International Journal of Population Data Science

Journal Website: www.ijpds.org

Mapping population vulnerability and community support during COVID-19: a case study from Wales

Nina H Di Cara[1,*], Jiao Song[2], Valerio Maggio[1], Christopher Moreno-Stokoe[1,3], Alastair R Tanner[1], Benjamin Woolf[1,3], Oliver SP Davis[1,4,†], and Alisha Davies[2,†]

Submission History

Submitted:	30/09/2020
Accepted:	19/02/2021
Published:	19/04/2021

[1]MRC Integrative Epidemiology Unit, University of Bristol, UK
[2]Research and Evaluation Division, Public Health Wales, UK
[3]School of Psychological Science, University of Bristol, UK
[4]The Alan Turing Institute, London, UK
[†] Joint Senior Authors

Abstract

Background
Disasters such as the COVID-19 pandemic pose an overwhelming demand on resources that cannot always be met by official organisations. Limited resources and human response to crises can lead members of local communities to turn to one another to fulfil immediate needs. This spontaneous citizen-led response can be crucial to a community's ability to cope in a crisis. It is thus essential to understand the scope of such initiatives so that support can be provided where it is most needed. Nevertheless, quickly developing situations and varying definitions can make the community response challenging to measure.

Aim
To create an accessible interactive map of the citizen-led community response to need during the COVID-19 pandemic in Wales, UK that combines information gathered from multiple data providers to reflect different interpretations of need and support.

Approach
We gathered data from a combination of official data providers and community-generated sources to create 14 variables representative of need and support. These variables are derived by a reproducible data pipeline that enables flexible integration of new data. The interactive tool is available online (www.covidresponsemap.wales) and can map available data at two geographic resolutions. Users choose their variables of interest, and interpretation of the map is aided by a linked bee-swarm plot.

Discussion
The novel approach we developed enables people at all levels of community response to explore and analyse the distribution of need and support across Wales. While there can be limitations to the accuracy of community-generated data, we demonstrate that they can be effectively used alongside traditional data sources to maximise the understanding of community action. This adds to our overall aim to measure community response and resilience, as well as to make complex population health data accessible to a range of audiences. Future developments include the integration of other factors such as well-being.

Keywords
coronavirus; public health; data visualisation; geospatial; community resilience

*Corresponding Author:
Email Address: nina.dicara@bristol.ac.uk (Nina H Di Cara)

https://doi.org/10.23889/ijpds.v5i4.1409

Background

Understanding the geographic distribution of need is crucial for localised and central agencies to provide relevant support. During a crisis this is particularly relevant as resources are likely to be overwhelmed. This process of vulnerability (or risk) mapping [1] is typically used in response to physical disasters, but the current COVID-19 pandemic has presented a global crisis in the field of public health. Whilst vulnerability to disease is a key risk to understand during a pandemic it is also crucial to consider that vulnerability to poor physical and mental health as a consequence of public actions (e.g. self-isolation) reflects existing social and economic inequalities such as financial security, and access to services and local support [2]. Evidence that the direct and indirect impacts of COVID-19 were greater amongst those already experiencing inequalities [3–6] was seen just months into the pandemic, including that these impacts reflected existing geographic distributions of inequality [7]. The challenges of meeting emerging needs in local communities can be somewhat mitigated by local resilience and citizen-led responses from existing or spontaneous community groups [8–10] which have the potential to improve the ability to withstand stress and survive adverse circumstances at both an individual and community level [11]. As such, it becomes crucial to understand which communities have the most need that cannot be mitigated by the available and emerging community support in each area [12, 13].

Strengthening community resilience is a global and national priority [14], set out in the United Nations Sustainable Development Goals and Well-being of Future Generations Act [15]. This emphasises the importance of curated and timely data that can capture the scale of community action. Data on the determinants of vulnerability, inequality, and community belonging is generally measured by annual government surveys and census data [16], but these methods are not often timely enough to capture a live assessment of localised well-being and support. During a crisis it is crucial for higher-level agencies and those organising support locally to have access to this information in order to enable more effective national *and* local action as well as to empower communities as partners in managing the impact of a disaster [17]. Citizen-led community support played a vital role towards the beginning of the UK lockdown, with the importance of digital communication quickly becoming apparent [18, 19]. In this situation, online platforms became hubs for spontaneous neighbourhood and community initiatives and provided a means to communicate and coordinate local resources; public support groups on *Facebook*, *NextDoor* and *WhatsApp* were being developed [20], alongside those led by existing third sector organisations, and community leaders [21, 22]. A survey by Supporting Communities in Northern Ireland [21] determined that 76% of community groups were communicating with local residents through social media.

Previous research into environmental disasters has shown evidence that sourcing community generated information about local action from social media and crowd-sourcing platforms is possible [23–26], and has been employed as a live data source in several natural disasters [27, 28]. Post-hoc analysis has also revealed that useful data can be drawn from these sources [23, 29, 30], including levels of community resilience [31]. These findings show simultaneously the power of the internet for connecting people and understanding the workings of communities, and subsequently the potentially dire consequences of digital exclusion that exacerbates the lack of available support for those who are most in need [32].

Aims

The aims of the COVID-19 Response Map project are two-fold: (I) collate data that represents the scale of unmet need during the pandemic across geographic areas; and (II) create a bespoke data platform that would facilitate the exploration of this complex population health data. The *need* is represented by the populations who are most vulnerable to poor health outcomes from COVID-19, and hence its fulfilment corresponds to the level of community *support*, and the resources available to mitigate the impacts of those needs. The data-driven approach also aims to include non-traditional sources of data (e.g. social media and community-generated data) to supplement administrative and publicly available data. In this case study paper we set out the steps we took to fulfil these aims with specific reference to the country of Wales, in the United Kingdom. We approached this problem with a multi-disciplinary team of public health experts, statisticians, data scientists and researchers in human-computer interaction.

Approach

In this section we will describe the systematic approach we took to identify, process and visualise community-level data. We first start in *Definitions* by clarifying our intended definitions for the *need*, *support*, and *vulnerability* of local communities. Then, in *Data sets and data providers*, we outline what data we identified to support these definitions, also outlining the data providers from whom we sourced the required information. *Data transformation pipeline* describes the subsequent data processing pipeline, focusing on how the design was developed to enable full reproducibility. Finally, in *Data visualisation*, the interactive mapping tool will be described, emphasising the choices we made for effective data visualisation, and easy data exploration.

Definitions

To approach the challenge of mapping the vulnerability of a local population across a specific geographic area, we first needed to define our interpretation of *need* and *support*.

Need could be defined and measured in many ways [33]. However, in the context of the COVID-19 pandemic we primarily focused on the clinical and social vulnerability of a geographical community, as divided into three main themes. These themes are designed to cover the existing features of a community alongside the changing risks presented by the pandemic: (1) *Health Vulnerabilities*, (2) *Transmission Risk*, and (3) *Deprivation and Exclusion*. The first represents the proportion of the population who are vulnerable to poor health outcomes from COVID-19 [34, 35]; transmission risk expresses the likelihood of becoming infected [36, 37]; the latter considers contextual socioeconomic factors [7].

Quantifying the resilience and the support in a community can be challenging, mainly due to the many possible

conceptualisations of resilience and its tendency to change over time [8]. Therefore, we defined resilience guided by known features that were likely to be expressed in available data. We again identified three themes of interest: (1) *Support Resources*, namely the known community assets or services that were supporting people in each area [8]; (2) *Community Cohesion*, the existing, measured cohesion of the community in each area; (3) *Reported Support on Social Media*, support being offered or reported on social media sites. Finally, we defined the vulnerability (or unmet needs [12]) of a specific area as the relative gap between local need and local support.

Datasets and data providers

After establishing the operational definitions of need and support, we underwent a process of scoping the data that were available to meet these definitions. To do so, we engaged in ongoing consultations with representatives from the public and the third sector, including the *Welsh Government (WG)*, *Third Sector Support Wales*, *Data Cymru*, the *County Voluntary Councils* and the *Wales Council for Voluntary Action (WCVA)*. After this process, we were able to characterise the list of *Data Providers*, as well as the individual *Variables* that could be captured from the data these providers could share.

Data providers here refers to the organisations providing access to data. These data were either available publicly, and released under the terms of open licenses (e.g. *GPL-v3* [38], *Creative Commons* [39]), or shared directly with us for the sake of the project. Identified data providers are (1) WCVA; (2) COVID-19 Mutual Aid UK; (3) National Health Service (NHS) Wales, through both the Informatics Service (NWIS) and Public Health Wales (PHW); (4) WG; (5) Office for National Statistics (ONS); (6) Secure Anonymised Information Linkage (SAIL); and (7) Twitter. As well as identifying available data, the scoping process also revealed challenges in sourcing data about online community support. The majority of online conversation about local support services was taking place either through private conversation channels (e.g. WhatsApp), on Facebook groups or on neighbourhood social media platforms [40]. Gathering data from some of these sources was undesirable (for instance, private messaging) or unsuccessful due to companies being unwilling to disclose commercially sensitive information, or not providing application programming interface (API) access to social media platforms.

The final set of Variables collected were 14 indicators of the concepts we sought to capture (eight for local need (N), and six for local support (S)): ($N1$) COVID-19 high risk; ($N2$) COVID-19 moderate risk; ($N3$) Over 65 age; ($N4$) COVID-19 cases; ($N5$) Population density; ($N6$) Welsh index of deprivation; ($N7$) No Internet access; ($N8$) No online GP registration; ($S1$) WCVA registered volunteers; ($S2$) WCVA increase in volunteers; ($S3$) Mutual aid community support group; ($S4$) Sense of community belonging; ($S5$) Symptoms tracker: can count on someone close; ($S6$) Twitter community support. Each Variable has been defined to match a specific theme of interest from the definitions of need and support. In terms of data architecture, each theme represents a single logical *Dataset* as composed by a group of Variables.

Figure 1 shows a comprehensive diagram mapping each Variable to the originating Data Provider. Each Variable is also grouped by the matched Dataset. A short description of each Dataset is reported below, along with a Summary table for each of the corresponding Variables. Further information on the details of each Variable (e.g. data frequency and geographic resolution) are reported in Supplementary Appendix A.

Local need - health and vulnerabilities

This dataset brings together existing clinical vulnerabilities in the population as expressed by the National Health Service (NHS) definitions for those at moderate risk (clinically vulnerable) and high risk (clinically extremely vulnerable) from COVID-19 [34].

While high-risk individuals are well defined (see Table 1, $N1$), the population of those at moderate risk is less specific. In order to index moderate risk (Table 1, $N2$), we constructed a proxy measure based on the NHS definition by finding the number of people who met one or more of the following criteria as reported in the National Survey for Wales 2018-19 [42]: (A) Aged 70+; (B) Asthma diagnosis; (C) Heart or circulatory illness; (D) Respiratory system illness; (E) Kidney complaints; (F) Other digestive complaints including stomach, liver, pancreas etc.; (G) Learning disability; (H) Diabetes (including hyperglycaemia). We also separately included those over age 65 (Table 1, $N3$), due to specific vulnerabilities around age that users of the tool may wish to explore independently from the coronavirus risk.

Local need – transmission risk

This dataset characterises the risk of transmission through contact with others [36], as represented by the number of cases in a given area [43] (Table 2, $N4$) relative to its population density (Table 2, $N5$), which directly affects contact rates [37].

Local need – deprivation and exclusion

This dataset comprises information characterising the deprivation of communities. The Welsh Index of Multiple Deprivation (IMD) is the official measure for relative deprivation in small areas, and ranks every Lower Super Output Area (LSOA) in Wales from the most to the least deprived, based on factors such as income, employment, access to services and community safety [2]. We chose to include it as a well-established and high-quality index of some important environmental determinants of health and well-being (Table 3, $N6$). Given the importance of digital connectivity in access to support and services we also considered digital exclusion. Digital exclusion can be represented in different ways [44]; here we were able to include data on the ability to access the internet from home (Table 3, $N7$), and the proportion of patients registered with online services at their GP surgery (Table 3, $N8$).

Local support – support resources

The WCVA is a national organisation with a central record of volunteers which, whilst not representing all forms of volunteerism, gives a measure of the distribution of registered volunteers (Table 4, $S1$). Their monthly reports of volunteer numbers also allowed us to derive the percentage increase

Figure 1: Diagram representing the mapping between *Data Providers*, and corresponding *Variables* for local need (*N*), and local support (*S*). Each box represents the *Dataset* each variable belongs to

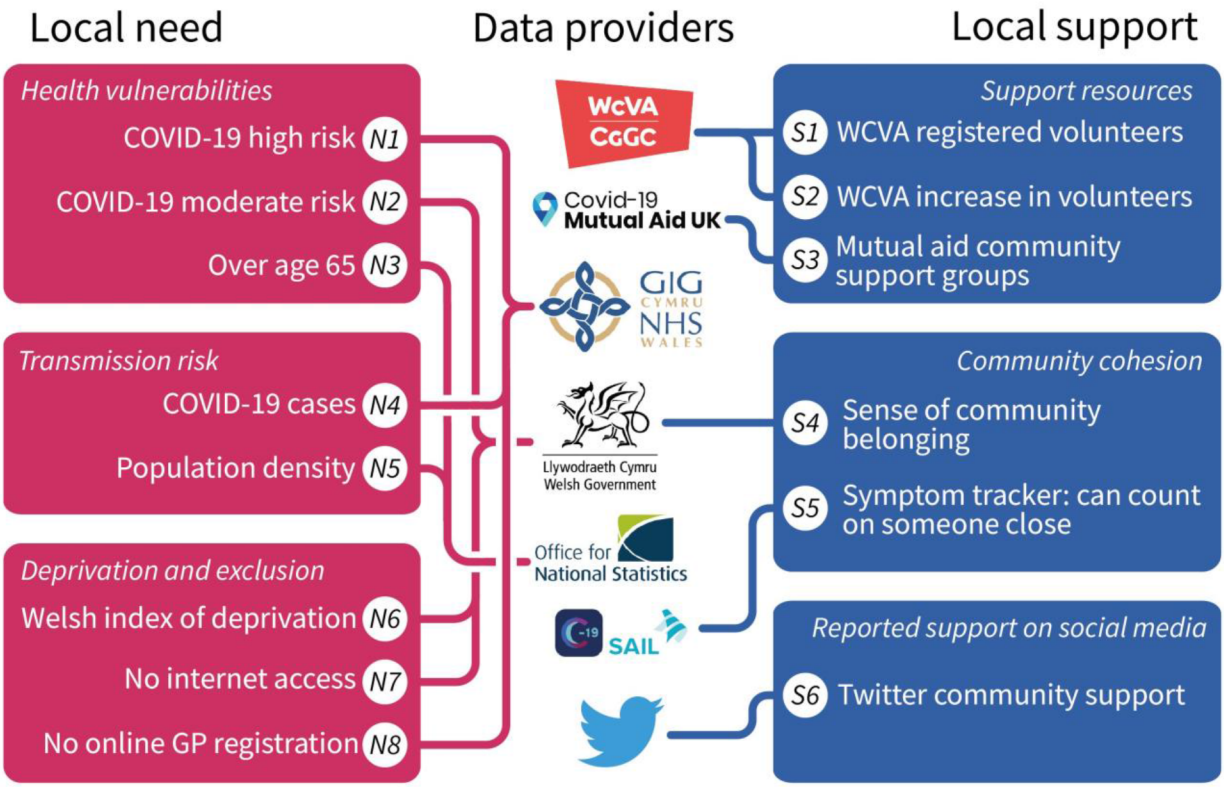


Table 1: Summary description of variables included in the *health and vulnerabilities* dataset

ID	Variable	Data provider	Short description	Benefits and limitations
*N*1	COVID-19 High Risk	Public Health Wales by request, with permission from Welsh Government	The percentage of the population who are high risk, also known as "shielding" or clinically extremely vulnerable.	This information was timely and provided by an official source, but does assume that records are correct and could miss those who are not in contact with services.
*N*2	COVID-19 Moderate Risk	National Survey for Wales via UK Data Service	Percentage of the population who are at moderate risk from coronavirus, based on responses to the National Survey for Wales 2018–19.	This is a proxy variable, and so not an exact measure. The response rate is 54.2%, and nationally representative [41]. However, results are one year old, and may not include some "in-need" groups, e.g. elderly not living at home.
*N*3	Over Age 65	ONS available on statswales.gov.wales	Percentage of the population who are aged 65 years or older.	The population over 65 is based on modelled projections by the ONS.

Table 2: Summary description of variables included in the *transmission risk* dataset

ID	Variable	Data provider	Short description	Benefits and limitations
*N*4	COVID-19 Cases	Public Health Wales publicly available	The cumulative number of confirmed cases.	Very timely data, but only includes confirmed cases, therefore an underestimate of the true no. of cases at any given time.
*N*5	Population Density	ONS available on statswales.gov.wales	No. people per square kilometre based on 2018 mid-year estimates.	Similarly to *N*3 (Table 1), these figures are based on projections made by the ONS

in volunteers between the beginning of the pandemic and June 2020 (Table 4, *S*2). To understand the distribution of community groups we turned to the open database collected by *Police Rewired*, which brings together COVID-19 Mutual Aid groups registered by community members [20], groups on the community networking app *LocalHalo*

Table 3: Summary description of variables included in the *deprivation and exclusion* dataset

ID	Variable	Data provider	Short description	Benefits and limitations
*N*6	Welsh Index of Multiple Deprivation (WIMD)	WG available on Statswales	At LSOA level this is a ranked list of all Welsh LSOAs by level of deprivation. At Local Authority (LA) level this is the percentage of LSOAs in each LA that are in the top 20% most deprived nationally.	The WIMD is measured at a small area level and is a high quality statistic of the multiple facets of deprivation.
*N*7	Digital Exclusion: No Internet Access	National Survey for Wales via UK Data Service	Percentage of the population without access to the internet as reported in the National Survey for Wales 2018–19.	Similarly to *N*2 (Table 1) the National Survey is over one year old, but was nationally representative at the time of the survey.
*N*8	Digital Exclusion: Not Registered with Online GP Services	NHS Wales Informatics Service by request	Percentage of total patients who are not registered with their GP's online patient service.	This data measures the uptake of digital services across the whole of Wales at a high geographic resolution. However, it does only include people registered with an NHS practice in Wales.

Table 4: Summary description of variables included in the *support resources* dataset

ID	Variable	Data provider	Short description	Benefits and limitations
*S*1	WCVA Registered Volunteers	WCVA by request	Number of volunteers who have signed up with the WCVA to provide voluntary support (per 100 people)	Covers the whole of Wales, but does not record volunteers registered with other organisations such as directly with charities.
*S*2	WCVA Increase in Volunteers	WCVA by request	The percentage increase in volunteers between 13th March 2020 and 18th May 2020.	As in (*S*1), this will not capture all volunteers registered through other organisations, or casual support (e.g. helping neighbours).
*S*3	Mutual Aid Community Support Groups	COVID-19 Mutual Aid and LocalHalo via Police Rewired available openly	Locations of local community support groups submitted by the public.	This provides exact locations for community groups, but not information about the size of the of organisation. Not all community groups will be registered online.

(www.localhalo.com), and council community hubs (see Table 4, *S*3).

Local support – community cohesion

This dataset includes two variables that index self-reported community cohesion. The first was a question in the National Survey for Wales 2018–19 that asks respondents how strongly they agree with the statement "I belong to my local area" (see Table 5, *S*4). The second was a question included as part of the sign-up process for the COVID-19 Symptom Tracker app [45], whose data was made available via the SAIL data bank (Table 5, *S*5). The question asks the user if they "could count on someone close to them if they need help".

Local support – reported support on social media

To quantify relative local levels of support reported on social media, we collected data from Twitter (www.twitter.com), a well-known social networking platform that allows users to share public updates of under 280 characters in length known as "tweets" (see Table 6, *S*6).

Publicly available data from Twitter were accessed via Twitter's Streaming API [46, 47] between 9th March and 15th June 2020, retrieving tweets whose Twitter place field was in Wales. The API returns a random sample of the total tweets from the specified area, up to a maximum of 1% of the total worldwide traffic [46]. The tweets returned by the API contain both the text of the tweet and associated meta-data. These meta-data allowed us to identify the Local Authority each tweet was most likely sent from using an automatic matching method based on the percentage overlap of a tweet's bounding box with Local Authority geographic boundaries [48], weighted by the approximate population of the overlapping areas.

To find tweets that were expressing community support we first used a keyword driven approach to obtain a shortlist of tweets that matched words relating to community (the full criteria are available in Supplementary Appendix B). We then qualitatively reviewed the shortlist of tweets to generate the set of tweets that we, as human coders, deemed to be indicative of positive community support. To test the

Table 5: Summary description of variables included in the *community cohesion* dataset

ID	Variable	Data provider	Short description	Benefits and limitations
*S*4	Sense of Community Belonging	National Survey for Wales via UK Data Service	The percentage of people who agreed, or strongly agreed with the statement "I belong to my local area" in the National Survey for Wales 2018–19.	As with *N*2 and *N*7 (see Tables 1 and 3), the survey results are nationally representative but now over a year old. The sample frame may also have missed some "in need" groups, such as the elderly not living at home.
*S*5	Symptom Tracker: Can Count on Someone Close	ZOE Symptom Tracker App via Secure Anonymised Information Linkage (SAIL) Databank	The percentage of people who agreed that they could count on someone close to them if they need help.	The sample is limited to those who have a smartphone and internet access. The response rate may not to be representative of the population the respondents are from.

Table 6: Summary description of variables included in the *reported support on social media* Dataset

ID	Variable	Data provider	Short description	Benefits and limitations
*S*6	Twitter Community Support	Twitter via the Streaming API	Number of Twitter users identified as having posted at least one tweet about community support since 9th March 2020, as a percentage of total users in each area.	If the underlying determinants of Twitter use are associated with levels of support then this variable could be misleading. Location of each tweet is not exact, so we matched the most likely LA, weighted by the approximate population.

effectiveness of human evaluation of community support indicators on Twitter, two researchers classified 3,215 tweets from the initial shortlist with an inter-rater reliability of 0.44 (Cohen's kappa) [49], which led to refinement of our inclusion criteria. Our final qualitative review criteria are listed in full in Supplementary Appendix B. The final data we included in the map was the percentage of total unique users in each area who had positively identified community support. In total 860,304 tweets from Wales were retrieved in the time frame, corresponding to 27,805 unique users. Of these, 6,640 tweets were shortlisted for coding using the keyword-based query, from which 972 tweets from 540 unique users were coded as being indicative of community support.

Data transformation pipeline

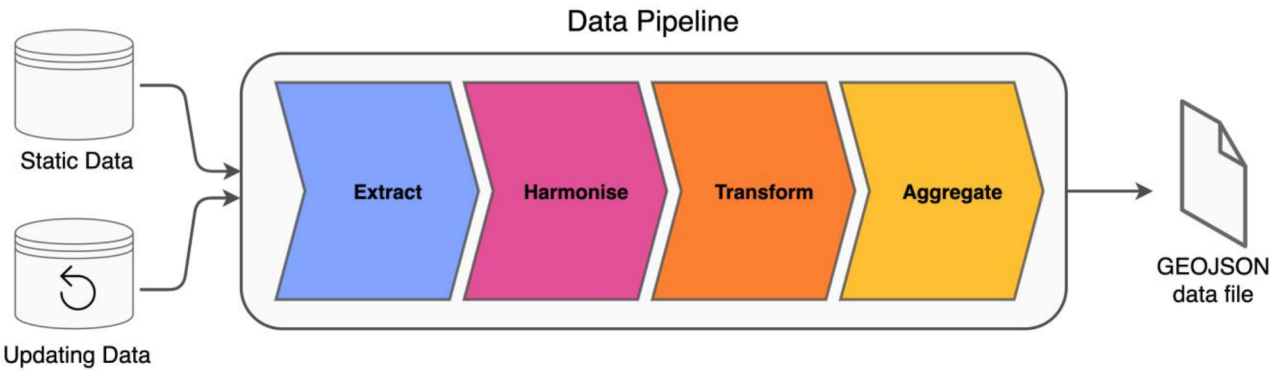
Considering the multitude of data sources needed to gather Variables, as well as their different formats (e.g. JSON, CSV, TSV, HTML), we defined a fully automated approach to harmonise and aggregate the data. This idea was originally motivated by our intention to guarantee a completely reproducible complex data pipeline, and transparent data documentation. Moreover, this systematic procedure favours our requirement for easy extensibility, both in terms of processing operations and of additional data sources. A sketch of the defined transformation pipeline is represented in Figure 2. The pipeline is composed by four main consecutive steps, aimed at extracting the target Variables from original data sources, aggregating them into the corresponding dataset, and finally preparing them in a format compliant with the interactive mapping tool. Most of this analysis has been carried out using the pandas library [50, 51], and the Python programming language (version 3.7.7). The source code and the technical documentation are publicly available on GitHub [52], along with specific instructions to recreate the development environment.

1. **Extract**: The input data source is processed in order to extract the data relating to the target Variable. This usually corresponds to grouping and filtering operations on the original data to retain only the information that is relevant to the target Variable. This is the only step of the whole pipeline that has to be customised and adapted to the specific format and layout of the original dataset. Nonetheless, the pipeline keeps tracks of all the applied transformations to the data so that they could be replicated and reproduced. The consistency of the extracted data is verified via automated testing procedures.

2. **Harmonise**: The aim of this step is to encapsulate extracted data into a tidy [53] and unified data layout. This step is crucial to allow generic transformation operations that can also be re-used regardless of the specific format of the original data. To do so, a generic Variable data abstraction is generated as an output of this step.

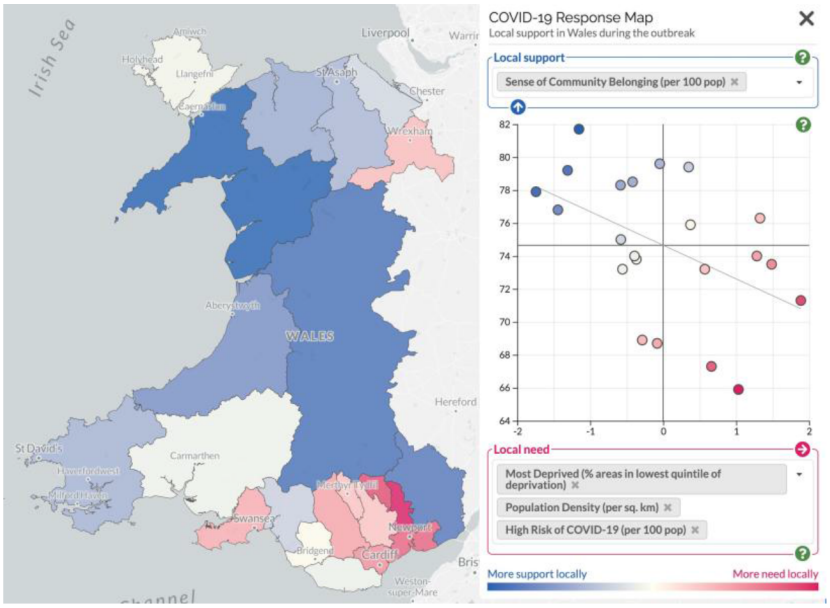
3. **Transform**: During the transformation steps, each collated Variable is subject to a series of transformations which is specific to the data at hand. The structure of the transformation pipeline for each Variable is dynamically defined via a series of generic and re-usable operators, leveraging on the harmonised data layout abstractions. Examples of these operations are *numeric format alignment*, *percentage calculation*, as well as *data*

Figure 2: A schematic of the data processing pipeline



Static data and regularly updating data are both handled by the pipeline, which takes a data file as input, extracts the relevant information, harmonises the data to a consistent format, makes any necessary transformations and then aggregates the data for output to the mapping tool as a GEOJSON file.

Figure 3: Illustration of mapping a composite need score using the number of people at high risk, population density and deprivation against an area's sense of community belonging



pivoting and *transposition*. Similarly to the extraction step, each applied transformation is logged for future replicability.

4. **Aggregate**: The last step of the pipeline aims to aggregate the multiple Variables into their corresponding *Dataset*, where they are matched by frequency and corresponding geographic resolution. Aggregated data are then formatted in GeoJSON to be integrated into the mapping tool.

Detailed information about the Variables themselves, including numeric transformations applied to them is given in Supplementary Appendix A, as well as being fully documented on our code repository [52].

Data visualisation

Vulnerability maps traditionally pinpoint the location of a natural disaster alongside information about the local area [54]. However, since our intention is to specifically identify unmet need, we adopted a bivariate approach that allowed us to combine indices of both need and support on a single map. Our approach to the visualisation is based on a choropleth map, a map whose colours represent a summary statistic relevant to each geographic area, in combination with a linked scatter plot. This plot displays the local need against local support for each considered area (see Figure 3).

Users are able to select the Variables of interest in relation to need and support from a drop-down list. Where users select more than one Variable to index need or support, each variable is then transformed to a z-score, summed, and normalised using a standard scaling procedure (that is to zero mean, and unit variance). This gives an equally weighted combined score that is fast enough to calculate in the browser, and that we considered accurate enough for visualisation. Furthermore, if a user removes all Variables from one dimension (need or support), the scatter plot automatically collapses to a univariate bee-swarm plot (see Figure 4).

Data points are coloured according to a scale based on the z-score for support minus the z-score for need. This gives a visual index of how close a data point falls to the bottom right

Figure 4: Illustration of mapping a composite need score using the number of people at high risk, population density and deprivation

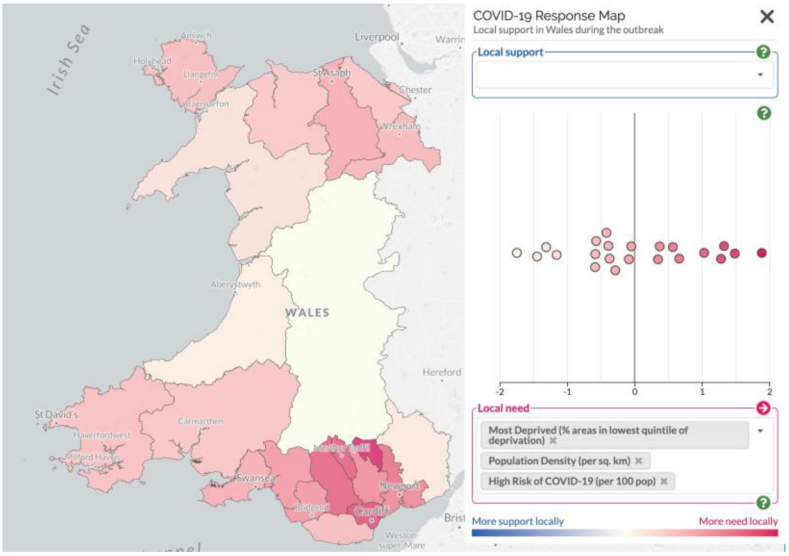
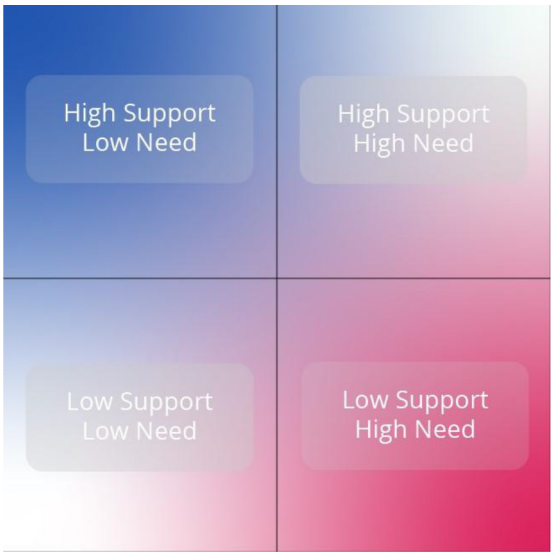


Figure 5: The colouring used to indicate the level of support or need for each area



quadrant of the plot, corresponding to an area with high need and low support, in contrast to the top left quadrant, referred to an area with low need and high support (Figure 5). These colours are mirrored on the accompanying choropleth map. We chose the colours so that red consistently represents areas of greater need, and blue consistently represents areas of greater support, with the intensity of the colours representing distance from the main bisecting line. Colours are interpolated in Hue-Chroma-Luminance colour space to maintain a perceptually constant colour scale.

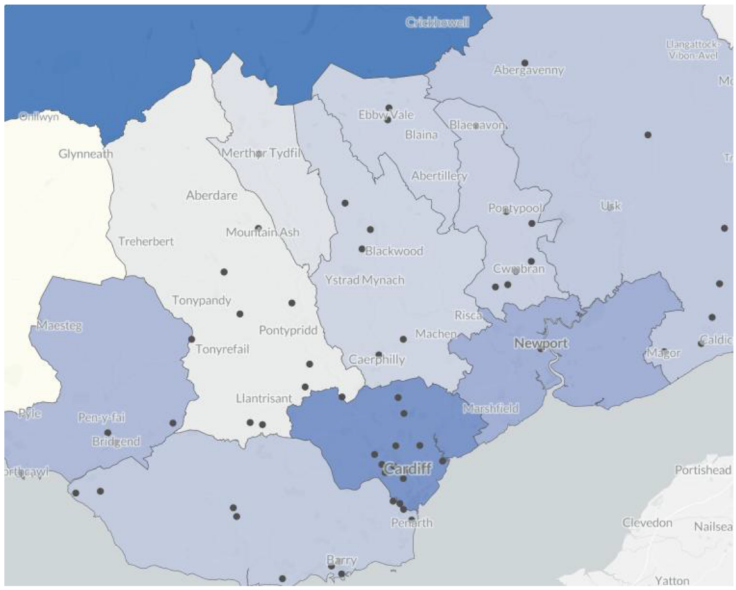
Since the Variables are available at different geographical resolutions, the results are presented at the highest resolution that is available for all the selected variables. Hovering the mouse over an area on the map, or over a data point in the scatter plot, highlights the area in both views, and labels the area in the scatter plot. Zooming in to the choropleth map reveals further geographical detail, including the location of specific community support groups (Figure 6). Clicking on one of these locations gives more information about the group, including a direct link to the group's web site.

The data visualisation was programmed in JavaScript using the D3 visualisation library (version 5, www.d3js.org) and the Mapbox API (version 1.11.0; www.mapbox.com). The tool is available online at www.covidresponsemap.wales or www.mapymatebcovid.cymru, and is supported by an explanatory web page and a comprehensive user guide.

Discussion

In response to the need for an understanding of how the citizen-led response to the COVID-19 pandemic was meeting the needs of local communities we have developed the COVID-19 Response Map project: an online interactive map, available in English (www.covidresponsemap.wales) and Welsh, (www.mapymatebcovid.cymru), that measures local levels of need and community action with a novel combination

Figure 6: Community groups are marked on the map at higher zoom levels as dark grey points



of data sources. The map uses a bespoke visualisation design that allows users to explore any combination of variables of interest to them, and makes it possible for non-specialists to derive meaningful insights from complex population data. This means that important information about local well-being and needs is available to everyone involved in disaster response, from community and third sector organisers to the government.

Although other efforts have created maps of the vulnerability of communities to COVID-19, notably the British Red Cross [55], we have approached this in a different way, allowing users to explore how flexibly-defined local need and support are related to each other, facilitating the identification of areas where the local need or vulnerability is not currently being met by local community support. Our approach also integrates non-traditional data sources such as Twitter and crowd-sourced data, which provide a unique perspective on how we can understand the workings of communities, both in a crisis situation and outside of it.

Whilst many of the data sources we used are open (available for anyone to download), the task of sourcing, and combining them is not trivial and requires access to key data owners, time and data-centric skills. This is due to the fact that all the original data are available in their own format and layout, which needed to be processed and harmonised in order to be integrated into a single output for comparison. In doing so, we have developed a systematic data processing strategy that ensures the reproducibility of our whole approach: every single operation to the data is recorded, whilst automated testing is used to verify data consistency.

With 42% of charities reporting that they are poor at managing, using and analysing data [56], having user-friendly tools available to combine and interpret population data is important. As well as creating the tool, sharing our documentation, data and code openly [52, 57] is a means of sharing this work with the public sector, so that organisations can reuse elements or refine it to their needs. In turn, we are continuing to work with local and national organisations to further adapt it to their requirements.

Development process

This community support tool was developed in collaboration with the Welsh Government, local councils, voluntary groups and the public sector. It has been received as a welcome contribution to the challenge of democratising access to data and mapping the complexities of communities. Local councils particularly wished to overlay their own data sources, which were sometimes not suitable for public dissemination, and to directly add lists of community groups. Feedback from community organisations and local charities has highlighted the value of better understanding what other local offers of support are so that they can work together to streamline their response. This feature of the map demonstrates the contribution to mapping and understanding community resilience more widely, as the ability to measure and visualise this complex concept will enable better support for communities who are struggling outside of the coronavirus pandemic. This is an area for further exploration going forwards.

Strengths and limitations

The strength of this tool lies in combining multiple data sources in an interpretable way, and bringing together sources of openly available data on community mobilisation and support groups to provide a novel perspective of community resilience and need. We identified a national register of community groups on a central database [20], which was helpful to provide local level information on community action, but there are limitations to community-generated data sources. Since the database relies on individuals to register their groups online it is not comprehensive, and many community groups were already known to residents through existing channels [19, 21]. It also relies on this information being maintained by individuals in order to remain up-to-date and reliable. Through the engagement exercises we undertook we also found that local authorities were holding databases of community groups that served their specific populations;

these were more likely to be up to date, but were not open data. To capture informal community mobilisation and support we also drew on social media data from Twitter. Social media has the potential to offer new insights in public health with the added benefit of being extremely timely; our approach to finding community support online did reveal many explicit examples of support being offered or received, or local support groups being advertised. However, it was challenging to rigidly classify community support, reflected in the inter-rater reliability of 0.44 that we achieved with human evaluators, which subsequently made it difficult to establish a reliable automated method for assessing tweets, which would have improved timeliness. There also remains the potential for such data to be misleading or incorrect [58].

Another challenge of combining multiple sources is the differing detail available in terms of timescales, and granularity. From Twitter data that is recorded to the millisecond, to census data that is collected once every ten years, the time-based variation means that the present-day accuracy of variables may be unknown. There are also differing degrees of geographic specificity available for mapping. The majority of data sources we have presented are available at a Local Authority District resolution, which in Wales corresponds to only 22 areas [48]. Data at Middle Super Output Area level would be the ideal resolution to aggregate relevant information and still have meaningful depth, but restrictions on granularity mean this is often not possible.

The last challenge we faced in combining data sources is the data that does not exist. The map shows that there are potential benefits of drawing on community generated data through Twitter and the COVID-19 Symptom Study, but these applications are inaccessible to the 13% of people in Wales who have no internet at home [32]. Of this population, over 70% are over 70 years old, and 25% have a low level of general health [32]; as such those without internet access represent some of the most vulnerable members of the population who are not being reached through these emerging data sources. It is for this reason that we deemed it especially important to provide information on digital exclusion.

Conclusion and future directions

Granular, localised and timely information on community resilience will help to direct support to those areas most in need, which is of significant importance given the contribution of communities to general population health and well-being. We have implemented an approach that allows people at all levels of community response to explore complex population data about the distribution of the citizen response to need. Our approach identified key datasets relevant to community support and community need which extended beyond traditional data collection methods for public health; these non-traditional data sources can be timelier than official datasets and add new dimensions to our understanding of communities but it is also important to understand the limitations in their accuracy. Future developments will include incorporating medium- to long-term impacts on communities such as mental well-being.

As the pandemic progresses and the research around its direct and indirect impact on health continues to evolve we aim to build on the approach we have developed by identifying new and existing data sources with a specific focus on community vulnerability and support. Given the need for more real time and longitudinal information we would like to use data from Twitter to measure mental health and mood in communities [59, 60]. We will also continue to evaluate the timeliness of our existing datasets, and intend to update the National Survey data with the 2019-20 collection when it becomes available.

Acknowledgements

We would first like to thank those who provided comments, assistance or practical support for this project, including Lucia Homolova (Public Health Wales), Elysha Rhys-Sambrook (Public Health Wales) and Professor Claire Haworth (MRC Integrative Epidemiology Unit at the University of Bristol). We kindly thank the Wales Council for Voluntary Action, COVID-19 Mutual Aid, Police Coders, Welsh Government Statistics and Research, and the Office for National Statistics for collecting and making the data we have used available. We would also like to thank the OpenStreetMap community, and the MapBox Community Team for discounted access to the MapBox API used in our maps of Wales.

This work uses data provided by participants of the COVID-19 Symptoms Study, developed, and set up by ZOE Global Limited with scientific and clinical input from King's College London. We wish to thank the participants, and acknowledge the collaborative partnership that enabled acquisition and access to the de-identified data, which led to this output. The collaboration was led by BREATHE The Health Data Research Hub for Respiratory Health, in partnership with SAIL Databank at Swansea University. We wish to acknowledge the input of ZOE Global Limited and King's College London in their development and sharing of the data, and their input into the understanding and contextualisation of data for COVID-19 research.

This work was supported in part by the UK Medical Research Council Integrative Epidemiology Unit at the University of Bristol (Grant ref: MC UU 12013/1). NHD is supported by an MRC PhD studentship (Grant ref: MR/N013794/1), CMS is supported by an ESRC PhD studentship (Grant ref: ES/P000630/1), as is BW. OSPD and CMAH are funded by the Alan Turing Institute under the EPSRC grant EP/N510129/1. This study was also supported by the National Institute for Health Research Biomedical Research Centre at the University Hospitals Bristol NHS Foundation Trust and the University of Bristol (BRC-1215-2011).

Author contributions

The contributions of each author to this project, as defined by the CRediT contributor roles taxonomy.
Conceptualisation: OSPD, AD
Data Curation: NHD, JS, VM, CMS, BW
Formal Analysis: NHD, VM, CMS, OSPD
Funding Acquisition: OSPD, AD
Investigation: NHD, JS, VM, CMS, ART, BW, OSPD, AD
Methodology: NHD, JS, VM, CMS, ART, BW, OSPD, AD

Project Administration: NHD, VM, CMS, OSPD, AD
Resources: JS, AD
Software: NHD, VM, CMS, ART, OSPD
Supervision: VM, OSPD, AD
Validation: NHD, VM, CMS, BW
Visualisation: VM, CMS, ART, OSPD
Writing - original draft: NHD, JS
Writing - reviewing and editing: NHD, JS, VM, CMS, ART, OSPD, AD

Statement on conflicts of interest

The authors declare no conflicts of interest.

Ethics statement

This research was reviewed and approved by the University of Bristol Faculty of Health Sciences Research Ethics Committee (ID 104005). This research was also completed under the permission and approval of SAIL independent Information Governance Review Panel (project 1095).

References

1. Morrow BH. Identifying and mapping community vulnerability. *Disasters*, 23(1):118, 1999. https://doi.org/10.1111/1467-7717.00102

2. Statistics for Wales. *Welsh index of multiple deprivation (WIMD) 2019*. Technical report, Welsh Government, 2019. https://gov.wales/sites/default/files/statistics-and-research/2020-02/welsh-index-multiple-deprivation-2019-technical-report.pdf

3. Abrams EM, and Szefler SJ. COVID-19 and the impact of social determinants of health. *The LancetRespiratory Medicine*,8(7):659-661, 2020. https://doi.org/10.1016/S2213-2600(20)30234-4

4. Public Health England. *Beyond the data: Understanding the impact of COVID-19 on BAME groups*. 2020 Public Health England: London.

5. Bibby J, Everest G, and Abbs I. *Report: Will COVID-19 be a watershed moment for health inequalities?* The Health Foundation, 2020. Available: www.health.org.uk/publications/long-reads/will-covid-19-be-a-watershed-moment-for-health-inequalities

6. All-party Parliamentary Group for left behind neighbourhoods. *Communities at risk: the early impact of covid-19 on 'left behind' neighbourhoods*. Report, Local Trust, 2020.

7. Office For National Statistics. *Deaths involving COVID-19 by local area and socioeconomic deprivation*. - Statistical bulletin, UK government; 2020.

8. Davies AR, Grey CNB, Homolova L, and Bellis MA (2019). *Resilience: Understanding the interdependence between individuals and communities*. Cardiff: Public Health Wales NHS Trust.

9. Cabinet Office (2019). *Community Resilience Development Framework*. Report, HM Government.

10. Curtis S, Congdon P, Aitkinson S, Corcoran R, MaGuire R, and Peasgood T. *Individual and local area factors associated with self-reported wellbeing, perceived social cohesion and sense of attachment to one's community: analysis of the Understanding Society Survey: Research Report*. 2019, available at: Durham Research Online, http://dro.dur.ac.uk Corresponding author: s.e.curtis@durham.ac.uk.

11. Ziglio E, Azzopardi-Muscat N, and Briguglio L. Resilience and 21st century public health. *European Journal of Public Health*, 27(5):789–790, 10 2017. https://doi.org/10.1093/eurpub/ckx116

12. Vreman RA, Heikkinen I, Schuurman A et al. Unmet medical need: an introduction to definitions and stakeholder perceptions. *Value in Health*, 22(11):1275–1282, 2019. https://doi.org/10.1016/j.jval.2019.07.007

13. Acheson RM. The definition and identification of need for health care. *Journal of Epidemiology & Community Health*, 32(1):10–15, 1978. https://doi.org/10.1136/jech.32.1.10

14. Public Health England. *Community-centred public health: Taking a whole system approach*. 2020; https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/857029/WSA_Briefing.pdf

15. National Assembly for Wales (2015) Well-being of Future Generations (Wales) Act, 2015 http://www.legislation.gov.uk/anaw/2015/2/contents/enacted

16. Wolkin A, Patterson JR, Harris S et al. Reducing public health risk during disasters: identifying social vulnerabilities. *Journal of homeland security and emergency management*, 12(4):809–822, 2015. https://doi.org/10.1515/jhsem-2014-0104

17. World Bank. 2014. Community mapping for disaster risk reduction and management: Harnessing local knowledge to build resilience. Washington, DC: World Bank. License: Creative Commons Attribution CC BY 3.0.

18. Public Health England. The community response to coronavirus (COVID-19). (2020). https://publichealthmatters.blog.gov.uk/2020/06/01/the-community-response-to-coronavirus-covid-19/

19. Bevan Foundation (2020). *Coronavirus: community responses in Merthyr Tydfil*. https://www.bevanfoundation.org/wp-content/uploads/2020/07/Community-responses-to-coronavirus-Merthyr-Tydfil.pdf

20. Covid-19 Mutual Aid UK (2020). website: https://covidmutualaid.org

21. Supporting Communities (2020). Community response to the coronavirus crisis. https://supportingcommunities.org/latest-news/2020/4/1/community-response-to-the-coronavirus-crisis

22. Welsh Assembly Government Written Statement: Support for the Third Sector and Volunteering, 6 April 2020. https://gov.wales/written-statement-coronavirus-covid-19-support-third-sector-and-volunteering

23. Reuter C, Keufhold MA. Fifteen years of social media in emergencies: A' retrospective review and future directions for crisis Informatics. *Journal of Contingencies and Crisis Management*, 26(1):41–57, 2018. https://doi.org/10.1111/1468-5973.12196

24. Ahmed A. Hypothesizing the aptness of social media and the information richness requirements of disaster management (2012). *ECIS 2012 Proceedings*.157. https://aisel.aisnet.org/ecis2012/157

25. Tarasconi F, Farina M, Mazzei A, and Bosca A. (2017). The Role of Unstructured Data in Real-Time Disaster-related Social Media Monitoring. Presented at the Data Science for Emergency Management, Co-located with IEEEE BigData 2017 (DSEM), Boston: Zenodo. http://doi.org/10.5281/zenodo.1149056

26. Rossi C, Acerbo FS, Ylinen K, et al. (2018). Early detection and information extraction for weather-induced floods using social media streams. http://doi.org/10.5281/zenodo.2553119

27. Ahmed A, and Sinnappan S. The role of social media during Queensland floods: An empirical investigation on the existence of multiple communities of practice (MCoPs). *Pacific Asia Journal of the Association for Information Systems*, 5(2):2, 2013. https://doi.org/10.17705/1pais.05201

28. Heinzelman J, and Waters C. 2010. *Crowdsourcing crisis information in disaster-affected Haiti*. Special Report. Washington: United States Institute of Peace.

29. Smith WR, Stephens KK, Robertson BW, Li J, and Murthy D. Social Media in Citizen-Led Disaster Response: Rescuer Roles, Coordination Challenges, and Untapped Potential. In *Proceedings of the 15th ISCRAM Conference*, pages 639–648, 2018.

30. Scholz S, Knight P, Eckle M, Marx S, and Zipf A. Volunteered geographic information for disaster risk reduction: The missing maps approach and its potential within the red cross and red crescent movement. *Remote Sensing*, 10(8):1239, 2018. https://doi.org/10.3390/rs10081239

31. Rachunok BA, Bennett JB, and Nateghi R. Twitter and disasters: a social resilience fingerprint. in *IEEE Access*, 7:58495–58506, 2019. https://doi.org/10.1109/ACCESS.2019.2914797

32. Davies AR, Sharp CA, Homolova L, and Bellis MA (2019). *Population health in a digital age: The use of digital technology to support and monitor health in Wales*. Public Health Wales and Bangor University.

33. Asadi-Lari M, Packham C, and Gray D. Need for redefining needs. *Health and quality of life outcomes*, 1(34):1–5, 2003. https://doi.org/10.1186/1477-7525-1-34

34. National Health Service. People at higher risk from coronavirus, 2020. https://www.nhs.uk/conditions/coronavirus-covid-19/people-at-higher-risk/

35. Office For National Statistics. Deaths involving COVID-19, England and Wales: deaths occurring in May 2020, UK government; 2020 https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/deathsinvolvingcovid19englandandwales/deathsoccurringinmay2020

36. World Health Organisation. Coronavirus disease (COVID-19): How is it transmitted? Q&A, 2020. https://www.who.int/news-room/q-a-detail/coronavirus-disease-covid-19-how-is-it-transmitted

37. Rocklov J, and Sjodin H. High population densities catalyse the spread of COVID-19. *Journal of travel medicine*, 27(3):taaa038, 2020. https://doi.org/10.1093/tm/taaa038. PMID: 32227186: PMCID: PMC7184409

38. GNU General Public License (June 29, 2007). Version 3. Free Software Foundation. URL: http://www.gnu.org/licenses/gpl.html

39. Creative Commons licenses (25^{th} November 2013) Version 4.0. Creative Commons. https://creativecommons.org/licenses/

40. Chamberlain J, Turpin T, Maged A, Chatsiou A, and O'Callaghan K. Designing for collective intelligence and community resilience on social networks, *Human Computation* (in-press 2020).

41. National Survey for Wales 2018-19. Technical report, Welsh Government, 2019. https://gov.wales/sites/default/files/statistics-and-research/2019-07/national-survey-for-wales-april-2018-to-march-2019-technical-report_0.pdf

42. Welsh Government, Office for National Statistics. (2019). *National Survey for Wales, 2018-2019*. [data collection]. UK Data Service. SN: 8591, http://doi.org/10.5255/UKDA-SN-8591-1

43. Public Health Wales. Rapid Covid-19 surveillance. https://public.tableau.com/profile/public.health.wales.health.protection#!/vizhome/RapidCOVID-19virology-Public/Headlinesummary

44. NHS Digital. What we mean by digital inclusion, 2020. https://digital.nhs.uk/about-nhs-digital/our-work/digital-inclusion/what-digital-inclusion-is

45. ZOE. Covid Symptom Study, 2020. https://covid.joinzoe.com/data

46. Kim Y, Nordgren R, and Emery S. The Story of Goldilocks and Three Twitters APIs: A Pilot Study on Twitter Data Sources and Disclosure. *International Journal of Environmental Research and Public Health*, 17(3):864, 2020. https://doi.org/10.3390/ijerph17030864

47. Alastair Tanner. Epicosm: Epidemiology of cohort social media (v1.0), 2020. https://github.com/DynamicGenetics/Epicosm

48. Office for National Statistics Open Geography Portal. Local Authority Districts (December 2019) Boundaries UK BUC, 2019. https://geoportal.statistics.gov.uk/datasets/3a4fa2ce68f642e399b4de70643eeed3_0

49. McHugh ML. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–82, 2012. PMID: 23092060; PMCID: PMC3900052

50. McKinney W. Data Structures for Statistical Computing in Python. In Stefan van der Walt' and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56–61, 2010. https://doi.org/10.25080/Majora-92bf1922-00a

51. The pandas development team. pandas-dev/pandas: Pandas 1.0.3, March 2020. https://pandas.pydata.org

52. Dynamic Genetics Lab. COVID-19 community response map (source code v1.1.1), 2020. https://github.com/DynamicGenetics/Covid-19-community-response

53. Wickham H. Tidy data. *Journal of Statistical Software*, 59(10):1–23, 2014. http://www.jstatsoft.org/

54. National Research Council. 2007. *Successful response starts with a map: improving geospatial support for disaster management*. Washington, DC: The National Academies Press. https://doi.org/10.17226/11793

55. British Red Cross COVID-19 Vulnerability Index, 2020. https://britishredcrosssociety.github.io/covid-19-vulnerability/

56. Skills Platform and Zoe Amar Digital. Charity digital skills report 2020. http://report.skillsplatform.org/charity-digital-report-2020/

57. COVID-19 Community Response Map Documentation, 2020. https://osf.io/c48hw/

58. Imran M, Castillo C, Diaz F, and Vieweg S. Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)*, 47(4):1–38, 2015. https://doi.org/10.1145/2771588

59. Daly M, Sutin A, and Robinson E. Longitudinal changes in mental health and the covid-19 pandemic: Evidence from the UK household longitudinal study. *Psychological Medicine*, pages 1–37, 2020. https://doi.org/10.1017/S0033291720004432

60. Jaidka K, Giorgi S, Schwartz HA, Kern ML, Ungar LH, and Eichstaedt JC. Estimating geographic subjective well-being from Twitter: A comparison of dictionary and data-driven language methods. *Proceedings of the National Academy of Sciences*, 117(19):10165–10171, 2020. https://doi.org/10.1073/pnas.1906364117

61. Kim Y, Huang J, and Emery S. Garbage in, Garbage out: data collection, quality assessment and reporting standards for social media data use in health research, infodemiology and digital disease detection. *Journal of medical Internet research*, 18(2):e41, 2016. https://doi.org/10.2196/jmir.4738

Abbreviations

API:	Application Programming Interface
IMD:	Index of Multiple Deprivation
LA:	Local Authority
LSOA:	Lower Super Output Area
NHS:	National Health Service
NWIS:	National Health Service Wales Informatics Service
ONS:	Office for National Statistics
PHW:	Public Health Wales
SAIL:	Secure Anonymised Information Linkage
WCVA:	Wales Council for Voluntary Action
WG:	Welsh Government

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

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

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/).