special scrutiny, then addresses could go unmatched which otherwise should match. Finally, secondary cities can be used to assist field interviewers with locating addresses in-person. While the benefit is not easily measurable, we already have evidence that field staff have been able to leverage the secondary city to locate at least one address.

While there are three ABS applications discussed in this paper we intend to pursue and consider additional applications of secondary cities to ABS. If we develop any more fruitful applications of the secondary cities then we will share our research with the Survey Research community.

Suggested Citation

Brown, D., & J. McMichael. 2018. *Give a Second Thought to the Secondary City: New applications of the USPS City State file.* In JSM Proceedings, Survey Research Methods Section. Alexandria: American Statistical Association.

A copy of this paper and the original presentation can also be found at: **abs.rti.org**

Or contact the authors directly:
Derick Brown – derickbrown@rti.org
Joe McMichael – mcmichael@rti.org

References

Esri. 2015. *StreetMap Premium for ArcGIS*. Documentation of StreetMap Premium for ArcGIS.

McMichael, J., & D. Brown. 2018. *PO Boxes on Address Based Sampling (ABS) frames: Under- or over-coverage or both?* In JSM Proceedings, Survey Research Methods Section. Alexandria: American Statistical Association.

United States Census Bureau. 2018. "ZIP Code Tabulation Areas (ZCTAs)," Available at https://www.census.gov/geo/reference/zctas.html, (Accessed Aug 8, 2018).

United States Postal Service. 2018a. "Address Information System (AIS) Products Pricing," Available at https://postalpro.usps.com/address-quality/AIS_Products_Pricing, (Accessed September 23, 2018).

United States Postal Service. 2018b. "City State Product," Available at https://postalpro.usps.com/address-quality/city-state-product, (Accessed September 19, 2018).

United States Postal Service. 2018c, "Address Information System Products Technical Guide," Available at https://postalpro.usps.com/address-quality/AIS_Products_Technical_Guide, (Accessed September 23, 2018).