The Networked Data Lab: Statistical analysis plan for a descriptive analysis of clinically extremely vulnerable people during COVID-19

December 2020

The Health Foundation

Authors	Affiliation
Sebastien Peytrignet, Karen Hodgson, Jorgen Engmann and Kathryn Dreyer	The Health Foundation
Jessica Butler, Katie Wilde, Nicola Beech, Claire Bell, Dimitra Blana, Corri Black, Adrian Martin, Graham Osler, Simon Sawhney, Bernhard Scheliga and Artur Wozniak	Aberdeen Centre for Health Data Science (ACHDS)
Alisha Davies and Jiao Song	Public Health Wales
Matthew Chisambi and Sara Sekelj	North West London (Imperial College Health Partners, Institute of Global Health Innovation and North West London Health and Care Partnership)
Ben Barr, Tim Caine, Simon Chambers, Helen Duckworth, Matthew Gilmore, Karen Jones, Michelle Jones, Lee Kirkham, David Knowles, Beverley Murray and Roberta Piroddi	NHS Liverpool CCG, NHS Wirral CCG and Wirral Council
Alex Brownrigg, Souheila Fox, Alison Phiri and Frank Wood	Leeds CCG and Leeds City Council

Contents

Summary

This Networked Data Lab analysis will focus on clinically extremely vulnerable (CEV) people, also known as the shielding population – the group of people most at risk of becoming seriously ill from COVID-19.* This group were asked to not leave their homes and to minimise all face-to-face contact up until the end of July 2020 in most of the UK. While the shielding guidance was paused over summer, people were once again asked to minimise their contact with others from November.

We are using a federated approach to data analysis and each partner will be contributing the same descriptive analysis based on their local population. These results will then be analysed, and aggregated where necessary, by the team at the Health Foundation.

The analysis will allow us to have a better understanding of who makes up this group of people and better understand their health care needs. It is structured around three outputs:

- the demographics of CEV people
- characterising the health and health care use of CEV people prior to the pandemic
- changes in secondary health care use during the shielding period for these patients.

Our main data sources for this analysis will be the shielded patient list (SPL), patient demographics databases and secondary health care records, from across our five partner sites.

This analysis is designed to provide insights on the CEV population and their use of secondary health care before and during the pandemic. This information will help local decision makers understand the needs of the CEV population and help arrange and design services that are able to meet these needs while still protecting them from COVID-19. This analysis is not an assessment of whether the shielding policy was effective. A limitation of this analysis is that it relies on secondary health care records and does not explore use of primary care or other health services.

Purpose of this document

This statistical analysis plan describes the analytical approach for the Networked Data Lab's central descriptive analysis of the shielding population.

It details the limitations of the analysis and how these should be considered when interpreting our findings.

This plan has been written before the analysis began, to allow coordination across network partners and ensure that all design and methodological choices are not influenced by what is found in the data. It may however be necessary to deviate from this plan if aspects of our analysis are not feasible.

* https://www.gov.uk/government/publications/guidance-on-shielding-and-protecting-extremely-vulnerable-persons-from-covid-19/guidance-on-shielding-and-protecting-extremely-vulnerable-persons-from-covid-19

Background

Networked Data Lab

The Networked Data Lab is a collaborative network of analytical teams across the UK working together on shared challenges and promoting the use of analytics in improving health and social care.

Using linked data, our aim is to work together to understand and try to solve the toughest health and care issues facing the UK today.

The Networked Data Lab was established by the Health Foundation and our five partners are:

- The Aberdeen Centre for Health Data Science (ACHDS) which includes NHS Grampian and the University of Aberdeen
- Public Health Wales, NHS Wales Informatics Service (NWIS), Swansea University (SAIL Databank) and Social Care Wales (SCW)
- Imperial College Health Partners (ICHP), Institute of Global Health Innovation (IGHI), Imperial College London (ICL), and North West London CCGs
- Liverpool CCG, Healthy Wirral Partnership and Citizens Advice Bureau
- Leeds CCG and Leeds City Council

Each of our five partners has linked health and social care data sets (see Table 1 in Annex).

Our first analysis will focus on clinically extremely vulnerable (CEV) people (also referred to as 'shielding'). This group of people are the ones deemed the most at risk of becoming seriously ill from COVID. Guidance was rolled out to advise people not to leave their homes and to minimise all face-to-face contact from 22 March 2020.* In most parts of the UK, this policy was paused by the end of July/early August.† During the second national lockdown in November in England, CEV people were once again asked to minimise their contact with other people.

One of the objectives of this first analysis is to test access to data and establish ways of analytical working. For this reason, we have chosen to pursue a descriptive analysis, rather than an analysis that requires detailed statistical modelling. Furthermore, we are basing this analysis on secondary care data. This will allow us to develop and test an approach to data standardisation on data sets where there are national data collections in place, and as a result a significant level of consensus across partners.

---

* https://www.gov.uk/government/news/major-new-measures-to-protect-people-at-highest-risk-from-coronavirus; https://www.sehd.scot.nhs.uk/publications/DC20200326letter.pdf

† https://gov.wales/shielding-wales-pause-16-august; https://www.gov.scot/news/shielding-to-be-paused/; https://www.england.nhs.uk/coronavirus/wp-content/uploads/sites/52/2020/06/C0624-shielding-letter-to-nhs.pdf

Structure for first analysis

This analysis will be structured around a central descriptive analysis, which is the focus of the present statistical analysis plan.

The central analysis will be supplemented by five satellite analyses: one per partner, which are related to the topic of clinically extremely vulnerable people but also uniquely reflect the local context and data availability. The satellite analyses will not be covered by this statistical analysis plan.

Topic selection process

This topic was identified through a process that involved consulting with local partners, as well as patients and the public. Partners were asked to identify priorities locally and we also held a patient focus group.

The topic was selected based on it being a priority across a majority of local areas. Patients highlighted that access to services was a major priority and agreed that shielded patients would be a useful cohort to examine for the first topic.

The research questions for each of the outputs have been developed through a series of workshops with our analytical leads within our local partners.

Aims

Overall, the descriptive central analysis will allow us to have a better understanding of who makes up this group of people and their health care needs.

Comparing results from our five partners will enable an exploration of the geographical variations in those findings.

The analysis will be split into three outputs, the aims of which are detailed below.

Output 1: Demographics of CEV people

CEV people were initially identified because they presented with a variety of conditions which places them at high risk from COVID-19. While these people are all previously known to the health service because of those conditions, they are a disparate group.

This output will allow us to quantify the size of the CEV population at each of the NDL locations, and to better understand who these people are.

Through characterising each cohort using both basic demographic information and information on how and why they were added to the shielded patient list, we will develop a better understanding of the needs of this group.

This descriptive analysis will highlight similarities and differences between local partners in the shielded populations they are serving and bring to the forefront the different approaches used to identify those who are most clinically vulnerable to COVID-19.

By developing a better understanding of heterogeneity of CEV people, services can be adjusted to better meet their needs.

Output 2: Characterising the health and health care use of CEV people prior to the pandemic

This output will aim to provide a picture of CEV people's health and pre-pandemic health care use, using information captured within secondary health care data.

People on the shielded patient list are clinically extremely vulnerable to COVID-19. By understanding the degree of multimorbidity for this group, the other long-term conditions that these people have, and quantifying how these long-term conditions relate to their health care use prior to the pandemic, we can build a more detailed picture of the 'usual' health of this group of particularly clinically vulnerable individuals, prior to the pandemic.

Understanding the health of this population, as well as their previous interactions with the health system, will enable policymakers and local services to better understand their health and health care needs.

Output 3: Changes in patterns of secondary health care use among CEV people

This output will describe the changes in secondary health care use for CEV people during the period of shielding between March and July 2020. We will also examine the direct impacts of COVID-19 on this group across each local area.

These findings will be useful in building an understanding of the impacts from both a patient and a service perspective. Findings can also help inform service planning for both future waves of the pandemic, and for exploring the impact of the shielding period on CEV people and the potential long-term impacts of unmet health care needs.

Data

Our analyses will rely on the following data sources:

- The shielded patient list (SPL).
- Patient demographics databases.
- External open data sources linked to a patient's LSOA (or other geography) of residence. These are the 2019 English/Welsh/Scottish Index of Multiple Deprivation quintiles and urban/rural indicators based on the 2011 Census.
- Secondary health care records (inpatient, outpatient and A&E) for the period 1 March 2018 – 31 July 2020 (although this time window is subject to change and is not final).

The underlying reasons for shielding will be drawn directly from the provided versions of the SPL, rather than deduced from local health records for the purpose of this analysis. We do not anticipate needing to determine why people were shielding from sources other than the SPL.

The central analysis will not rely on primary care health records.

Cohorts

Defining the shielding cohort

Since its creation at the onset of the pandemic, the SPL and its underlying algorithm have been subject to regular changes, with patients being added and removed regularly. In order to conduct this analysis, a cohort of people who were told to shield needs to be defined.

The shielding cohort will be defined as any person who was on the SPL at any point in the period prior to 31 July 2020.

As local areas have different approaches to managing those who were asked to shield, we expect that each partner may need to take a slightly different approach to defining this cohort, depending on the data they have available. However, the final cohort should be maximally inclusive to reflect the above definition as far as possible.

Other inclusion criteria

After identifying those people who were on the shielding patient list at any point prior to 31 July 2020, there are a number of other inclusion criteria that need to be met.

For Output 1 on demographics of CEV people, individuals are included regardless of whether they have historical health care data.

For Outputs 2 and 3, included individuals should have historical health care data available for the period between 1 March 2018 and 29 February 2020 (two years pre-COVID). This enables the identification of pre-existing long-term conditions, as well as general use of secondary health care services prior to the pandemic, to be examined.

Only those individuals who were alive at the point they were added to the SPL should be included within the cohort.

Methods

For all analyses the total size of the cohort and the number of individuals with missing or unknown information will also be reported.

Output 1: Demographics of CEV people

The following demographic characteristics will be summarised (number of individuals per category, as defined in provided tables):

- reason(s) for shielding (provided in the SPL)
- method of addition to the SPL
- sex
- age (categorised into bands)
- ethnicity
- place-of-residence characteristics for each patient including:
  - IMD-based deprivation level
  - urban/rural classification

Interactions of interest are:

- reason for shielding by age
- reason for shielding by deprivation
- age by sex
- deprivation by sex
- age by deprivation

Output 2: Characterising the health and health care use of CEV people prior to the pandemic

We will identify long-term conditions by looking back at hospital admission health care records in the period between 1 March 2018 and 29 February 2020.

The long-term conditions identified for this analysis are the conditions which comprise the Elixhauser Comorbidity Index. In some cases, there may be overlap with the conditions on the shielded patient list.

We will aim to identify each of the following conditions by identifying inpatient records that match the ICD-10 codes below. In short, an individual is assumed to have had the condition if any record exists during this period matching the relevant ICD-10 codes corresponding to this condition (regardless of whether this was the primary diagnosis).

A coding algorithm is set out by Quan et al (2005).* Please note that the conditions and ICD-10 codes reported here may not be the final version used in the analysis – and some adjustment based on local coding practices may be necessary.†

* Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi JC, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Med Care. 2005; 43(11): 1130–9. doi: 10.1097/01.mlr.0000182534.19832.83. PMID: 16224307.

† https://pubmed.ncbi.nlm.nih.gov/21764557/

Conditions	ICD 10 Codes
Alcohol abuse	F10, E52, G62.1, I42.6, K29.2, K70.0, K70.3, K70.9, T51.x, Z50.2, Z71.4, Z72.1
Blood loss anaemia	D50.0
Cardiac arrhythmias	I44.1 - I44.3, I45.6, I45.9, I47.x - I49.x, R00.0, R00.1, R00.8, T82.1, Z45.0, Z95.0
Chronic pulmonary disease	I27.8, I27.9, J40.x - J47.x, J60.x - J67.x, J68.4, J70.1, J70.3
Coagulopathy	D65 - D68.x, D69.1, D69.3 - D69.6
Congestive heart failure	I09.9, I11.0, I13.0, I13.2, I25.5, I42.0, I42.5 - I42.9, I43.x, I50.x, P29.0
Deficiency anaemia	D50.8, D50.9, D51.x - D53.x
Depression	F20.4, F31.3 - F31.5, F32.x, F33.x, F34.1, F41.2, F43.2
Diabetes (combined uncomplicated and complicated)	E10.0, E10.1, E10.9, E11.0, E11.1, E11.9, E12.0, E12.1, E12.9, E13.0, E13.1, E13.9, E14.0, E14.1, E14.9, E10.2 - E10.8, E11.2 - E11.8, E12.2 - E12.8, E13.2 - E13.8, E14.2 - E14.8
Drug abuse	F11.x - F16.x, F18.x, F19.x, Z71.5, Z72.2
Fluid and electrolyte disorders	E22.2, E86.x, E87.x
Hypertension (combined uncomplicated and complicated)	I10.x, I11.x - I13.x, I15.x, I11.x - I13.x, I15.x
Hypothyroidism	E00.x - E03.x, E89.0
Liver disease	B18.x, I85.x, I86.4, I98.2, K70.x, K71.1, K71.3 - K71.5, K71.7, K72.x - K74.x, K76.0, K76.2 - K76.9, Z94.4
Lymphoma	C81.x - C85.x, C88.x, C96.x, C90.0, C90.2
Metastatic cancer	C77.x - C80.x
Obesity	E66.x
Other neurological disorders	G10.x - G13.x, G20.x - G22.x, G25.4, G25.5, G31.2, G31.8, G31.9, G32.x, G35.x - G37.x, G40.x, G41.x, G93.1, G93.4, R47.0, R56.x
Paralysis	G04.1, G11.4, G80.1, G80.2, G81.x, G82.x, G83.0 - G83.4, G83.9
Peptic ulcer disease, excluding bleeding	K25.7, K25.9, K26.7, K26.9, K27.7, K27.9, K28.7, K28.9
Peripheral vascular disorders	I70.x, I71.x, I73.1, I73.8, I73.9, I77.1, I79.0, I79.2, K55.1, K55.8, K55.9, Z95.8, Z95.9
Psychoses	F20.x, F22.x - F25.x, F28.x, F29.x, F30.2, F31.2, F31.5
Pulmonary circulation disorders	I26.x, I27.x, I28.0, I28.8, I28.9
Renal failure	I12.0, I13.1, N18.x, N19.x, N25.0, Z49.0 - Z49.2, Z94.0, Z99.2
Rheumatoid arthritis/collagen vascular diseases	L94.0, L94.1, L94.3, M05.x, M06.x, M08.x, M12.0, M12.3, M30.x, M31.0 - M31.3, M32.x - M35.x, M45.x, M46.1, M46.8, M46.9
Solid tumour without metastasis	C00.x - C26.x, C30.x - C34.x, C37.x - C41.x, C43.x, C45.x - C58.x, C60.x - C76.x, C97.x
Valvular disease	A52.0, I05.x - I08.x, I09.1, I09.8, I34.x - I39.x, Q23.0 - Q23.3, Z95.2 - Z95.4
Weight loss	E40.x - E46.x, R63.4, R64

The number of patients with each Elixhauser long-term condition will be described (where the number of people is sufficient for statistical disclosure purposes).

For each person, a single comorbidity score will be computed using as inputs:

- the presence/absence of each condition
- a single weight for each condition, based on an algorithm as set out by van Walraven et al (2009)*.

We will present a frequency table of these patient-level comorbidity scores, grouped into the following classes: <0, 0, 1–4 and >=5.

We will also report the number of people with 0, 1 or 2+ of these Elixhauser long-term conditions.

These two comorbidity measures will capture the burden of multiple long-term conditions.

We will report the number of hospital admissions with a primary diagnosis relating to each Elixhauser long-term condition between 1 March 2018 and 29 February 2020 (where the number of people is sufficient for statistical disclosure purposes). These will be used to calculate the admission rate per person with that condition.

The burden of multiple long-term conditions will also be summarised by:

- reason for addition to the SPL (using a compressed list of conditions: a level for each of the top three most frequent followed by other, and unknown)
- age (with compressed age bands based on those used in the patient demographics)
- sex
- deprivation quintile

Output 3: Changes in patterns of secondary health care use among CEV people

We will report summarised data on five types of secondary health care:

- hospital admissions – emergencies
- hospital admissions – electives or other
- outpatient attendance
- A&E attendance
- any secondary health care (ie any of the above)

For each of the above five types of health care contact, we will describe use with two different metrics:

- The number of people with any contacts each month – this will be used to calculate the percentage of people with any contacts per month.
- The number of contacts each month – this will be used to compute the average number of contacts per person per month.

* van Walraven C, Austin PC, Jennings A, Quan H, Forster AJ. A modification of the Elixhauser comorbidity measures into a point system for hospital death using administrative data. Med Care. 2009; 47(6): 626–33. doi: 10.1097/MLR.0b013e31819432e5. PMID: 19433995. https://pubmed.ncbi.nlm.nih.gov/19433995/

We will produce these monthly secondary health care use statistics for each month spanning the period 1 March 2018 – 31 July 2020.

If sample sizes are sufficiently large for statistical disclosure processes, we may also present these values by:

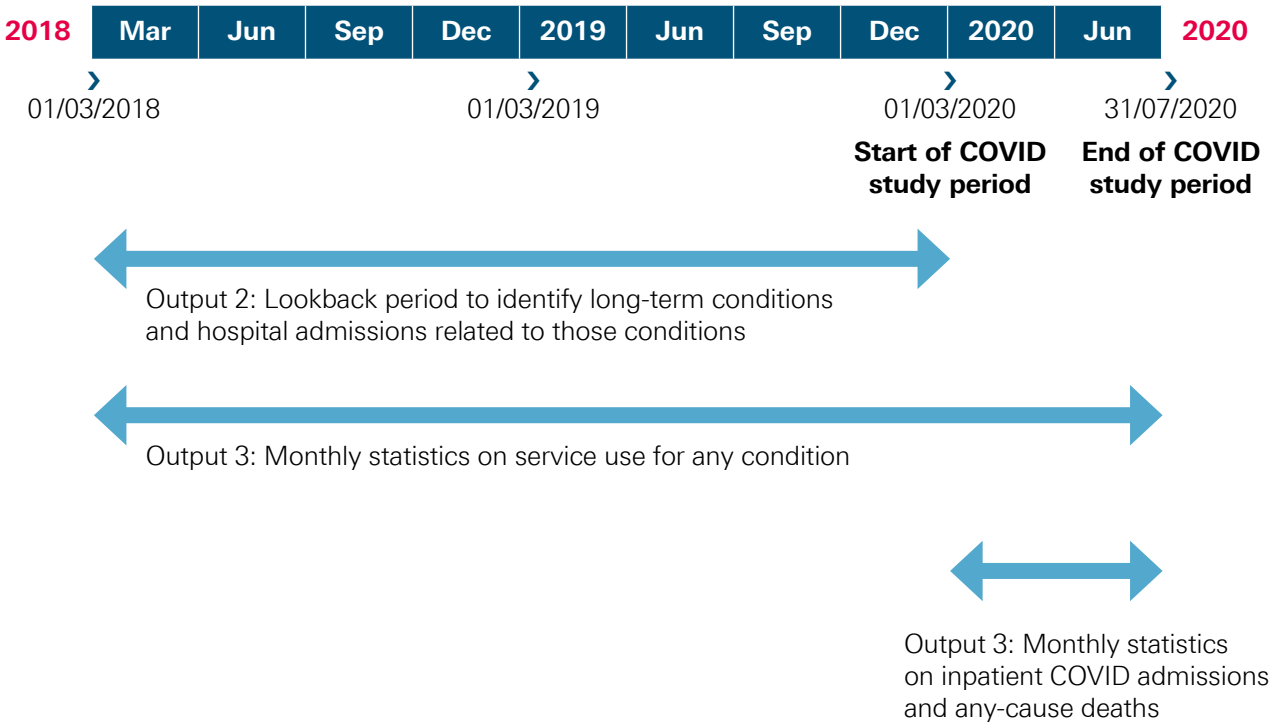
- age (with compressed age bands based on those used in the patient demographics)
- sex
- deprivation quintile.

We will also summarise the number of COVID-related hospital admissions over the period 1 March 2020 – 31 July 2020. COVID admissions will be identified using ICD-10 codes as per WHO guidelines, including cases when the virus was and wasn't identified.* If the number of COVID-related admissions is small, to avoid disclosive results these figures will be aggregated over the entire period rather than reported monthly.

If this data is available, we will also summarise the number of any-cause deaths over the period 01 March 2020 – 31 July 2020, regardless of where they occurred. We will also report as a subset the number of deaths that occurred in hospital. If the number of deaths is small, to avoid disclosive results these figures will be aggregated over the entire period rather than reported monthly.

Additional information

A summary of the time periods used to identify outcomes in Outputs 2 and 3 is provided below. Please refer to the sections above for the full study methodology.



* https://www.who.int/classifications/icd/COVID-19-coding-icd10.pdf

Limitations

This analysis does not draw on primary care records and relies exclusively on secondary health care records to identify health conditions and measure service use. Therefore, conditions (eg diabetes) that have been managed in a primary care setting and have not resulted in a hospital admission will not be identified. The result is that we are underestimating the burden of disease in this cohort.

This analysis identifies people who shielded based on their inclusion in the SPL. There may be people who were not on this list but who are at risk of becoming seriously ill from COVID-19. This analysis was not able to include these individuals. Furthermore, we are unable to identify whether individuals followed shielding guidelines or not.

This analysis takes March 2020 – July 2020 as the relevant period of study to understand the impact of the pandemic and shielding advice on people on the SPL. While the national changes made to health care delivery and the development of the SPL occurred during March 2020, there may have been some changes in interactions with health care services pre-dating this. For context, the reported first case in the UK occurred on 31 January 2020 and the first reported death was on 5 March – these events and others may have impacted both individual and clinician behaviours.

This analysis is descriptive and does not aim to answer causal questions relating to whether the shielding policy was effective at preventing adverse events among high-risk people.

Reporting

Our analysis will be published online by the Health Foundation on their website, using a variety of formats. These may be:

- blogs
- charts or long charts
- long reads
- briefing papers.

A suggested timeline for these outputs is as follows:

Output	Format	Date
Demographics of shielded patients	Long chart	January 2021
Previous secondary health care use for conditions	Long chart	February 2021
Current and previous secondary health care use among shielded patients	Long chart	March 2021
Lessons learned from the central and satellite analyses	Long read or blog	March 2021

Annex

Table 1. Linked data sets across Networked Data Lab partners

Data	North West London	Wales (Public Health Wales, NHS Wales Informatics Service NWIS, Social Care Wales & SAIL Databank)	Liverpool & Healthy Wirral Partnership	NHS Grampian and University of Aberdeen	Leeds City Council
Primary care					
GP contacts	✔	✔	✔		✔
Prescriptions	✔	✔	✔	✔	✔
Secondary care					
A&E	✔	✔	✔	✔	✔
Outpatient	✔	✔	✔	✔	✔
Inpatients	✔	✔	✔	✔	✔
Specialist mental health	✔	✔	✔	✔	✔
Social care					
Adult social care	✔		✔	✔	✔
Children's social care		✔		✔	
Community care	✔	✔		✔	✔
Other					
	Electronic frailty index	COVID-19 linked data	Citizens Advice Bureau	Maternity	Community beds
	Vulnerable patient watch lists	Survey data	Infant mortality risk	Cancer	
	High cost drugs	Specialist services and screening data	Complex lives	Aberdeen children of the 1950s	
	Asthma radar	Electronic frailty index and care homes		GenScotland	
	Patient capacity to self-care	Other data sets in SAIL			