Journal of
Epidemiology &
Community Health

Clinical trial results in context: comparison of baseline characteristics and outcomes of 38,510 RECOVERY trial participants versus a reference population of 346,271 people hospitalised with COVID-19 in England

Journal:	*Journal of Epidemiology & Community Health*
Manuscript ID	Draft
Article Type:	Original research
Date Submitted by the Author:	n/a
Complete List of Authors:	Pessoa-Amorim, Guilherme; University of Oxford, Clinical Trial Service Unit, Oxford Population Health; University of Oxford, Medical Research Council Population Health Research Unit Goldacre, Raph; University of Oxford, Big Data Institute, Oxford Population Health Crichton, Charles; University of Oxford, Big Data Institute, Oxford Population Health Stevens, Will; University of Oxford, Clinical Trial Service Unit, Oxford Population Health; University of Oxford, Medical Research Council Population Health Research Unit Nunn, Michelle; University of Oxford, Clinical Trial Service Unit, Oxford Population Health; University of Oxford, Medical Research Council Population Health Research Unit King, Andrew; University of Oxford, National Perinatal Epidemiology Unit, Oxford Population Health Murray, Dave; University of Oxford, National Perinatal Epidemiology Unit, Oxford Population Health Welsh, Richard; University of Oxford, National Perinatal Epidemiology Unit, Oxford Population Health Pinches, Heather; NHS England, NHS DigiTrials Rees, Andrew; NHS England, NHS DigiTrials Morris, Eva JA; University of Oxford, Big Data Institute, Oxford Population Health; Oxford University Hospitals NHS Foundation Trust, NIHR Oxford Biomedical Research Centre Landray, Martin; University of Oxford, Clinical Trial Service Unit, Oxford Population Health, University of Oxford; University of Oxford, Medical Research Council Population Health Research Unit, Oxford Population Health; University of Oxford, Big Data Institute, Oxford Population Health; Oxford University Hospitals NHS Foundation Trust, NIHR Oxford Biomedical Research Centre Haynes, Richard; University of Oxford, Clinical Trial Service Unit, Oxford Population Health; University of Oxford, Medical Research Council Population Health Research Unit, Oxford Population Health Horby, Peter; University of Oxford, Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine; University of Oxford, International Severe Acute Respiratory and emerging Infections Consortium (ISARIC); University of Oxford, Pandemic Sciences Centre;

	Oxford University Hospitals NHS Foundation Trust, Department of Infectious Diseases and Microbiology Wallendszus, Karl; University of Oxford, Clinical Trial Service Unit, Oxford Population Health; University of Oxford, Medical Research Council Population Health Research Unit, Oxford Population Health Peto, Leon; University of Oxford, Clinical Trial Service Unit, Oxford Population Health; University of Oxford, Medical Research Council Population Health Research Unit, Oxford Population Health; Oxford University Hospitals NHS Foundation Trust, Department of Infectious Diseases and Microbiology Campbell, Mark; University of Oxford, Clinical Trial Service Unit, Oxford Population Health; University of Oxford, Medical Research Council Population Health Research Unit, Oxford Population Health; Oxford University Hospitals NHS Foundation Trust, Department of Infectious Diseases and Microbiology Harper, Charlie; University of Oxford, Clinical Trial Service Unit, Oxford Population Health; University of Oxford, Medical Research Council Population Health Research Unit, Oxford Population Health Mafham, Marion; University of Oxford, Clinical Trial Service Unit, Oxford Population Health, University of Oxford; University of Oxford, Medical Research Council Population Health Research Unit, Oxford Population Health
Keywords:	COVID-19, RANDOMIZED CONTROLLED TRIAL, EPIDEMIOLOGY, METHODS, RECORD LINKAGE

Title page

Title:

Clinical trial results in context: comparison of baseline characteristics and outcomes of 38,510 RECOVERY trial participants versus a reference population of 346,271 people hospitalised with COVID-19 in England

Short title:

Setting the RECOVERY trial results in context

Authors:

Guilherme Pessoa-Amorim[ab]†*, Raphael Goldacre[c]†, Charles Crichton[c]†, Will Stevens[ab], Michelle Nunn[ab], Andy King[d], Dave Murray[d], Richard Welsh[d], Heather Pinches[e], Andrew Rees[e], Eva JA Morris[cf], Martin Landray[abcd], Richard Haynes[ab], Peter Horby[ghij], Karl Wallendszus[ab], Leon Peto[abj], Mark Campbell[abj]‡, Charlie Harper[ab]‡, Marion Mafham[ab]‡

†Joint first-authors

‡Joint senior authors

Affiliations:

[a] Clinical Trial Service Unit, Oxford Population Health, University of Oxford, Oxford, United Kingdom

[b] Medical Research Council Population Health Research Unit, Oxford Population Health, University of Oxford, Oxford, United Kingdom

[c] Big Data Institute, Oxford Population Health, University of Oxford, Oxford, United Kingdom

[d] National Perinatal Epidemiology Unit, Oxford Population Health, University of Oxford, Oxford, United Kingdom

[e] NHS DigiTrials, Leeds, United Kingdom

[f] NIHR Oxford Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, Oxford, United Kingdom

[g] Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom

[h] International Severe Acute Respiratory and emerging Infections Consortium (ISARIC), University of Oxford, Oxford, United Kingdom

[i] Pandemic Sciences Centre, University of Oxford, Oxford, United Kingdom

[j] Department of Infectious Diseases and Microbiology, Oxford University Hospitals NHS Foundation Trust, Oxford, United Kingdom

***Corresponding author:**

Guilherme Pessoa-Amorim. Clinical Trial Service Unit, Richard Doll Building, Old Road Campus, University of Oxford, Roosevelt Drive, Oxford OX37LF, United Kingdom. Email: guilherme.pessoa-amorim@ndph.ox.ac.uk

Word count: 2858 (excluding abstract, key messages, and declarations)

Abstract

Background:

Randomised trials are essential to reliably assess medical interventions. Nevertheless, interpretation of such studies, particularly when considering absolute effects, is enhanced by understanding how the trial population may differ from the populations it aims to represent.

Methods:

We compared baseline characteristics and mortality of RECOVERY participants recruited in England (n= 38,510) with a reference population hospitalised with COVID-19 in England (n = 346,271) from March 2020 - November 2021. We used hospitalisation and mortality data for both cohorts to extract demographics, comorbidity/frailty scores, and crude and age- and sex-adjusted 28-day all-cause mortality.

Results:

Demographics of RECOVERY participants were broadly similar to the reference population, but RECOVERY participants were younger (mean age [standard deviation]: RECOVERY 62.6 [15.3] vs reference 65.7 [18.5] years) and less frequently female (37% vs 45%). Comorbidity and frailty scores were lower in RECOVERY, but differences were attenuated after age stratification. Age- and sex-adjusted 28-day mortality declined over time but was similar between cohorts across the study period (RECOVERY 23.7% [95% confidence interval: 23.3%-24.1%]; vs reference 24.8% [24.6%-24.9%]), except during the first pandemic wave in the UK (March-May 2020) when adjusted mortality was lower in RECOVERY.

Conclusion:

Adjusted 28-day mortality in RECOVERY was similar to a nationwide reference population of patients admitted with COVID-19 in England during the same period but varied substantially over time in both cohorts. Therefore, the absolute effect estimates from RECOVERY were broadly applicable to the target population at the time, but should be interpreted in the light of current mortality estimates.

Key messages

What is already known on this topic:

- Relative treatment effect sizes from randomised controlled trials (RCTs) are used to estimate the likely absolute impact of an intervention at a population level, but absolute event rates (and therefore expected absolute risk reductions) in trials may differ from the target population.

What this study adds:

- Using linked healthcare systems data, we compared baseline characteristics and outcomes of participants recruited in England to the RECOVERY trial of in-hospital COVID-19 treatments with those from a reference population of patients admitted with COVID-19 in England.
- RECOVERY participants had similar characteristics to those in the reference population, but were slightly younger and less frequently female.
- After adjustment for age and sex, 28-day mortality in RECOVERY was similar to the reference population throughout the study period, except March-May 2020 when it was lower in RECOVERY, with mortality rates in both cohorts falling during the period studied.

How this study might affect research, practice or policy

- The use of absolute event rates from healthcare systems data can be used to assess the likely impact of the relative treatment effects from clinical trials at a population level, but secular trends in event rates should be considered.

Introduction

Randomised controlled trials (RCTs) are essential to reliably evaluate safety and efficacy of health interventions.[1,2] The use of randomisation (with allocation concealment) minimises the risk of bias but, inevitability, due to eligibility criteria, trial participants are rarely representative of the populations whose treatment they aim to inform. Nonetheless, the proportional estimates of treatment effects from the trial are usually generalisable to the broader population, unless there are good grounds for believing there may be systematic differences in the effectiveness of the intervention or in the biology of the target disease outside of the trial setting (e.g. the advent of a new variant that renders a pathogen resistant to the particular drug that was studied).[3] However, the estimates of *absolute* harm and benefit generated by such trials may not be directly generalisable, and assessment of the absolute rates of the relevant outcomes in the target population is useful to understand the likely absolute effects of the intervention in clinical practice.[4,5]

The Randomised Evaluation of COVID-19 Therapy (RECOVERY) trial is an ongoing, randomised, controlled, open-label, pragmatic, platform trial of potential therapies for patients hospitalised with COVID-19.[6] Eligibility criteria are broad and simple (i.e. hospitalisation for suspected or confirmed COVID-19), and trial procedures are streamlined to be feasible in local practice. Data collection by trial staff, using dedicated case-report forms (CRF), focuses on the minimum information needed, and is complemented with extensive linkage to several healthcare systems data sources in the UK. The trial is taking place in all acute UK National Health Service (NHS) hospitals, and in several other countries globally.

Here, we aimed to compare the baseline characteristics (demographics and comorbidities) and all-cause 28-day mortality (the trial primary outcome) for RECOVERY participants with a reference population hospitalised with COVID-19, within England.

Methods

RECOVERY cohort

The RECOVERY trial design has been described previously.[6] Briefly, RECOVERY recruits patients admitted to hospital with confirmed or suspected COVID-19 who are considered suitable for inclusion by their attending clinical team. Recruitment was not targeted to any particular subgroups or aimed at achieving a representative sample of the target population; the aim was to recruit a large number of participants rapidly. Randomisation is performed via a short online CRF in which essential baseline data are collected. Follow-up data are collected using a simple CRF upon death, hospital discharge, or at 28 days from randomisation (whichever occurs sooner). In the UK, these data are complemented with linkage to national healthcare systems data sources. The protocol, data analysis plan, baseline characteristics and outcome derivation documentation, and published results are openly available at www.recoverytrial.net, and the trial is registered with ISRCTN (50189673) and ClinicalTrials.gov (NCT04381936). Written informed consent was obtained from all the patients or from a legal representative if they were unable to provide consent. The RECOVERY trial has been approved by the UK Medicines and Healthcare products Regulatory Agency and the Cambridge East Research Ethics Committee (reference 20/EE/0101).

For this analysis we included all RECOVERY participants recruited in England who had not withdrawn consent and had available healthcare systems data on hospital admissions (Hospital Episode Statistics [HES]),[7] with or without mortality data from official death records (Civil Registrations).[8] We excluded children aged <16 years due to difficulties in accessing linked healthcare systems data in this group in RECOVERY. HES data contained information on admissions to all NHS hospitals in England (using standardised coding practices since the 1990s), namely admission and discharge dates and relevant diagnostic and procedure codes. Diagnostic codes are recorded using the *International Classification of Diseases and Related Health Problems, Tenth Revision* (ICD-10) clinical terminology, and can be assigned a position from 1 to 20; codes in position 1 usually indicate the primary cause of admission (or main cause of extension of hospital stay).[9] Civil Registrations included information on date of death and

underlying and contributing causes of death (also coded using ICD-10). HES and Civil Registrations were linked and supplied by NHS England.[10]

Reference population

To derive a reference population of people hospitalised with COVID-19 in England (thus potentially eligible for RECOVERY), we used an anonymised database covering the entirety of England which includes linked HES and Civil Registrations data continuously collected since 1999. These data were linked and supplied by NHS England, and are analysed at the University of Oxford.[11] More information can be found in the NHS England Data Uses Register at http://digital.nhs.uk/services/data-access-request-service-dars/data-uses-register (reference: DARS-NIC-315419-F3W7K). Approval for the use of the datasets was provided by the Central and South Bristol Research Ethics Committee (ref 04/Q2006/176).

The reference population was ascertained based on the presence of a COVID-19 ICD-10 code (U071 - "COVID-19, Virus identified", or U072 - "COVID-19, Virus not identified").[12] This approach was informed by preliminary cross-validation work (Annex III) using linked HES and SARS-CoV-2 testing data for RECOVERY participants, which showed 92% of RECOVERY participants recruited in England with a positive SARS-CoV-2 test (as captured in NHS England's COVID-19 Second Generation Surveillance System – SGSS dataset)[13] had an admission in the HES data which included one of these codes in the primary diagnostic position. We therefore restricted our reference population to individuals with relevant ICD-10 codes in the primary position to avoid inclusion of people in whom COVID-19 was not the main reason for care. The RECOVERY cohort largely overlaps the reference population, but given the anonymised nature of the national datasets it was not possible to identify them.

Analysis period

For each individual in RECOVERY and the reference population, we assigned an index date as the start of the earliest HES episode with U071/U072 in the first diagnostic position. For RECOVERY participants with index dates before 1st March 2020 (indicating long episodes before inclusion in the study; n=22), or no COVID-19 codes in their HES records (n=1465) we used randomisation date as the index date. We then restricted our analysis period to index dates between 1st March 2020 and 30th

November 2021 inclusive. These analyses were not extended beyond this time-point as the launch of the high-dose dexamethasone comparison in the UK (only suitable to patients with oxygen or ventilation requirements) resulted in more selected patient populations being included in the trial.[14]

Baseline characteristic and outcomes

We used HES data in both cohorts to extract baseline clinical characteristics and demographics including age, sex, ethnicity, deprivation (quintile of Index of Multiple Deprivation 2019),[15] geographical location, Charlson Comorbidity Score [16,17] and its components, and Hospital Frailty Risk Score.[18] Comorbidities were defined as the presence of a relevant ICD-10 code in any diagnostic position recorded within 5 years before the index date (i.e. excluding the index episode). Further methodological details, including the ICD-10 codes used, are provided in Annex I. Geographical location data (including for deprivation assessments) were extracted from HES records and ascertained from full postcode in the RECOVERY HES data and lower-super output area of the postcode in the national HES data.

For outcomes, we calculated all-cause mortality within 28 days using linked HES and Civil Registrations data. Ascertainment of fact and date of death was based on these linked data sources (derivation methodology described elsewhere).[19] We considered death records occurring in either healthcare systems data source. We ignored reports of deaths of RECOVERY participants recorded only on the CRF data as there were no CRF data for the reference population.

Statistical analyses

This analysis is limited to RECOVERY participants in England with available HES data. To assess how this selection may have affected the cohort characteristics we first compared the characteristics of those recruited in England with those recruited in other UK nations (using CRF data for all characteristics except ethnicity, and healthcare systems data in each nation for ethnicity). We then compared the characteristics of RECOVERY participants recruited in England who had available HES data with those who did not (using CRF data for all characteristics except ethnicity, for which we used healthcare systems data from primary care).

We compared baseline characteristics and 28-day mortality of the RECOVERY cohort with those of the reference population, in each case restricted to England only. Age was stratified into 4 groups: <60, 60-69, 70-79, and ≥80. We presented continuous parameters as mean with standard deviation (SD) or median with interquartile range (IQR) as appropriate (with visual assessment of frequency distribution for normality), and frequency counts and percentage distribution for categorical parameters. We compared age, sex, and region of residence by calculating a representativeness ratio - defined as the proportion of people within RECOVERY in each category divided by the proportion of people within the reference population in the same category – and presented these along with 95% confidence intervals [95% CI].[20] We also calculated a recruitment ratio defined as number of individuals included in RECOVERY divided by the number of individuals in the reference population. We then aggregated individuals in each cohort into three-month periods and conducted the same calculations as above for each time period separately.

The primary RECOVERY trial outcome of 28-day all-cause mortality was calculated starting from the index date in both cohorts, overall and over time (by three-month periods). We presented crude and age- and sex-adjusted mortality rates with 95% CI,[20] with adjustment performed using direct standardisation methods[20] (i.e. applying RECOVERY mortality rates to the reference population age and sex composition using the age groups mentioned above). Further methodological details are provided in Annex I.

We used Stata v17/MP to derive baseline characteristics and outcomes in HES and Civil Registrations data in both cohorts, and R v4.2.1 for all subsequent data management, statistical analysis, and plotting (further details are provided in Annex I).

Ethics committee approval

The RECOVERY trial has been approved by the UK Medicines and Healthcare products Regulatory Agency and the Cambridge East Research Ethics Committee (reference 20/EE/0101). The dataset used to build the reference population by the University of Oxford has been approved by the Central and South Bristol Research Ethics Committee (reference 04/Q2006/176).

1

Results

Baseline characteristics

Up until 1st September 2022, RECOVERY recruited 46,010 participants, of which 44,766 in the UK and 39,952 in England. Of these, 39,304 (98.4%) had available HES data, and 38,780 were recruited within the analysis period (1st March 2020 – 30th November 2021). After excluding participants aged below 16 at the index date, a total of 38,510 participants were finally included in our analysis (Figure 1). RECOVERY participants recruited in other UK nations had generally similar characteristics to those recruited in England (**Error! Reference source not found.**). People with no HES data available were younger, less frequently of white ethnicity, and had generally lower comorbidity burden and need for respiratory support at randomisation (Supplementary Table S2). The reference population included 346,271 individuals (Figure 1); for every 100 people admitted with COVID-19 in England, 11 participants were recruited to RECOVERY. When considering geographical region, the proportion of relevant patients recruited to RECOVERY in London, West Midlands, and Yorkshire and The Humber was lower than in the other England regions (Figure 2).

Table 1 shows the baseline characteristics of both cohorts. RECOVERY participants were less frequently female (RECOVERY 37% vs reference population 45%) and were on average slightly younger than the reference population (mean age [SD]: 62.6 [15.3] vs 65.7 [18.5] years), with people aged 80+ and women underrepresented in RECOVERY throughout the analysis period (**Error! Reference source not found.** and **Error! Reference source not found.**3). RECOVERY participants were more frequently of White background (83% vs 79%) (Table 1 and **Error! Reference source not found.**), but had similar deprivation status overall and throughout the study period (**Error! Reference source not found.**).

With respect to clinical conditions, RECOVERY participants had a lower prevalence of comorbidity (median Charlson Comorbidity Score [IQR]: RECOVERY 3.0 [1.0-5.0] vs reference population 4.0 [1.0-6.0]) and were less frail (median Hospital Frailty Risk Score [IQR]: 5.1 [1.8-11.4] vs 6.3 [1.8-16.3]) These differences were largely explained by the age structure of the two cohorts, with small differences

remaining in the prevalence of some comorbidities, including cardiovascular disease, congestive heart failure, and dementia, after accounting for age (Supplementary Figures S4-S6).

Outcomes

Overall, the crude all-cause 28-day mortality in RECOVERY was 20.6% (95% CI: 20.2%-21.0%) and 24.8% (95% CI: 24.6%-25.0%) in the reference population, with mortality decreasing substantially in both cohorts from March 2021 onwards. After standardising the RECOVERY cohort to the age-sex composition of the national reference population, 28-day mortality in RECOVERY was similar to the reference population (23.7%, 95% CI: 23.3%-24.1%; **Error! Reference source not found.**). Age-stratified mortality rates were similar between the two cohorts, with the exception of March-May 2020 where mortality was lower in RECOVERY (Supplementary Figures S7-S8 and Supplementary Table S4). When mortality was assessed separately by comorbidity level and age, the difference in 28-day mortality between the two cohorts in March-May 2020 appeared to be mostly driven by older and more comorbid patients (Supplementary Figure S9).

Discussion

This study compared the characteristics of RECOVERY trial participants with people admitted to hospital due to COVID-19 in England. Our main findings were that RECOVERY participants were generally similar, but slightly younger, less frequently female, and had an overall lower comorbidity and frailty burden, much of which attributable to age differences. After adjustment for age and sex, 28-day mortality in the RECOVERY cohort was similar to that in the wider population of patients admitted to hospital with COVID-19 in England. This pattern was observed throughout the period studied, with the exception of March – May 2020 (corresponding to the first COVID-19 wave in the UK) when, even after adjusting for age and sex, 28-day mortality in RECOVERY was slightly lower than the reference population. The reasons for this are not fully explained by differences in measured frailty or comorbidity as assessed in our analyses and may be attributable to factors not captured in the datasets available in this study.

Older adults are frequently underrepresented in trials,[22] and have been excluded from over half of COVID-19 clinical trials and all major vaccine trials.[23] Although RECOVERY does not have an upper age limit (and some participants were aged over 100 years old), in our study RECOVERY participants were on average 3 years younger, with underrepresentation of people aged ≥80. RECOVERY participants were also less frequently female (37% vs 45%) but it is not possible to identify the possible reasons for this in the available data. We also found that comorbidity and frailty scores were lower in the RECOVERY cohort compared with the reference population. Most of these differences were attributable to age composition, but within older age groups comorbidities and the overall frailty risk scores remained slightly higher in the reference population. Clinical decision making about eligibility for randomised trials will inevitably result in differences between the trial cohort and the target population; however, the proportional estimates of treatment effect from trials are usually generalisable, unless there are substantial differences in the biology of the target disease or the effectiveness of the intervention in the non-trial context.[4,5]

While crude 28-day all-cause mortality was lower in RECOVERY, age- and sex-adjusted mortality were generally similar, with similar trends in both cohorts over time. The reduction seen from March 2021

onwards, consistent with previous reports,[24] may represent the effect of SARS-CoV-2 vaccination uptake, which greatly reduced the likelihood not only of hospital admission but also of death following hospitalisation.[25,26] Overall, the absolute effect estimates generated by RECOVERY were generalisable to the national population during the period studied. However, secular trends in mortality rates should be considered and the best estimate of the likely absolute effect size in current clinical practice requires application of the proportional treatment effect from the RECOVERY trial to current absolute event rates among patients hospitalised with COVID-19.[4,5]

Our study has a number of limitations. We were not able to determine baseline respiratory status (which has been shown to be an important determinant of the proportional and absolute benefits of corticosteroid treatment)[21] in our reference cohort, since there was low agreement between respiratory support status extracted from HES alone and that collected in the trial (based on a larger number of linked data sources) and used in published analyses (Annex IV). We also cannot be certain whether our reference population had clinically significant COVID-19, although we have mitigated this by including only people with a relevant ICD-10 code in the primary diagnostic position. Finally, our analysis was restricted to people admitted in England. Baseline characteristics were similar when comparing RECOVERY participants recruited across all UK nations, but may differ from non-UK countries. Finally, our analysis was restricted to the period from March 2020 to November 2021, due to changes to trial eligibility which could not be replicated in the reference population with the available data. However, recruitment to RECOVERY declined significantly from December 2021 onwards (along with national COVID-19 admissions), so that extending the analysis period to the time of writing (mid-2023) would add only a small number of additional deaths (~4%), which were unlikely to meaningfully influence interpretation of our results.

Conclusion

The RECOVERY trial recruited a broad patient population that was generally representative of people admitted to hospital due to COVID-19 in England during the same period, with respect to both baseline

characteristics and subsequent mortality. 28-day mortality declined substantially in both the RECOVERY and reference populations throughout the period studied. Estimates of current mortality rates from healthcare systems data combined with the proportional treatment effects from trials are needed to estimate the likely absolute effects of the treatments tested within current practice.

Declarations

Data availability

The RECOVERY trial protocol, consent form, statistical analysis plan, definition and derivation of clinical characteristics and outcomes, training materials, regulatory documents, and other relevant study materials are available online at www.recoverytrial.net. Data will be made available in line with the Nuffield Department of Population Health policy and procedures. Those wishing to request access should complete the form available at http://www.ndph.ox.ac.uk/data-access and email it to data.access@ndph.ox.ac.uk. Nationwide anonymised English mortality (Civil Registrations) and hospitalisations (HES) data used to derive the reference population can be obtained upon application to NHS England at www.digital.nhs.uk. The statistical programming code used in this work is available for inspection and reuse at http://gitlab.ndph.ox.ac.uk/guilhermep/recovery-generalizability-representativeness.

Author contributions

This manuscript was initially drafted by GPA, RG, MC, CH, and MM, and further developed and approved by all authors. GPA, RG, MC, CH, and MM conceptualised and designed this analysis. GPA and RG performed data derivation and analysis. GPA performed data visualisation. GPA, RG, CC, WS, AK, DM, RW, MN, KW, and MC contributed to data acquisition, management, and processing, together with the broader RECOVERY Collaborative Group. All authors contributed to data interpretation and critical review and revision of the manuscript. MM is the guarantor for this study.

Funding

This work was supported by grants to the University of Oxford from UK Research and Innovation and NIHR (MC_PC_19056), the Wellcome Trust (grant reference 222406/Z/20/Z) through the COVID-19 Therapeutics Accelerator, the Oxford BHF Centre of Research Excellence (grant reference RE/18/3/34214), and Health Data Research UK, and by core funding provided by the NIHR Oxford

Biomedical Research Centre, the Wellcome Trust, the Bill and Melinda Gates Foundation, the Foreign, Commonwealth and Development Office, the Medical Research Council Population Health Research Unit, the NIHR Health Protection Unit in Emerging and Zoonotic Infections and NIHR Clinical Trials Unit Support Funding; and by PhD funding from the Medical Research Council Population Health Research Unit at the University of Oxford (to GPA and MC) and by a PhD studentship (grant reference MR/L004933/2) funded by the MRC Network of Hubs for Trials Methodology Research (HTMR) (to CH). For the purpose of open access, the author(s) has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

Acknowledgements

The authors wish to thank the many thousands of RECOVERY participants, as well as the doctors, nurses, pharmacists, other allied health professionals, and research administrators at National Health Service hospital organisations across the whole of the UK, supported by staff at the National Institute for Health and Care Research (NIHR) Clinical Research Network, NHS DigiTrials, Public Health England, Department of Health & Social Care, the Intensive Care National Audit & Research Centre, Public Health Scotland, National Records Service of Scotland, the Secure Anonymised Information Linkage at the University of Swansea, and the NHS in England, Scotland, Wales, and Northern Ireland. This work uses data provided by patients and collected by the NHS as part of their care and support.

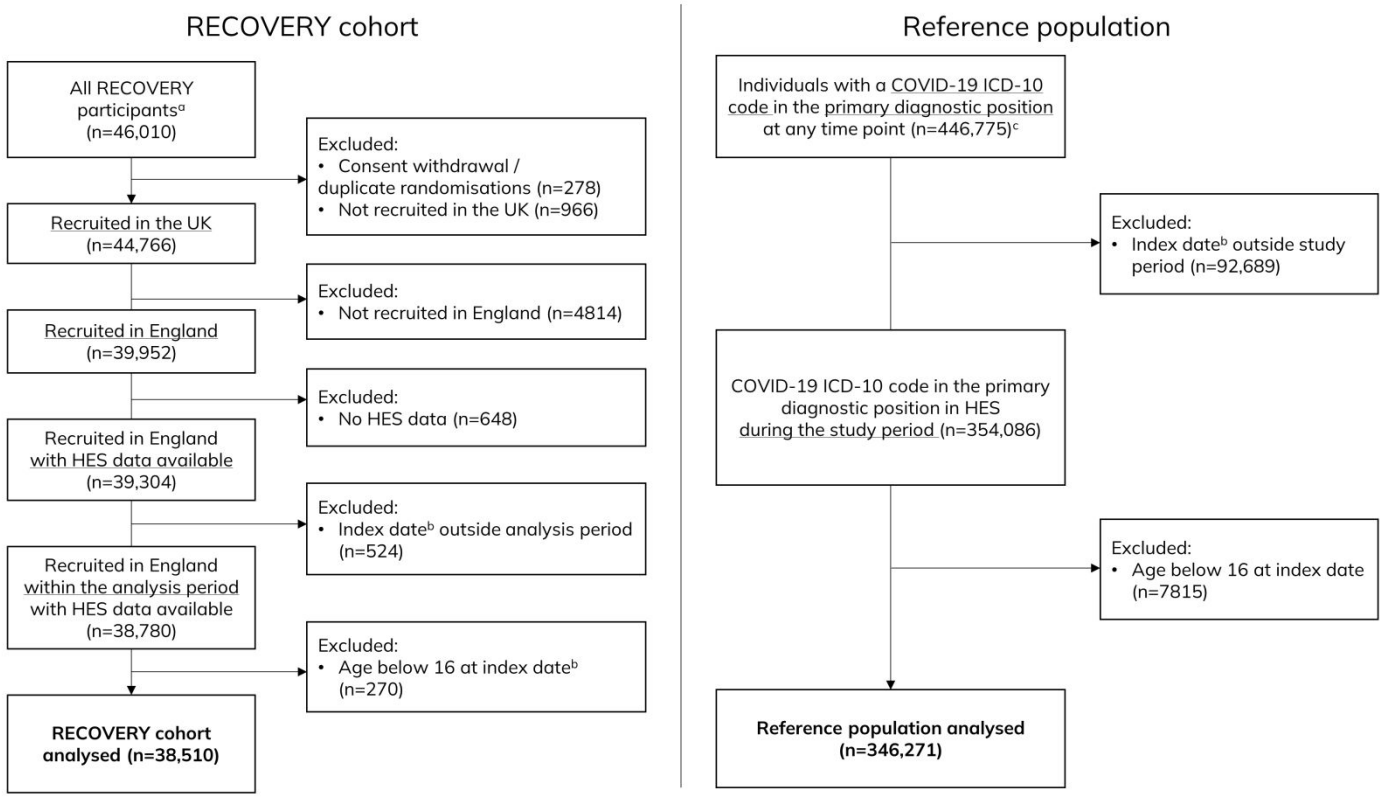
Conflict of interest

Roche, AbbVie, Regeneron, and GSK have provided study drugs for evaluation in the RECOVERY trial. MM is an applicant on research grants from Novartis and Novo Nordisk (unrelated to this work). ML is in receipt of grants to University of Oxford from Novartis and Boehringer Ingelheim and grants to Protas from Regeneron, Sanofi, Moderna, FluLab, Google Ventures and Schmidt Futures (all unrelated to this work).

1

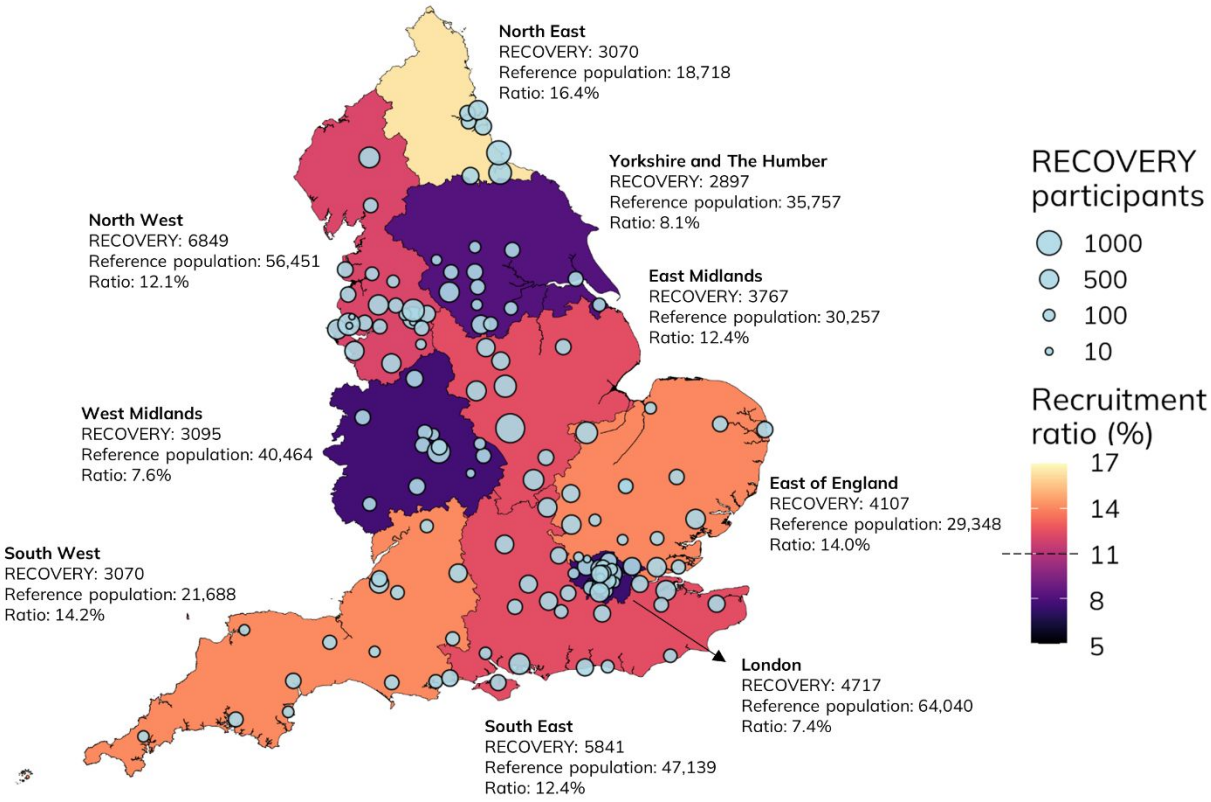
Figures

Figure 1 - CONSORT diagram depicting the cohort derivation process



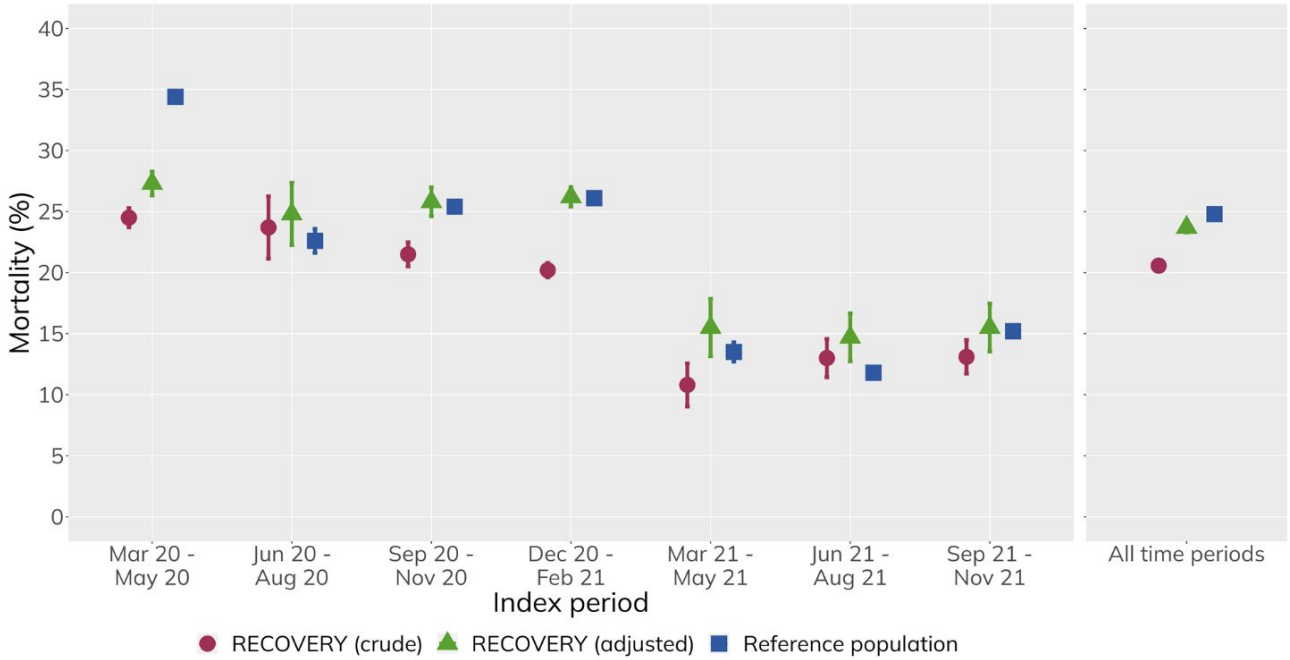
[a] Randomised up until the 1st September 2022; [b] Index date is the episode start date for the earliest episode with a COVID-19 ICD-10 code in the primary diagnostic position; [c] Up to June 2022 (latest data included in the raw extract)

Figure 2 - Geographical representativeness of the RECOVERY trial cohort in comparison with the national reference population



Number of RECOVERY participants plotted at the location of the recruiting NHS Trust hospital site. Recruitment ratios were calculated by dividing the number of RECOVERY participants recruited in each region by the number of individuals in the reference population in the same region, and are presented by region. The average recruitment ratio across all English regions was 11.1%. There were 1097 and 2409 individuals with missing residential area in HES data in the RECOVERY and the reference population cohort, respectively.

Figure 3 - All-cause 28-day mortality over time in RECOVERY and the reference population



Period		Mar 20 – May 20	Jun 20 – Aug 20	Sep 20 – Nov 20	Dec 20 – Feb 21	Mar 21 –May 21	Jun 21 – Aug 21	Sep 21 – Nov 21	All time periods
Number of individuals	RECOVERY	9634	1002	6208	16,448	1160	1905	2153	38,510
	Reference population	78,180	6,939	49,431	127,750	9,319	32,464	42,188	346,271
Recruitment ratio (%)		12.3	14.4	12.6	12.9	12.4	5.9	5.1	11.1
Deaths	RECOVERY	2364	237	1335	3330	125	248	283	7922
	Reference population	26,855	1570	12,557	33,361	1255	3841	6405	85,844
28-day mortality (95% CI)	RECOVERY, crude	24.5 (23.7 - 5.3)	23.7 (21.2 -26.2)	21.5 (20.5 -22.5)	20.2 (19.6 -20.8)	10.8 (9.0 -12.6)	13.0 (11.4 -14.6)	13.1 (11.7 -14.5)	20.6 (20.2 -21.0)
	RECOVERY, adjusted	27.3 (26.3 -28.3)	24.8 (22.3 -27.3)	25.8 (24.6 -27.0)	26.2 (25.4 -27.0)	15.5 (13.1 -17.9)	14.7 (12.7 -16.7)	15.4 (13.4 -17.4)	23.7 (23.3 -24.1)
	Reference population	34.4 (34.0 - 34.8)	22.6 (21.6 - 23.6)	25.4 (25.0 - 25.8)	26.1 (25.9 - 26.3)	13.5 (12.7 - 14.3)	11.8 (11.4 - 12.2)	15.2 (14.8 - 15.6)	24.8 (24.6 – 25.0)

28-day mortality is the proportion of people with death recorded within 28 days of their index date (with 95% confidence intervals included). Adjustment performed by applying RECOVERY 28-day mortality to an age- (5-year bands) and sex-standardised population using the reference population, in a rolling basis within each time period (for 28-day mortality and age and sex breakdown).

Tables

Table 1 - Baseline cohort characteristics

Characteristic	RECOVERY, N = 38,510	Reference population, N = 346,271
Age, mean (SD)	62.6 (15.3)	65.7 (18.5)
<60	16,121 (41.9%)	123,790 (35.7%)
60-69	8,906 (23.1%)	56,452 (16.3%)
70-79	7,871 (20.4%)	69,107 (20.0%)
80+	5,612 (14.6%)	96,922 (28.0%)
Sex		
Female	14,060 (36.5%)	155,441 (44.9%)
Male	24,424 (63.5%)	190,748 (55.1%)
Geographical region[a]		
London	4717 (12.2%)	64,040 (18.5%)
North West	6849 (17.8%)	56,451 (16.3%)
South East	5841 (15.2%)	47,139 (13.6%)
West Midlands	3095 (8%)	40,464 (11.7%)
Yorkshire and The Humber	2897 (7.5%)	35,757 (10.3%)
East Midlands	3767 (9.8%)	30,257 (8.7%)
East of England	4107 (10.7%)	29,348 (8.5%)
South West	3070 (8%)	21,688 (6.3%)
North East	3070 (8%)	18,718 (5.4%)
Unknown/not resident in England	1097 (2.9%)	2409 (0.7%)
Ethnicity[a]		
White	29,595 (83.3%)	253,842 (78.9%)
Black	1,171 (3.3%)	16,909 (5.3%)
Asian	3,263 (9.2%)	35,785 (11.1%)
Other	1,146 (3.2%)	11,853 (3.7%)
Mixed	351 (1.0%)	3,532 (1.1%)
Unknown	2,984 (7.7%)	24,350 (7.0%)
Index of multiple deprivation (quintile)		
1 (Most deprived)	9,821 (25.5%)	94,487 (27.3%)
2	8,284 (21.5%)	78,400 (22.6%)

3	7,466 (19.4%)	65,082 (18.8%)
4	6,910 (17.9%)	57,203 (16.5%)
5 (Least deprived)	5,797 (15.1%)	48,650 (14.0%)
Unknown	232 (0.6%)	2,449 (0.7%)
Charlson score, median (IQR)	3.0 (1.0, 5.0)	4.0 (1.0, 6.0)
Myocardial infarction	2,941 (7.6%)	34,895 (10.1%)
Congestive heart failure	3,158 (8.2%)	44,007 (12.7%)
Peripheral vascular disease	2,052 (5.3%)	24,327 (7.0%)
Cerebrovascular disease	2,603 (6.8%)	41,812 (12.1%)
Chronic pulmonary disease	8,160 (21.2%)	80,492 (23.2%)
Rheumatic disease	1,579 (4.1%)	17,365 (5.0%)
Dementia	1,234 (3.2%)	30,314 (8.8%)
Peptic ulcer disease	710 (1.8%)	7,693 (2.2%)
Liver disease (mild)	1,515 (3.9%)	14,674 (4.2%)
Liver disease (moderate-severe)	175 (0.5%)	2,490 (0.7%)
Diabetes mellitus (without chronic complications)	6,056 (15.7%)	59,244 (17.1%)
Diabetes mellitus (with chronic complications)	1,354 (3.5%)	16,111 (4.7%)
Chronic kidney disease	3,800 (9.9%)	54,019 (15.6%)
Solid tumour	2,463 (6.4%)	26,844 (7.8%)
Metastatic cancer	589 (1.5%)	8,287 (2.4%)
Lymphoma	389 (1.0%)	3,238 (0.9%)
Leukaemia	320 (0.8%)	2,775 (0.8%)
AIDS/HIV[b]	0 (0.0%)	0 (0.0%)
Hospital frailty score, median (IQR)	5.1 (1.8, 11.4)	6.3 (1.8, 16.3)
High-risk (>15)	6,737 (17.5%)	94,191 (27.2%)
Intermediate risk (5-15)	12,751 (33.1%)	99,402 (28.7%)
Low risk (<5)	19,022 (49.4%)	152,678 (44.1%)
Other comorbidities/demographics		
Renal replacement therapy	439 (1.1%)	4,922 (1.4%)
Immunosuppression	1,471 (3.8%)	15,531 (4.5%)
Obesity	6,147 (16.0%)	48,749 (14.1%)
Severe mental illness	4,268 (11.1%)	43,969 (12.7%)
Alcohol-attributable diseases	945 (2.5%)	10,909 (3.2%)

HES - Hospital Episode Statistics; IQR - interquartile range; SD - standard deviation

Data are shown as Mean (SD); n (%); or Median (IQR)

[a]Proportions for people with known and unknown geographical region and ethnicity were calculated separately, using the entire cohort as denominator for each calculation

[b]ICD-10 codes for AIDS/HIV are censored from HES data

References

1. Concato J, Shah N, Horwitz RI. Randomized, Controlled Trials, Observational Studies, and the Hierarchy of Research Designs. *N Engl J Med*. 2000 Jun 22;**342**(25):1887–1892.

2. Good Clinical Trials Collaborative. Good Clinical Trials Collaborative. Guidelines for Good Randomized Clinical Trials. [Internet]. 2023 [cited 2023 Apr 25]. Available from: www.goodtrials.org

3. Collins R, Reith C, Emberson J, et al. Interpretation of the evidence for the efficacy and safety of statin therapy. *The Lancet*. 2016 Nov;**388**(10059):2532–2561.

4. MacMahon S, Collins R. Reliable assessment of the effects of treatment on mortality and major morbidity, II: observational studies. *The Lancet*. 2001 Feb;**357**(9254):455–462.

5. Collins R, MacMahon S. Reliable assessment of the effects of treatment on mortality and major morbidity, I: clinical trials. *The Lancet*. 2001 Feb;**357**(9253):373–380.

6. Pessoa-Amorim G, Campbell M, Fletcher L, et al. Making trials part of good clinical care: lessons from the RECOVERY trial. *Future Healthc J*. 2021 Jul;**8**(2):e243–e250.

7. Herbert A, Wijlaars L, Zylbersztejn A, Cromwell D, Hardelid P. Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC). *Int J Epidemiol*. 2017 Aug 1;**46**(4):1093–1093i.

8. NHS Digital. Civil Registration – Deaths [Internet]. 2022 [cited 2022 Oct 11]. Available from: https://digital.nhs.uk/services/data-access-request-service-dars/dars-products-and-services/data-set-catalogue/civil-registration-deaths

9. NHS Digital. National Clinical Coding Standards ICD-10 [Internet]. 5th ed. 2021. Available from: https://classbrowser.nhs.uk/ref_books/ICD-10_2021_5th_Ed_NCCS.pdf

10. NHS Digital. NHS Digital [Internet]. [cited 2022 Oct 11]. Available from: https://digital.nhs.uk/

11. Big Data Institute, University of Oxford. Unit of Health Care Epidemiology [Internet]. 2022 [cited 2022 Oct 11]. Available from: https://www.bdi.ox.ac.uk/research/unit-of-health-care-epidemiology

12. World Health Organization. Emergency use ICD codes for COVID-19 disease outbreak (who.int) [Internet]. 2021 [cited 2022 Oct 11]. Available from: https://www.who.int/standards/classifications/classification-of-diseases/emergency-use-icd-codes-for-covid-19-disease-outbreak

13. NHS England. COVID-19 Second Generation Surveillance System (SGSS) [Internet]. 2023 [cited 2023 Apr 25]. Available from: https://digital.nhs.uk/services/data-access-request-service-dars/dars-products-and-services/data-set-catalogue/covid-19-second-generation-surveillance-system-sgss

14. RECOVERY Collaborative Group. Higher dose corticosteroids in patients admitted to hospital with COVID-19 who are hypoxic but not requiring ventilatory support (RECOVERY): a randomised, controlled, open-label, platform trial. *The Lancet*. 2023 Apr;S014067362300510X.

15. Ministry of Housing, Communities & Local Government. English indices of deprivation 2019 [Internet]. 2019 [cited 2022 Oct 11]. Available from: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019

16. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis*. 1987;**40**(5):373–383.

17. Quan H, Sundararajan V, Halfon P, et al. Coding Algorithms for Defining Comorbidities in ICD-9-CM and ICD-10 Administrative Data: *Med Care*. 2005 Nov;**43**(11):1130–1139.

18. Gilbert T, Neuburger J, Kraindler J, et al. Development and validation of a Hospital Frailty Risk Score focusing on older people in acute care settings using electronic hospital records: an observational study. *Lancet Lond Engl*. 2018 May 5;**391**(10132):1775–1782.

19. NHS Digital. A Guide to Linked Mortality Data from Hospital Episode Statistics and the Office for National Statistics [Internet]. 2015 [cited 2022 Dec 6]. Available from: https://nhs-prod.global.ssl.fastly.net/binaries/content/assets/website-assets/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics/hes-ons_linked_mortality_data_guide.pdf

20. Kirkwood BR, Sterne JAC. Essential Medical Statistics. 2nd ed. Malden, Massachusetts, USA: Blackwell Science; 2003.

21. RECOVERY Collaborative Group, Horby P, Lim WS, et al. Dexamethasone in Hospitalized Patients with Covid-19. *N Engl J Med*. 2021 Feb 25;**384**(8):693–704.

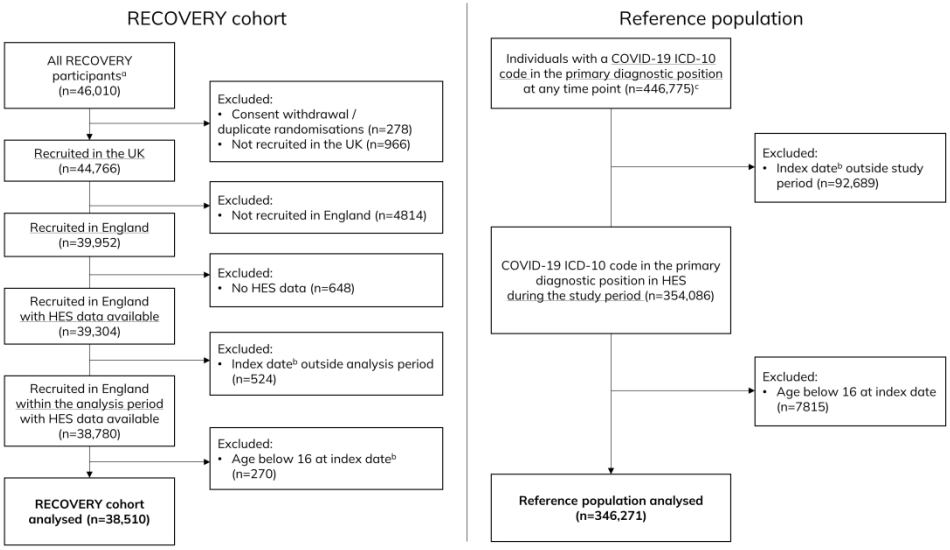
22. Marum RJ. Underrepresentation of the elderly in clinical trials, time for action. *Br J Clin Pharmacol*. 2020 Oct;**86**(10):2014–2016.

23. Helfand BKI, Webb M, Gartaganis SL, Fuller L, Kwon C-S, Inouye SK. The Exclusion of Older Persons From Vaccine and Treatment Trials for Coronavirus Disease 2019—Missing the Target. *JAMA Intern Med*. 2020 Nov 1;**180**(11):1546.

24. Gray WK, Navaratnam AV, Day J, Wendon J, Briggs TWR. COVID-19 hospital activity and in-hospital mortality during the first and second waves of the pandemic in England: an observational study. *Thorax*. 2021 Nov 24;thoraxjnl-2021-218025.

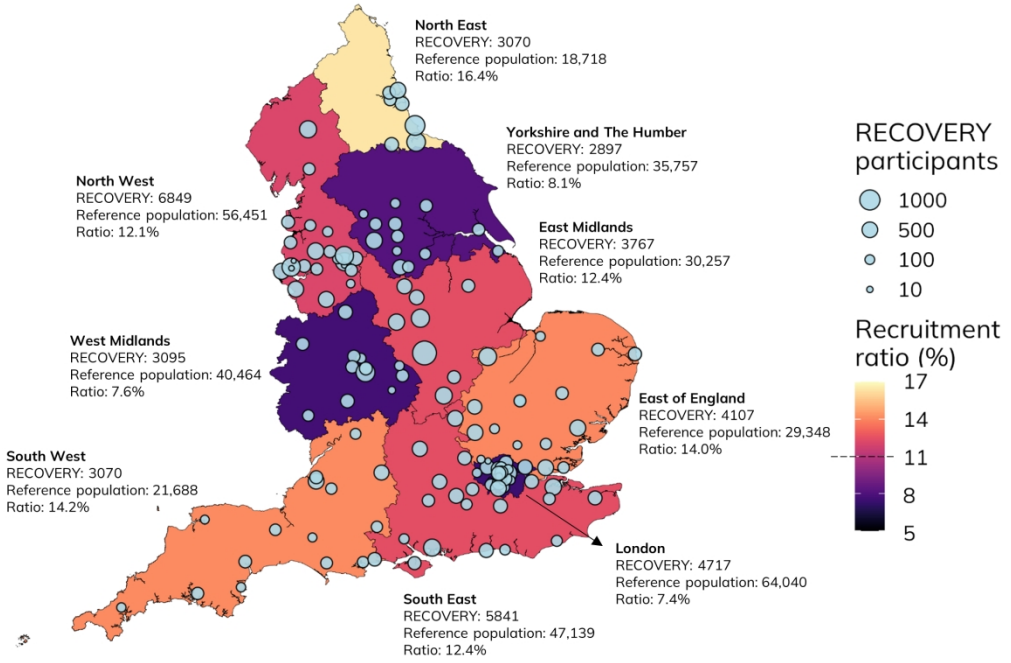
25. Au WY, Cheung PP-H. Effectiveness of heterologous and homologous covid-19 vaccine regimens: living systematic review with network meta-analysis. *BMJ*. 2022 May 31;**377**:e069989.

26. Lopez Bernal J, Andrews N, Gower C, et al. Effectiveness of the Pfizer-BioNTech and Oxford-AstraZeneca vaccines on covid-19 related symptoms, hospital admissions, and mortality in older adults in England: test negative case-control study. *BMJ*. 2021 May 13;n1088.



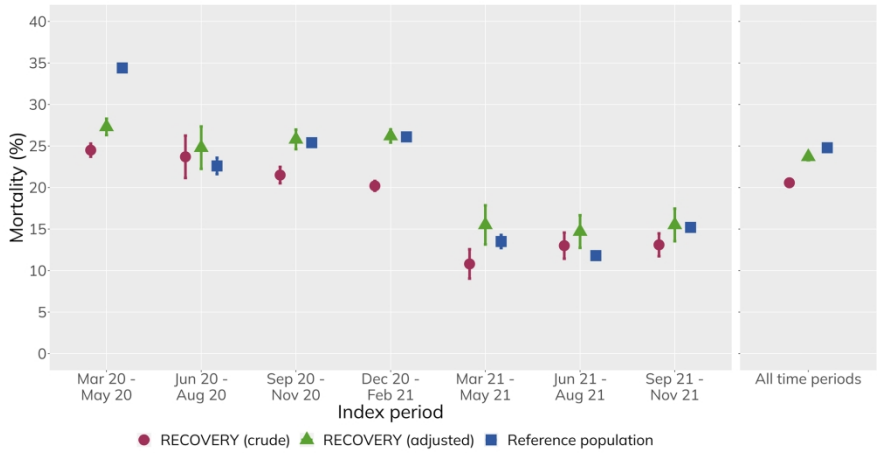
[a] Randomised up until the 1st September 2022; [b] Index date is the episode start date for the earliest episode with a COVID-19 ICD-10 code in the primary diagnostic position; [c] Up to June 2022 (latest data included in the raw extract)

861x484mm (118 x 118 DPI)



Number of RECOVERY participants plotted at the location of the recruiting NHS Trust hospital site. Recruitment ratios were calculated by dividing the number of RECOVERY participants recruited in each region by the number of individuals in the reference population in the same region, and are presented by region. The average recruitment ratio across all English regions was 11.1%. There were 1097 and 2409 individuals with missing residential area in HES data in the RECOVERY and the reference population cohort, respectively.

508x359mm (118 x 118 DPI)



Period		Mar 20 – May 20	Jun 20 – Aug 20	Sep 20 – Nov 20	Dec 20 – Feb 21	Mar 21 –May 21	Jun 21 – Aug 21	Sep 21 – Nov 21	All time periods
Number of individuals	RECOVERY	9634	1002	6208	16,448	1160	1905	2153	38,510
	Reference population	78,180	6,939	49,431	127,750	9,319	32,464	42,188	346,271
Recruitment ratio (%)		12.3	14.4	12.6	12.9	12.4	5.9	5.1	11.1
Deaths	RECOVERY	2364	237	1335	3330	125	248	283	7922
	Reference population	26,855	1570	12,557	33,361	1255	3841	6405	85,844
28-day mortality (95% CI)	RECOVERY, crude	24.5 (23.7 - 5.3)	23.7 (21.2 -26.2)	21.5 (20.5 -22.5)	20.2 (19.6 -20.8)	10.8 (9.0 -12.6)	13.0 (11.4 -14.6)	13.1 (11.7 -14.5)	20.6 (20.2 -21.0)
	RECOVERY, adjusted	27.3 (26.3 -28.3)	24.8 (22.3 -27.3)	25.8 (24.6 -27.0)	26.2 (25.4 -27.0)	15.5 (13.1 -17.9)	14.7 (12.7 -16.7)	15.4 (13.4 -17.4)	23.7 (23.3 -24.1)
	Reference population	34.4 (34.0 - 34.8)	22.6 (21.6 - 23.6)	25.4 (25.0 - 25.8)	26.1 (25.9 - 26.3)	13.5 (12.7 - 14.3)	11.8 (11.4 - 12.2)	15.2 (14.8 - 15.6)	24.8 (24.6 – 25.0)

28-day mortality is the proportion of people with death recorded within 28 days of their index date (with 95% confidence intervals included). Adjustment performed by applying RECOVERY 28-day mortality to an age- (5-year bands) and sex-standardised population using the reference population, in a rolling basis within each time period (for 28-day mortality and age and sex breakdown).

483x330mm (118 x 118 DPI)

Supplementary data

Annex I – Supplementary methods

ANALYSIS POPULATIONS AND TIMEFRAMES

1. RECOVERY cohort
    - All RECOVERY participants recruited in England for whom linked HES data is available (regardless of the presence of COVID codes)

2. Reference population
    - All individuals in England with a relevant COVID-19 ICD-10 code in the primary diagnostic position (U071 or U072), in any episode during the study period

Data linkage for RECOVERY was performed by NHS Digital using a unique patient identifier (NHS number) along with name, gender, and date-of-birth where necessary. Data linkage (between sources) for the reference population was also performed by NHS Digital, and data were provided in an anonymised fashion.

We did not take COVID-19 ICD-10 coding characteristics into consideration when defining the RECOVERY cohort (i.e. matching these to the rules used to derive the reference population) as the purpose of this analysis was not to create perfectly matched populations, but rather highlight any potential differences between the RECOVERY cohort and a reference population.

An index date was defined for each individual as the *epistart* (in HES) for the first episode with an ICD-10 COVID-19 code (U071/U072) in the first diagnostic position in HES.

The analysis populations were restricted to individuals with an index date before the 1st December 2021 (so that a 28 day follow-up period will have been completed by the end of December 2021). This allowed at least a two-month interval between reception of linked datasets in the first quarter of 2022 and the end of the follow-up period. Mortality was assessed within 28 days after the index date, with

randomisation counted as day 1 and including day 28 (i.e. last index date is 30/11/21 and last day for outcome assessments is 27/12/21, inclusive).

We used the presence of a single COVID-19 ICD-10 code to define the reference population, following preliminary work exploring cross-coding of ICD-10 codes (in HES) and SARS-CoV-2 testing (from the Second Generation Surveillance System – SGSS data) in the RECOVERY population - which may be interpreted as a population with a proven or suspected clinical diagnosis of COVID-19. This work showed high agreement between COVID-19 coding in HES and a positive COVID-19 test in SGSS (see Annex IV)

No sampling or matching method was used to define the reference population as the purpose of this study is to compare how representative the RECOVERY population is in comparison with the national population (sampling would reduce these differences). However, a quantitative assessment of the generalisability of the trial results was performed by presenting age- and sex-adjusted all-cause 28 day mortality (i.e. the RECOVERY trial primary outcome), overall and over time.

Due to ethical considerations surrounding the ongoing collection of linkage data on people below the age of 16 in RECOVERY, we excluded individuals aged below 16 at the index date from all cohorts.

DATA SOURCES AND DATA CUTS

1. RECOVERY:
    a. Hospital Episode Statistics + Civil Registrations data (RECOVERY extract 79, received September 2022)

2. Reference population:
    a. Hospital Episode Statistics + Civil Registrations data - these data are part of an existing anonymised data flow provided to the University of Oxford by NHS Digital for public health research and held by the Unit of Healthcare Epidemiology at the University of Oxford (reference: DARS-NIC-315419-F3W7K); see the NHS Digital data uses register at https://digital.nhs.uk/services/data-access-request-service-dars/data-uses-register?msclkid=480bee5ab0e111ec99ed4c48b4e33bc8 for more details on the data used

BASELINE CHARACTERISTICS

1. Clinical conditions and general demographics

Clinical conditions and demographics were derived from HES data. When comparing RECOVERY participants recruited in England vs other nations, and those with available HES data vs those without, we also used CRF recorded date. These conditions were defined based on recorded answers in the baseline CRF data, or the presence/absence of a relevant clinical code recorded within 5 years before the index date in HES. Respiratory status at baseline for RECOVERY trial participants was derived using a combination of data sources as described in the RECOVERY trial website (http://www.recoverytrial.net/results)

The exceptions to these rules are the following:

a. Immunosuppression, defined based on the presence of any of the COVID-19 Greenbook Immunisation Criteria[1] as specified below:

- Cancer undergoing active chemotherapy: Cancer diagnosis ICD-10 code in any diagnostic position in the 5 years prior to randomisation AND chemotherapy ICD-10 or OPCS-4 code in any diagnostic position in the 6 months prior to randomisation.
- Haematological malignancy or bone marrow transplant: Haematological malignancy or transplant ICD-10 or OPCS-4 code in any diagnostic position in the 2 years prior to randomisation
- Solid organ transplant: solid organ transplant ICD-10 or OPCS-4 code in any diagnostic position in the 5 years prior to randomisation
- Hyposplenism: hyposplenism ICD-10 or splenectomy OPCS-4 code in any diagnostic position in the 5 years prior to randomisation
- Other immunosuppression: Other long-term immunosuppression ICD-10 in any diagnostic position in the 5 years prior to randomisation or other short-term immunosuppression ICD-10 or OPCS-4 code in any diagnostic position in the 6 months prior to randomisation

b. Renal replacement therapy, defined based on the presence of any of the criteria below in HES data only (using an adaptation of a previously-published algorithm):[2]

- Peritoneal dialysis: occurrence of any admission with a peritoneal dialysis code (without diagnosis of acute kidney injury) in the 5 years before index date (epistart of the index spell)
- Maintenance haemodialysis: Occurrence of a dialysis code in the 5 years before index date (*epistart* of the index spell) of the index spell in a patient who has had:
  (a) a diagnostic code for ESKD any prior time, or within 365 days; OR
  (b) the insertion of an arteriovenous fistula or graft any prior time, or within 365 days.
- Probable maintenance haemodialysis: Occurrence of at least two episodes containing a dialysis code, with at least 90 days between the start of the first recorded dialysis, and the start of any subsequent dialysis (without diagnosis of acute kidney injury) in the 5 years before index date (*epistart* of the index spell)

c. **Charlson Comorbidity Score (CCS):** scoring was derived using previously-published methodology[3]

d. **Hospital Frailty Risk Score:** the score wase assessed using HES data only, and only records with *epistart* in the 2 years preceding index date (following the published methodology);[4] scoring of each item will be based on the same paper. NB: although the score was developed and validated in a population of over 75 years, it will be assessed here in the entire HES cohorts regardless of age

e. **Ethnicity:** ethnicity was derived in line with methodology used in the main RECOVERY publications (http://www.recoverytrial.net/results) and categorised according to the UK Department of Health categories (http://www.ethnicity-facts-figures.service.gov.uk/style-guide/ethnic-groups); in short, we used HES data only and selected the most frequent code in the ethnicity field. For comparisons of RECOVERY participants recruited in England versus other UK nations, we used all available data sources (which include HES and primary care data from the General Practice Extraction Service Data for Pandemic Planning and Research [GDPPR] in England, and datasets similar to HES in Scotland and Wales; no linkage data was available for Northern Ireland). In GDPPR, we selected the most frequent ethnicity SNOMED code in the journals table was used where available; if not, the most frequent record in the *ethnic* field in the patients table was used; for HES, the most frequent code in the ethnicity field was used as described above. All records were considered regardless of when they were recorded. If there was a tie, ethnic groups were prioritised in the following order (based on ascending recording frequency in the overall population): Mixed, Other, Black, Asian, White. Where there was a discrepancy between the best estimate from GDPPR and HES, the GDPPR group was prioritised.

2. Codelist derivation

Where possible, we used publicly-available clinical codelists to derive baseline characteristics using ICD-10 codes. For the remaining characteristics, original codelists were generated by one clinician by manually searching the ICD-10 terminology and reviewed by a second clinician. ICD-10 codelists to derive CCS,[3] HFS,[4] renal-replacement therapy,[2] alcohol-attributable diseases,[5] diabetes,[6] severe mental illness,[7] and obesity[8] were taken from previously-published reports. Chronic heart disease was extracted with ICD-10 codes used to assign cardiac cause of death in the RECOVERY trial. Chronic liver disease was defined by merging codelists for mild and moderate-to-severe liver disease components of the Charlson score. The table below provides details on each HES-derived condition. All codelists used in this project are available for reuse and inspection at http://gitlab.ndph.ox.ac.uk/guilhermep/recovery-generalizability-representativeness/Tools.

Hospital Episode Statistics derivation methodology and codelists

Note – the HES APC data dictionary is available at http://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics/hospital-episode-statistics-data-dictionary#download-hes-data-dictionaries

Characteristic	HES field	ICD-10 / OPCS-4 codes (NB: all codes are ICD-10 unless otherwise specified)	Derivation rule	Notes	References
Age	admiage	N/A	Value recorded in index spell		HES APC data dictionary
Gender	sex	N/A	Value recorded in index spell		HES APC data dictionary
Ethnicity	ethnos	N/A	Value recorded in index spell		HES APC data dictionary
COVID-19 coding	diag_01 to diag_20 (any position)	U071, U072	If present in index spell (NB: the presence of these codes is used to define the national HES cohort; for	WHO documentation	https://www.who.int/standards/classifications/classification-of-diseases/emergency-

			the RECOVERY HES cohort fwe have used presence of a COVID-19 code, or else date of randomisation)		use-icd-codes-for-covid-19-disease-outbreak
Index of multiple deprivation (score)	lsoa11	N/A	Value recorded in index spell	Derived from the lsoa11 (lower-super-output-area) field rather than IMD as IMD provided in HES at the moment refers to the 2010 indices (most recent are 2019)	https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019
Index of multiple deprivation (quintiles)	lsoa11	N/A	Value recorded in index spell		
Charlson score (aggregated)	diag_01 to diag_20	(see individual parameters)	Any record in the 5 years before index date (epistart	Charlson score calculated from	10.1097/01.mlr.0000182534.19832.83

Myocardial infarction (Charlson score)	(any position)	I21, I22, I25.2	of the index spell)	ICD-10 codes using published and validated methodology and codelists	
Congestive heart failure (Charlson score)		I09.0, I11.0, I13.0, I13.2, I25.5, I42.0, I42.5, I42.6, I42.7, I42.8, I42.9, I43, I50, P29.0			
Peripheral vascular disease (Charlson score)		I70, I71, I73.1, I73.8, I73.9, I77.1, I79.0, I79.2, K55.1, K55.8, K55.9, Z95.8, Z95.9			
Cerebrovascular disease (Charlson score)		G45			
Chronic pulmonary disease (Charlson score)		I27.8, I27.9, J40, J41, J42, J43, J44, J45, J46, J47, J60, J61, J62, J63, J64, J65, J66, J67, J68.4, J70.1, J70.3			
Rheumatic disease (Charlson score)		M05, M06, M31.5, M32, M33, M34, M35.1, M35.3, M36.0			

Dementia (Charlson score)		F00, F01, F02, F03, F05.1, G30, G31.1			
Peptic ulcer disease (Charlson score)		K25, K26, K27, K28			
Liver disease, mild (Charlson score)		B18, K70.0, K70.1, K70.2, K70.3, K70.9, K71.3, K71.4, K71.5, K71.7, K73, K74, K76.0, K76.2, K76.3, K76.4, K76.8, K76.9, Z94.4			
Liver disease, moderate-severe (Charlson score)		I85.0, I85.9, I86.4, I98.2, K70.4, K71.1, K72.9, K76.5, K76.6, K76.7			
Diabetes mellitus, with chronic complications (Charlson score)		E10.2, E10.3, E10.4, E10.5, E10.7, E11.2, E11.3, E11.4, E11.5, E11.7, E12.2, E12.3, E12.4, E12.5, E12.7, E13.2,			

		E13.3, E13.4, E13.5, E13.7, E14.2, E14.3, E14.4, E14.5, E14.7			
Diabetes mellitus, without chronic complications (Charlson score)		E10.0, E10.1, E10.6, E10.8, E10.9, E11.0,E11.1, E11.6, E11.8, E11.9, E12.0, E12.1, E12.6, E12.8, E12.9, E13.0, E13.1, E13.6, E13.8, E13.9, E14.0, E14.1, E14.6, E14.8, E14.9			
Renal disease (Charlson score)		I21.0, I13.1, N03.2, N03.3, N03.4, N03.5, N03.6, N03.7, N05.2, N05.3, N05.4, N05.5, N05.6, N05.7, N18, N19, N25.0, Z49.0, Z49.1, Z49.2, Z94.0, Z99.2			

Any solid tumour, excluding malignant neoplasm of skin (Charlson score)		C00-C26, C30-C34, C37-C41, C43, C45-C58, C60—C76, C88.2, C88.7, C88.9, C90.0, C90.2, C90.3, C96-C97			
Metastatic solid tumour (Charlson score)		C77-C80			
Hemiplegia or paraplegia (Charlson score)		G04.1, G11.4, G80.1, G80.2, G81, G82, G83.0, G83.1, G83.2, G83.3, G83.4, G83.9			
AIDS /HIV (Charlson score)		B20, B21, B22, B24			
Lymphoma (Charlson score)		B21.1, C81-C86, C88.0, C88.3, C88.4		Codelists built based on clinical review of the ICD-10 terminology by two clinicians (no publicly available	Bespoke
Leukemia (Charlson score)		C90.1, C91-C95, M36.1, D47.1, N16.1, D47.5			

				codelists for these criteria)	
Heart disease (any)	diag_01 to diag_20 (any position)	I08, I09, I11, I13, I20, I21, I22, I23, I24, I25, I27.1, I27.8, I27.9, I30, I31, I32, I33, I34, I35, I36, I37, I38, I39, I40, I41, I42, I43, I44, I45, I46, I47, I48, I49, I50, I51, I52	Any record in the 5 years before index date (epistart of the index spell)	RECOVERY trial cause-of-death derivation ("cardiac" category)	https://www.recoverytrial.net/results
Diabetes (any)		E10-E14		Merged both Charlson diabetes codelists (codes above), manual review of the ICD-10 terminology, and literature review	https://bmjopen.bmj.com/content/6/8/e009952.long

Chronic liver disease		B18, K70.0, K70.1, K70.2, K70.3, K70.9, K71.3, K71.4, K71.5, K71.7, K73, K74, K76.0, K76.2, K76.3, K76.4, K76.8, K76.9, Z94.4, I85.0, I85.9, I86.4, I98.2, K70.4, K71.1, K72.9, K76.5, K76.6, K76.7		Merged mild and moderate-severe liver disease lists	N/A
Severe mental illness		F20, F21, F22, F23, F24, F25, F28, F29, F30, F31, F32.2, F32.3 F32.8, F32.9, F33.2, F33.3 F33.4, F33.8, F33.9		Publicly-available codelist used for the QCOVID NHS risk calculator	https://www.datadictionary.nhs.uk/Covid19PRA/Severe_Mental_Illness.html
Alcohol-attributable diseases		F10.3-F10.9, F10.0, F10.1, F10.2, G62.1, G31.2, G72.1, I42.6, K29.2, K70.0-K70.4, K70.9, K85.2, K86.0, Q86.0,		Publicly-available codelist published by the US CDC	https://www.cdc.gov/alcohol/ardi/alcohol-related-icd-codes.html

		P04.3			
Obesity		E66	Any record in the 5 years before index date (epistart of the index spell)	ICD-10 terminology review, literature review	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6748036/
Immunosuppression (cancer with active chemotherapy)	diag_01 to diag_20 for ICD-10 codes (any position) opertn_01 to opertn _21 for OPCS-4 codes (any position)	**Cancer (ICD-10):** Check full list in codelist repository file (spreadsheet) **Chemotherapy (ICD-10):** Y43.1, Y43.3, Z51.1 **Chemotherapy (OPCS-4):** Check full list in codelist repository file (spreadsheet)	Cancer diagnosis recorded in the 5 years prior to index date AND chemotherapy recorded in 6 months prior to index date	Definition according to the UK Green Book criteria (and the Chief Medical Officer's recommendations for the shielded-patient list)	https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1057798/Greenbook-chapter-14a-28Feb22.pdf https://digital.nhs.uk/coronavirus/shielded-patient-

Immunosuppression (haematological malignancy or bone marrow transplant)		**Haematological malignancy (ICD-10):** Check full list in codelist repository file (spreadsheet) **Haematological transplant (ICD-10):** T86.0 **Haematological transplant (OPCS-4):** W34, X33.6	Haematological malignancy or haematological transplant recorded in the 2 years prior to index date		list/methodology/rule-logic
Immunosuppression (solid organ transplant)		**Solid organ transplant (ICD-10):** N16.5, T86.1, T86.2, T86.3, T86.4, T86.8, T86.9, Y83.0, Z94	Solid organ transplant recorded in the 5 years prior to index date		

		Solid organ transplant (OPCS-4): Check full list in codelist repository file (spreadsheet)			
Immunosuppression (hyposplenism)		**Hyposplenism (ICD-1):** D73.0, Q89.0, D57.0, D57.1 **Splenectomy (OPCS-4):** J69.2	Hyposplenism or splenectomy recorded in the 5 years prior to index date		
Immunosuppression (other long-term)		D829, D83, D830, D831, D832, D838, D839, D84, D840, D841, D848, D849, D71	Any record within 5 years prior to index date	Manual terminology review and code selection	N/A
Immunosuppression (other short-term)		**ICD-10:** D70, Y43.3 **OPCS:** X92.1, X95.1	Any record within 6 months prior to index date		
Renal replacement therapy (peritoneal dialysis)		**Peritoneal dialysis (ICD-10):** Z49.2	Occurrence of any admission with a peritoneal dialysis code (without	Previously-published algorithm	https://www.kidney-international.org/article/S0085-

		Acute kidney injury (ICD-10): N17 **Peritoneal dialysis (OPCS-4):** X41.1, X40.2, X40.5, X40.6	diagnosis of acute kidney injury) in the 5 years before index date (epistart of the index spell)		2538(17)30856-6/fulltext
Renal replacement therapy (maintenance haemodialysis)		**Dialysis (ICD-10):** E85.3, Y60.2, Y61.2, Y62.2, Y84.1, Z99.2, T82.4, Z49.1 **Dialysis (OPCS-4):** X40.1, X40.3, X40.4	Occurrence of a dialysis code in the 5 years before index date (epistart of the index spell) of the index spell in a patient who has had:		
		ESKD (ICD-10): N18.0, N18.5, Q60.1	(a) a diagnostic code for ESKD any prior time; OR		
		Fistula/graft (OPCS-4): L74.1, L74.2, L74.6, L74.8, L74.9	(b) the insertion of an arteriovenous fistula or graft any prior time.		

Renal replacement therapy (probable maintenance haemodialysis)		**Dialysis (ICD-10):** E85.3, Y60.2, Y61.2, Y62.2, Y84.1, Z99.2, T82.4, Z49.1 **Dialysis (OPCS-4):** X40.1, X40.3, X40.4	Occurrence of at least two episodes containing a dialysis code, with at least 90 days between the start of the first recorded dialysis, and the start of any subsequent dialysis (without diagnosis of acute kidney injury) in the 5 years before index date (epistart of the index spell)		
Hospital frailty score	diag_01 to diag_20 for ICD-10 codes (any position);	Codes and scoring used as per reference publication	Last 2 years including index admission	https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(18)30668-8/fulltext	https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(18)30668-8/fulltext

3. Geographical location

We produced a geographical map depicting the ratio ("representativeness ratio") between relative frequencies of the respective cohort (RECOVERY HES over All-England HES) in each geographical area (based on the lower-super output area/LSOA field in HES). We also included counts of RECOVERY participants recruited in each NHS Trust (based on the Trust's postcode). The geographical areas presented cover England only and were depicted at England region level.

Mapping from the geographical location recorded in HES (postcode or LSOA) to the geographical areas presented was performed using publicly-available UK government files (available at https://geoportal.statistics.gov.uk/datasets/postcode-to-output-area-to-lower-layer-super-output-area-to-middle-layer-super-output-area-to-local-authority-district-may-2021-lookup-in-the-uk/about and https://geoportal.statistics.gov.uk/datasets/local-authority-district-to-county-april-2021-lookup-in-england/explore)

4. Deprivation

Deprivation was assessed using the aggregated English Index of Multiple Deprivation 2019 (and presented using quintiles). Although HES provides deprivation data in the *imd* fields, these are calculated using the 2004 indices and are therefore no longer up to date. Hence, deprivation will be derived from home residence postcodes included in the RECOVERY HES and national HES records, and mapped using the publicly-available lookup tables at https://imd-by-postcode.opendatacommunities.org/imd/2019.

5. Inclusion in RECOVERY over time

The total number of individuals included in RECOVERY and the reference population was calculated and aggregated at monthly level (based on the index date). Then, we calculated the monthly proportions of individuals in the Al-England HES cohort who were recruited to RECOVERY (i.e. ratio between absolute counts in each cohort, not on actual matching since the all-England HES cohort is anonymised).

OUTCOME DEFINITION

1. All-cause death within 28 days of admission

This outcome was only assessed in RECOVERY and the reference population. A timeframe of 28 days from index date was used in both cohorts (rather than 28 days from randomisation in RECOVERY) to allow a more approximate comparison. In both populations, the outcome was derived using HES + Civil Registrations (death record) data only as this is considered the definitive source of mortality records and allowed a meaningful comparison between both cohorts.

R STUDIO PACKAGES

Below is a list of all R packages used in this report (generated using the "grateful" package) and an R session info report:

R packages used

Package	Version	Citation
base	4.2.1	R Core Team (2022)
beepr	1.3	Bååth (2018)
broom.helpers	1.8.0	Larmarange and Sjoberg (2022)
DataExplorer	0.8.2	Cui (2020)
directlabels	2021.1.13	Hocking (2021)
flextable	0.7.2	Gohel (2022a)
forestploter	0.2.3	Dayimu (2022)
fs	1.5.2	Hester, Wickham, and Csárdi (2021)
ggh4x	0.2.3	van den Brand (2022)
ggpattern	1.0.1	FC, Davis, and ggplot2 authors (2022)
ggrepel	0.9.1	Slowikowski (2021)
ggvenn	0.1.9	Yan (2021)
grateful	0.1.11	Rodríguez-Sánchez, Jackson, and Hutchins (2022)
gtsummary	1.6.1	Sjoberg et al. (2021)
modelsummary	1.0.2	Arel-Bundock (2022)
officer	0.4.3	Gohel (2022b)
patchwork	1.1.1	Pedersen (2020)
raster	3.5.21	Hijmans (2022)
RColorBrewer	1.1.3	Neuwirth (2022)
remotes	2.4.2	Csárdi et al. (2021)
rgdal	1.5.32	Bivand, Keitt, and Rowlingson (2022)
scales	1.2.0	Wickham and Seidel (2022)
tidyverse	1.3.2	Wickham et al. (2019)
viridis	0.6.2	Garnier et al. (2021)

Package citations

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Bååth, Rasmus. 2018. *Beepr: Easily Play Notification Sounds on Any Platform*. https://CRAN.R-project.org/package=beepr.

Bivand, Roger, Tim Keitt, and Barry Rowlingson. 2022. *Rgdal: Bindings for the 'Geospatial' Data Abstraction Library*. https://CRAN.R-project.org/package=rgdal.

Csárdi, Gábor, Jim Hester, Hadley Wickham, Winston Chang, Martin Morgan, and Dan Tenenbaum. 2021. *Remotes: R Package Installation from Remote Repositories, Including 'GitHub'*. https://CRAN.R-project.org/package=remotes.

Cui, Boxuan. 2020. *DataExplorer: Automate Data Exploration and Treatment*. https://CRAN.R-project.org/package=DataExplorer.

Dayimu, Alimu. 2022. *Forestploter: Create Flexible Forest Plot*. https://github.com/adayim/forestploter.

FC, Mike, Trevor L Davis, and ggplot2 authors. 2022. *Ggpattern: 'Ggplot2' Pattern Geoms*. https://CRAN.R-project.org/package=ggpattern.

Garnier, Simon, Ross, Noam, Rudis, Robert, Camargo, et al. 2021. *viridis - Colorblind-Friendly Color Maps for r*. https://doi.org/10.5281/zenodo.4679424.

Gohel, David. 2022a. *Flextable: Functions for Tabular Reporting*. https://CRAN.R-project.org/package=flextable.

———. 2022b. *Officer: Manipulation of Microsoft Word and PowerPoint Documents*. https://CRAN.R-project.org/package=officer.

Hester, Jim, Hadley Wickham, and Gábor Csárdi. 2021. *Fs: Cross-Platform File System Operations Based on 'Libuv'*. https://CRAN.R-project.org/package=fs.

Hijmans, Robert J. 2022. *Raster: Geographic Data Analysis and Modeling*. https://CRAN.R-project.org/package=raster.

Hocking, Toby Dylan. 2021. *Directlabels: Direct Labels for Multicolor Plots*. https://CRAN.R-project.org/package=directlabels.

Larmarange, Joseph, and Daniel D. Sjoberg. 2022. *Broom.helpers: Helpers for Model Coefficients Tibbles*. https://CRAN.R-project.org/package=broom.helpers.

Neuwirth, Erich. 2022. *RColorBrewer: ColorBrewer Palettes*. https://CRAN.R-project.org/package=RColorBrewer.

Pedersen, Thomas Lin. 2020. *Patchwork: The Composer of Plots*. https://CRAN.R-project.org/package=patchwork.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rodríguez-Sánchez, Francisco, Connor P. Jackson, and Shaurita D. Hutchins. 2022. *Grateful: Facilitate Citation of r Packages*. https://github.com/Pakillo/grateful.

Sjoberg, Daniel D., Karissa Whiting, Michael Curry, Jessica A. Lavery, and Joseph Larmarange. 2021. "Reproducible Summary Tables with the Gtsummary Package." *The R Journal* 13: 570–80. https://doi.org/10.32614/RJ-2021-053.

Slowikowski, Kamil. 2021. *Ggrepel: Automatically Position Non-Overlapping Text Labels with 'Ggplot2'*. https://CRAN.R-project.org/package=ggrepel.

van den Brand, Teun. 2022. *Ggh4x: Hacks for 'Ggplot2'*. https://CRAN.R-project.org/package=ggh4x.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, and Dana Seidel. 2022. *Scales: Scale Functions for Visualization*. https://CRAN.R-project.org/package=scales.

Yan, Linlin. 2021. *Ggvenn: Draw Venn Diagram by 'Ggplot2'*. https://CRAN.R-project.org/package=ggvenn.

R session info report:

```
session_info()

Session info ─────────────────────────────────────────────────────────────────────────────

 setting  value

 version  R version 4.2.1 (2022-06-23 ucrt)

 os       Windows 10 x64 (build 19044)

 system   x86_64, mingw32

 ui       RStudio

 language (EN)

 collate  English_United Kingdom.utf8

 ctype    English_United Kingdom.utf8

 tz       Europe/London

 date     2022-11-11

 rstudio  2022.07.2+576 Spotted Wakerobin (desktop)

 pandoc   2.19.2 @ C:/Program Files/RStudio/bin/quarto/bin/tools/ (via rmarkdown)

─ Packages ─────────────────────────────────────────────────────────────────────────────────

 package       * version   date (UTC) lib source

 assertthat      0.2.1     2019-03-21 [1] CRAN (R 4.2.1)

 audio           0.1-10    2021-11-25 [1] CRAN (R 4.2.0)

 backports       1.4.1     2021-12-13 [1] CRAN (R 4.2.0)

 base64enc       0.1-3     2015-07-28 [1] CRAN (R 4.2.0)

 beepr         * 1.3       2018-06-04 [1] CRAN (R 4.2.1)

 bit             4.0.4     2020-08-04 [1] CRAN (R 4.2.1)

 bit64           4.0.5     2020-08-30 [1] CRAN (R 4.2.1)

 broom           1.0.0     2022-07-01 [1] CRAN (R 4.2.1)

 broom.helpers   1.8.0     2022-07-05 [1] CRAN (R 4.2.1)

 cachem          1.0.6     2021-08-19 [1] CRAN (R 4.2.1)

 callr           3.7.1     2022-07-13 [1] CRAN (R 4.2.1)

 cellranger      1.1.0     2016-07-27 [1] CRAN (R 4.2.1)
```

```
checkmate      2.1.0    2022-04-21 [1] CRAN (R 4.2.1)
class          7.3-20   2022-01-16 [2] CRAN (R 4.2.1)
classInt       0.4-8    2022-09-29 [1] CRAN (R 4.2.1)
cli            3.3.0    2022-04-25 [1] CRAN (R 4.2.1)
codetools      0.2-18   2020-11-04 [2] CRAN (R 4.2.1)
colorspace     2.0-3    2022-02-21 [1] CRAN (R 4.2.1)
commonmark     1.8.0    2022-03-09 [1] CRAN (R 4.2.1)
crayon         1.5.1    2022-03-26 [1] CRAN (R 4.2.1)
data.table     1.14.2   2021-09-27 [1] CRAN (R 4.2.1)
DBI            1.1.3    2022-06-18 [1] CRAN (R 4.2.1)
dbplyr         2.2.1    2022-06-27 [1] CRAN (R 4.2.1)
devtools       2.4.5    2022-10-11 [1] CRAN (R 4.2.2)
digest         0.6.29   2021-12-01 [1] CRAN (R 4.2.1)
directlabels * 2021.1.13 2021-01-16 [1] CRAN (R 4.2.1)
dplyr        * 1.0.9    2022-04-28 [1] CRAN (R 4.2.1)
dtplyr       * 1.2.1    2022-01-19 [1] CRAN (R 4.2.1)
e1071          1.7-11   2022-06-07 [1] CRAN (R 4.2.1)
ellipsis       0.3.2    2021-04-29 [1] CRAN (R 4.2.1)
evaluate       0.15     2022-02-18 [1] CRAN (R 4.2.1)
fansi          1.0.3    2022-03-24 [1] CRAN (R 4.2.1)
farver         2.1.1    2022-07-06 [1] CRAN (R 4.2.1)
fastmap        1.1.0    2021-01-25 [1] CRAN (R 4.2.1)
flextable    * 0.7.2    2022-06-12 [1] CRAN (R 4.2.1)
forcats      * 0.5.1    2021-01-27 [1] CRAN (R 4.2.1)
forestploter * 0.2.3    2022-11-10 [1] Github (adayim/forestploter@8bf12b2)
fs           * 1.5.2    2021-12-08 [1] CRAN (R 4.2.1)
gargle         1.2.0    2021-07-02 [1] CRAN (R 4.2.1)
gdtools        0.2.4    2022-02-14 [1] CRAN (R 4.2.1)
generics       0.1.3    2022-07-05 [1] CRAN (R 4.2.1)
```

```
ggh4x        * 0.2.3    2022-11-09 [1] CRAN (R 4.2.2)
ggpattern    * 1.0.1    2022-11-09 [1] CRAN (R 4.2.2)
ggplot2      * 3.4.0    2022-11-04 [1] CRAN (R 4.2.2)
ggrepel      * 0.9.1    2021-01-15 [1] CRAN (R 4.2.1)
ggvenn       * 0.1.9    2021-06-29 [1] CRAN (R 4.2.1)
glue           1.6.2    2022-02-24 [1] CRAN (R 4.2.1)
googledrive    2.0.0    2021-07-08 [1] CRAN (R 4.2.1)
googlesheets4  1.0.0    2021-07-21 [1] CRAN (R 4.2.1)
gridExtra      2.3      2017-09-09 [1] CRAN (R 4.2.1)
gridpattern    1.0.2    2022-11-07 [1] CRAN (R 4.2.2)
gt             0.6.0    2022-05-24 [1] CRAN (R 4.2.1)
gtable         0.3.1    2022-09-01 [1] CRAN (R 4.2.2)
gtsummary    * 1.6.1    2022-06-22 [1] CRAN (R 4.2.1)
haven        * 2.5.0    2022-04-15 [1] CRAN (R 4.2.1)
hms            1.1.1    2021-09-26 [1] CRAN (R 4.2.1)
htmltools      0.5.3    2022-07-18 [1] CRAN (R 4.2.1)
htmlwidgets    1.5.4    2021-09-08 [1] CRAN (R 4.2.1)
httpuv         1.6.5    2022-01-05 [1] CRAN (R 4.2.1)
httr           1.4.3    2022-05-04 [1] CRAN (R 4.2.1)
jsonlite       1.8.0    2022-02-22 [1] CRAN (R 4.2.1)
KernSmooth     2.23-20  2021-05-03 [2] CRAN (R 4.2.1)
knitr          1.39     2022-04-26 [1] CRAN (R 4.2.1)
labeling       0.4.2    2020-10-20 [1] CRAN (R 4.2.0)
later          1.3.0    2021-08-18 [1] CRAN (R 4.2.1)
lattice        0.20-45  2021-09-22 [2] CRAN (R 4.2.1)
lifecycle      1.0.3    2022-10-07 [1] CRAN (R 4.2.2)
lubridate    * 1.8.0    2021-10-07 [1] CRAN (R 4.2.1)
magrittr     * 2.0.3    2022-03-30 [1] CRAN (R 4.2.1)
memoise        2.0.1    2021-11-26 [1] CRAN (R 4.2.1)
```

```
mime          0.12     2021-09-28 [1] CRAN (R 4.2.0)
miniUI        0.1.1.1  2018-05-18 [1] CRAN (R 4.2.1)
modelr        0.1.8    2020-05-19 [1] CRAN (R 4.2.1)
modelsummary  * 1.0.2    2022-07-17 [1] CRAN (R 4.2.1)
munsell       0.5.0    2018-06-12 [1] CRAN (R 4.2.1)
officer       * 0.4.3    2022-06-12 [1] CRAN (R 4.2.1)
patchwork     * 1.1.1    2020-12-17 [1] CRAN (R 4.2.1)
pillar        1.8.0    2022-07-18 [1] CRAN (R 4.2.1)
pkgbuild      1.3.1    2021-12-20 [1] CRAN (R 4.2.2)
pkgconfig     2.0.3    2019-09-22 [1] CRAN (R 4.2.1)
pkgload       1.3.0    2022-06-27 [1] CRAN (R 4.2.1)
prettyunits   1.1.1    2020-01-24 [1] CRAN (R 4.2.1)
processx      3.7.0    2022-07-07 [1] CRAN (R 4.2.1)
profvis       0.3.7    2020-11-02 [1] CRAN (R 4.2.2)
promises      1.2.0.1  2021-02-11 [1] CRAN (R 4.2.1)
proxy         0.4-27   2022-06-09 [1] CRAN (R 4.2.1)
ps            1.7.1    2022-06-18 [1] CRAN (R 4.2.1)
purrr         * 0.3.4    2020-04-17 [1] CRAN (R 4.2.1)
quadprog      1.5-8    2019-11-20 [1] CRAN (R 4.2.0)
R6            2.5.1    2021-08-19 [1] CRAN (R 4.2.1)
ragg          1.2.4    2022-10-24 [1] CRAN (R 4.2.2)
raster        3.5-21   2022-06-27 [1] CRAN (R 4.2.1)
RColorBrewer  * 1.1-3    2022-04-03 [1] CRAN (R 4.2.0)
Rcpp          1.0.9    2022-07-08 [1] CRAN (R 4.2.1)
readr         * 2.1.2    2022-01-30 [1] CRAN (R 4.2.1)
readxl        * 1.4.0    2022-03-28 [1] CRAN (R 4.2.1)
remotes       2.4.2    2021-11-30 [1] CRAN (R 4.2.1)
reprex        2.0.1    2021-08-05 [1] CRAN (R 4.2.1)
rgdal         * 1.5-32   2022-05-09 [1] CRAN (R 4.2.1)
```

```
rlang         1.0.6   2022-09-24 [1] CRAN (R 4.2.2)
rmarkdown     2.14    2022-04-25 [1] CRAN (R 4.2.1)
rstudioapi    0.13    2020-11-12 [1] CRAN (R 4.2.1)
rvest         1.0.2   2021-10-16 [1] CRAN (R 4.2.1)
sass          0.4.2   2022-07-16 [1] CRAN (R 4.2.1)
scales      * 1.2.0   2022-04-13 [1] CRAN (R 4.2.1)
sessioninfo   1.2.2   2021-12-06 [1] CRAN (R 4.2.2)
sf            1.0-8   2022-07-14 [1] CRAN (R 4.2.1)
shiny         1.7.2   2022-07-19 [1] CRAN (R 4.2.1)
sp          * 1.5-0   2022-06-05 [1] CRAN (R 4.2.1)
stringi       1.7.8   2022-07-11 [1] CRAN (R 4.2.1)
stringr     * 1.4.0   2019-02-10 [1] CRAN (R 4.2.1)
systemfonts   1.0.4   2022-02-11 [1] CRAN (R 4.2.1)
tables        0.9.6   2020-09-22 [1] CRAN (R 4.2.1)
terra         1.6-3   2022-07-25 [1] CRAN (R 4.2.1)
textshaping   0.3.6   2021-10-13 [1] CRAN (R 4.2.2)
tibble      * 3.1.8   2022-07-22 [1] CRAN (R 4.2.1)
tidyr       * 1.2.0   2022-02-01 [1] CRAN (R 4.2.1)
tidyselect    1.1.2   2022-02-21 [1] CRAN (R 4.2.1)
tidyverse   * 1.3.2   2022-07-18 [1] CRAN (R 4.2.1)
tzdb          0.3.0   2022-03-28 [1] CRAN (R 4.2.1)
units         0.8-0   2022-02-05 [1] CRAN (R 4.2.1)
urlchecker    1.0.1   2021-11-30 [1] CRAN (R 4.2.2)
usethis       2.1.6   2022-05-25 [1] CRAN (R 4.2.2)
utf8          1.2.2   2021-07-24 [1] CRAN (R 4.2.1)
uuid          1.1-0   2022-04-19 [1] CRAN (R 4.2.0)
vctrs         0.5.0   2022-10-22 [1] CRAN (R 4.2.2)
viridis     * 0.6.2   2021-10-13 [1] CRAN (R 4.2.1)
viridisLite * 0.4.0   2021-04-13 [1] CRAN (R 4.2.1)
```

```
vroom       1.5.7   2021-11-30 [1] CRAN (R 4.2.1)
withr       2.5.0   2022-03-03 [1] CRAN (R 4.2.1)
xfun        0.31    2022-05-10 [1] CRAN (R 4.2.1)
xml2        1.3.3   2021-11-30 [1] CRAN (R 4.2.1)
xtable      1.8-4   2019-04-21 [1] CRAN (R 4.2.1)
zip         2.2.0   2021-05-31 [1] CRAN (R 4.2.1)
```

References (Annex I)

1. UK Health Security Agency. COVID-19: the green book, chapter 14a. 2022.

2. Storey BC, Staplin N, Harper CH, et al. Declining comorbidity-adjusted mortality rates in English patients receiving maintenance renal replacement therapy. *Kidney Int*. 2018 May;**93**(5):1165–1174.

3. Quan H, Sundararajan V, Halfon P, et al. Coding Algorithms for Defining Comorbidities in ICD-9-CM and ICD-10 Administrative Data: *Med Care*. 2005 Nov;**43**(11):1130–1139.

4. Gilbert T, Neuburger J, Kraindler J, et al. Development and validation of a Hospital Frailty Risk Score focusing on older people in acute care settings using electronic hospital records: an observational study. *Lancet Lond Engl*. 2018 May 5;**391**(10132):1775–1782.

5. Centers for Disease Control and Prevention. Alcohol-Related ICD Codes [Internet]. 2021 [cited 2022 Oct 11]. Available from: https://www.cdc.gov/alcohol/ardi/alcohol-related-icd-codes.html

6. Khokhar B, Jette N, Metcalfe A, et al. Systematic review of validated case definitions for diabetes in ICD-9-coded and ICD-10-coded data in adult populations. *BMJ Open*. 2016 Aug;**6**(8):e009952.

7. NHS Digital. NHS Data Model and Dictionary - Severe Mental Illness [Internet]. 2022 [cited 2022 Oct 11]. Available from: https://www.datadictionary.nhs.uk/Covid19PRA/Severe_Mental_Illness.html

8. Gribsholt SB, Pedersen L, Richelsen B, Thomsen RW. Validity of ICD-10 diagnoses of overweight and obesity in Danish hospitals. *Clin Epidemiol*. 2019;**11**:845–854.

Annex II – Supplementary tables and figures

Supplementary table S1 - Baseline characteristics of the RECOVERY population recruited in the UK, grouped by nation (using data from the case report form only, except where stated)

Characteristic	England, N = 39,952[a]	Northern Ireland, N = 985 [a]	Scotland, N = 2496 [a]	Wales, N = 1333 [a]
Age, mean (SD)	62 (16)	58 (15)	62 (15)	61 (15)
<60	17,063 (43%)	505 (51%)	1036 (42%)	602 (45%)
60-69	9154 (23%)	256 (26%)	657 (26%)	350 (26%)
70-79	8047 (20%)	156 (16%)	473 (19%)	257 (19%)
80-89	4709 (12%)	60 (6.1%)	268 (11%)	105 (7.9%)
90+	979 (2.5%)	8 (0.8%)	62 (2.5%)	19 (1.4%)
Female	14,616 (37%)	340 (35%)	958 (38%)	487 (37%)
Ethnicity[b]				
White	32,205 (82%)	N/A	2040 (97%)	574 (93%)

Black	1491 (3.8%)	N/A	13 (0.6%)	3 (0.5%)
Asian	3964 (10%)	N/A	44 (2.1%)	23 (3.7%)
Other	948 (2.4%)	N/A	7 (0.3%)	8 (1.3%)
Mixed	516 (1.3%)	N/A	5 (0.2%)	6 (1.0%)
Unknown	828 (2.1%)	N/A	387 (16%)	719 (54%)
Respiratory support status[c]				
Invasive mechanical ventilation or ECMO	2674 (6.7%)	46 (4.7%)	180 (7.2%)	233 (17%)
Non-invasive mechanical ventilation or supplementary oxygen	32,420 (81%)	851 (86%)	1955 (78%)	913 (68%)
None	4858 (12%)	88 (8.9%)	361 (14%)	187 (14%)
Chronic lung disease	8784 (22%)	200 (20%)	572 (23%)	290 (22%)
Diabetes	10,434 (26%)	209 (21%)	564 (23%)	313 (23%)
Chronic heart disease	9120 (23%)	174 (18%)	506 (20%)	250 (19%)

Severe liver disease	471 (1.2%)	10 (1.0%)	33 (1.3%)	19 (1.4%)
Severe renal impairment	2238 (5.6%)	19 (1.9%)	76 (3.0%)	76 (5.7%)

Cohorts restricted to participants randomised before the 1st January 2022, with no age restrictions. ECMO - extracorporeal membrane oxygenation; HES - Hospital Episode Statistics; SD – standard deviation;

[a]Mean (SD); n (%)

[b]Ethnicity extracted from either primary care data (General Practice Extraction Service Data for Pandemic Planning and Research - GDPPR) or HES data for people in England, and data sources equivalent to HES Scotland and Wales; no such data was available in Northern Ireland; proportions for people with known and unknown ethnicity were calculated separately and using the entire cohort as denominator

[c]Respiratory status derived from the case-report form combined with linked data sources such as hospital admissions and intensive care data (except for Northern Ireland where no linkage data was available)

Supplementary table S2 - Baseline characteristics of the RECOVERY population recruited in England, grouped by HES linkage status (using data from the case report form only, except where stated)

Characteristic	HES data available, N = 38,510[a]	HES data unavailable, N = 648[a]

Age, mean (SD)	63 (15)	58 (15)
<60	16,121 (42%)	349 (54%)
60-69	8906 (23%)	146 (23%)
70-79	7872 (20%)	107 (17%)
80-89	4644 (12%)	37 (5.7%)
90+	967 (2.5%)	9 (1.4%)
Female	14,068 (37%)	245 (38%)
Ethnicity[2]		
White	31,326 (83%)	330 (64%)
Black	1367 (3.6%)	71 (14%)
Asian	3809 (10%)	67 (13%)
Other	877 (2.3%)	25 (4.9%)

Mixed	465 (1.2%)	19 (3.7%)
Unknown	666 (1.7%)	136 (21%)
Respiratory support status[3]		
Invasive mechanical ventilation or ECMO	2562 (6.7%)	72 (11%)
Non-invasive mechanical ventilation or supplementary oxygen	31,363 (81%)	504 (78%)
None	4585 (12%)	72 (11%)
Chronic lung disease	8569 (22%)	130 (20%)
Diabetes	10,193 (26%)	152 (23%)
Chronic heart disease	8936 (23%)	108 (17%)
Severe liver disease	448 (1.2%)	19 (2.9%)
Severe renal impairment	2185 (5.7%)	39 (6.0%)

Restricted to RECOVERY participants aged >=16 years and recruited in England within the analysis period. ECMO - extracorporeal membrane oxygenation; HES - Hospital Episode Statistics

[a]Mean (SD); n (%)

[b]Ethnicity extracted from either primary care data (General Practice Extraction Service Data for Pandemic Planning and Research - GDPPR) or HES data. Proportions for people with known and unknown ethnicity were calculated separately, using the entire cohort as denominator

[c]Respiratory status derived from the case-report form combined with linked data sources such as hospital admissions and intensive care data

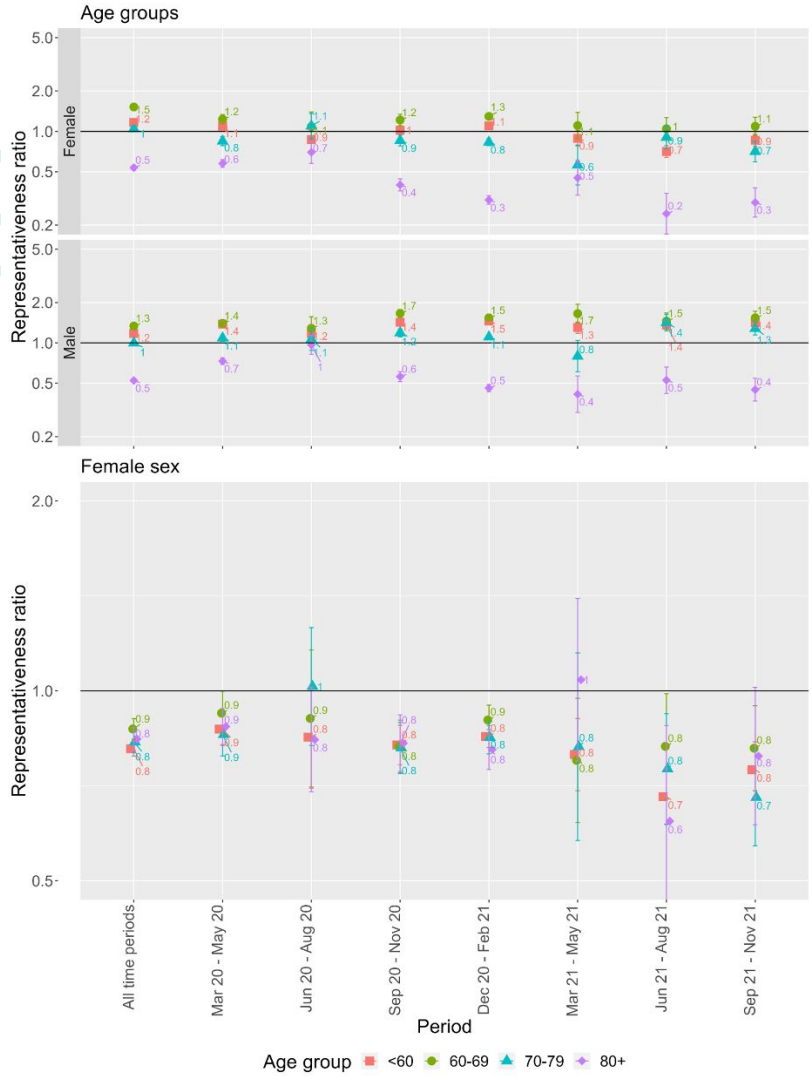
Supplementary table S3 - Number of individuals included in RECOVERY versus reference population over time, by age groups

Period	Age bands											Aggregate		
	<60			60-69			70-79			80+					
	RECOVERY (% of all ages total)	Reference population (% of all ages total)	Proportion included in RECOVERY (%)	RECOVERY (% of all ages total)	Reference population (% of all ages total)	Proportion included in RECOVERY (%)	RECOVERY (% of all ages total)	Reference population (% of all ages total)	Proportion included in RECOVERY (%)	RECOVERY (% of all ages total)	Reference population (% of all ages total)	Proportion included in RECOVERY (%)	RECOVERY (% of all ages total)	Reference population (% of all ages total)	Proportion included in RECOVERY (%)
All time periods	16,121 (100)	123,790 (100)	13	8906 (100)	56,452 (100)	15.8	7871 (100)	69,107 (100)	11.4	5612 (100)	96,922 (100)	5.8	38,510	346,271	11.1
Mar 20 - May 20	3337 (20.7)	21,514 (17.4)	15.5	1992 (22.4)	12,082 (21.4)	16.5	2068 (26.3)	16,915 (24.5)	12.2	2237 (39.9)	27,669 (28.5)	8.1	9634	78,180	12.3
Jun 20 - Aug 20	311 (1.9)	2095 (1.7)	14.8	188 (2.1)	1084 (1.9)	17.3	218 (2.8)	1396 (2)	15.6	285 (5.1)	2364 (2.4)	12.1	1002	6939	14.4
Sep 20 - Nov 20	2263 (14)	14,494 (11.7)	15.6	1512 (17)	8082 (14.3)	18.7	1488 (18.9)	11,252 (16.3)	13.2	945 (16.8)	15,603 (16.1)	6.1	6208	49,431	12.6
Dec 20 - Feb 21	7280 (45.2)	43,526 (35.2)	16.7	4143 (46.5)	22,371 (39.6)	18.5	3241 (41.2)	25,535 (36.9)	12.7	1784 (31.8)	36,318 (37.5)	4.9	16,448	127,750	12.9
Mar 21 - May 21	716 (4.4)	5129 (4.1)	14	264 (3)	1488 (2.6)	17.7	93 (1.2)	1086 (1.6)	8.6	87 (1.6)	1616 (1.7)	5.4	1160	9319	12.4
Jun 21 - Aug 21	1162 (7.2)	19,016 (15.4)	6.1	321 (3.6)	4266 (7.6)	7.5	313 (4)	4453 (6.4)	7	109 (1.9)	4729 (4.9)	2.3	1905	32,464	5.9
Sep 21 - Nov 21	1052 (6.5)	18,016 (14.6)	5.8	486 (5.5)	7079 (12.5)	6.9	450 (5.7)	8470 (12.3)	5.3	165 (2.9)	8623 (8.9)	1.9	2,153	42,188	5.1

Supplementary table S4 - 28-day mortality along time in the RECOVERY HES and All-England HES cohorts (split by age groups)

Period	Age bands											
	<60				60-69				70-79				80+			
	RECOVERY		Reference population		RECOVERY		Reference population		RECOVERY		Reference population		RECOVERY		Reference population	
	Deaths/ Population	Mortality (%)	Deaths/ Population	Mortality (%)	Deaths/ Population	Mortality (%)	Deaths/ Population	Mortality (%)	Deaths/ Population	Mortality (%)	Deaths/ Population	Mortality (%)	Deaths/ Population	Mortality (%)	Deaths/ Population	Mortality (%)
All time periods	1144/ 16,109	7.1	6901/ 123,429	5.6	1789/ 8897	20.1	11,031/ 56,435	19.5	2584/ 7868	32.8	22,481/ 69,097	32.5	2400/ 5610	42.8	45,296/ 96,665	46.9
Mar 20 - May 20	309/ 3,333	9.3	2103/ 21,490	9.8	465/ 1990	23.4	3250/ 12,077	26.9	681/ 2067	32.9	6925/ 16,914	40.9	907/ 2236	40.6	14,574/ 27,665	52.7
Jun 20 - Aug 20	14/ 311	4.5	105/ 2076	5.1	37/ 188	19.7	182/ 1084	16.8	68/ 217	31.3	387/ 1395	27.7	118/ 285	41.4	896/ 2364	37.9
Sep 20 - Nov 20	133/ 2260	5.9	710/ 14,492	4.9	291/ 1510	19.3	1438/ 8079	17.8	505/ 1488	33.9	3481/ 11,252	30.9	404/ 945	42.8	6925/ 15,600	44.4
Dec 20 - Feb 21	547/ 7276	7.5	2538/ 43,512	5.8	826/ 4138	20.0	4342/ 22,364	19.4	1115/ 3240	34.4	8602/ 25,528	33.7	841/ 1783	47.2	17,871/ 36,316	49.2
Mar 21 - May 21	22/ 716	3.1	127/ 5128	2.5	39/ 264	14.8	192/ 1488	12.9	22/ 93	23.7	280/ 1086	25.8	42/ 87	48.3	617/ 1542	40.0
Jun 21 - Aug 21	60/ 1161	5.2	612/ 19,012	3.2	64/ 321	19.9	607/ 4266	14.2	82/ 313	26.2	1021/ 4452	22.9	42/ 109	38.5	1521/ 4555	33.4
Sep 21 - Nov 21	59/ 1052	5.6	706/ 17,719	4.0	67/ 486	13.8	1020/ 7077	14.4	111/ 450	24.7	1785/ 8470	21.1	46/ 165	27.9	2892/ 8623	33.5

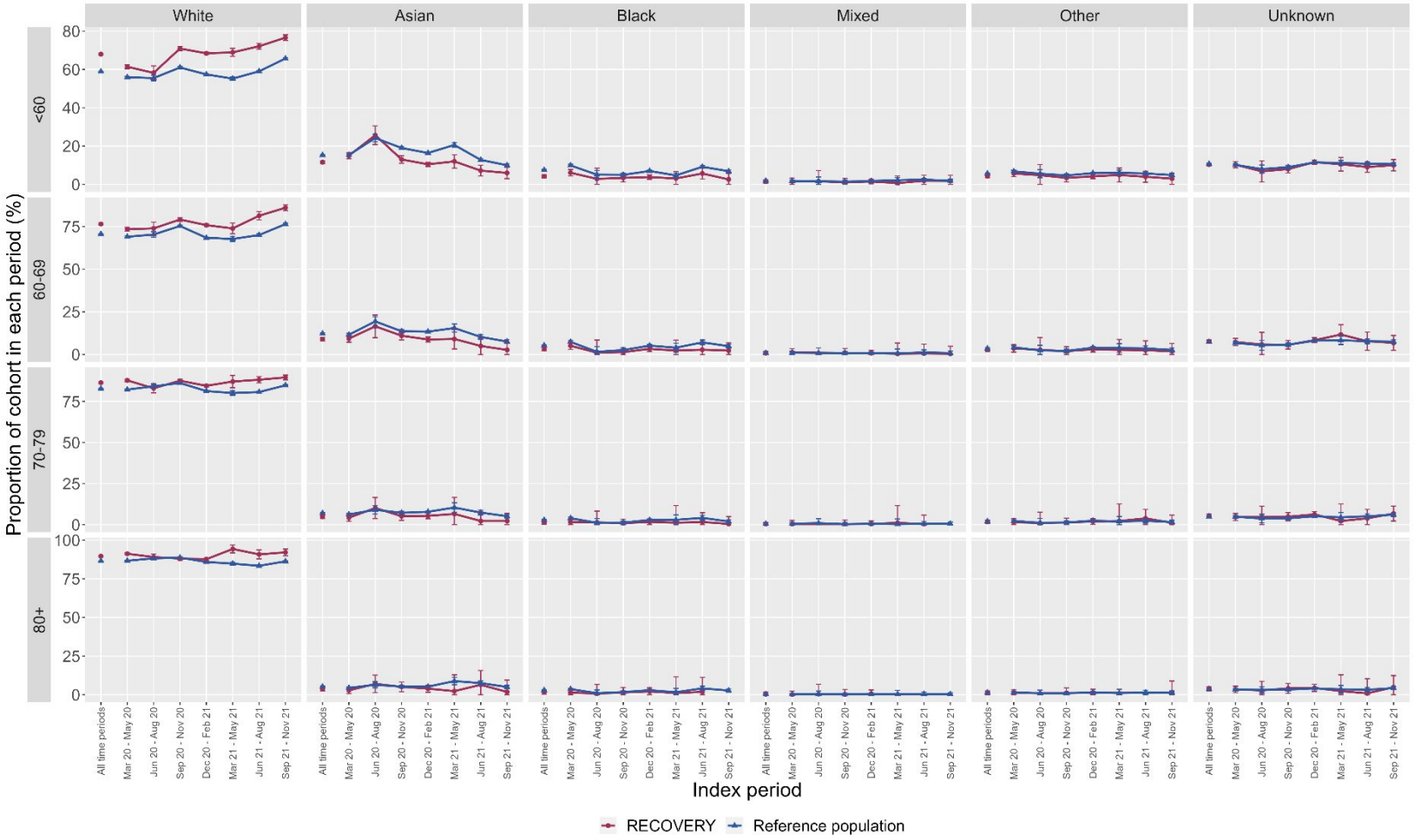
Supplementary figure S1 - Age and sex representativeness of RECOVERY in comparison with the reference population



	All time periods	Mar 20 - May 20	Jun 20 - Aug 20	Sep 20 - Nov 20	Dec 20 - Feb 21	Mar 21 - May 21	Jun 21 - Aug 21	Sep 21 - Nov 21
RECOVERY (individuals)	38,510	9626	1001	6203	16,437	1160	1904	2153
Reference population (individuals)	346,271	78,180	6,939	49,431	127,750	9,319	32,464	42,188
Recruitment ratio (%)	11.1	12.3	14.4	12.6	12.9	12.4	5.9	5.1

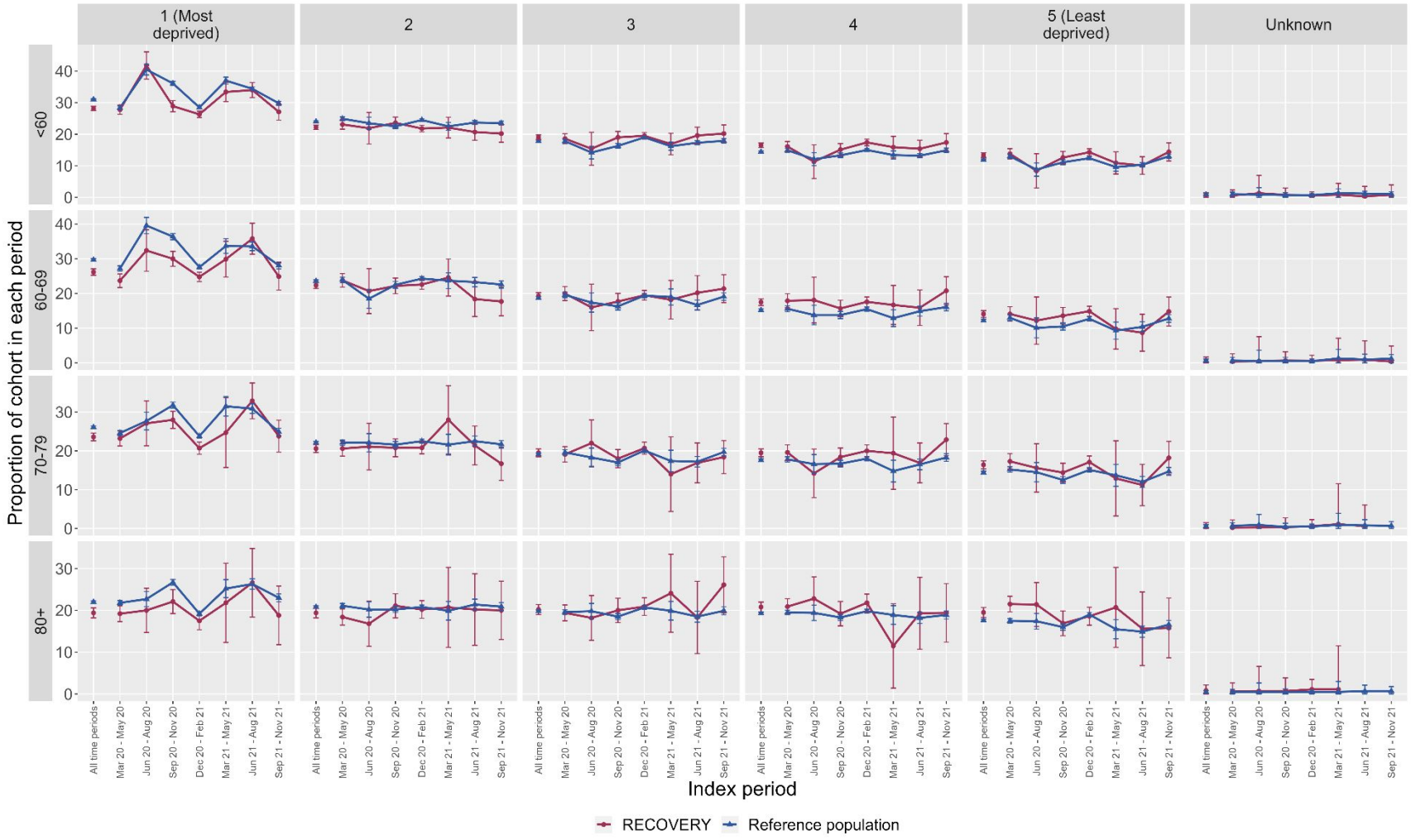
Vertical axes plotted using a log2 scale. Representativeness ratio calculated as the ratio between the proportion of individuals in each age group in RECOVERY over the reference population (i.e. ratio > 1 indicates overrepresentation in RECOVERY, ratio <1 indicates underrepresentation). Error bars plotted using 95% confidence intervals.

Supplementary figure S2 - Ethnicity over time in the RECOVERY cohort in comparison with the reference population, by age



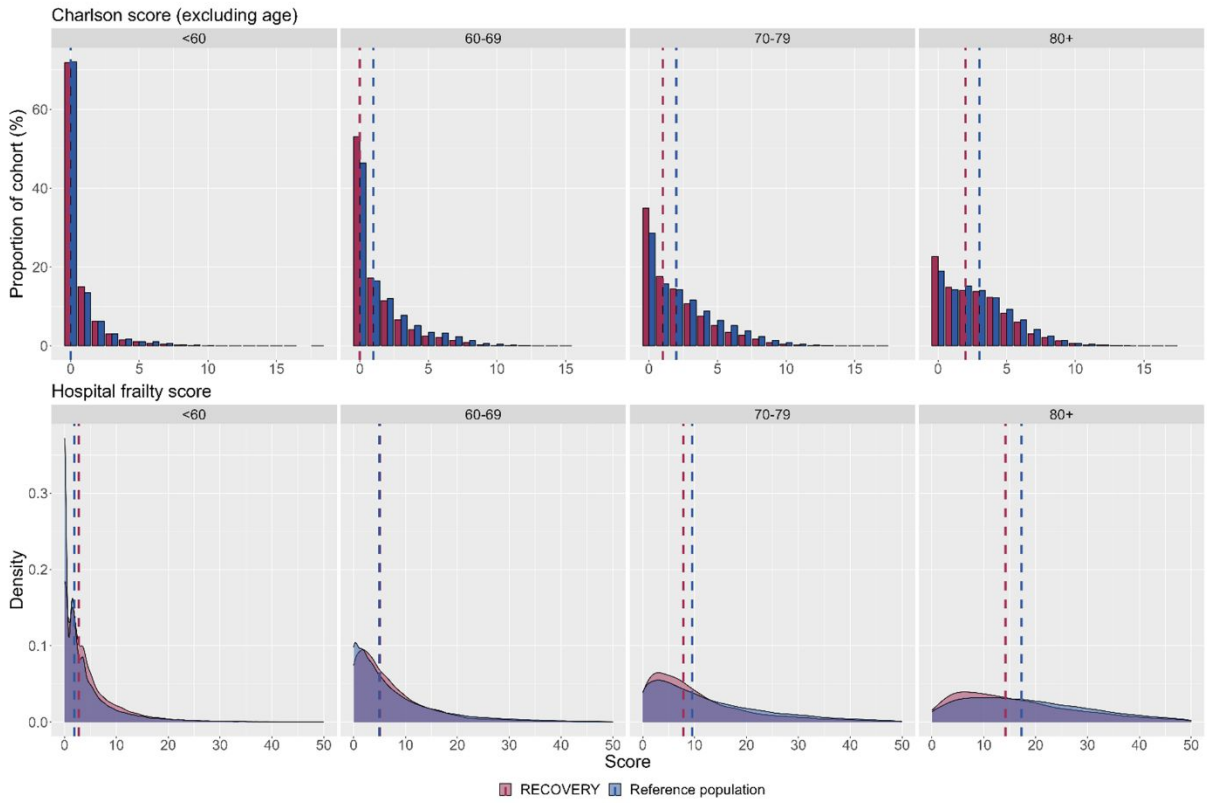
Error bars depict standard errors

Supplementary figure S3 - Deprivation over time in the RECOVERY cohort in comparison with the reference population, by age



Error bars depict standard errors

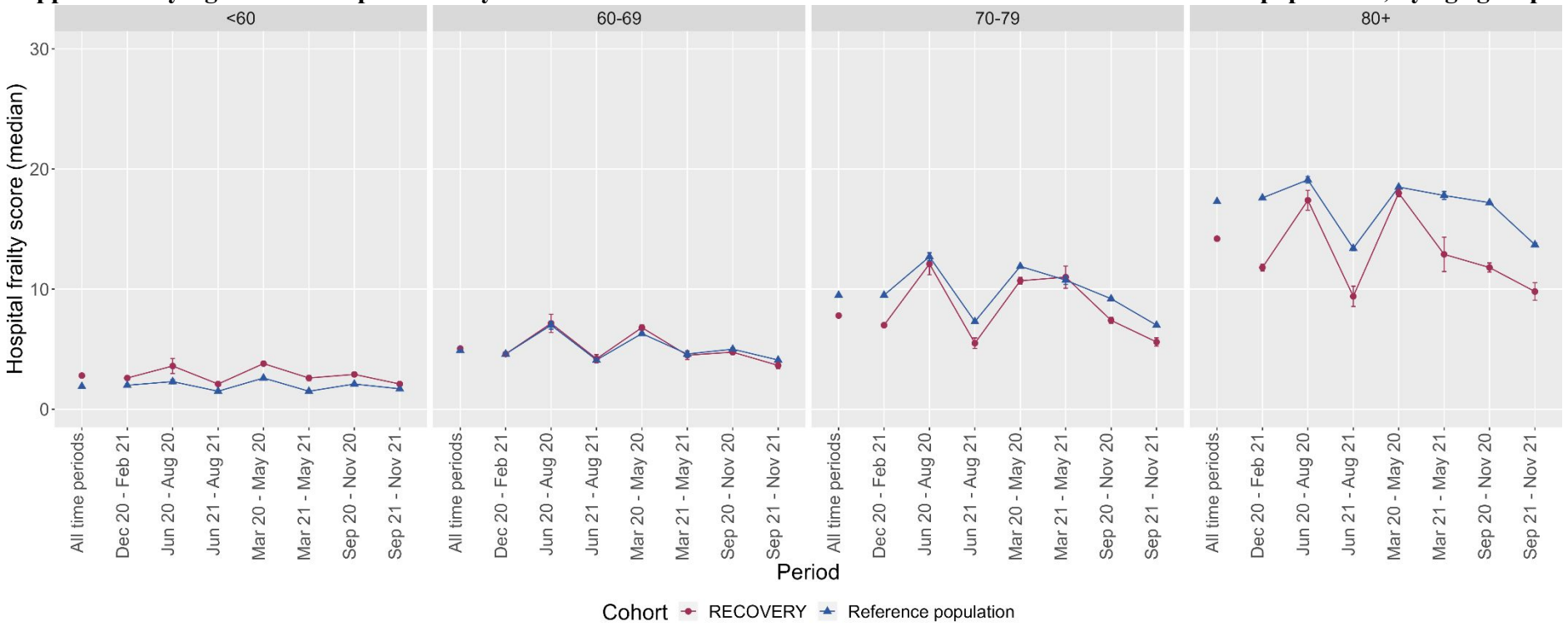
Supplementary figure S4 - Charlson Comorbidty Score (excluding age) and Hospital Frailty Risk Score in the RECOVERY cohort and the reference population, by age groups



Age groups	Individuals (proportion of cohort)		Hospital frailty score (IQR)		Charlson score, excluding age (IQR)	
	RECOVERY cohort	Reference population	RECOVERY cohort	Reference population	RECOVERY cohort	Reference population
<60	16,121 (41.9%)	123,790 (35.7%)	3 (1 - 6)	2 (0 - 5)	0 (0 - 1)	0 (0 - 1)
60-69	8906 (23.1%)	56,452 (16.3%)	5 (2 - 10)	5 (2 - 11)	0 (0 - 2)	1 (0 - 3)
70-79	7871 (20.4%)	69,107 (20.0%)	8 (4 - 15)	10 (4 - 19)	1 (0 - 3)	2 (0 - 4)
80+	5612 (14.6%)	96,922 (28.0%)	14 (7 - 24)	17 (9 - 28)	2 (1 - 4)	3 (1 - 5)

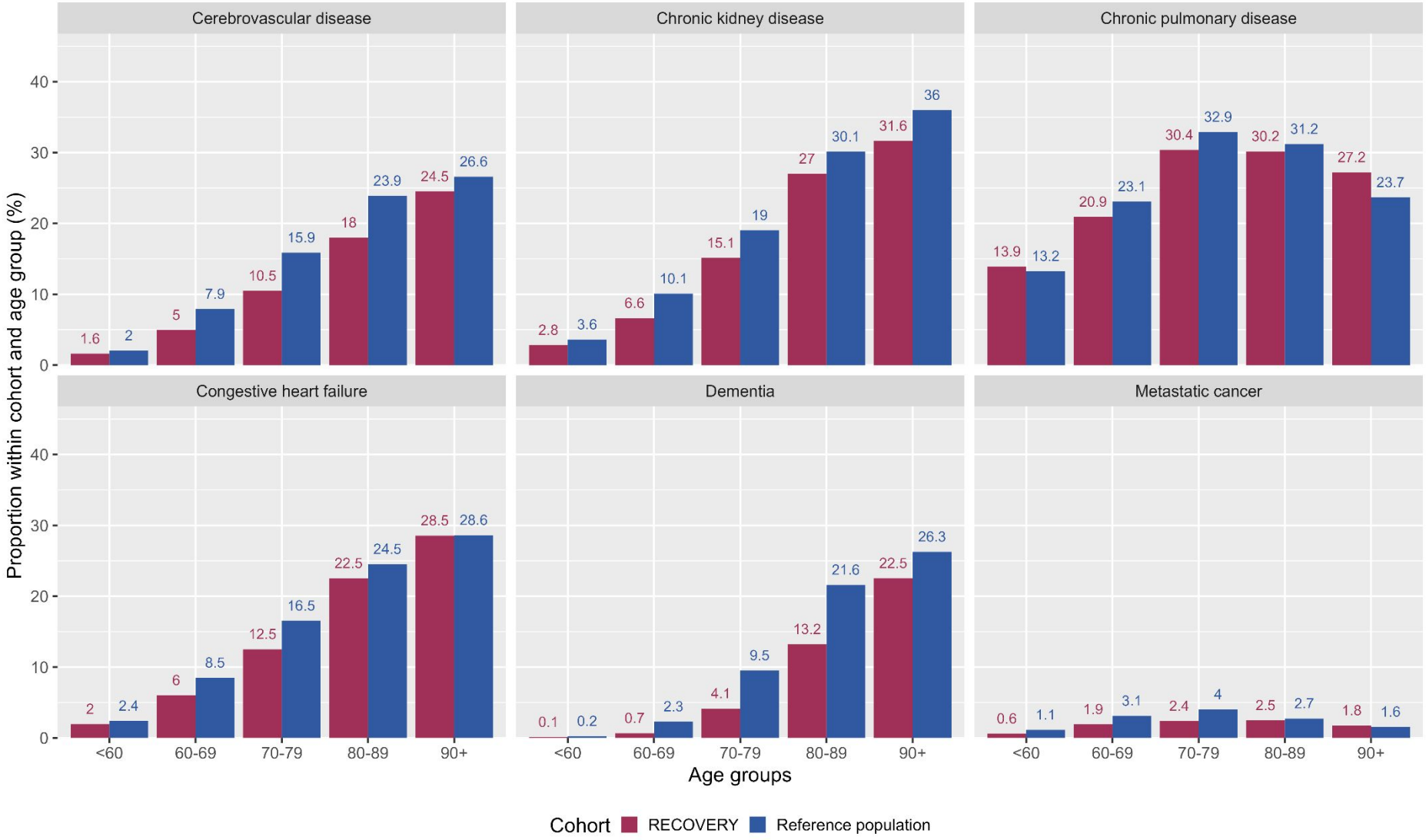
Vertical dashed bars depict the median in each cohort

Supplementary figure S5 - Hospital Frailty Risk Score over time in the RECOVERY cohort and the reference population, by age groups

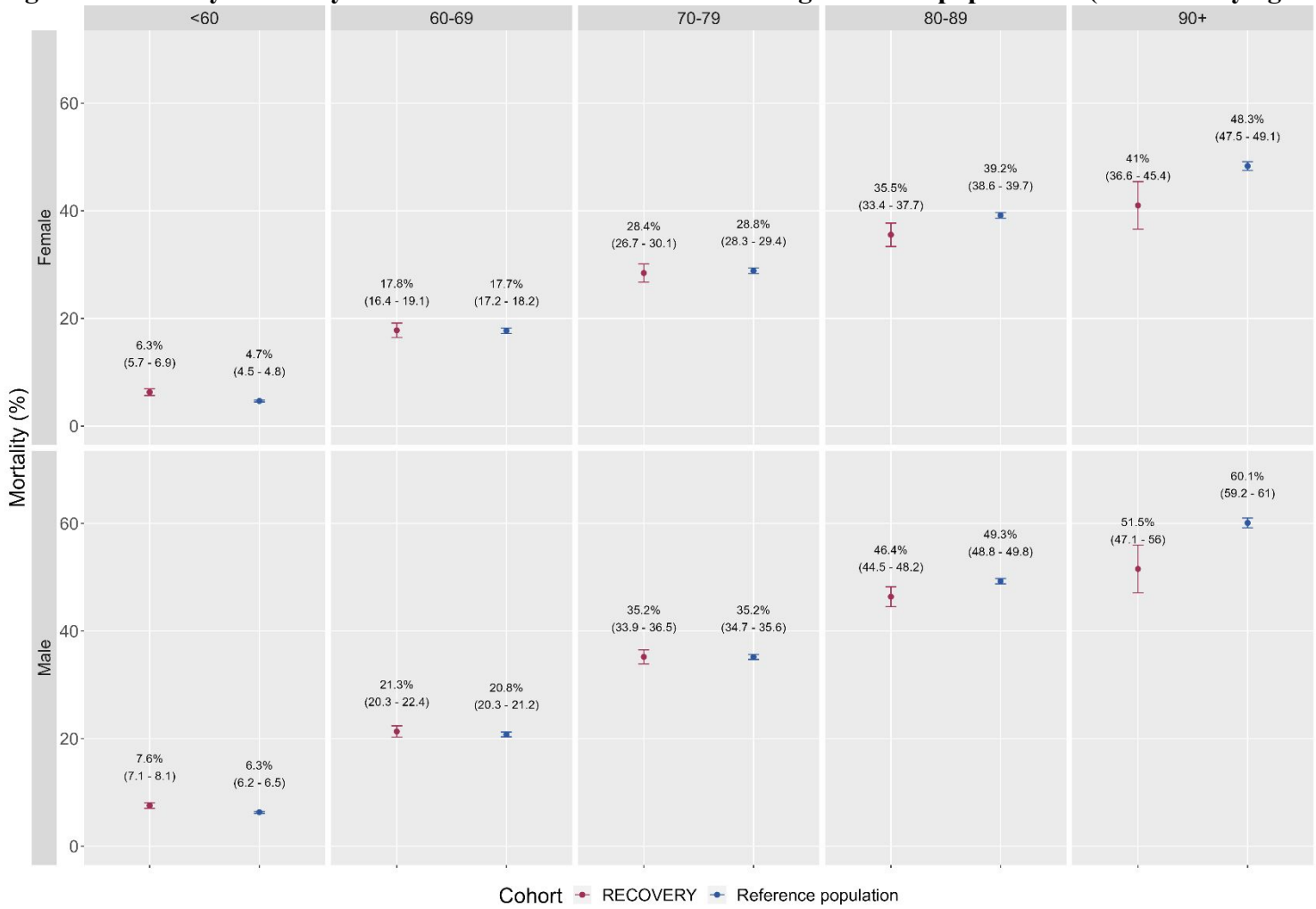


Error bars depict standard errors

Supplementary figure S6 - Prevalence of select comorbidities in the RECOVERY cohort and reference population, by age groups

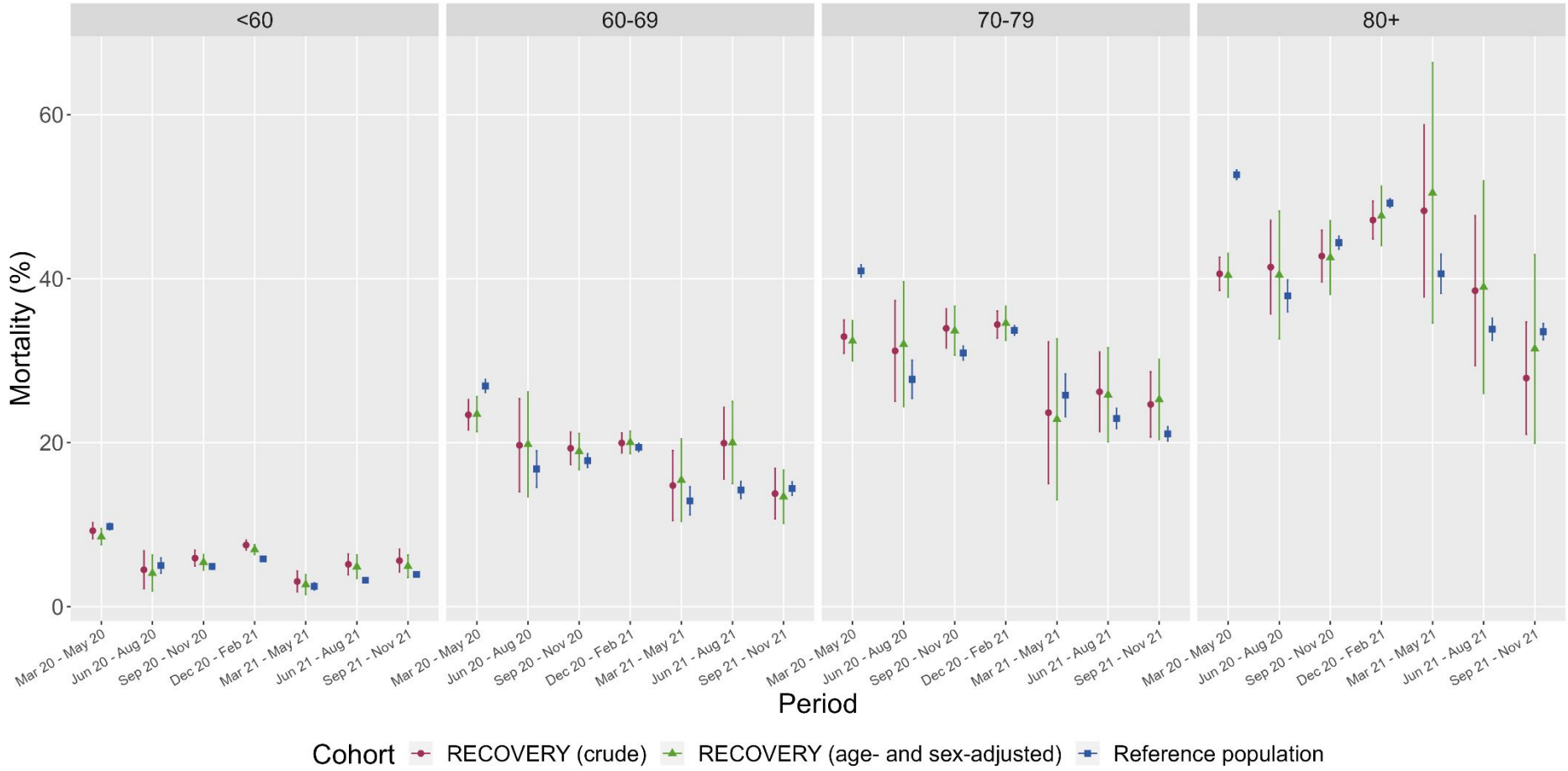


Supplementary figure S7 - 28-day mortality in the RECOVERY HES and All-England HES populations (stratified by age and sex)



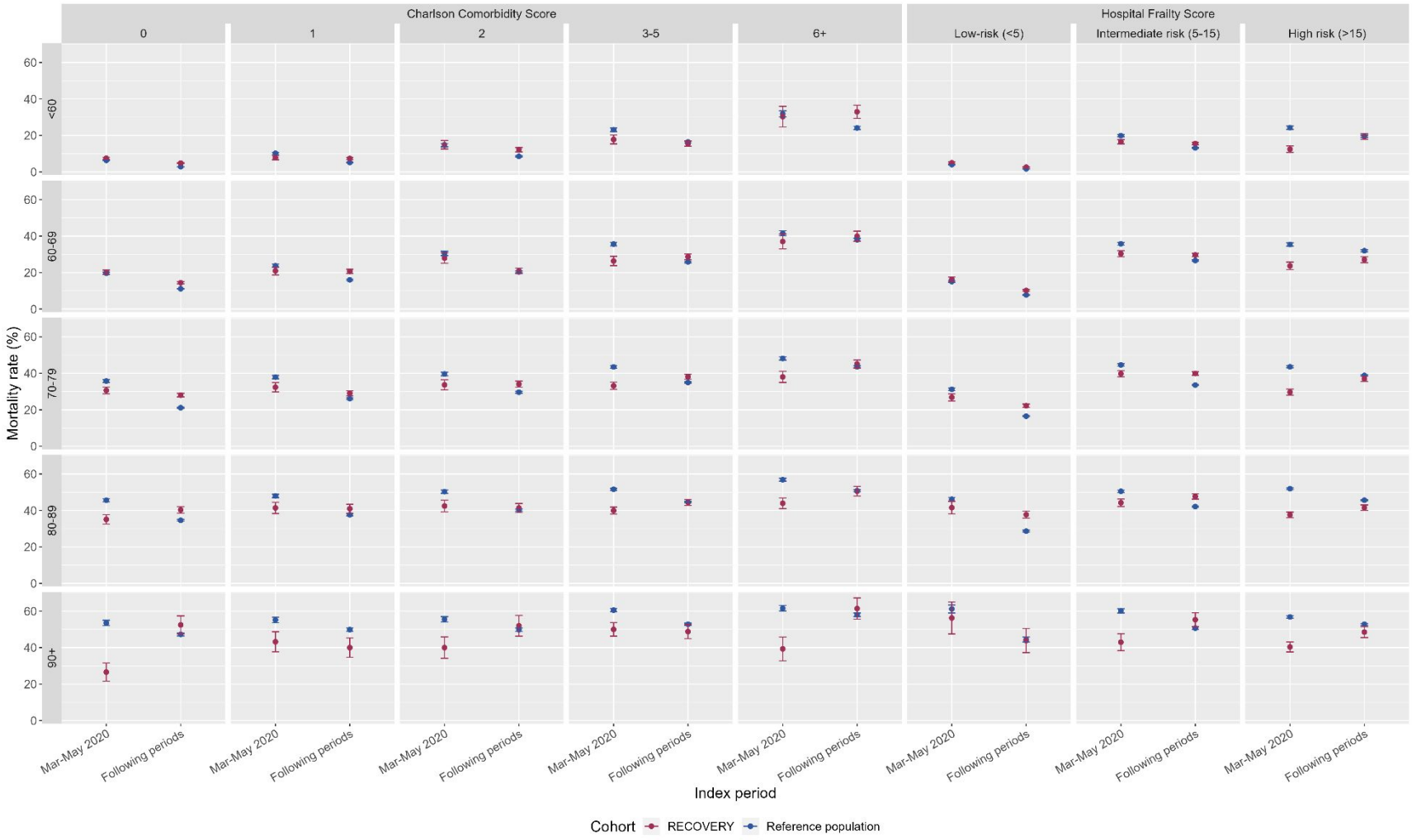
Error bars show 95% confidence intervals

Supplementary figure S8 - 28-day mortality over time in RECOVERY and the reference population (by age groups)



28-day mortality presented as the proportion of people with death recorded within 28 days of their index date along with 95% confidence intervals.

Supplementary figure S9 - Mortality rates over time in RECOVERY and the reference population, by age, comorbidity, and frailty



Error bars depict standard errors

Annex III - Cross-coding of COVID-19 coding in Hospital Episode Statistics versus testing data in SGSS within the RECOVERY population

For this project, a wealth of linked datasets was available for the RECOVERY population (including admissions data from Hospital Episode Statistics – HES, and in-hospital COVID-19 testing from Second Generation Surveillance System data - SGSS). However, for the reference population (anonymised All-England cohort), only HES/ONS data were available. Therefore, preliminary validation work was undertaken to validate the proposed approach of building a nationwide cohort of patients admitted to hospital with COVID-19 using HES data only by exploring cross-coding of ICD-10 codes (in HES) and SARS-CoV-2 testing (from the SGSS data) in the RECOVERY population (which may be seen as a population with a proven or suspected clinical diagnosis of COVID-19). This analysis was based on RECOVERY participants recruited in England between March 2020 and November 2021, using HES data received in September 2022.

1. COVID-19 coding in HES (any code) vs SGSS testing records

Of 38,920 RECOVERY participants with HES data available:

1) 98.7% (n=38,412) had a HES spell straddling randomisation date;
2) 96.1% (n=37,047) had a HES spell containing COVID-19 codes in any diagnostic position at any time point, and 91.2% (n=35,493) had a HES spell containing COVID-19 codes in the primary diagnostic position at any time point;
3) 94.8% (n=36,943) had a HES spell straddling randomisation date and containing COVID-19 codes in any diagnostic position;
4) 89.7% (n=34,929) had a HES spell straddling randomisation date and containing COVID-19 codes in the primary diagnostic position; of these, 93.8% (n=32,770) had a positive test recorded in SGSS, and 6.2% (n=2159) did not;

5) 91.5% (n=35,614) had a positive test recorded in SGSS - of these, 92.0% (n=32,770) had a COVID-19 code in the primary diagnostic position in the HES record straddling randomisation date, and 8.0% (n=2844) did not;
6) 8.7% (n=3391) had no COVID-19 codes in the primary diagnostic position in the HES record straddling randomisation – of these, 63.7% (n=2159) had a positive test recorded in SGSS, and 36.3% (n=1232) did not;
7) 8.5% (n=3306) had no SGSS records – of these, 71.3% (n=2356) had a COVID-19 code in HES, and 28.7% (n=950) did not;

Supplementary table S5 - Cross-coding of COVID-19 in HES and SGSS data within the RECOVERY population

HES COVID-19 coding and SGSS status in RECOVERY (March 2020 – November 2021)		COVID-19 coding in the primary diagnostic position in the HES spell straddling randomisation (U071 or U072)		
		Y	N	Total
SGSS status (Y = positive test; N = negative test or no record)	Y	32,770 (84.2%)	2844 (7.3%)	35,614 (91.5%)
	N	2159 (5.5%)	547 (1.4%)	3306 (8.5%)
	Total	34,929 (89.7%)	3391 (10.3%)	38,920 (100%)*

*all proportions calculated with 38,920 (row and column total, corresponding to the total number of participants with available HES data) as denominator

These data show that, in a population admitted to hospital with clinically-proven or suspected COVID-19 (RECOVERY trial participants), using HES as a single source to identify a potential

COVID-19 cohort (and restricting to the primary diagnostic position) would capture ~91% of the intended population, with very high agreement between HES COVID-19 coding and the presence of a positive test in SGSS. Nonetheless, this would lead to the inclusion of 8.5% of people with no record of a positive test in SGSS (although the absence of a record in SGSS could be due to either a negative test or the absence of a test result, and SGSS only covers in-hospital testing). Conversely, this approach would lose 10.3% of patients recruited to RECOVERY (and 8.0% of those with a positive in-hospital COVID-19 test). Overall, HES seems to be a reliable way of identifying a population admitted to hospital with COVID-19 (even if not using testing data from SGSS).

2. U071 vs U072 coding in HES

COVID-19 can be coded in ICD-10 using either U071 ("COVID-19, virus identified") or U072 ("COVID-19, virus not identified"); of note, other COVID-19 related codes are available in ICD-10 but these are unrelated to acute diagnosis (see https://www.who.int/standards/classifications/classification-of-diseases/emergency-use-icd-codes-for-covid-19-disease-outbreak)

When comparing different COVID-19 coding in HES with SGSS records:

1) The majority of RECOVERY participants with HES records (n=38,920) are coded using the U071 code with or without U072 (92.0%, n=35,812), with only a minority recorded with the U072 code only (2.9%, n=1125);
2) Of those people with no SGSS record (8.5% of those with HES data available, n= 3306), a significant proportion (54.1%; n=1789) had a record of U071 (with or without U072), 30.6% (n=1012) had a record of U072 only, and 20.1% (n=690) had no COVID-19 coding;
3) Of those people with a U072 record only (n=1251), 239 (19.1%) still had a positive SGSS record.

The value of using U072 is in identifying possible COVID-19 patients versus those with no COVID-19 codes. If this code were employed as intended by its definition, it would be expected that its use, when compared with people with no COVID-19 codes, would be more frequent in SGSS-negative patients (those for whom a test was negative or not performed) versus those SGSS-positive. In the RECOVERY population, the odds of having a U072 code recorded in HES versus no code recorded is 0.3 (239:779) in SGSS-positive patients, and 1.28 (886:690) in SGSS-negative patients (odds-ratio: 4.3); this shows an increased likelihood of U072 use in RECOVERY patients in the absence of a positive test (versus

those with a positive test recorded), suggesting a correct use of U072 in cases of suspected COVID-19 with no positive test recorded.

Supplementary table S6 - Cross-coding of COVID-19 in HES (disaggregated across different ICD-10 codes) and SGSS data within the RECOVERY population

HES COVID-19 coding and SGSS status in RECOVERY (March 2020 – November 2021)		COVID-19 coding in any diagnostic position in HES, at any time point (U071 or U072)				
		U071 or U072	U071 (+/- U072)	U072 only	None	Total
SGSS status (Y = positive test; N = negative test or no record)	Y	34,466 (88.6%)	34,023 (87.4%)	239 (0.6%)	779 (2.0%)	35,614 (91.5%)
	N	2477 (6.4%)	1789 (4.6%)	886 (2.3%)	690 (1.8%)	3306 (8.5%)
	Total	36,943 (94.9%)	35,812 (92.0%)	1125 (2.9%)	1469 (3.8%)	38,920 (100%)*

*all proportions calculated with 38,920 (corresponding to the total number of participants with available HES data) as denominator; NB: columns are not mutually exclusive

These observations suggest that U071 is the most widely used code, with only a minority of people coded using U072 only; of these, there is still evidence of a positive COVID-19 test in a significant proportion; moreover, the relative frequency of U072 use in SGSS-positive and SGSS-negative patients points towards correct use by clinical coders. In conclusion, these calculations support the use of U072 alongside U071 in a population admitted to hospital with COVID-19.

Hence, we have defined the national reference population based on the presence of a U071 or U072 code, further restricted to the primary diagnostic position to avoid inclusion of people in whom COVID-19 is not the main reason for care.

Annex IV - Cross-coding of invasive and non-invasive mechanical ventilation in the RECOVERY case report form versus Hospital Episode Statistics data

Methodology:

- RECOVERY participants randomised in England on or before the 30-11-2021 (analysis period for the current study)
- People with HES data available for whom the most recent episode included in the HES data starts on or after randomisation (i.e. HES data covers the period of randomisation)
- HES records restricted to OPCS codes recorded (procedure date) on or before the day of randomisation (inclusive), and using a lookback period of 15 or 30 days before randomisation in the HES data

Population:

- 47,029 distinct participant records in the CRF data
- 46,827 after removing withdrawals or duplicates
- 40,725 recruited in England
- 39,450 randomised within the study period
- 38,025 included in the HES data (all of which recruited in England)

Supplementary table S7 - IMV coding cross tabulation (HES vs CRF) - 15 days before randomisation

HES data		Case report form			Total
		IMV/ECMO	NIV/oxygen	None	All
IMV	N	**1023**	808	12	1843
	% col	**39.9**	2.6	0.3	4.8
None	N	1542	30062	4578	36182
	% col	60.1	97.4	99.7	95.2
All	N	2565	30870	4590	38025
	% col	100.0	100.0	100.0	100.0

Supplementary table S8 - IMV coding cross tabulation (HES vs CRF) - 30 days before randomisation

HES data		Case report form			Total
		IMV/ECMO	NIV/oxygen	None	All
IMV	N	**1125**	942	17	2084
	% col	**43.9**	3.1	0.4	5.5
None	N	1440	29928	4573	35941
	% col	56.1	96.9	99.6	94.5
All	N	2565	30870	4590	38025
	% col	100.0	100.0	100.0	100.0

Supplementary table S9 - NIV coding cross tabulation (HES vs CRF) - 15 days before randomisation

HES data		Case report form			Total
		IMV/ECMO	NIV/oxygen	None	All
NIV	N	1305	8222	129	9656
	% col	50.9	26.6	2.8	25.4
None	N	1260	22648	4461	28369
	% col	49.1	73.4	97.2	74.6
All	N	2565	30870	4590	38025
	% col	100.0	100.0	100.0	100.0

Supplementary table S10 - NIV coding cross tabulation (HES vs CRF) - 30 days before randomisation

HES data		Case report form			Total
		IMV/ECMO	NIV/oxygen	None	All
NIV	N	1438	8948	150	10536
	% col	56.1	29.0	3.3	27.7
None	N	1127	21922	4440	27489
	% col	43.9	71.0	96.7	72.3
All	N	2565	30870	4590	38025
	% col	100.0	100.0	100.0	100.0