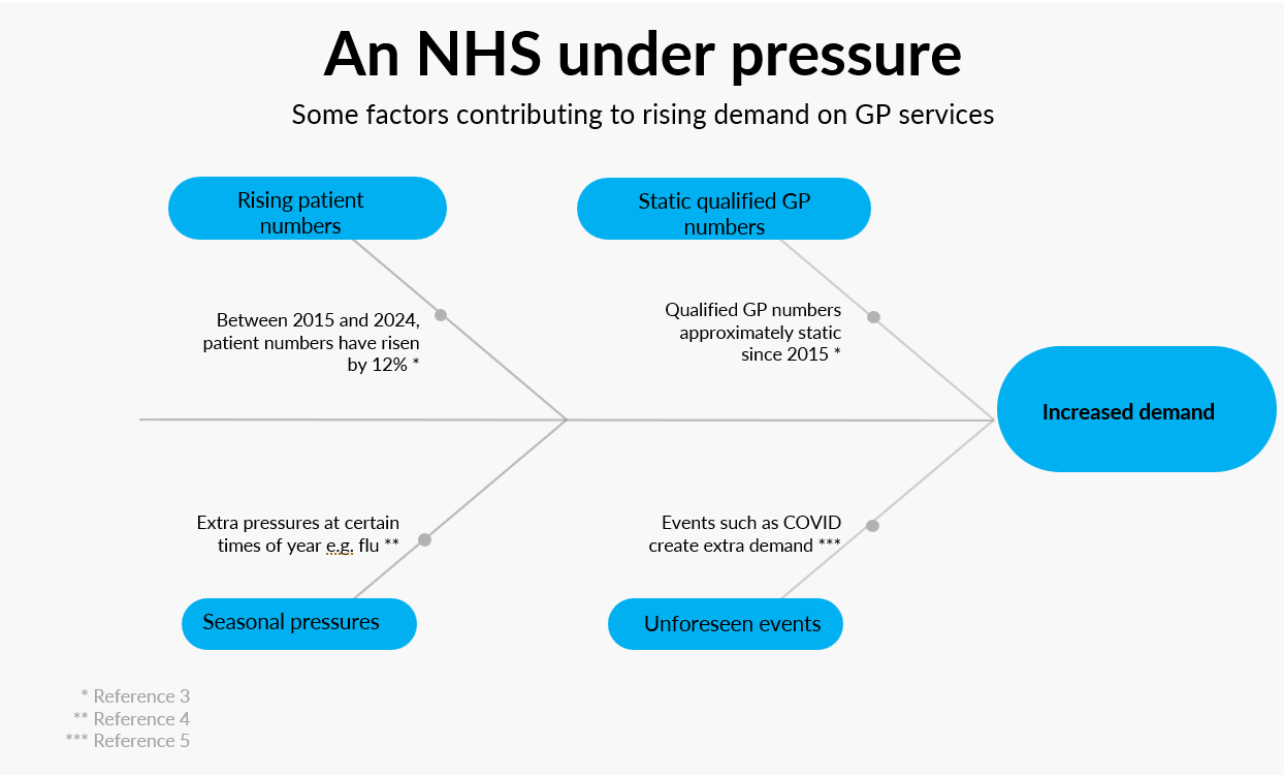
Data Analytics using Python – Assignment submission

The NHS is facing increased demand (reference 1). This diagram shows some factors responsible.



Our goal is to answer these questions

1. Does the NHS have sufficient capacity for appointments?
2. What is the impact of missed appointments?
3. What can trending hashtags from X tell us about views on the NHS?

How will we determine whether there is sufficient capacity?

By comparing the number of appointments in a certain period with the planned capacity for that period (reference 2)

How will we determine the impact of missed appointments?

By examining the percentage of appointments missed and how this varies by time and appointment mode.

Jonathan Shields

Analysis of appointment data

Python was selected because it is an industry-standard platform providing popular libraries for analysis and visualization.

Data characteristics

The data are grouped differently in the three files, with each grouping having a count of appointments. The number of distinct values of the categories was examined to inform further analysis.

Category	Number of distinct values
NHS Region	7
ICB	42
Sub-ICB	106
Service settings	5 (Primary Care Network, Other, General Practice, Unmapped, Extended Access Provision)
Context types	3 (Care Related Encounter, Unmapped, Inconsistent Mapping)
National Category	18
Appointment Status	3 (Attended, DNA, Unknown)

Data quality

The actual appointment data contained no missing values or duplicates.

```
# Determine whether there are missing values in the actual appointments data
ad_na=ad[ad.isna().any(axis=1)]
ad_na.shape
```

```
(0, 8)
```

```
#Perform a similar check for nulls
ad_null=ad[ad.isnull().any(axis=1)]
ad_null.shape
```

```
(0, 8)
```

```
#Check for duplicates
ad_dupes=ad[ad.duplicated()]
ad_dupes.shape
```

```
(0, 8)
```

The regional appointment data contained no missing values. There were records with the same grouping of categories but differing appointment counts. As they have differing appointment counts they have been included and not treated as duplicates.

```
# Determine whether there are missing values in the regional appointments data set
ar_na=ar[ar.isna().any(axis=1)]
ar_na.shape
```

```
(0, 7)
```

```
#Perform a similar check for nulls
ar_null=ar[ar.isnull().any(axis=1)]
ar_null.shape
```

```
(0, 7)
```

```
#Check for duplicates
ar_dupes=ar[ar.duplicated()]
ar_dupes.shape
```

```
(21604, 8)
```

The national categories data did not contain any missing values or duplicates.

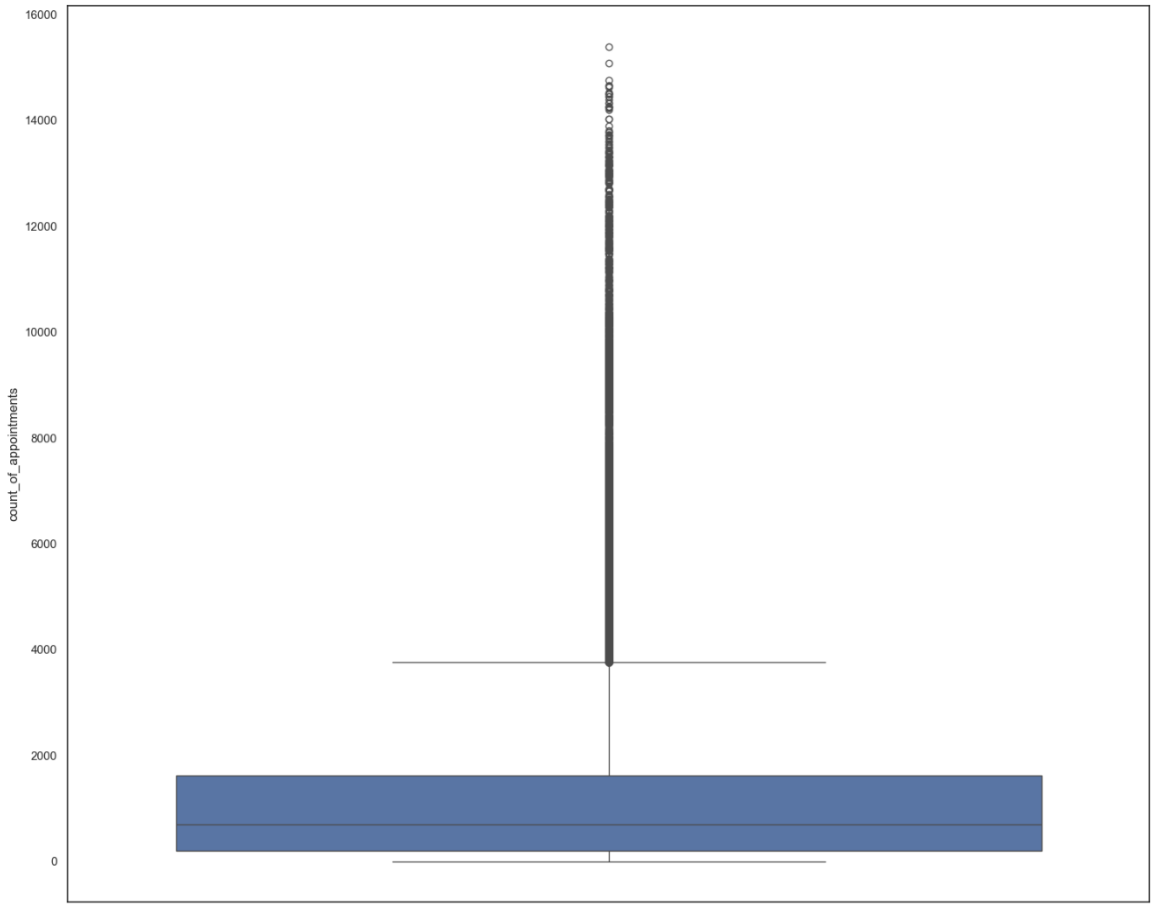
Jonathan Shields

```
[0]: # Determine whether there are missing values in the national categories dataset.
     nc_na=nc[nc.isna().any(axis=1)]
     nc_na.shape

[0]: (0, 8)

[1]: # Perform a similar check for nulls
     nc_null=nc[nc.isnull().any(axis=1)]
     nc_null.shape

[1]: (0, 8)

[5]: #Check for duplicates
     nc_dupes=nc[nc.duplicated()]
     nc_dupes.shape
```

Outliers

All 3 datasets showed a lot of outliers above 1.5 x IQR(Interquartile Range) for the count of appointments. No justification for removing these is apparent so they were allowed to remain.

The actual appointments (count of appointments) exploratory boxplot is shown below as an example.



Jonathan Shields

Date ranges covered by the data

The date ranges in the three data sources differ. Actual appointments covers December 2021 to June 2022 inclusive. Regional appointments covers January 2020 to June 2022 inclusive, and national categories covers August 2021 to June 2022 inclusive.

The largest period covered by all 3 datasets is December 2021 to June 2022 inclusive.

```
#Actual appointments
#Add a column to the dataframe to convert the appointment_date object into a python datetime using 01-Dec-21 format
ad['appointment_date_datetime']=pd.to_datetime(ad['appointment_date'],format='%d-%b-%y')
print(f"Appointment date in actual appointments file covers range {ad['appointment_date_datetime'].min()} to {ad['appointment_date_datetime'].max()}")
```

```
Appointment date in actual appointments file covers range 2021-12-01 00:00:00 to 2022-06-30 00:00:00
```

```
#Regional appointments
#Add a new column converting using 2021-01 format, leaving day to default to 1st
ar['appointment_month_datetime']=pd.to_datetime(ar['appointment_month'],format='%Y-%m')
print(f"Appointment date in regional appointments file covers range {ar['appointment_month_datetime'].min()} to {ar['appointment_month_datetime'].max()}"
```

```
Appointment date in regional appointments file covers range 2020-01-01 00:00:00 to 2022-06-01 00:00:00
```

```
#National categories date range
#Add a new column converting using dd/mm/yyyy format
nc['appointment_date_datetime']=pd.to_datetime(nc['appointment_date'],format='d%/m%/%Y')
print(f"Appointment date in national categories file covers range {nc['appointment_date_datetime'].min()} to {nc['appointment_date_datetime'].max()}")
```

```
Appointment date in national categories file covers range 2021-08-01 00:00:00 to 2022-06-30 00:00:00
```

What are the descriptive statistics of the number of appointments by month?

Data source	Mean	Pop. standard deviation	Range
Actual	23,997,242	1,605,356	5,221,188
Regional	24,760,151	3,414,396	14,397,189
National categories	26,913,343	2,346,375	6,552,899

As the regional file includes the initial COVID period a greater range and standard deviation is perhaps expected.

Specific initial investigations

Do the files show the same number of appointments for the dates they overlap?

The regional and national categories file have the same number of appointments (182,963,194) for December 2021 to June 2022 inclusive.
The actual appointments file has fewer appointments (167,980,692) but it is assumed this is due to missed appointments (status DNA) not being recorded.

Jonathan Shields

Comparison of the number of appointments in each file in the overlapping range

```
#Create variables for the start and end of the date range for which the 3 files overlap
start=datetime.strptime("2021-12-01","%Y-%m-%d")
end=datetime.strptime("2022-06-30","%Y-%m-%d")

#Find the number of appointments from the actual appointments file in the range
ad_date_subset=ad[(ad['appointment_date']>=start) & (ad['appointment_date']<=end)]
ad_appts=ad_date_subset["count_of_appointments"].sum()

#Find the number of appointments from the regional appointments file in the range
ar_date_subset=ar[(ar['appointment_month']>=start) & (ar['appointment_month']<=end)]
ar_appts=ar_date_subset["count_of_appointments"].sum()

#Find the number of appointments from the national categories file in the range
nc_date_subset=nc[(nc['appointment_date']>=start) & (nc['appointment_date']<=end)]
nc_appts=nc_date_subset["count_of_appointments"].sum()

print(f"In the range {start.strftime("%Y-%m-%d")} to {end.strftime("%Y-%m-%d")} the actual appointments file has {ad_appts} appointments,\n"
        + f"the regional appointments file has {ar_appts} and the national categories file has {nc_appts}")
```

```
In the range 2021-12-01 to 2022-06-30 the actual appointments file has 167980692 appointments,
the regional appointments file has 182963194 and the national categories file has 182963194
```

Which month had the highest number of appointments?

November 2021 had the largest number of appointments.

```
#group by appointment date month and year
nc_month_group=nc.groupby([nc.appointment_date.dt.month_name(),nc.appointment_date.dt.year])
#sum the count of appointments for each group
nc_month_group["count_of_appointments"].sum().sort_values(ascending=False)
```

```
appointment_date  appointment_date
November          2021.0              30405070
October           2021.0              30303834
March             2022.0              29595038
September         2021.0              28522501
May               2022.0              27495508
June              2022.0              25828078
January           2022.0              25635474
February          2022.0              25355260
December          2021.0              25140776
April             2022.0              23913060
August            2021.0              23852171
Name: count_of_appointments, dtype: int64
```

How many appointments are missed in the regional data?

4.16% of appointments were missed between January 2020 and June 2022.

Additional question: how many missed appointments were there by month and what is the percentage of all appointment missed?

```
#Calculate the percentage of missed (DNA) appointments as sum of missed/all
all_appointments=ar["count_of_appointments"].sum()
missed_appointments=ar[ar["appointment_status"]=="DNA"]["count_of_appointments"].sum()

missed_percentage=(missed_appointments/all_appointments)*100

print(f"{round(missed_percentage,2)}% of appointments were missed (DNA), representing {missed_apppointments} appointments \
from a total of {all_appointments}")
```

```
4.16% of appointments were missed (DNA), representing 30911233 appointments from a total of 742804525
```

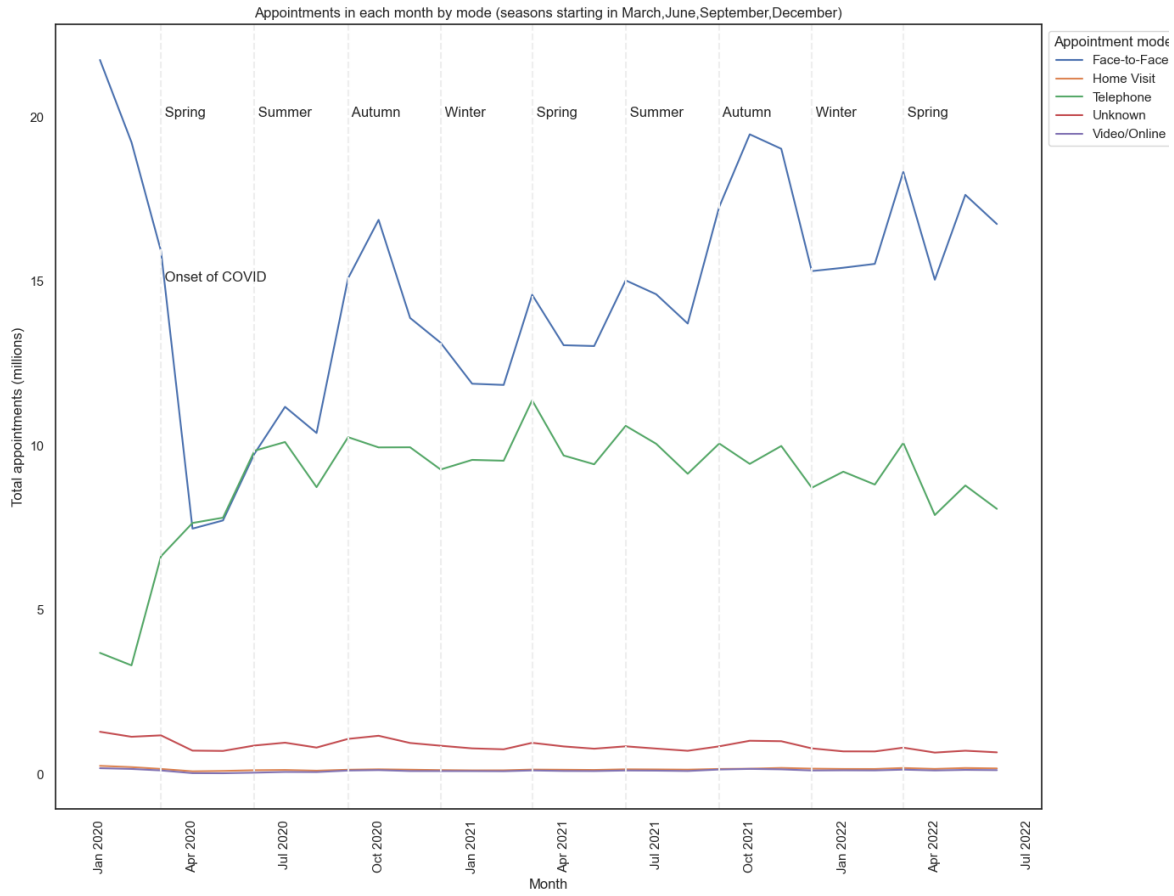
Analysis and Visualisations

The visualizations focus on the seasonal variation of appointments for different categorical variables. Seaborn was used as it creates visuals with little code. Matplotlib was used where finer control was required. Lineplots are used as they are suited to displaying variation over time.

Jonathan Shields

Appointments by mode (January 2020 to June 2022)

The regional file allows us to explore a longer period as an explanatory visual. The number of appointments is plotted against the month below. A different coloured line is shown for each appointment mode.

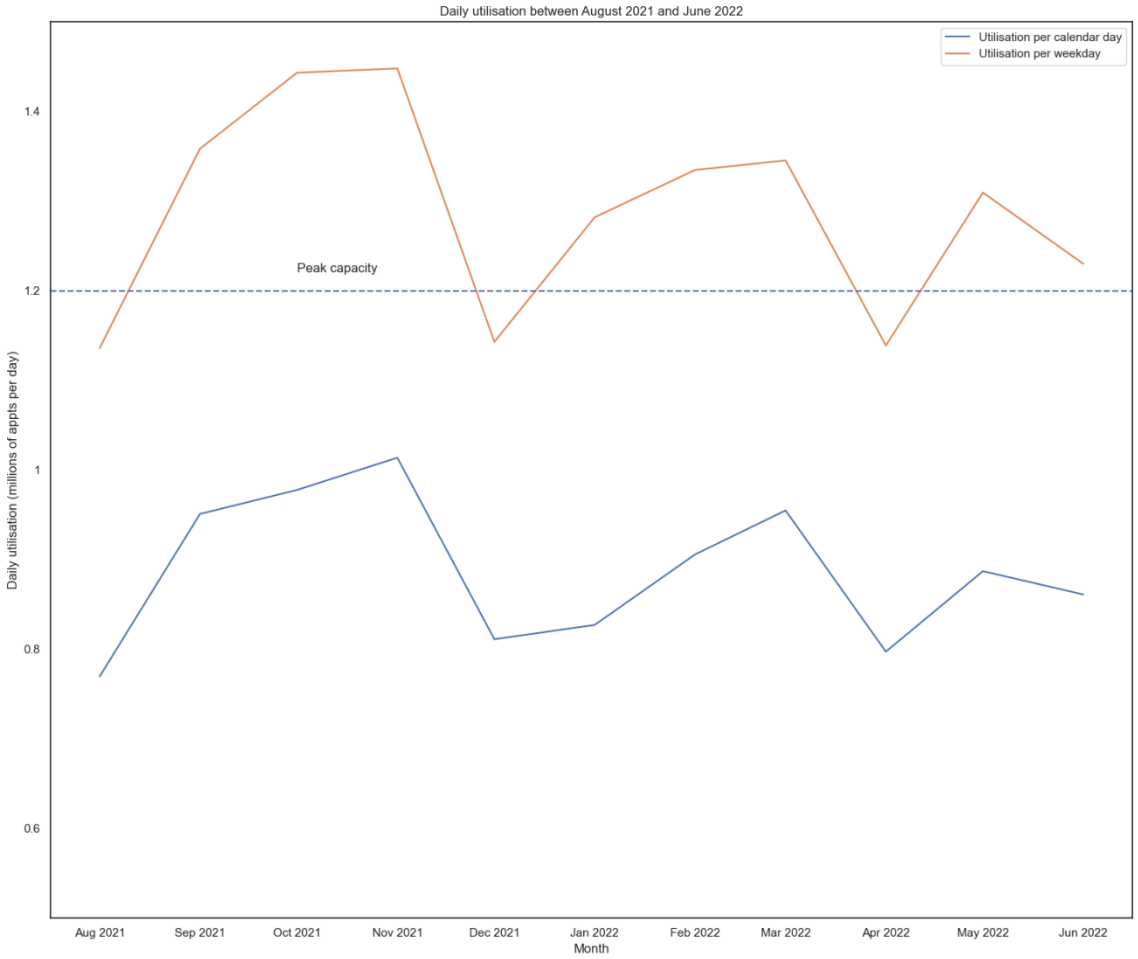


This shows a large dip in face-to-face and a rise in telephone appointments due to the COVID "lockdown" in Spring 2020. Post-COVID we see spikes in face-to-face appointments in autumn but we do not see a similar spike in telephone appointments. Telephone and face-to-face appointments show a spike in spring.

The rising trend in patient numbers (reference 3) and seasonal pressures due to respiratory illness (references 4,7) may explain these seasonal peaks.

Jonathan Shields

Utilisation against capacity between August 2021 and June 2022



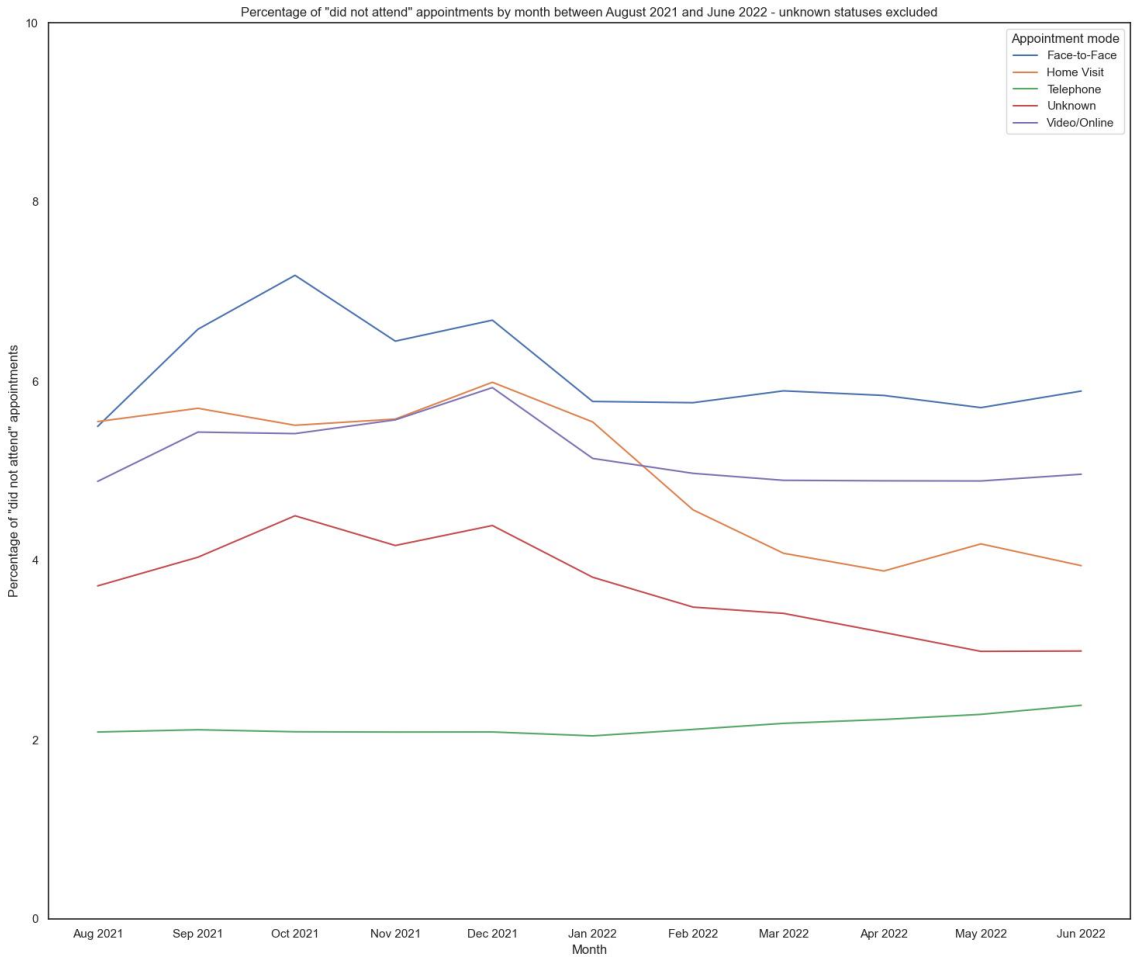
The chart above shows utilisation per month between August 2021 and June 2022.

We have defined utilisation as the **appointments per weekday**. As only 0.65% of the actual appointments were on weekends we felt this more accurate than using appointments per calendar day. However, the chart also plots utilisation based on calendar days for reference.

Missed appointments between August 2021 and June 2022

The "Unknown" status was excluded as it cannot be said to be either attended or unattended.

Jonathan Shields



This shows the percentage of missed appointments by month and appointment mode. We see that face-to-face appointments are missed the most and telephone the least. There is an Autumn spike for face-to-face appointments while telephone appointments do not show any seasonal change.

Jonathan Shields

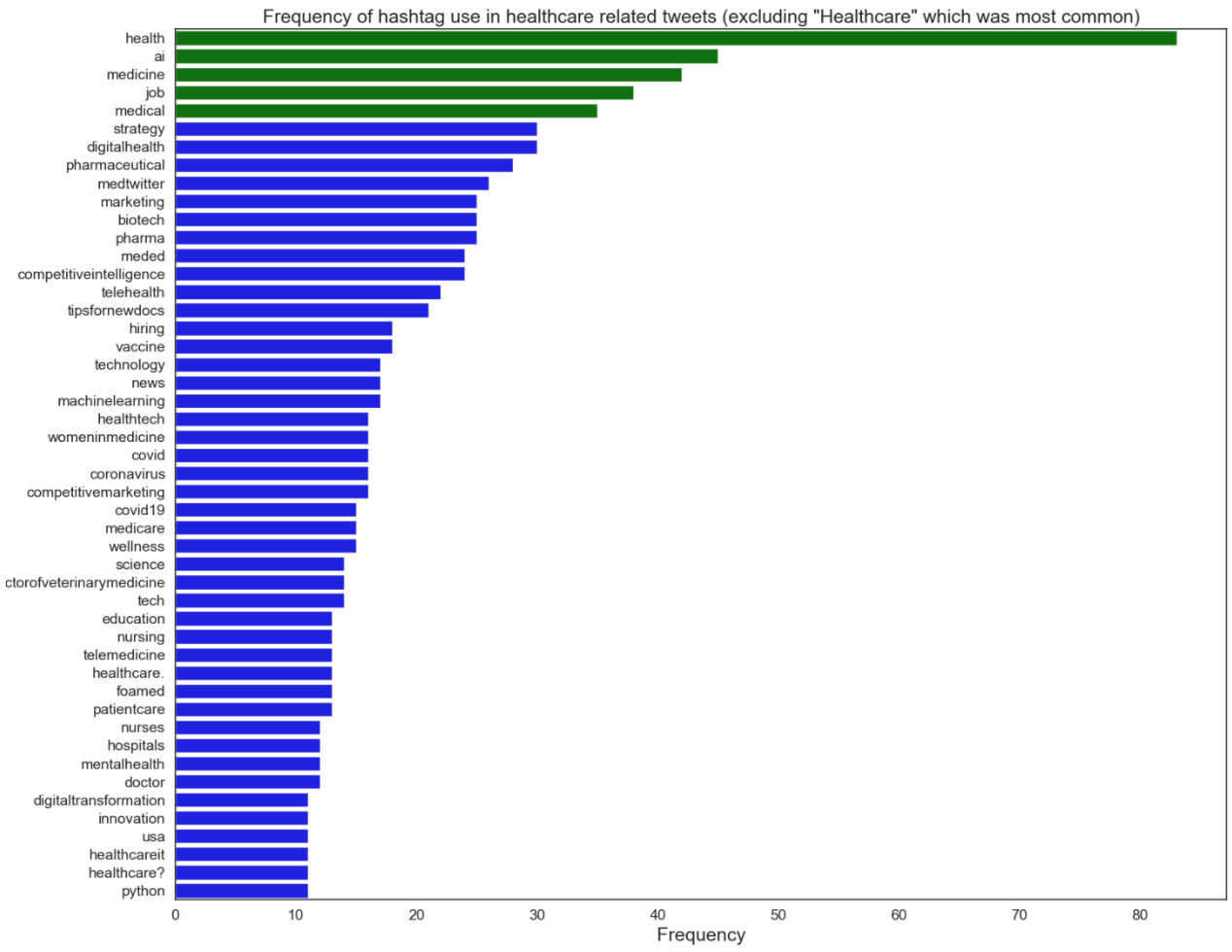
Demand across appointment modes

This chart shows the variation in face-to-face and telephone appointments only for clarity as these are the most common modes by far. Face-to-face shows a pronounced increase in October/November. There is less variation in telephone appointments although there is a peak in March.

Analysis of tweets on X

Analysis of tweets to identify the most common hashtags did not yield anything relevant to the resourcing. The six most common hashtags were “healthcare”, “health”, ”ai”, ”medicine”, “job” and “medical” as illustrated below. This barplot clearly shows the distribution across the hashtags.

Jonathan Shields



Conclusions

Does the NHS have sufficient capacity for appointments?

Planned capacity was exceeded in 8 of the 11 months between August 2021 and June 2022. This was most pronounced in October, November, and March. This is calculated using the 1.2 million figure per day and the **weekdays** per month.

This suggests that appointments are exceeding planned capacity at present.

What is the impact of missed appointments?

Between August 2021 and June 2022, 4.69% of appointments were missed. In November 2021, the busiest month, the figure was 4.9%. This is a concern and equates to around 1 in 20 appointments being wasted. There were more face-to-face missed appointments (6.14%) than telephone appointments (2.15%).

Jonathan Shields

What can trending hashtags from X tell us about views on the NHS?

The tweets data supplied did not reveal anything relevant to utilisation or missed appointments.

Recommendations

1) Provide more resources, prioritising October, November, and March.

2) Demand for face-to-face has peaks in these months but telephone appointments less so. Prioritise resources for face-to-face appointments.

3) Measures to reduce missed appointments should be targeted at face-to-face as almost three times as many of these are missed versus telephone appointments. Some ideas could be:

- text reminders at one day and one hour before
- contacting "repeat DNA" patients sympathetically, asking them if they would prefer telephone appointments

We are aware that the NHS plans to tackle this issue (reference 6) and would be keen to work with the NHS to analyse the effectiveness of these measures in the future.

4) Telephone appointments show less seasonal demand and a lower ratio of “DNA”’s. Continuing to offer these when justified could help with resourcing.

Limitations

- The supplied data is classed as experimental
- Calculation of utilisation using weekdays may require further discussion with the NHS

Appendix

Further investigations that are less relevant to the final analysis

These are included for information in case they are of interest.

Which service setting was the most popular for NHS North West London from 1 January to 1 June 2022?

In this ICB "General Practice" has the highest number of booked appointments.

```
#Group the data by service setting and sum the count of appointments for each setting. Sort the resulting series descending.
nc_nwlondon_subset.groupby("service_setting")["count_of_appointments"].sum().sort_values(ascending=False)
```

```
service_setting
General Practice             10432225
Unmapped                       904234
Other                          343642
Primary Care Network           240283
Extended Access Provision      222006
Name: count_of_appointments, dtype: int64
```

Jonathan Shields

What was the total number of records per month?

The number of records for each month in the actual appointments file is shown below. Mean 19685, range 2262 and standard deviation 729.

```
# Determine the total records per month in the actual appointments dataset
ad_month_group=ad.groupby([ad.appointment_date.dt.month_name(),ad.appointment_date.dt.year])

#Calculate and display the population standard deviation of each group size rounded to 2 decimal places
std_dev_ad=round(ad_month_group.size().std(ddof=0),2)
print(f"The population standard deviation of the number of actual appointment records per month is {std_dev_ad}")

#Calculate and display the range
range_ad=ad_month_group.size().max()-ad_month_group.size().min()
print(f"The range is {range_ad}")

#Display group sizes (records per month)
ad_month_group.size()
```

```
The population standard deviation of the number of actual appointment records per month is 728.9
The range is 2262
appointment_date  appointment_date
April             2022              19078
December          2021              19507
February          2022              18974
January           2022              19643
June              2022              19227
March             2022              21236
May               2022              20128
dtype: int64
```

Similarly, in the regional appointments file, the mean was 19895, range 3012 and standard deviation 668.

```
The population standard deviation of the number of regional appointment records per month is 668.23
The range is 3012
The mean is 19894.03
```

```
128]:  appointment_month  appointment_month
       April              2020                19124
                          2021                19452
                          2022                20073
       August             2020                19247
                          2021                19786
       December           2020                19394
                          2021                20393
       February           2020                20689
                          2021                18949
                          2022                20133
       January            2020                20889
                          2021                19319
                          2022                20225
       July               2020                19502
                          2021                19899
       June               2020                18844
                          2021                19814
                          2022                20231
       March              2020                21350
                          2021                19369
                          2022                20532
       May                2020                18338
                          2021                19384
                          2022                20276
       November           2020                19675
                          2021                20766
       October            2020                20122
                          2021                20562
       September          2020                20043
                          2021                20441
```

Finally for the national categories file the mean was 74308, range 12823 and standard deviation 3642.

Jonathan Shields

```python
# Determine the total records per month in the national categories dataset
nc_month_group=nc.groupby([nc.appointment_date.dt.month_name(),nc.appointment_date.dt.year])

#Calculate and display the population standard deviation of each group size rounded to 2 decimal places
std_dev_nc=round(nc_month_group.size().std(ddof=0),2)
print(f"The population standard deviation of the number of national category records per month is {std_dev_nc}")

#Calculate the range
range_nc=nc_month_group.size().max()-nc_month_group.size().min()
print(f"The range is {range_nc}")

#Calculate and display the mean
mean_nc=round(nc_month_group.size().mean(),2)
print(f"The mean is {mean_nc}")

#Display group sizes (records per month)
nc_month_group.size()
```

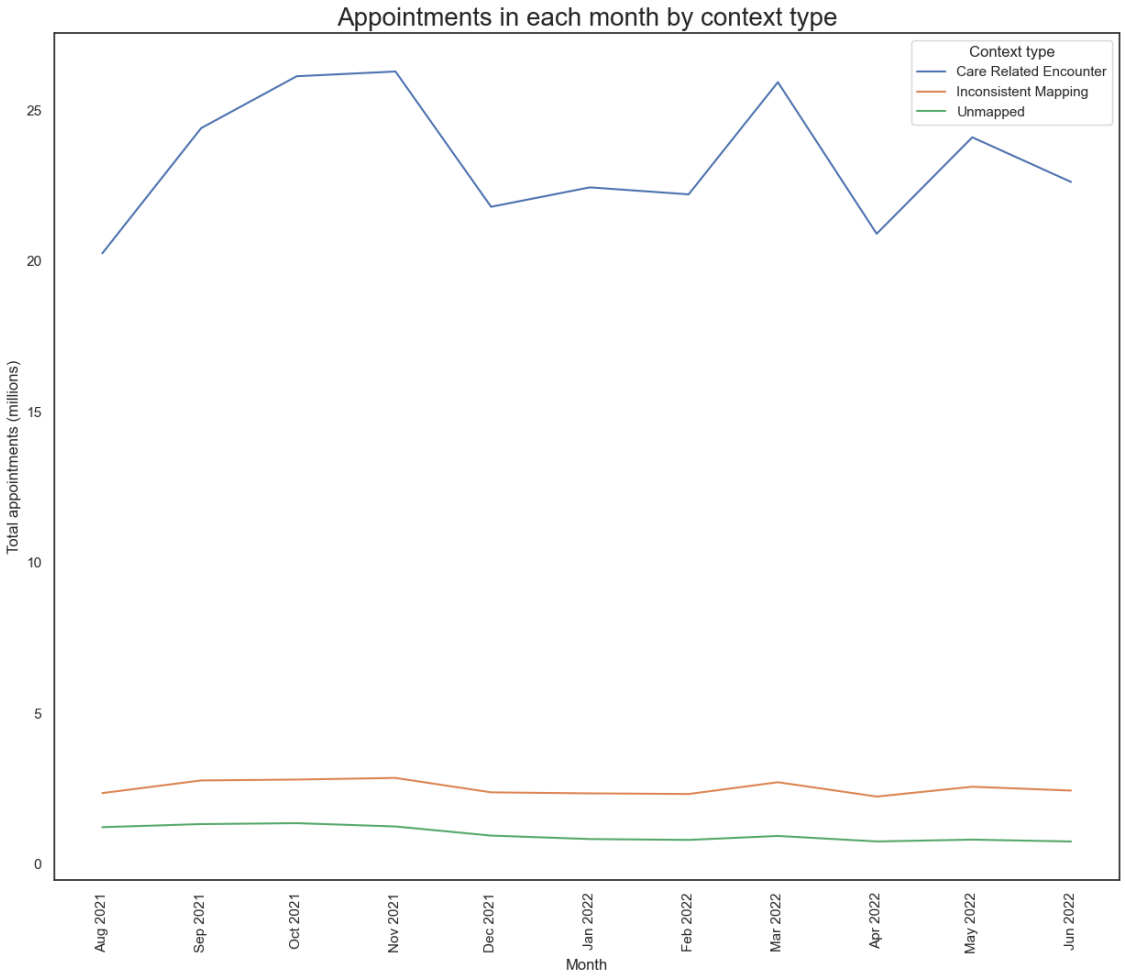
```
The population standard deviation of the number of national category records per month is 3641.54
The range is 12823
The mean is 74308.55
```

appointment_date	appointment_date	
April	2022.0	70012
August	2021.0	69999
December	2021.0	72651
February	2022.0	71769
January	2022.0	71896
June	2022.0	74168
March	2022.0	82822
May	2022.0	77425
November	2021.0	77652
October	2021.0	74078
September	2021.0	74922

The grouping of appointments in each file is different so this data reflects the spread of records across months as opposed to appointments.

Seasonal variation by context types

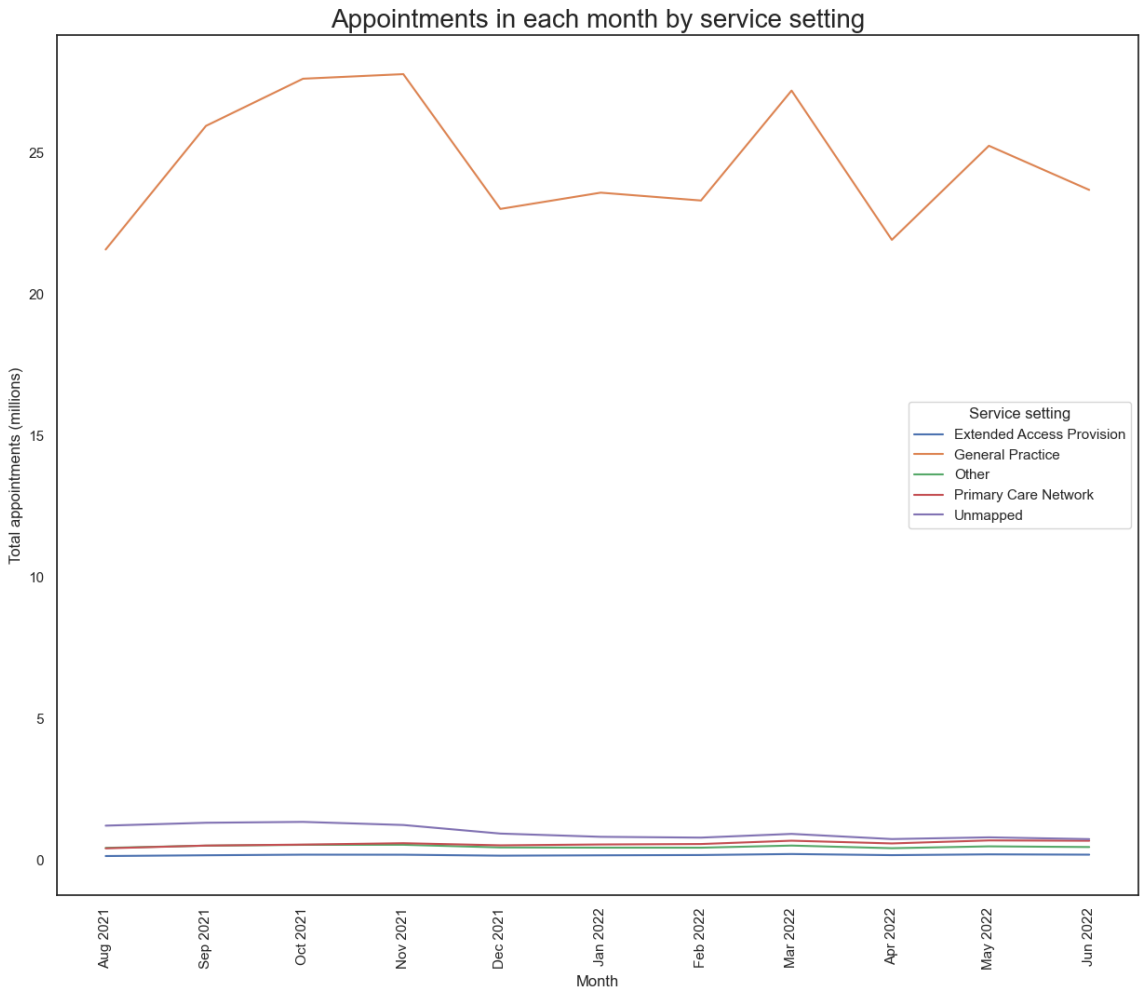
The line chart below shows the total appointments in each month with a line for each context type. “Care Related Encounter” is by far the most common type. There are again seasonal variations with a spike in autumn and again in spring.



Service settings

The line chart b shows appointments per month by service setting. The vast majority of appointments are in the "General Practice" setting with particular peaks in autumn and spring.

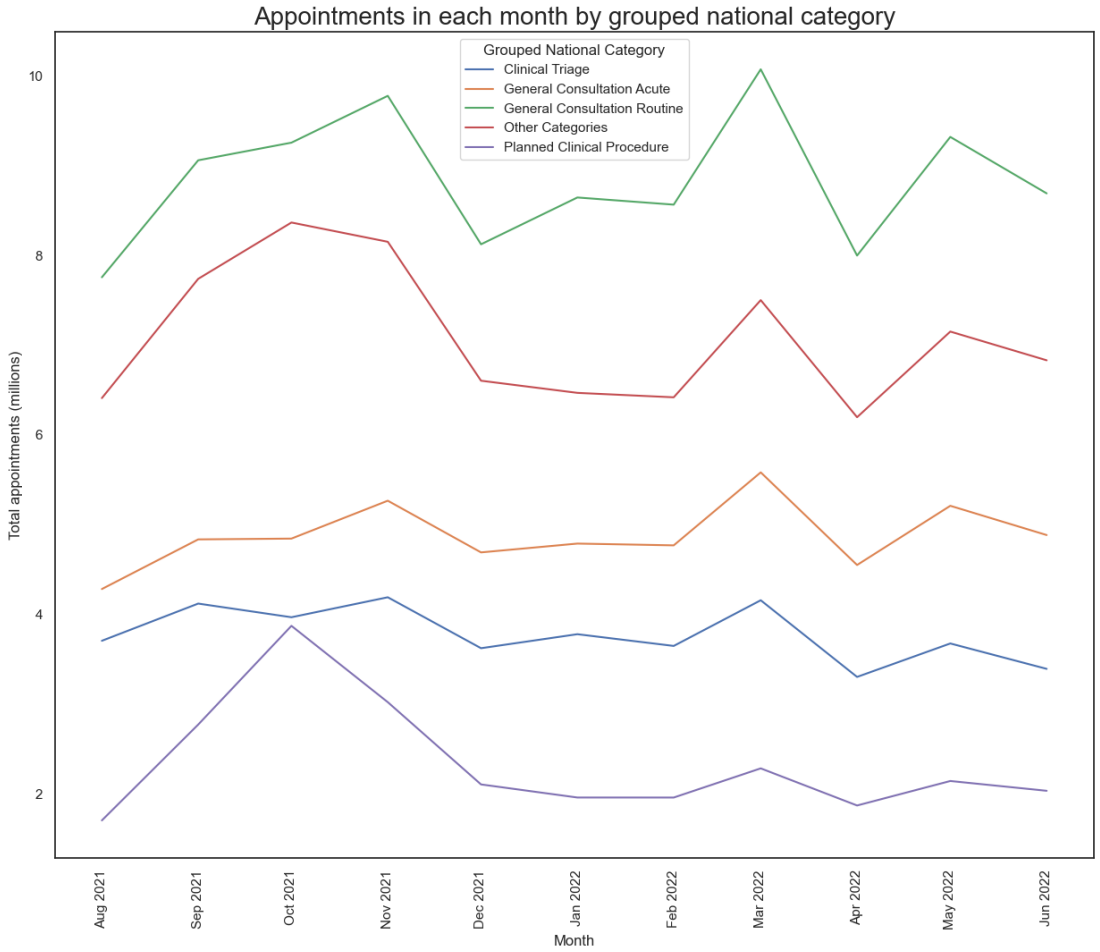
Jonathan Shields



National categories

“General Consultation Acute", "General Consultation Routine", "Planned Clinical Procedure" and "Clinical Triage" have the most appointments. For clarity, these are shown separately and others are grouped under "Other Categories" below.

Jonathan Shields



There is a spike in General Consultation (Routine), Planned Clinical Procedures and "Other" categories in Autumn. All of the categories have a spike in March, although this is minor in Planned Clinical Procedures.

References

1. NHS demand https://www.nhsconfed.org/topic/capacity-performance/system-under-pressure#:~:text=The%20NHS%20is%20facing%20unsustainable%20pressures%2C%20with%20record,that%20pressures%20on%20NHS%20staff%20are%20becoming%20unbearable.
2. A planned capacity of 1.2 million appointments per day across the NHS in England has been assumed in line with the brief.
3. Rising patient numbers https://www.bma.org.uk/advice-and-support/nhs-delivery-and-workforce/pressures/pressures-in-general-practice-data-analysis
4. Seasonal NHS pressures https://www.bbc.co.uk/news/articles/c047ky5qv4ro
5. Pressure due to COVID https://www.bma.org.uk/advice-and-support/nhs-delivery-and-workforce/pressures/an-nhs-under-pressure
6. https://www.england.nhs.uk/2023/01/nhs-drive-to-reduce-no-shows-to-help-tackle-long-waits-for-care/ missed appointment measures
7. https://fullfact.org/health/nhs-winter-crisis-worse-than-usual/ seasonal NHS pressures

Jonathan Shields

Jonathan Shields