**Data sources for modelling:**

Overview:

Below are brief descriptions of the data sheets used as input for the optimal schedule dynamic programming model as well as the processes and methodologies used to produce those data. This document should be read in conjunction with the technical appendix of the report (https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-40-hospital-scheduling/) and the dictionaries of the different datasets used namely: HES, 2020 NHS Electronic Staff Record dataset (<https://digital.nhs.uk/binaries/content/assets/website-assets/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics/hes-data-dictionary/dd-apc_v12.pdf>) ; March 2020 Quarterly Bed Availability and Occupancy Dataset (KH03 dataset) (<https://www.england.nhs.uk/statistics/wp-content/uploads/sites/2/2020/06/KH03-Guidance-June-2020-FINAL.pdf>) ; February 2020 Critical Care Monthly Situation Reports dataset (<https://digital.nhs.uk/binaries/content/assets/legacy/pdf/h/t/msitreps_guidance.pdf> ; Organisational Reference Cost Schedule (RCS), and the National RCS (<https://www.england.nhs.uk/national-cost-collection/>) .

Data processing and release were in accordance with the Code of Practice at Imperial College London (<https://www.imperial.ac.uk/admin-services/secretariat/information-governance/data-protection/our-policy/codes-of-practice/code-of-practice-1---handling-of-personal-data/>)

Data used by Health Foundation staff for this analysis were anonymised in line with the ICO's Anonymisation Code of Practice. The data was accessed in The Health Foundation's Secure Data Environment; a secure data analysis facility (accredited with the ISO-27001 information security standard, and recognised for the NHS Digital Data Security and Protection Toolkit). No information that could directly identify a patient or other individual was used. Statistical results were released in accordance with the [Handbook for Statistical Disclosure Control for Outputs](https://ukdataservice.ac.uk/media/622521/thf_datareport_aw_web.pdf).

Datasets used and samples

Data on hospital activity are estimated from HES APC and CC data extracted for the period 1 January 2015 – 29 February 2020 from NHS Acute Trusts with an organisational code beginning with R. Due to the fundamental differences in operations and management, we excluded any NHS Mental Health Acute Trusts. An overview of the variables used can be found in Table 1.

Exclusion criteria for APC include

* Method of admissions was not elective or emergency
* Admissions to mental health trusts
* Patient age, encrypted HES ID, discharge date or primary diagnosis was missing
* Discharge method was stillborn baby or patient was still in hospital
* Episode ended before it started
* Negative waiting time
* Waiting time > 365 days.
* Exclusions of specific age/diagnosis combinations (ICD15 patients > 64 years old and ICD16 patients > 24 years old, both of which are clinically unlikely)

Exclusion criteria for CC include

* Episode ended before it started
* Ccunitfun is 1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 90 or 91
* Epistart or epiend was missing
* Episode did not overlap with any APC episode

The resulted in a dataset with N= 39,2328,605 observations over the period 01/01/2015 - 29/02/2020.

The other datasets were used on its entirety with no further exclusion criteria beyond the year cut offs as explained above.

Data release:

For data prepared using the *time\_series\_forecast.R* script (phi1-3, pi\_z, ICDprop, Frailty), the values are projections generated through linear trend models, rather than aggregates of HES data, hence there are no minimum group-size limits for reporting. For elective patient groups, data was forecast from 31st March 2019, as opposed to 29th February 2020 for emergency patient groups. This was to ensure forecasts started from a representative baseline of waiting patients, given the waiting time of patients is only recorded in HES upon discharge.

For the pi\_y transition probabilities, for Emergency patients these were calculated as proportions of each transitions occurring. Namely patients could remain in G&A, move to CC (& vice versa for CC patients), be discharged or die, with the proportions of each of these transitions occurring per week calculated. As such the minimum group size required for a given week’s patient group (a given ICD and age group combination) was 10. Any patient group/week combination that contained <10 observations was not released. For Elective transitions, these were derived using a multinomial logistic regression, with a minimum required group size of 5. For the pi\_x transitions, these were calculated from an aggregated survival function model, with patient groups with <10 observations excluded (no patient groups fell below this threshold).

For the stocks of patients, these were calculated by patient group, with any patient group with <10 observations excluded.

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 1. Input variables calculated via HES data and used for projections of non-COVID emergency admissions and new patients waiting for elective care** | | | |
| **Variable name** | **Description** | **Type** | **What HES variable is this linked to?** |
| GA\_LoS | Length of stay in G&A  bed | integer | Ccdisdate, ccstartdate, epistart and epiend |
| cc\_LoS | Length of stay in cc bed | integer | Ccdisdate and ccstartdate |
| admidate\_MDY | Admission date (DD/MM/YYY) | String “%Y-%m-%d” (i.e “2020-07-23") | admidate |
| epistart\_YYYY | Year of episode start (e.g. 2015) | integer | admidate |
| epistart\_week | Week of episode start date (1-52), using ISO8601 definition | integer | admidate |
| admimeth\_C | Admission method category (1 if elective, 2 if emergency) | integer | admimeth |
| hesid | Unique patient identifier (pseudonymised) | string | hesid |
| diag\_01 | Primary ICD-10 code (4-character code) | string | diag\_01 |
| rttstart | Start date for referral to treat (DD/MM/YYY) | string "%Y-%m-%d" (i.e “2020-07-23") | rttperstart |
| rttstart\_week | Start week of referral to treat as 1-52 for week of year, using ISO8601 definition | integer | rttperstart |
| rttstart\_YYYY | Start calendar year of referral | integer | rttperstart |
| ICD | Our ICD groupings for  elective and emergencies with separate bundles for both | integer | diag\_01 and admimeth |
| agegrp\_v3 | Age groups (1 if 0-24-year-olds, 2 if 25-64-year-olds; 3 if 65+-year-olds) | integer | startage |
| WaitingTime | Length that an elective waiting for referral (rtt) to admission for treatment | integer | This is calculated as the difference between admidate\_MDY and rttstart\_MDY |
| MainICD10Cat | ICD-10 grouping number for main cause of admission | integer | diag\_01 |
| disdate\_MDY | Discharge date for episode | string "%Y-%m-%d" (i.e “2020-07-23") | disdate |
| cc | Is patient admitted into CC. (1 if there is a cc record attached determined by the presence of ccstartdate; 0 otherwise). | binary | ccstartdate |
| Frail | Is patient frail based on Custom Imperial Frailty Score (1 if frail; 0 otherwise) | binary | diag\_01-diag\_20 |
| ccdisdest | Destination of patient following discharge from critical care (see HES data dictionary for destination details specific to each HES year) | integer | ccdisdest |
| cc\_dis\_flg | Flag if discharge date from critical care is equal to discharge date on episode line (1 if ccdisdate = disdate, 0 otherwise) | binary | ccdisdate and disdate |
| cc\_start\_flg | Flag if admission date to critical care is equal to admission date on episode line (1 if admidate = ccstartdate, 0 otherwise) | binary | ccstartdate and admidate |

**Optimization model data inputs found in Optimization\_model\_input\_data.xlsx:**

**Non-COVID forecasts**

Sheet names: phi1, phi2, phi3

Relevant R Script: *time\_series\_creator.R* and *time\_series\_forecast.R*

Data sources: HES APC and CC datasets. Variables used are derived from the following original HES variables:

* epistart
* startage
* diag01-diag20
* rttperstart
* disdate
* admidate
* admimeth
* hesid

Data description: Phi1, phi2, and phi3 provide data on the weekly number of forecasted non-COVID emergency admissions and newly waiting elective patients based on reproduction numbers equal to 1.0, 1,1, and 1,2, respectively. These estimates were calculated using time-series forecasting techniques (see Appendix C1 of the report).

Columns: Each sheet has the following columns:

* *t* is the forecast week number, ranging from 0 (March 2, 2020) to 77 (August 23rd, 2021)
* *a* is the admission type, where:
  + N = electives
  + E = emergencies
* *ICDXX\_AGEYY* refers to the patient group’s forecasted weekly number of admissions or newly waiting elective patients. Each patient group refers to a specific ICD group XX (Table 2) and age group YY, where:
  + AGE01, 02, and 03 refer to the following age-groups: 0-24, 25-64, 65+
  + e.g., ICD02\_AGE01 refers to Neoplasms for 0-24-year-olds

|  |  |  |
| --- | --- | --- |
| **Table 2. ICD root group identification** | | |
| **ICD Group Number** | **ICD-10 Chapter** | **Disease Category** |
| 01 | [A00–B99](https://en.wikipedia.org/wiki/ICD-10_Chapter_I:_Certain_infectious_and_parasitic_diseases) | Certain infectious and parasitic diseases |
| 02 | [C00–D48](https://en.wikipedia.org/wiki/ICD-10_Chapter_II:_Neoplasms) | Neoplasms |
| 03 | [D50–D89](https://en.wikipedia.org/wiki/ICD-10_Chapter_III:_Diseases_of_the_blood_and_blood-forming_organs,_and_certain_disorders_involving_the_immune_mechanism) | Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism |
| 04 | [E00–E90](https://en.wikipedia.org/wiki/ICD-10_Chapter_IV:_Endocrine,_nutritional_and_metabolic_diseases) | Endocrine, nutritional and metabolic diseases |
| 05 | [F00–F99](https://en.wikipedia.org/wiki/ICD-10_Chapter_V:_Mental_and_behavioural_disorders) | Mental and behavioural disorders |
| 06 | [G00–G99](https://en.wikipedia.org/wiki/ICD-10_Chapter_VI:_Diseases_of_the_nervous_system) | Diseases of the nervous system |
| 07 | [H00–H59](https://en.wikipedia.org/wiki/ICD-10_Chapter_VII:_Diseases_of_the_eye,_adnexa) | Diseases of the eye and adnexa |
| 08 | [H60–H95](https://en.wikipedia.org/wiki/ICD-10_Chapter_VIII:_Diseases_of_the_ear_and_mastoid_process) | Diseases of the ear and mastoid process |
| 09 | [I00–I99](https://en.wikipedia.org/wiki/ICD-10_Chapter_IX:_Diseases_of_the_circulatory_system) | Diseases of the circulatory system |
| 10 | [J00–J99](https://en.wikipedia.org/wiki/ICD-10_Chapter_X:_Diseases_of_the_respiratory_system) | Diseases of the respiratory system |
| 11 | [K00–K93](https://en.wikipedia.org/wiki/ICD-10_Chapter_XI:_Diseases_of_the_digestive_system) | Diseases of the digestive system |
| 12 | [L00–L99](https://en.wikipedia.org/wiki/ICD-10_Chapter_XII:_Diseases_of_the_skin_and_subcutaneous_tissue) | Diseases of the skin and subcutaneous tissue |
| 13 | [M00–M99](https://en.wikipedia.org/wiki/ICD-10_Chapter_XIII:_Diseases_of_the_musculoskeletal_system_and_connective_tissue) | Diseases of the musculoskeletal system and connective tissue |
| 14 | [N00–N99](https://en.wikipedia.org/wiki/ICD-10_Chapter_XIV:_Diseases_of_the_genitourinary_system) | Diseases of the genitourinary system |
| 15 | [O00–O99](https://en.wikipedia.org/wiki/ICD-10_Chapter_XV:_Pregnancy,_childbirth_and_the_puerperium) | Pregnancy, childbirth and the puerperium |
| 16 | [P00–P96](https://en.wikipedia.org/wiki/ICD-10_Chapter_XVI:_Certain_conditions_originating_in_the_perinatal_period) | Certain conditions originating in the perinatal period |
| 17 | [Q00–Q99](https://en.wikipedia.org/wiki/ICD-10_Chapter_XVII:_Congenital_malformations,_deformations_and_chromosomal_abnormalities) | Congenital malformations, deformations and chromosomal abnormalities |
| 18 | [R00–R99](https://en.wikipedia.org/wiki/ICD-10_Chapter_XVIII:_Symptoms,_signs_and_abnormal_clinical_and_laboratory_findings) | Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified |
| 19 | [S00–T98](https://en.wikipedia.org/wiki/ICD-10_Chapter_XIX:_Injury,_poisoning_and_certain_other_consequences_of_external_causes) | Injury, poisoning and certain other consequences of external causes |
| 20 | [V01–Y98](https://en.wikipedia.org/wiki/ICD-10_Chapter_XX:_External_causes_of_morbidity_and_mortality) | External causes of morbidity and mortality |
| 21 | [Z00–Z99](https://en.wikipedia.org/wiki/ICD-10_Chapter_XXI:_Factors_influencing_health_status_and_contact_with_health_services) | Factors influencing health status and contact with health services |
| 50\* | - | Bundled ICD Group for electives |
| 51\*\* | - | Bundled ICD Group for emergencies |
| \* ICD01, 04, 05, 08, 15, 16, and 17 were bundled for electives  \*\* ICD03, 07, 08, 16, 17, and 21 were bundled for emergencies | | |

**Transition probabilities**

Sheet name: pi\_x

Relevant R Script: *regression\_analyses.R*

Data sources: HES APC datasets. Variables used are derived from the following original HES variables:

* rttperstart
* diag01
* admidate
* admimeth
* hesid

Data description: Pi\_x contains the probabilities (*π\_e\_wp*) that an individual within patient group *p* who is not admitted as an elective in a certain week may be admitted as an emergency in that week conditional on having already waited for a certain time period. These probabilities were calculated using survival analysis methods (see Appendix E1).

Columns:

* *a* is the admission type, where N refers to electives and E refers to emergencies
* *p* is the patient group (ICDXX\_AGEYY)
* *pi\_x* is the probability that a patient switches from waiting for elective care to being admitted as an emergency in the following week

Sheet name: pi\_y

Relevant R Script: *regression\_analyses.R* and *transistions\_coding.R*

Data sources: HES APC & CC datasets. Variables used are derived from the following original HES variables:

* rttperstart
* diag01
* admidate
* admimeth
* hesid
* startage
* ccstartdate
* ccdisdate
* epistart
* epiend
* dismeth
* disdate
* ccdisdest

Data description: Pi\_y contains the transition probabilities (*π\_SS’\_y,ap*) patients can undergo once admitted to the hospital. Patients can either be admitted to G&A and transition to being discharged (I.e., recovery), CC, or die, or can also be admitted to CC and transition into recovery, (return) to G&A, or die. These probabilities were calculated using multinomial logit regressions (see Appendix E2).

Columns:

* *a* is the admission type, where N refers to electives and E refers to emergencies
* *p* is the patient group (ICDXX\_AGEYY)
* *s* is the start state of the patient, where:
  + G = G&A patients who can transition to CC
  + C = CC patients
  + G\_STAR = G&A patients who cannot transition to CC
* *sbar* is the end state of the patient, where:
  + H = patients who are discharged/recovery
  + C = CC patients
  + D = patients who die
  + G = G&A patients who can transition to CC
  + G\_STAR = G&A patients who cannot transition to CC
* *pi\_y* is the probability that a patient transition from *s* to *sbar*
* *week* is the week at which each transition probability was estimated. Week values are not contiguous: *either* the week sequence [0.5,1.5,2.5,3.5,4.5] *or* the sequence [1,2,3,4.5] must be used.

**Forecasted proportions**

Sheet name: pi\_z

Relevant R Script: *time\_series\_creator.R* and *time\_series\_forecast.R*

Data sources: HES APC and CC datasets. Variables used are derived from the following original HES variables:

* epistart
* startage
* diag01
* rttperstart
* disdate
* admidate
* admimeth
* hesid

Data description: Pi\_z contains the proportion of patients who are directly admitted to CC. These proportions are forecasted using a local linear trend model with trigonometric seasonality (Appendix C1). They are used to estimate the number of elective and emergency patients in each patient group split by G&A vs. CC entry point.

Columns:

* *t* is the forecast week number
* *a* is the admission type
* *ICDXX\_AGEYY* refers to the patient group’s forecasted proportion of patients directly admitted to CC

Sheet name: ICD\_prop

Relevant R Script: *time\_series\_creator.R* and *time\_series\_forecast.R*

Data sources: HES APC and CC datasets. Variables used are derived from the following original HES variables:

* epistart
* startage
* diag01
* rttperstart
* disdate
* admidate
* admimeth
* hesid

Data description: ICD\_prop contains the proportion of patients belonging to each patient group within each ICD bundle (grouped due to small sample size, see Appendix B and Table B2). These proportions are forecasted using a local linear trend model with trigonometric seasonality (Appendix C1). They are used to estimate the number of patients belonging to the electives bundle who switch from waiting for elective care to being admitted as an emergency (Appendix E1).

Columns:

* *t* is the forecast week number
* *a* is the admission type
* *ICDXX\_AGEYY* refers to the forecasted proportion of patients belonging to the specific patient group within the electives bundle

Sheet name: frailty

Relevant R Script: *time\_series\_creator.R* and *time\_series\_forecast.R*

Data sources: HES APC and CC datasets. Variables used are derived from the following original HES variables:

* epistart
* startage
* diag01-diag20
* rttperstart
* disdate
* admidate
* admimeth
* hesid

Data description: Frailty provides data on the weekly number of frail patients, as defined in forecasted non-COVID emergency admissions and newly waiting elective patients based on reproductive number. These estimates were calculated using time-series forecasting techniques (see Appendix C1).

Columns:

* *t* is the forecast week number
* *a* is the admission type
* *ICDXX\_AGEYY* is the forecasted proportion of frail patients per patient group

**Stock of patients**

Excel sheets names: x0, y0

Relevant R Script: *time\_series\_creator.R*

Data sources: HES APC and CC. Variables used are derived from the following original HES variables:

* diag01
* admidate
* epistart
* admimeth
* disdate
* rttperstart
* hesid

Data description: X0 refers to the stock of patients waiting for care (*w\_0p*) at the beginning of week zero (Appendix A2.2). Y0refers to the stock of patients already in hospital care in state *s* (y0), admitted prior to week zero but not yet discharged by week one.

Columns: Each sheet has the following columns:

* *a* is the admission type, where N refers to electives and E refers to emergencies
* *p* is the patient group (ICDXX\_AGEYY)
* *x0* is the stock of patients waiting for care (*x0* sheet only)
* *s* is the start state of the patient (*y0* sheet only)
* *y0* is the stock of patients already in hospital care (*y0* sheet only)

**Costing**

Sheet name: costs

Relevant R Script: *cost\_merging\_HF.R*

Data Sources: HES APC and CC datasets, Organisational Reference Cost Schedule (RCS), and the National RCS.

Variables used are derived from the following original HES variables:

* procode
* SUSHRG
* admimeth
* epiend
* startage
* diag\_01
* rttperstart
* admidate

Variables used from the RCS datasets:

* unitcost
* currencycode
* orgcode

Data description: Costs contains the average unit cost for each patient group and admission type (Appendix F1). These were calculated by individually costing each HES patient using their Healthcare Resource Group (HRG) code. Non-COVID patients were each assigned an HRG and were costed using publicly available data from NHS Digital. Since COVID HRGs were not yet published, we generated COVID HRGs using the HRG 4+ 2020/21 Local Payment Grouper, publicly available from NHS Digital.

Columns:

* *a* is the admission type
* *p* is the patient group (ICDXX\_AGEYY)
* *avg\_cost* is the average unit cost for each patient group/admission type
* *sd\_cost* is the standard deviation for the unit cost

**Hospital Capacity**

Sheet names: staff, staff\_to\_bed, extra\_capacity

Relevant R Script: *N/A*

Data Sources: 2020 NHS Electronic Staff Record (ESR) dataset; March 2020 Quarterly Bed Availability and Occupancy Dataset (KH03 dataset); February 2020 Critical Care Monthly Situation Reports dataset.

Data description: Staff refers to the number of G&A and CC senior doctor, junior doctor, and nurse FTE equivalents employed at all NHS hospitals during the month of April 2020 and an average using monthly data from January to April 2020. Beds\_to\_staff refers to the maximum beds to staff ratio allowances for senior doctors, junior doctors, and nurses across G&A and CC. These ratios were taken from guidelines set by the Royal College of Physicians and the Royal College of Nursing. Extra\_capacity refers to the estimated number of units increased (e.g. CC beds, junior doctors in G&A, etc) due to various policies implemented by the English government.

Columns:

staff

* *occupation* is the type of healthcare worker (Senior Doctor, Junior Doctor, or Nurse)
* *ward* is the admission setting (G&A or CC)
* *FTE\_april* is the number of FTEs working during April 2020
* *FTE\_avg* is the number of monthly FTEs working averaged between January and April 2020

beds\_to\_staff

* *occupation* is the type of healthcare worker (Senior Doctor, Junior Doctor, or Nurse)
* *ward* is the admission setting (G&A or CC)
* *beds\_to\_staff* is the maximum number of beds managed by one unit of staff member

extra\_capacity

* *policy* is the type of policy related to each type of capacity increase (e.g. Field Hospitals, Return of Staff, etc)
* *ward* is the admission setting (G&A or CC)
* *capacity\_type* is the type of capacity (e.g. bed, Nurse, Junior Doctor, or Senior Doctor)
* *increase* is the amount increased of the capacity type