Supplementary appendices

Supplementary appendix A

Supplementary Table 1 provides further detail about each variable available for mapping. Further documentation for all data involved in this project is also available through the project's online data documentation (https://osf.io/c48hw/wiki/Data%20Records/).

Supplementary appendix B

The following gives further detail on the search queries and coding methodology we used to process data from Twitter [61].

Data acquisition

Twitter data were collected from the Twitter Streaming API, using the Epicosm software [47]. We did not use search terms to source tweets from the API, but instead searched by geography. This search strategy returns a sample of tweets with a Twitter 'place' (that is, an associated geographic bounding box) that fall within the given geographic search boundary which was the country of Wales in this instance. Due to the nature of the Twitter Streaming API there is no indication of what proportion of all tweets are retrieved, and so it is not possible to know how representative the retrieved is. It is known that Twitter will limit the data returned if the tweets matching the query exceed 1% of the total traffic on Twitter at that time [46]. Our Twitter data collection phase ran between 9th March and 15th June 2020 (with a 3 day down-period on 20th, 21st, 22nd March) and returned a total of 860,304 tweets with associated Twitter 'places' from 27,805 unique users. The tweets that we collected are available on our open code and data repository [52], shared in the form of Twitter IDs that can be used to reproduce the full tweet objects.

Development of search queries

Given the data collected from the API we sought to develop a set of search terms and queries that adequately shortlisted the tweets we were interested in. This was an interactive process that involved:

1. Identifying key search terms to produce a broad subset.
2. Human coding the broad subset.
3. Testing the precision of the different terms, and refining queries based on their precision.
4. Refining terms by repeating steps 2 and 3.

Our original intention was to derive a set of dictionary-based rules for tweet classification, but our process of developing

Supplementary Table 1: Further detail about each variable

ID	Variable name	LA or LSOA	Update time	Data provider (Source No.)	Data type	Numeric transformation
N1	Shielding Population	LA	None	PHW (2)	Count	Percentage of LA population
N2	Vulnerable Population	LA	Annual	Welsh Gov. (3)	Percentage	None
N3	Population Over 65	LSOA	Annual	ONS (1)	Count	Sum of those age 6590+, then percentage of LA population
N3	Population Over 65	LA	Annual	ONS (2)	Count	Percentage of LA population
N4	COVID-19 Known Cases	LA	Daily	PHW (1)	Per 100,000 people	Percentage of LA population
N5	Population Density	LSOA	Annual	ONS (3)	Density	None
N5	Population Density	LA	Annual	ONS (4)	Density	None
N6	Deprivation (WIMD)	LSOA	3–5 years	Welsh Gov. (1)	Rank	Numeric direction inverted
N6	Deprivation (WIMD)	LA	3–5 years	Welsh Gov. (2)	Percentage	Most deprived 20% / Total LSOAs
N7	No Internet Access	LA	Annual	Welsh Gov. (4)	Percentage	Numeric direction inverted
N8	Not Using GP Online Services	LA	None	NWIS (1)	Percentage	Numeric direction inverted. Then percentage of total patients.
S1	WCVA Registered Volunteers	LA	None	WCVA (1)	Count	Percentage of LA population
S2	WCVA Volunteer Increase	LA	None	WCVA (1)	Percentage	100*(new vols/(total new vols))
S3	Community Support Groups	LA	Live	Police Rewired (1)	Count	Percentage of LA population
S4	Community Cohesion	LA	Annual	Welsh Gov. (3)	Percentage	Agree + Strong Agree) / Total Responses
S5	Can Count on Someone Close	LA	None	SAIL Databank (1)	Count	Percentage of LA population
S6	Support Related Tweets	LA	Live	Twitter (1)	Count	Percentage of total tweets by LA

Each Variable's geographic resolution, update time, data provider, data type and any numeric transformations made are detailed. Data providers are numbered to indicate separate data sources from the same provider.

Supplementary Table 2: Summary of the regex rules used to find tweets relating to community support

<table>
<tr><th>Query name</th><th>Regex</th><th>Tweets retrieved (N)</th><th>True positives (N)</th><th>Positives (N) false</th><th>Precision</th></tr>
<tr><td>isolate</td><td>(\w{4})?\s?-?isolat |(\w{6})?\s?- ?dist</td><td>363</td><td>94</td><td>269</td><td>0.258953</td></tr>
<tr><td>groups</td><td>community support |support group|community group</td><td>136</td><td>85</td><td>51</td><td>0.63</td></tr>
<tr><td>help</td><td>help |support| need any ?thing</td><td>1551</td><td>364</td><td>1185</td><td>0.23</td></tr>
<tr><td>shop</td><td>shop |food| medic |pharmac</td><td>513</td><td>145</td><td>368</td><td>0.28</td></tr>
<tr><td>comm</td><td>street |neighbour |road |village |community |next ?door</td><td>590</td><td>183</td><td>407</td><td>0.31</td></tr>
<tr><td>social</td><td>facebook |whatsapp |next ?door</td><td>2663</td><td>447</td><td>2185</td><td>0.17</td></tr>
<tr><td>vol</td><td>volunt</td><td>661</td><td>271</td><td>388</td><td>0.41</td></tr>
</table>

Each row defines a rule, the number of tweets this rule retrieved, and the precision of this rule based on the final annotated dataset.

these queries showed that it was unlikely that the concept we were seeking to measure could be adequately described by dictionary-based rules. Based on this we made the decision to only use tweets that had been human-coded in our final data, and used the search queries described in Supplementary Table 2 to shortlist the tweets for human-coding. As a result, we were willing to accept a set of queries with lower precision, in return for higher recall.

The final rule we used to subset data for human coding was: *groups* OR *social* OR (*iso* AND *shop*) OR (*iso* AND *vol*) OR (*help* AND *vol*) OR (*shop* AND *vol*). Ultimately 972 tweets were coded as positive for community support from the subset of 6,640 tweets this query generated by 15th June 2020, which gave it a precision of 0.15.

Assessment of accuracy

Given the variable, and potentially personal, definition of what is considered to show 'community support' we also sought to test our human coding process. We used a method that is common in qualitative research for assessing coding quality that compares the labels attributed to the data separately by two coders, and uses Cohen's Kappa to see how similar they are [49]. Two researchers classified the first broad search query result, which returned 3,215 tweets and resulted in a Cohen's Kappa of 0.439. These tweets, and their annotations, are available in our code and data repository [52]. The initial parameters used were that:

- Tweets should indicate support that is particular to the current COVID-19 situation. This usually means that they indicate they are helping their local community or neighbours in some way. (e.g. not "Congratulations X for being the best volunteer this year"). This could include dropping off food or prescriptions for neighbours, or being a recipient of support from a local person or business.
- Online support should not be included (e.g. sharing of online resources).

Following this assessment, and a review of the similarities and differences after the dual coding exercise we refined our coding guide so that we included the following:

- Online events to combat loneliness or generate a sense of online community such as quizzes or religious services.
- Individuals (either the user, or someone named in a tweet) having done things to help their community, or offering help. This includes reports of receiving support such as "my neighbour dropped off some meals for us yesterday".
- Voluntary groups tweeting to recruit volunteers to help them with community support.
- Voluntary groups tweeting about their offers of community support, or work they are already doing (e.g. delivering food parcels).
- Tweets naming or advertising a community support group.

We did not include:

- Ambivalent tweets about volunteering such as "if I could join the NHS volunteers in Wales then I would", or "I'd like to volunteer when I am no longer shielding" (these are not exact tweets).
- Charitable work that is not related to a community cause, such as international work.

Data access statement

Underlying data for this project are openly available on the project GitHub page (https://github.com/DynamicGenetics/COVID-19-Community-Response). Documentation for all data sources included can be found on the Open Science Framework (https://osf.io/c48hw/wiki/Data%20Records/).