

## Suggested Scrape Process

This document details the “suggested” process to follow to get the most relevant data out of the scrape functionality in the ConnectedMD application. The goal of the application is to provide enough search flexibility and features that this document will become unnecessary.

1. **Do an initial search.** Use the application to do an initial search for doctors in a specific city/state. <http://connectedmd.quantumconnect.com/doctors/bulkadd>

## SEARCH DOCTORS

**Location**

Crestview Hills, KY

**Doctors**

Daniel G. Fage|, M.D.  
Chadwick Hatfield, M.D.  
Michael E. Jones, M.D.  
Scott C. Leverage, M.D.  
Ross McHenry, M.D.  
Karlina M. Patton, M.D.  
Gregory L. Salzman, M.D.  
Joel M. Warren, M.D.  
Tri-State Gastroenterology Practice

Find Doctors

NOTE: to get more profiles returned from this initial search try adding the doctor names without “M.D.” or middle initials

2. **Review the search results.** Take the time to “Remove” irrelevant profiles, leaving only the profiles you want to include in the initial scrape.

## Google

Remove

Daniel G. Fagel, MD: Tri-State Gastroenterology  
doctor, health, point\_of\_interest, es  
[Profile Link](#)

---

## HealthGrades

Remove

Dr. Daniel Fagel, MD  
Gastroenterology  
[Profile Link](#)

---

Remove

Dr. Daniel Vogel, MD  
Child & Adolescent Psychiatry

3. **Run the scrape on the chosen profiles.** When you have reviewed the profiles you want to include in the initial scrape, click on the “Scrape Profiles” button.

#### 4. Review the files created during the scrape.

### CSV Files:

- [scrape\\_results-20170503-131221.csv](#)
- [scrape\\_results-review-text-20170503-131221.csv](#)
- [Profile Report](#)
- [Subjects.csv](#)
- [Subject\\_Site\\_Identifiers.csv](#)
- [Sites.csv](#)
- [scrape\\_debug.txt](#)

- **Profile Report:** this csv file will look similar to the Subject\_Site\_Identifier.csv except it will contain a “status” column containing success/failure information. You can use this file to determine the success/failure of your scrape. It can also be used to identify what profiles you will need to manually find.

	A	B	C	D	E	F	G
1	subject	site	ident	status			
2	daniel-g--fag	healthgrades		No Profile Supplied			
3	daniel-g--fag	vitals		No Profile Supplied			
4	daniel-g--fag	ratemds		No Profile Supplied			
5	daniel-g--fag	yelp		No Profile Supplied			
6	daniel-g--fag	google	ChIJ5zJfjtO5QYgReInzCi4OY9M	OK			
7	daniel-g--fag	facebook		Facebook profiles are skipped at this time			
8	shadwick-h	healthgrades		No Profile Supplied			

5. **Download the Subjects.csv and Subject\_Site\_Identifiers.csv.** When you run a scrape from the search functionality, it will automatically create a correctly formatted Subjects.csv and Subject\_Site\_Identifier.csv file. The Subject\_site\_identifier.csv file is guaranteed to have at least 1 line for every subject/site combination. You will easily be able to tell what is missing and what you need to do search for manually.
6. **Manually add profiles if necessary.** If you need to add profiles manually, you will need to add a line to the Subject\_Site\_Identifiers.csv file. You will need to make sure you add the following on each line:
  - a. A valid “subject\_key” that matches a “subject\_key” in the Subjects.csv file.
  - b. A valid “site\_key” from the list (ratemds, healthgrades, vitals, google, yelp, and facebook)

- c. A properly formatted “site\_subject\_id” value for the site profile you are adding. In all cases except for Google, the “site\_subject\_id” value is the portion of the URL after the main site url. For example, the site\_subject\_id value for the RateMds profile “<https://www.ratemds.com/doctor-ratings/3640635/Dr-Tejas-Patel-Ahmedabad-GJ.html>” is “/doctor-ratings/3640635/Dr-Tejas-Patel-Ahmedabad-GJ.html”. Examples for each site are listed below:

- i. RateMds: /doctor-ratings/3640635/Dr-Tejas-Patel-Ahmedabad-GJ.html
- ii. Healthgrades: /physician/dr-aaron-jarrett-y9pvx4z?cc=MTPSR-1
- iii. Vitals: /doctors/Dr\_Joe\_Howell.html
- iv. Yelp: /biz/pinball-jones-fort-collins
- v. Google: SEE BELOW
- vi. Facebook: /focopinball/

7. **Adding Google Profiles.** Adding google profiles manually requires a little more effort. When you are going a search for something in google, you will notice sometimes on the right side that a “profile” of sorts shows up pertaining to the search you are conducting.

tes ...

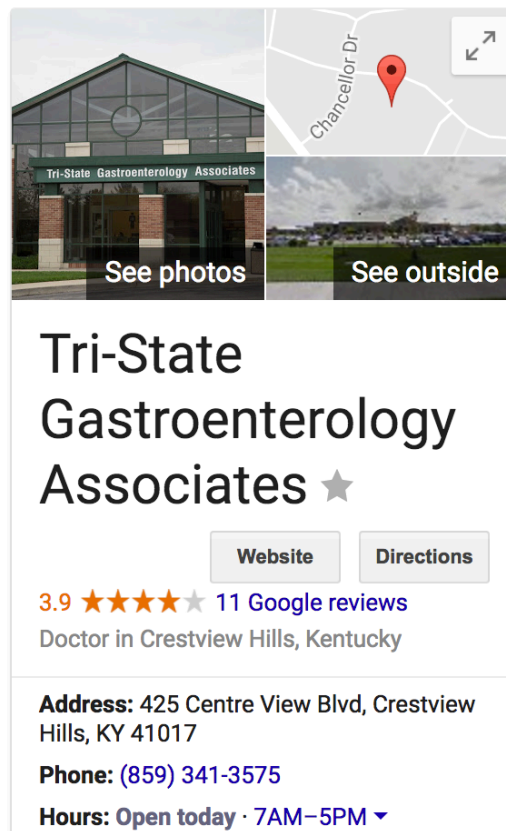
tive

stive

inati

3 ▼

Dr. Warren is



The image shows a Google Maps snippet for 'Tri-State Gastroenterology Associates'. It includes a main photo of the building, a map view with a red pin, and a smaller photo of the building's exterior. Below the photos are buttons for 'See photos' and 'See outside'. The business name 'Tri-State Gastroenterology Associates' is displayed with a star rating. Below the name are buttons for 'Website' and 'Directions'. The rating is 3.9 stars with 11 Google reviews. The address is 425 Centre View Blvd, Crestview Hills, KY 41017. The phone number is (859) 341-3575. The hours are Open today, 7AM-5PM.

See photos See outside

**Tri-State Gastroenterology Associates** ★

Website Directions

3.9 ★★★★★ 11 Google reviews

Doctor in Crestview Hills, Kentucky

**Address:** 425 Centre View Blvd, Crestview Hills, KY 41017

**Phone:** (859) 341-3575

**Hours:** Open today · 7AM–5PM ▼

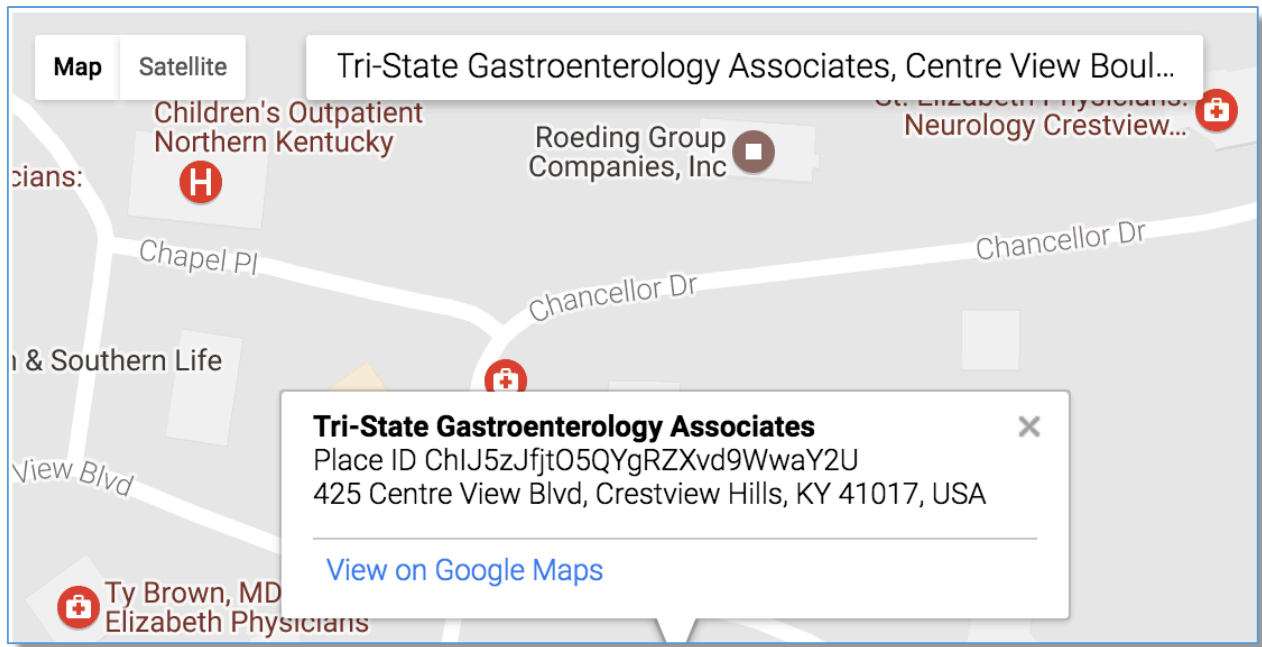
In google terms, this is a “place”. It is a location (business, person, etc) that has been added to the google maps database. When a place has been added, google users are

then able to add comments and review the “place”. These are the comments/reviews that we are going after.

In order to scrape these google “places”, we need to add the “place\_id” to the Subject\_Site\_Identifiers.csv file.

How to find the google Place ID:

- a. Do a normal google search for the doctor. Modify the search terms until you find the correct google place on the google map or the correct side-bar profile information shows up on the search results page.
- b. Use the Google Place ID finder: <https://developers.google.com/places/place-id>
- c. You will want to type the name of the place you found in the step above the exactly. You will then see your place on the map with an info box containing the Place ID.



- d. Put ONLY the value after “Place ID” in the above image into the Subject\_site\_Identifier.csv file.

**8. Use the “Upload CSV Files” to re-scrape the updated files**

**9. Review the ProfileReport.csv**