British Heart Foundation Data Science Centre

Led by Health Data Research UK

NEW `token_pseudo_id_lookup` table

A table to determine the type and quality of pseudonymised patient identifiers in the NHS England SDE

Tom Bolton
on behalf of the Health Data Science Team

CVD-COVID-UK/COVID-IMPACT Consortium Update Meeting

July 11, 2024

HDRUK
Health Data Research UK

bhfdsc@hdruk.ac.uk | @BHFDataScience | YouTube BHF Data Science Centre

Pseudonymised patient identifiers within the NHS England SDE

British Heart Foundation Data Science Centre
Led by Health Data Research UK

NHS_NUMBER_DEID

deaths	DEC_CONF_NHS_NUMBER_CLEAN_DEID
epma_administration	NHS_NUMBER_DEID
epma_prescription	NHS_NUMBER_DEID
gdppr	NHS_NUMBER_DEID
icnarc	NHS_NUMBER_DEID
nicor_congenital	1_03_NHS_NUMBER_DEID
nicor_hf	1_03_NHS_NUMBER_DEID
nicor_minap	1_03_NHS_NUMBER_DEID
nicor_pci	1_03_NHS_NUMBER_DEID
sus	NHS_NUMBER_DEID

PERSON_ID_DEID

chess	PERSON_ID_DEID
covid_antibody_testing_pillar3	PERSON_ID_DEID
covid_antigen_testing_pillar2	PERSON_ID_DEID
hes_ae	PERSON_ID_DEID
hes_ae_otr	PERSON_ID_DEID
hes_apc_acp	PERSON_ID_DEID
hes_apc	PERSON_ID_DEID
hes_apc_mat	PERSON_ID_DEID
hes_apc_otr	PERSON_ID_DEID
hes_cc	PERSON_ID_DEID
hes_cc_otr	PERSON_ID_DEID
hes_op	PERSON_ID_DEID
hes_op_otr	PERSON_ID_DEID
iapt_v2_1_care_activities	Person_ID_DEID
iapt_v2_1_care_cluster	Person_ID_DEID
iapt_v2_1_coded_scored_assessments	Person_ID_DEID
iapt_v2_1_demographics_and_referral	Person_ID_DEID
iapt_v2_1_employment_status	Person_ID_DEID
iapt_v2_1_internet_enabled_therapies	Person_ID_DEID
iapt_v2_1_mental_and_physical_health_conditions	Person_ID_DEID
iapt_v2_1_onward_referrals	Person_ID_DEID
iapt_v2_1_waiting_time_pauses	Person_ID_DEID
lowlat_apc_all_years	PERSON_ID_DEID
lowlat_cc_all_years	PERSON_ID_DEID
lowlat_ecds_all_years	PERSON_ID_DEID
lowlat_op_all_years	PERSON_ID_DEID

Definitions

NHS_NUMBER_DEID
Tokenized (de-identified [“de-id”]) version of the NHS Number.

PERSON_ID_DEID
Tokenized version of Person_ID, which is comprised of **three** levels:

1) NHS number
From Personal Demographic Service (PDS) records - the collection of all NHS numbers and patients’ demographic information

Linkable to all other tables

2) Master Person Service (MPS) ID
From the MPS bucket of previously unmatched records that could not be identified as records with an NHS number in PDS. If sufficient demographic information is provided a new MPS ID can be created and added to the MPS bucket.

Only linkable to other PERSON_ID_DEID tables

3) One-time-use ID
If neither an NHS number or an MPS ID could be assigned.

Not linkable within the table or across other tables

The Person_ID Handbook

Date Published: 24 January 2024

Jump to overview

Current Chapter
The Person_ID Handbook
View all

Next Chapter →
Introduction

Page contents

Summary and outline

The Person_ID is a unique patient identifier used by NHS England with the objective of standardising the approach to patient-level data linkage across different data sets.

This handbook aims to provide users of the Person_ID in the Hospital Episode Statistics (HES) databases with supporting documentation on what the Person_ID is, how it is derived via the Master Person Service (MPS), how the data flows between services (Data Processing Services (DPS) and Spine), and how to interpret the output information associated with the Person_ID.

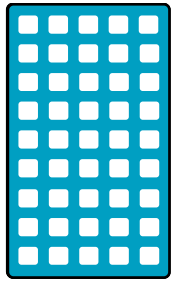
Example of Data Linkage Behaviour

British Heart Foundation Data Science Centre
Led by Health Data Research UK

- NHS number
- MPS ID
- One-time-use ID

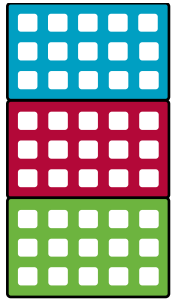
NHS_NUMBER_DEID

GDPPR



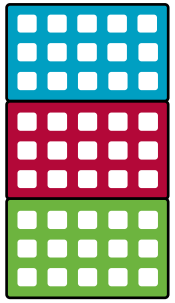
PERSON_ID_DEID

HES APC



PERSON_ID_DEID

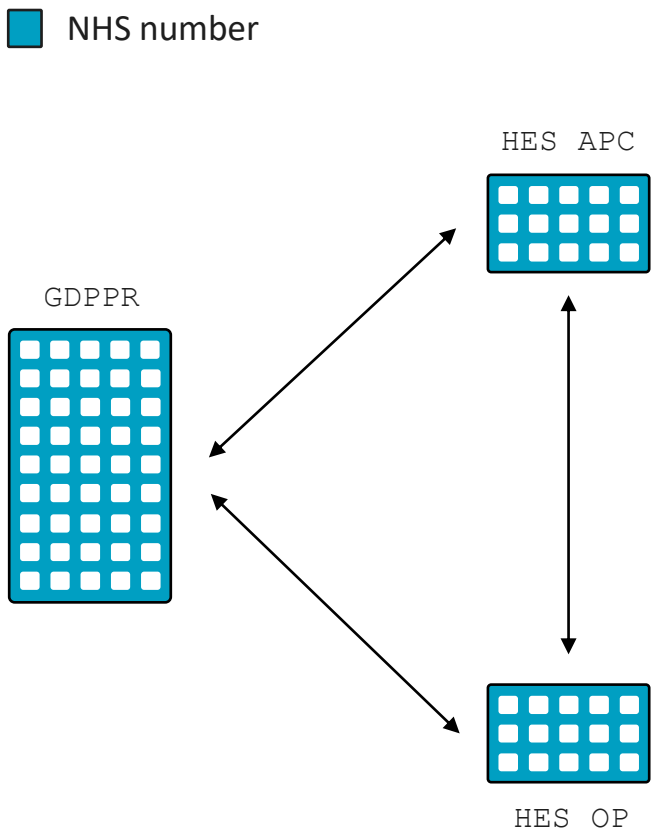
HES OP



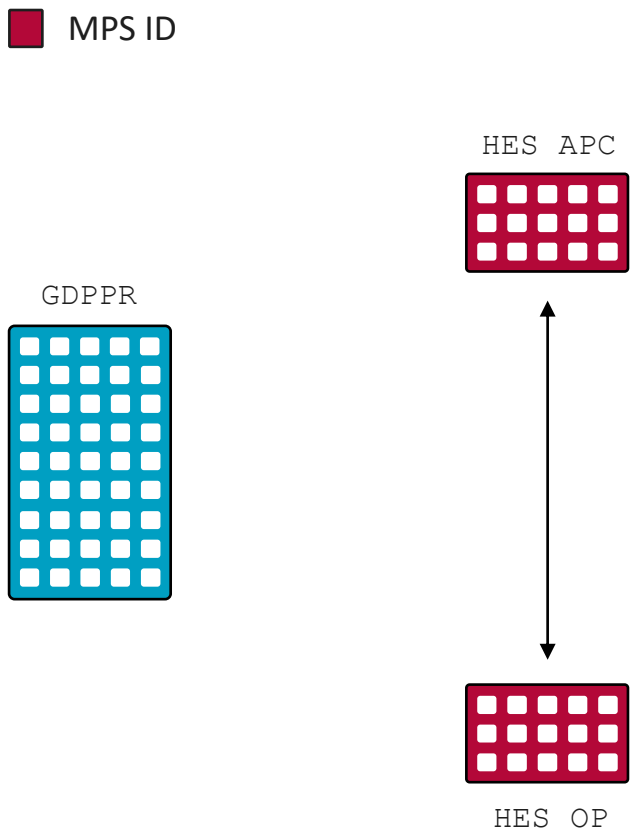
GDPPR: General Practice Extraction Service (GPES) Data for Pandemic Planning and Research
HES: Hospital Episode Statistics
APC: Admitted Patient Care
OP: Outpatient

Example of Data Linkage Behaviour

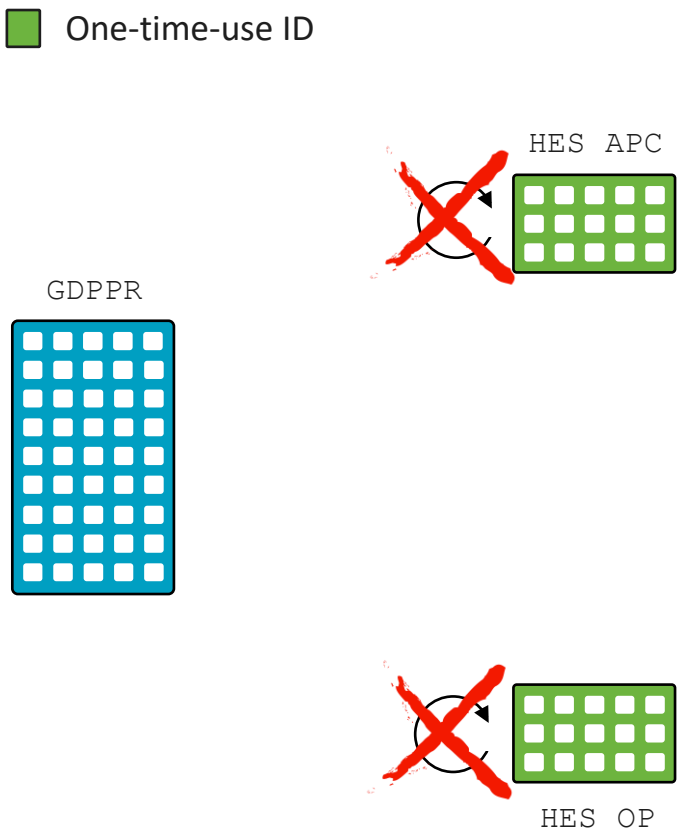
British Heart Foundation Data Science Centre
Led by Health Data Research UK



NHS number IDs are linkable across both NHS_NUMBER_DEID and PERSON_ID_DEID tables.



MPS IDs are only linkable across PERSON_ID_DEID tables. MPS IDs are not linkable to NHS_NUMBER_DEID tables.



One-time-use IDs are not linkable across NHS_NUMBER_DEID or PERSON_ID_DEID tables. Multiple one-time-use IDs might relate to the same individual.

token_pseudo_id_lookup table

British Heart Foundation Data Science Centre
Led by Health Data Research UK

Path:
dars_nic_391419_j3w9t.token_pseudo_id_lookup

England SECURE DATA ENVIRONMENT

Documentation:

The token_pseudo_id_lookup table provides indicator columns for the type of pseudonymised identifier.

This table determines whether the pseudonymised identifier corresponds to a valid NHS number, Master Person Service (MPS) ID, or a one-time-use ID. Please refer to the NHS England Person_ID handbook for further information about these types of Person_ID.

The table covers all pseudonymised identifiers (e.g., NHS_NUMBER_DEID, PERSON_ID_DEID) that feature in the data sharing agreement.

The pseudo_id column uniquely identifies each row in the table. As at 2024-06-04, the token_pseudo_id_lookup table included ~450 million rows (i.e., distinct pseudo_id).

The table is partitioned on the first_char column (the first character of the pseudo_id column) and this column can be used in addition to the pseudo_id column when joining the token_pseudo_id_lookup table to other tables to improve the efficiency of the join by reducing the shuffling required.

The token_pseudo_id_lookup table will be updated each month by the NHS England Data Wrangler team inline with monthly batch provisioning and updates, with any new pseudonymised identifiers inserted into the table, which will be stored in the live (read-only) database (dars_nic_391419_j3w9t) for the data sharing agreement.

Schema:

Column name	Data type	Description
first_char	String	The first character of the pseudo_id column (partition column)
pseudo_id	String	The pseudonymised version of the identifier (primary key)
valid_nhs_number	Boolean	An indicator for whether the pseudo_id is a valid NHS number (passes the checksum)
mps_id	Boolean	An indicator for whether the pseudo_id is an MPS ID (length 10 and first character "A/B")
single_use_id	Boolean	An indicator for whether the pseudo_id is a one-time-use ID (length 10 and first character "U")

Sample Data:

first_char	pseudo_id	valid_nhs_number	mps_id	single_use_id
1	189LXT3VMG	TRUE	FALSE	FALSE
9	9PA9U0XUKE	FALSE	TRUE	FALSE
A	A29MFXEY2D	FALSE	FALSE	TRUE
E	EQSQ5B1W8N	FALSE	FALSE	FALSE

Note: The sample data provided above is fictitious and is presented for illustrative purposes only.

Acknowledgements:

Shoaib Ali Ajaib and the NHS England Data Wrangler Team

Using the `token_pseudo_id_lookup` table

first_char	pseudo_id	valid_nhs_number	mps_id	single_use_id
1	189LXT3VMG	TRUE	FALSE	FALSE
9	9PA9U0XUKE	FALSE	TRUE	FALSE
A	A29MFXEY2D	FALSE	FALSE	TRUE
E	EQSQ5B1W8N	FALSE	FALSE	FALSE

→

first_char	pseudo_id	pseudo_id_type
1	189LXT3VMG	1: NHS number
9	9PA9U0XUKE	2: MPS ID
A	A29MFXEY2D	3: One-time-use ID
E	EQSQ5B1W8N	4: None

```python
# Create a dataFrame 'id_lookup' by transforming data from the 'token_pseudo_id_lookup' table
id_lookup = (
  spark.table('dars_nic_391419_j3w9t.token_pseudo_id_lookup')
  # Add a new column 'pseudo_id_type' based on conditions
  # If 'valid_nhs_number' is true, set 'pseudo_id_type' to 1
  # If 'mps_id' is true, set 'pseudo_id_type' to 2
  # If 'single_use_id' is true, set 'pseudo_id_type' to 3
  # For all other cases, set 'pseudo_id_type' to 4
  .withColumn(
    'pseudo_id_type',
    F.when(F.col('valid_nhs_number'), F.lit(1))
    .when(F.col('mps_id'), F.lit(2))
    .when(F.col('single_use_id'), F.lit(3))
    .otherwise(F.lit(4))
  )
  .drop(['valid_nhs_number', 'mps_id', 'single_use_id'])
)

# Save the dataframe with the new column
id_lookup.write.partitionBy('first_char').mode('overwrite').option('overwriteSchema', 'true').saveAsTable
('dsa_391419_j3w9t_collab.ccu005_01_id_lookup')
```

```python
# Read the saved dataframe
id_lookup = spark.read('dsa_391419_j3w9t_collab.ccu005_01_id_lookup')

# Create a DataFrame 'hes_apc' by selecting a monthly batch from the 'hes_apc_all_years_archive' table to
ensure reproducibility
hes_apc = (
  spark.table('dars_nic_391419_j3w9t_collab.hes_apc_all_years_archive')
  # Filter records where 'archived_on' date is '2024-06-04'
  .where(F.col('archived_on') == '2024-06-04')
)

# Create a dataFrame 'hes_apc_id_lookup' by joining 'hes_apc' with 'id_lookup' on specified conditions
hes_apc_id_lookup = (
  hes_apc
  # Add a new column 'first_char' containing the first character of 'PERSON_ID_DEID'
  .withColumn('first_char', F.substring(F.col('PERSON_ID_DEID'), 1, 1))
  # Perform a left join with 'id_lookup', renaming 'pseudo_id' to 'PERSON_ID_DEID' in the lookup table
  for matching, joining on 'first_char' and 'PERSON_ID_DEID' to to improve the efficiency of the join by
  reducing shuffling.
  .join(id_lookup.withColumnRenamed('pseudo_id', 'PERSON_ID_DEID'), on=['first_char', 'PERSON_ID_DEID'],
  how='left')
)
```

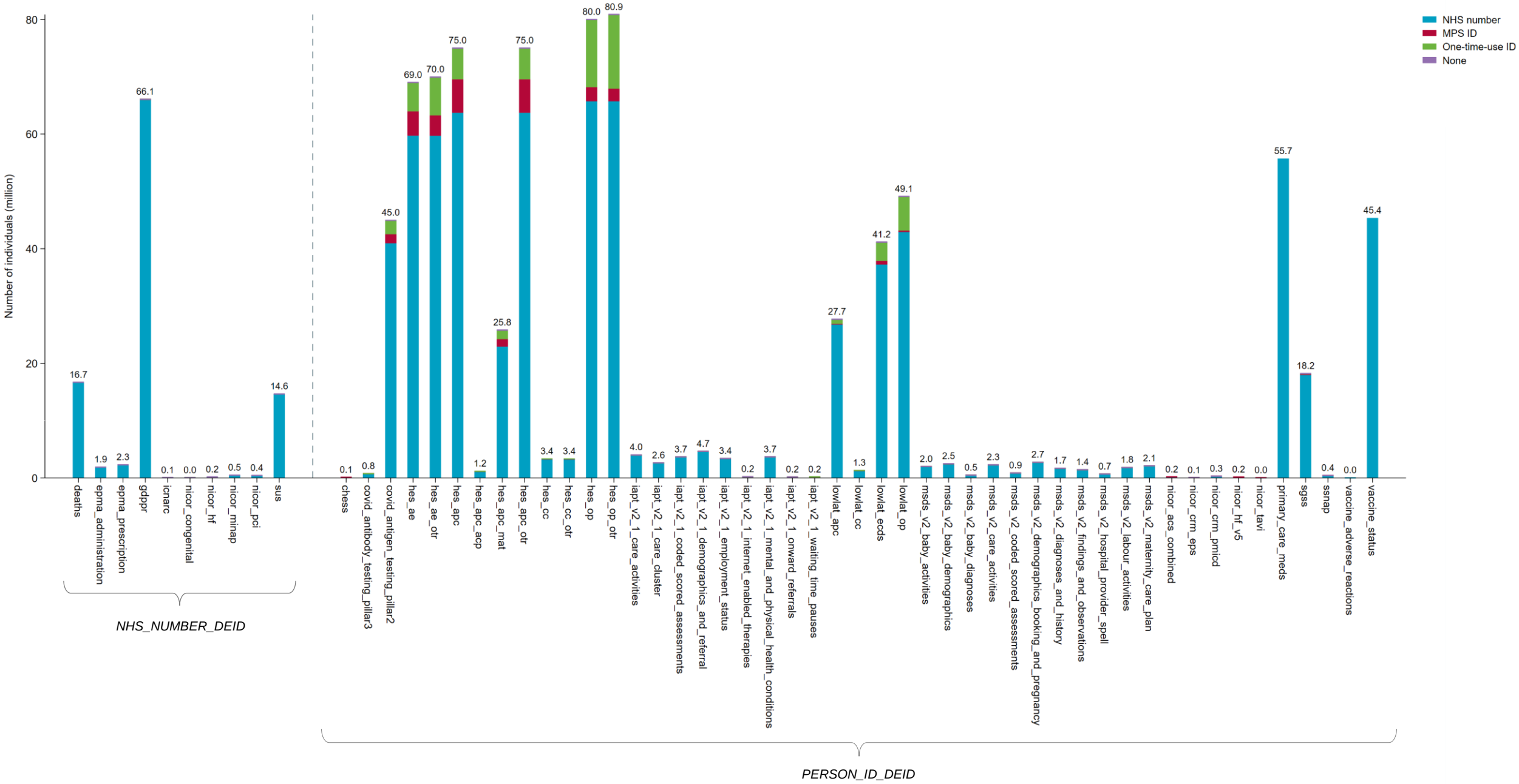


***Figure*: Number of individuals by type of pseudonymised identifier in the provisioned tables within the NHS England SDE.**

Data not shown for 59 tables relating to the Mental Health Services Dataset (MHSDS).

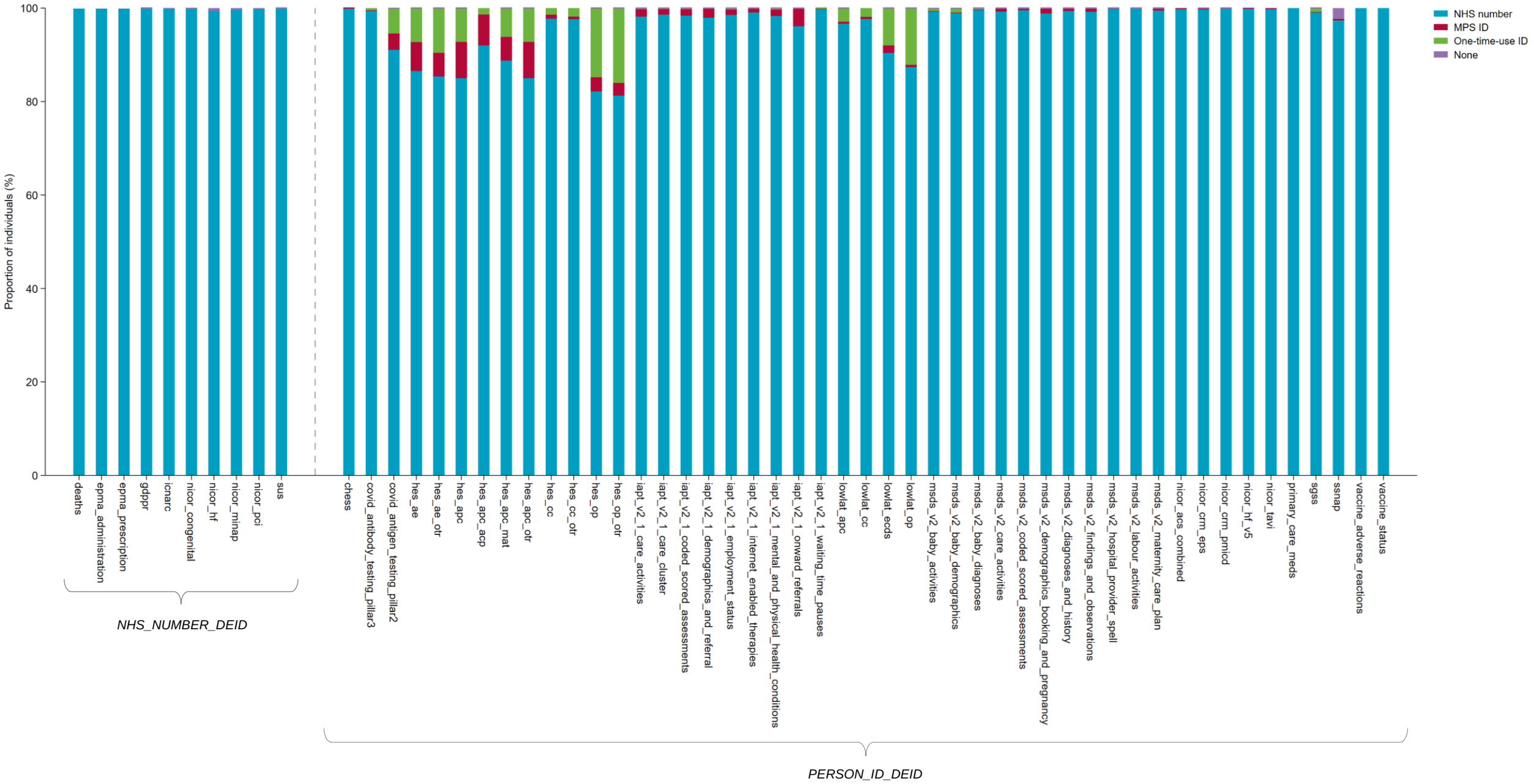


***Figure*: Proportion of individuals by type of pseudonymised identifier in the provisioned tables within the NHS England SDE.**

Data not shown for 59 tables relating to the Mental Health Services Dataset (MHSDS).

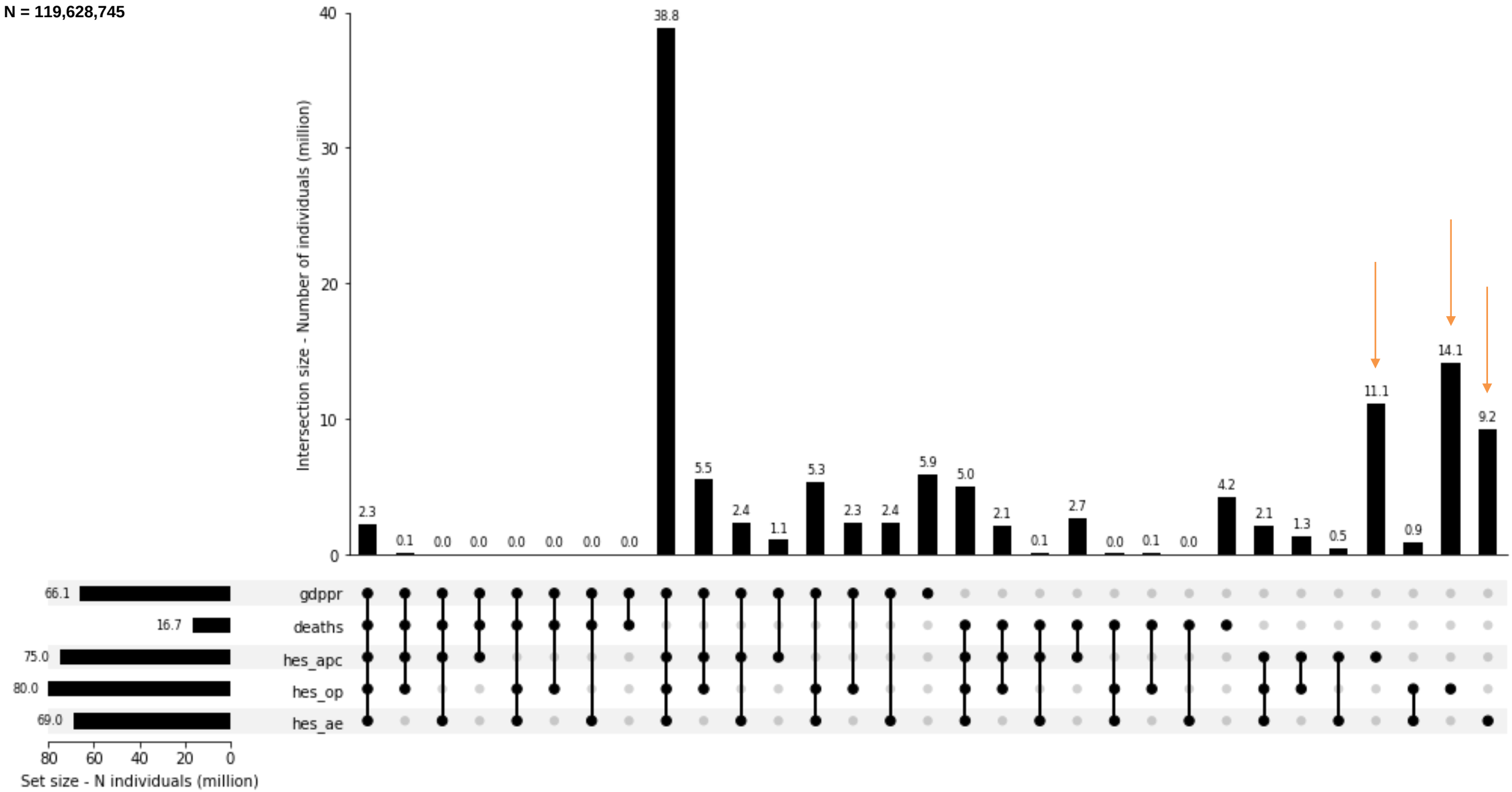


***Figure*: Upset plot of individuals across primary and secondary care and deaths datasets within the NHS England SDE.**

Vertical bars report unique individuals in the intersection denoted by the intersection matrix below. Horizontal bars report unique individuals identified from each dataset. Datasets were GDPPR (primary care), ONS Civil …

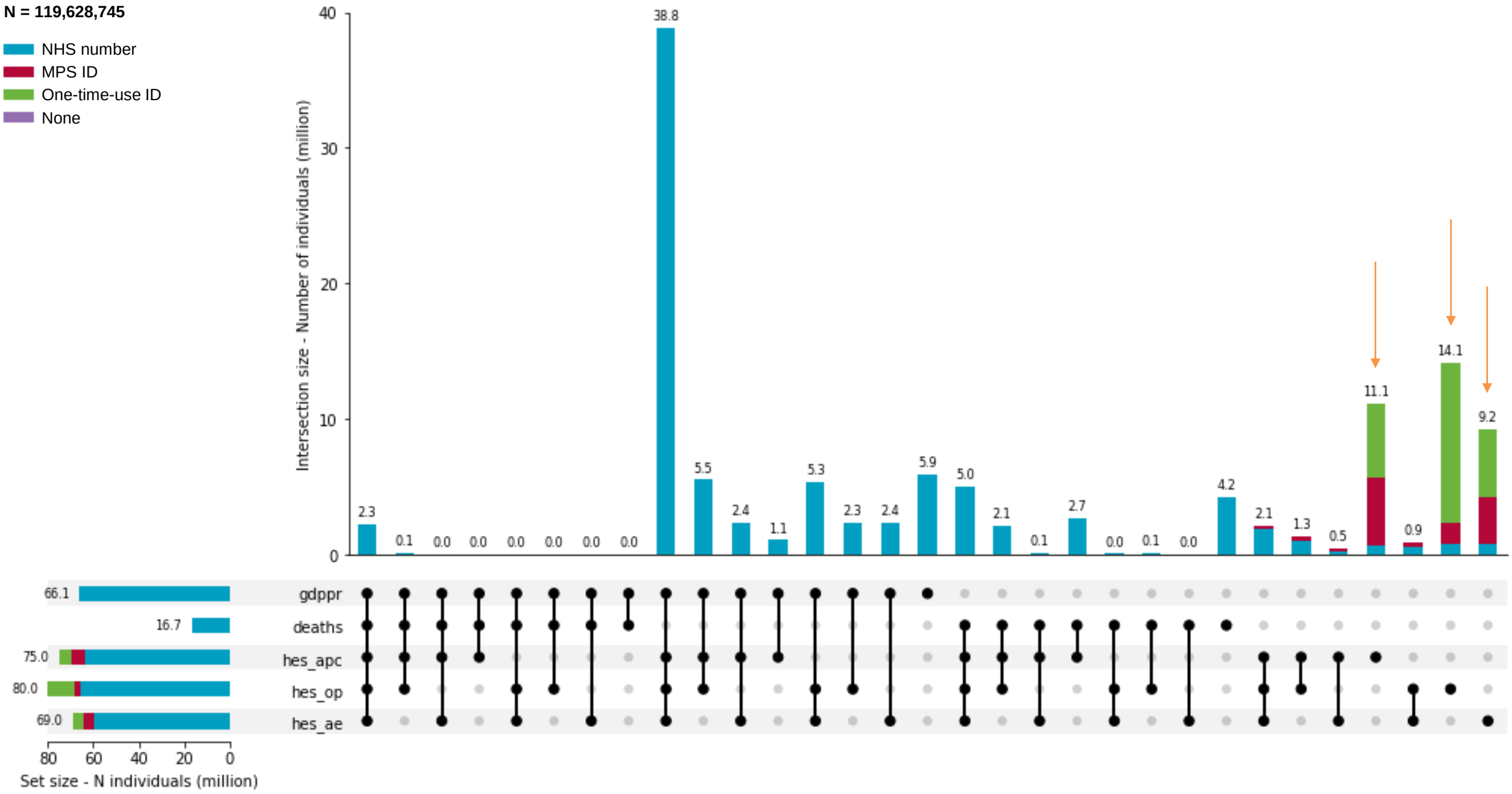


***Figure*: Upset plot of individuals across primary and secondary care and deaths datasets by type of pseudonymised identifier within the NHS England SDE.**

Vertical bars report unique individuals in the intersection denoted by the intersection matrix below. Horizontal bars report unique individuals identified from each dataset. Datasets were GDPPR (primary care), ONS Civil …

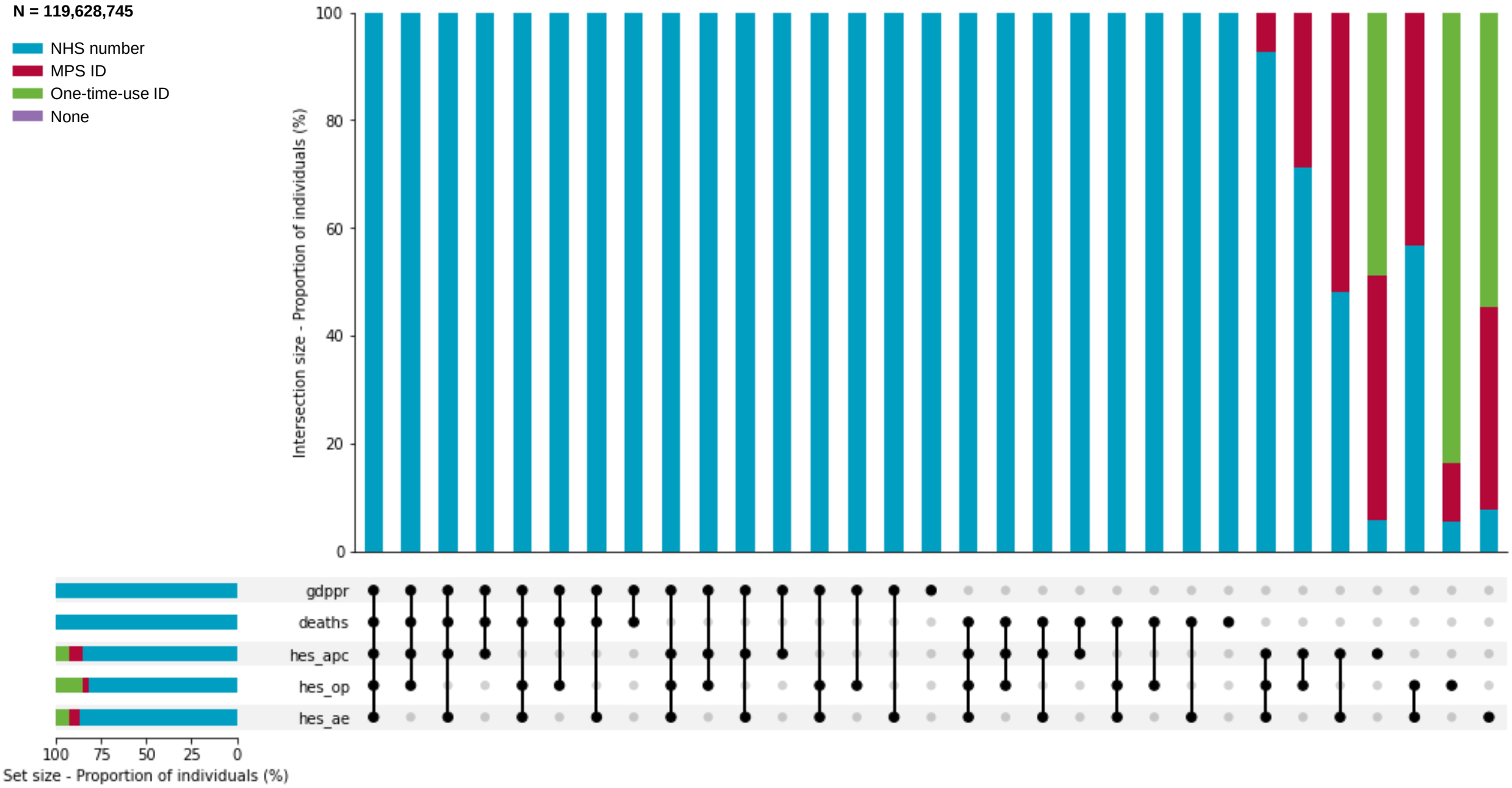


***Figure*: Upset plot of individuals across primary and secondary care and deaths datasets by type of pseudonymised identifier within the NHS England SDE.**

Vertical bars report the proportion of unique individuals in the intersection denoted by the intersection matrix below. Horizontal bars report the proportion of unique individuals identified from each dataset. Datasets were GDPPR

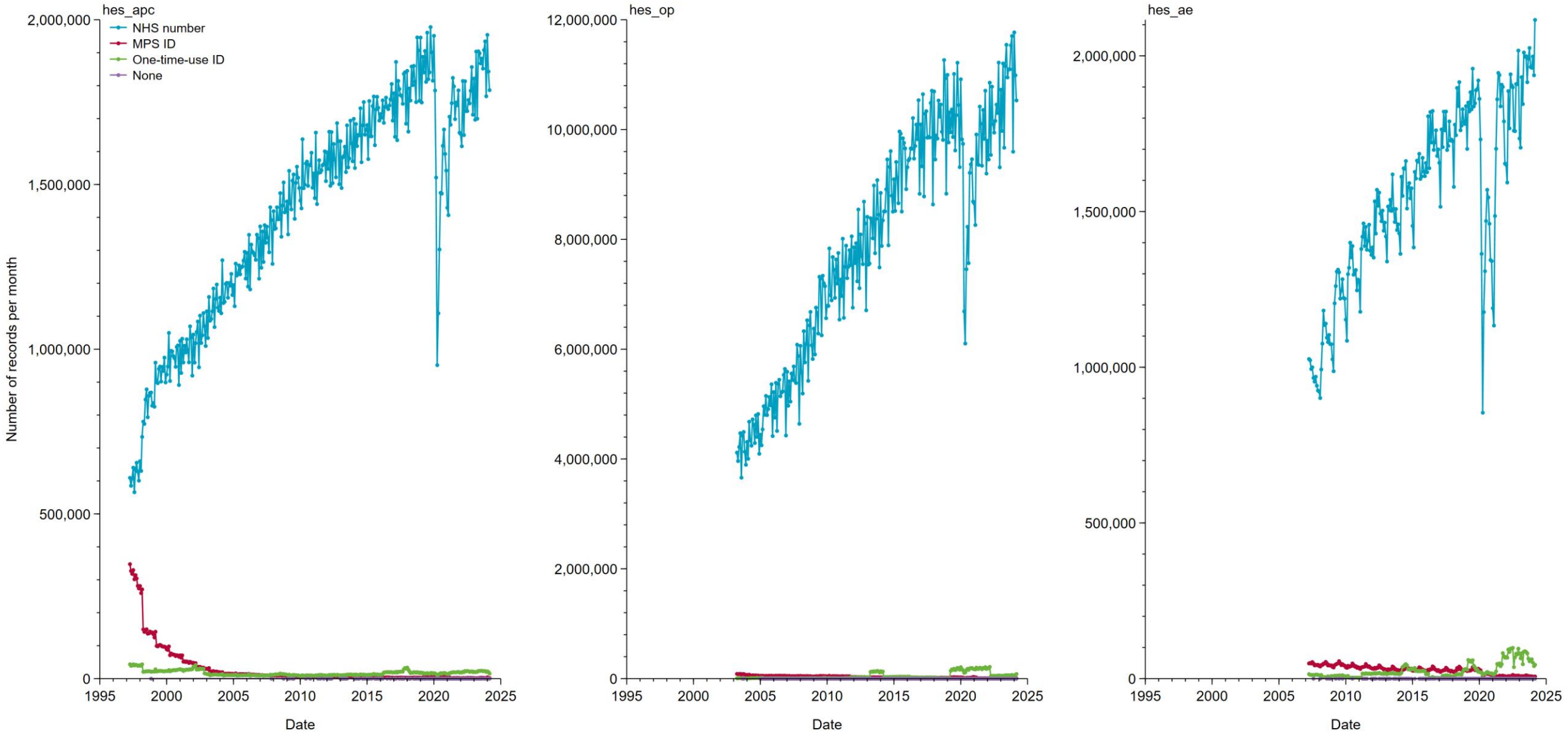


***Figure*: Number of records within Hospital Episode Statistics during 1997-2024 by table and type of pseudonymised identifier within the NHS England SDE.**

Hospital Episode Statistics (HES): Admitted Patient Care (APC), Outpatients (OP), and Accident and Emergency (AE).

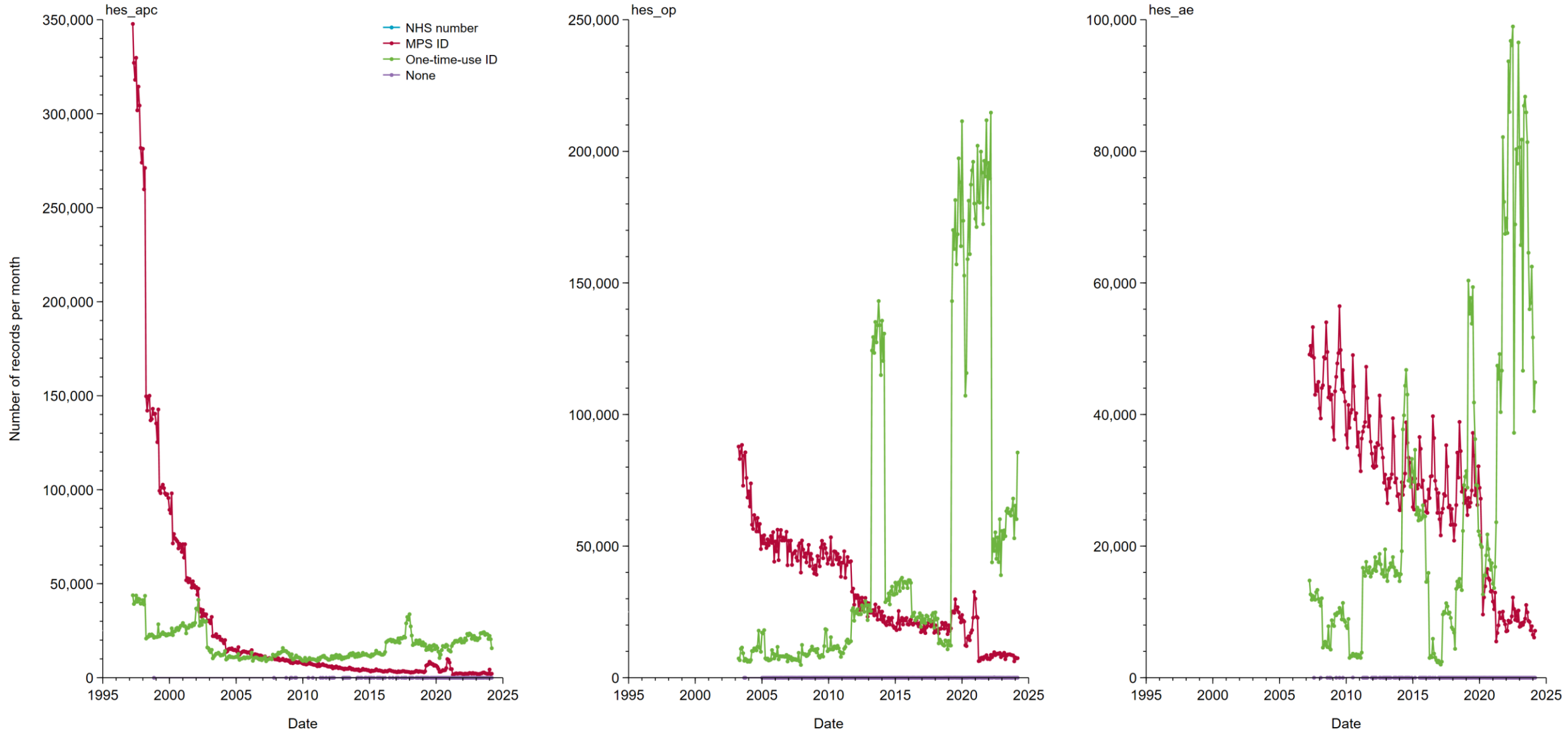


***Figure*: Number of records within Hospital Episode Statistics during 1997-2024 by table and type of pseudonymised identifier within the NHS England SDE.**

Hospital Episode Statistics (HES): Admitted Patient Care (APC), Outpatients (OP), and Accident and Emergency (AE). Excluding type of pseudonymised identifier == “NHS number”.

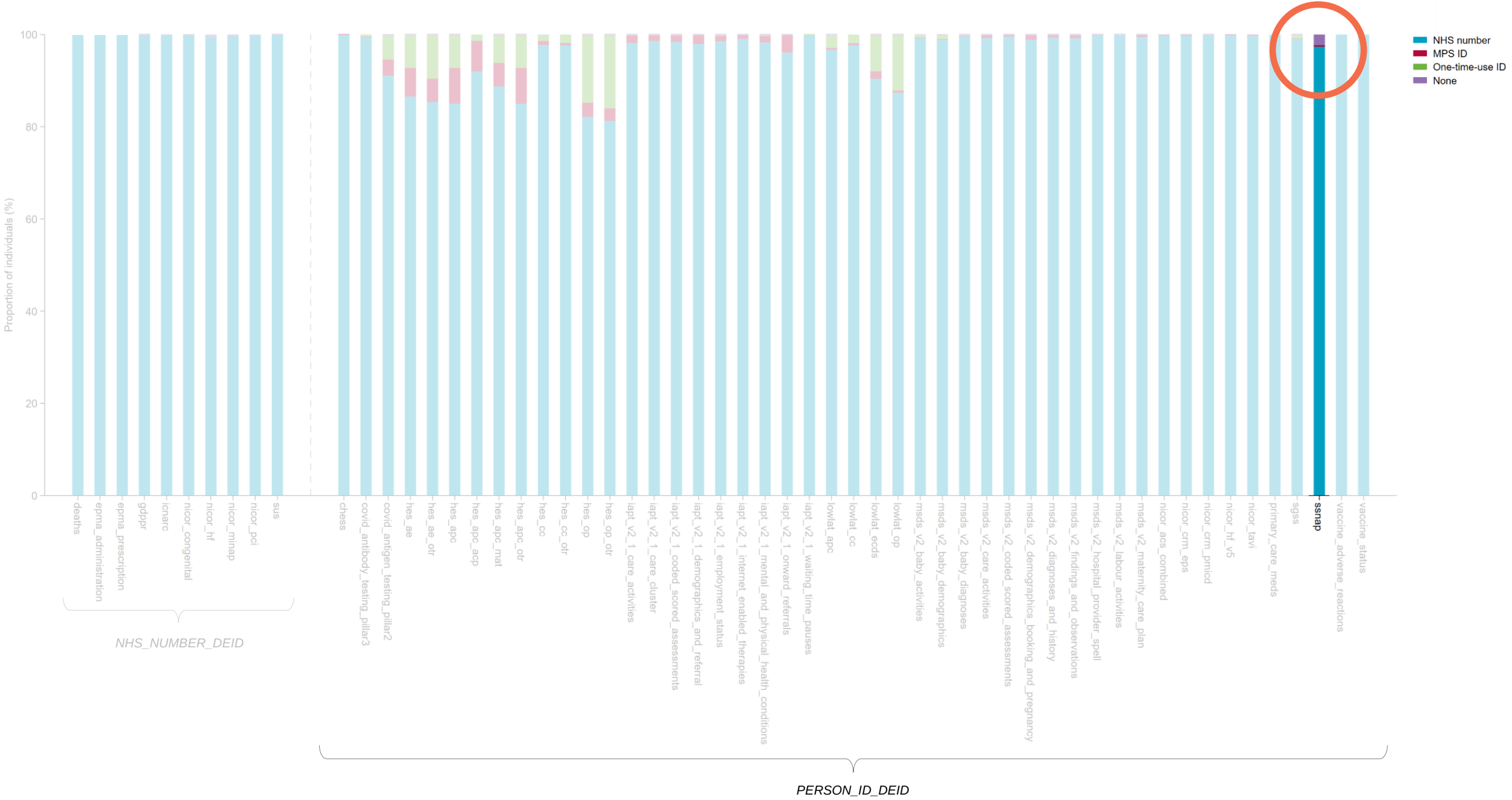


***Figure 2*: Proportion of individuals by type of pseudonymised identifier in the provisioned tables within the NHS England SDE.**

Data not shown for 59 tables relating to the Mental Health Services Dataset (MHSDS).

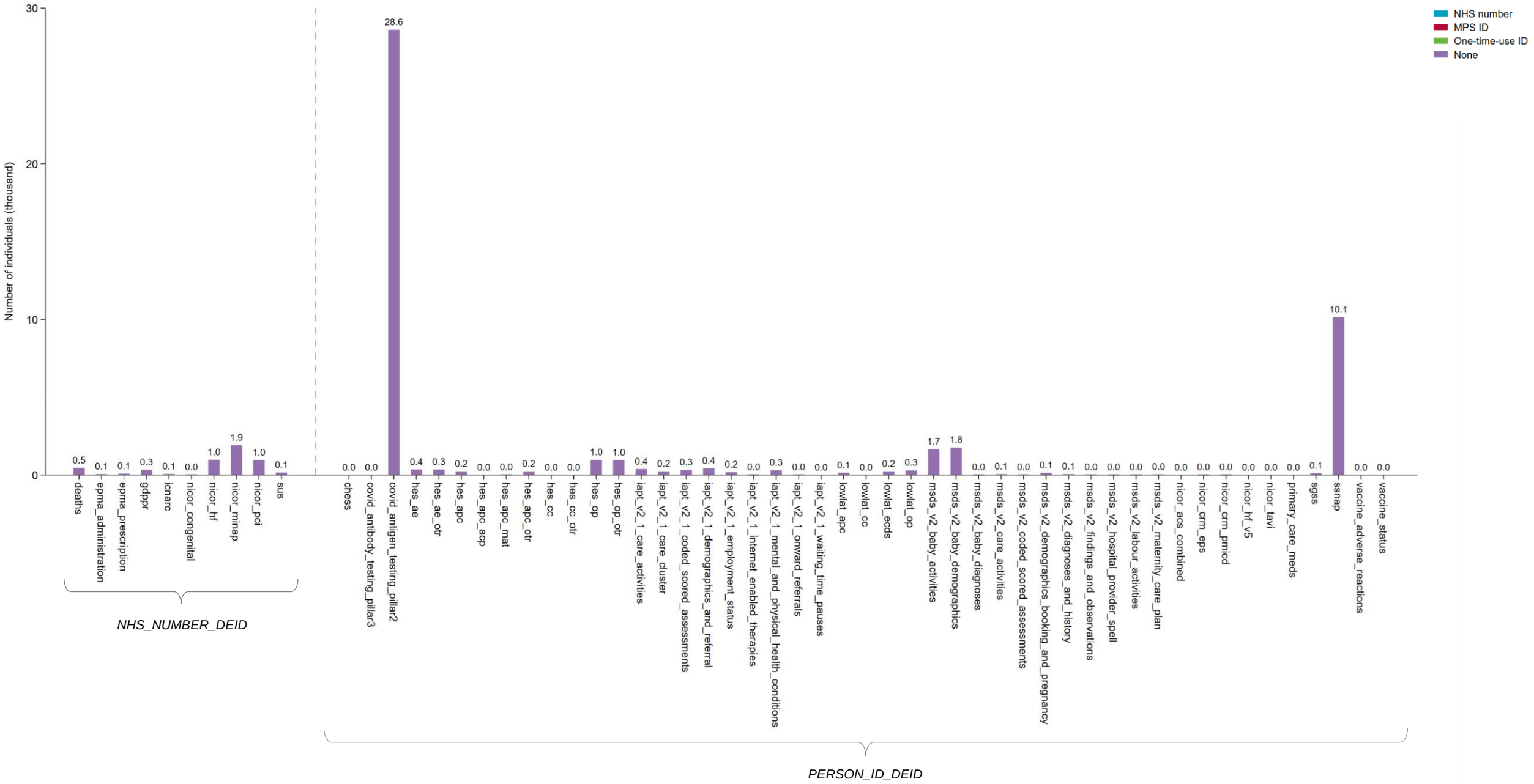


***Figure*: Number of individuals by type of pseudonymised identifier in the provisioned tables within the NHS England SDE.**

Data not shown for 59 tables relating to the Mental Health Services Dataset (MHSDS). Including type of pseudonymised identifier == “None” only.

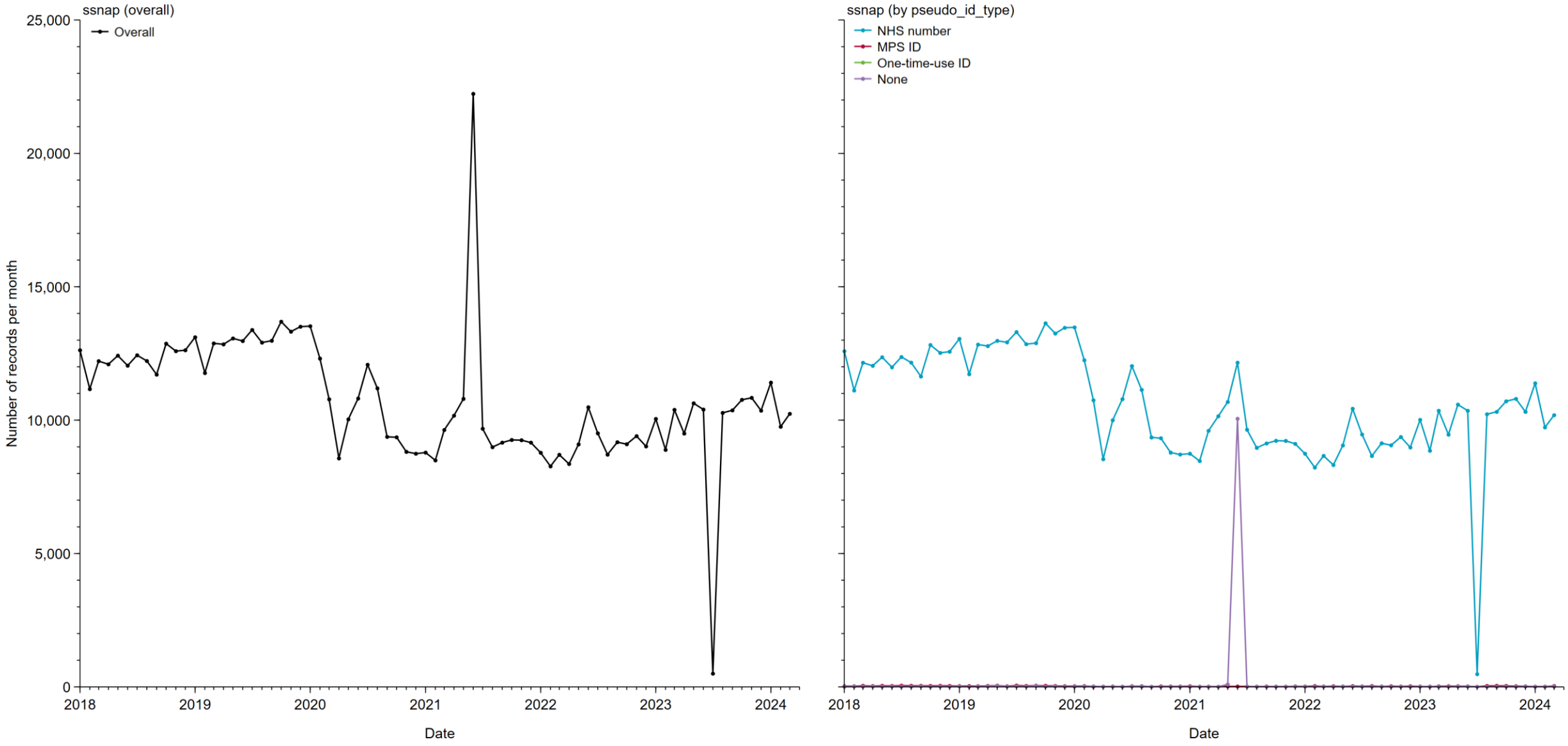


***Figure*: Number of records within Sentinel Stroke National Audit Programme during 2018-2024 by type of pseudonymised identifier within the NHS England SDE.**

First panel provides overall number of records. Second panel provides number of records by type of pseudonymised identifier.

Summary

British Heart Foundation Data Science Centre
Led by Health Data Research UK

Background

- NHS_NUMBER_DEID and PERSON_ID_DEID tables
- Definitions
- Data linkage behaviour
- Please refer to the NHS England Person_ID handbook for further information https://digital.nhs.uk/services/personal-demographics-service/master-person-service/the-person_id-handbook

Methods

- Description of the token_pseudo_id_lookup table and how this can be used

Findings from initial explorations

- All identifiers in the SDE are available in the token_pseudo_id_lookup table
- In general, the behaviour of `pseudo_id_type` is as expected in terms of the:
  - Distribution across different types of tables
  - Intersection of (selected) tables
- A few small queries to work through with the NHS England Data Wrangler Team
- Information on patterns over time, particularly one-time-use IDs, might be helpful

Conclusions

- Anchoring on GDPPR has had the desired effect of excluding lower quality IDs
- token_pseudo_id_lookup table will be a useful resource for future projects
- Potentially include a `pseudo_id_type` column in the `hds_curated_assets_demographics` table

Path:

`dars_nic_391419_j3w9t.token_pseudo_id_lookup`

England SECURE DATA ENVIRONMENT

Sample Data:

first_char	pseudo_id	valid_nhs_number	mps_id	single_use_id
1	189LXT3VMG	TRUE	FALSE	FALSE
9	9PA9U0XUKE	FALSE	TRUE	FALSE
A	A29MFXEY2D	FALSE	FALSE	TRUE
E	EQSQ5B1W8N	FALSE	FALSE	FALSE

Note: The sample data provided above is fictitious and is presented for illustrative purposes only.

Acknowledgements:

Shoaib Ali Ajaib and the NHS England Data Wrangler Team

Thank you for listening

✉ thomas.bolton@hdruk.ac.uk
✉ bhfdsc_hds@hdruk.ac.uk