

# Overview of cleaning - primary care data

## Summary

The output file `/dbfs/mnt/analysis-outputs/primary-care-oids-tidied.csv` is a csv with the following columns:

- PseudonymisedID - the pseudonymised patient id
- Date - the date of the opioid
- OpioidName - the opioid name
- Category - a brief description of the underlying prescription, more detail below
- DoseInMg - the most likely total amount of opioid that day in mg
- MME - the total MME of that opioid
- DoseToExhaustInMg - the daily dose if the patient completely used up the prescription before their next one
- ExhaustMME - the MME based on the DoseToExhaustInMg

An example is as follows:

PseudonymisedID	Date	OpioidName	Category	DoseInMg	MME	DoseToExhaustInMg	ExhaustMME
1001	2021-05-07	codeine	Last Rx - known length	40	6	NA	NA
1001	2021-05-07	morphine	Last Rx - known length	10	10	NA	NA
2886	2019-03-09	morphine	Next Rx too soon	20	20	40	40
2886	2019-03-10	morphine	Last Rx - known length	20	20	NA	NA
2886	2019-03-10	fentanyl	Last Rx - known length	0.5	65	NA	NA

**NB1:** The total MME for a patient is calculated by adding up each MME for the patient on each day. E.g. patient 1001 has a total MME of 16 on 2021-05-07, and patient 2886 has a total MME of 85 on 2019-03-10.

**NB2:** A patient can have multiple rows for a given date and OpioidName, if they have multiple prescriptions falling into multiple categories.

## Date range

The data is filtered to just administrations in the range "2010-01-01" to "2021-09-30" inclusive.

## Data

The underlying data is from the Salford Integrated Records. Each row represents a clinical event or a medication prescription. For prescriptions (which is all we're interested in for this data cleaning) the columns, and their meanings, are:

- PseudonymisedID - a unique id for each patients
- EntryDate - the date of the prescription

- ReadCode - the clinical code (either Read v2 or EMIS) for the medication
- Rubric - a description of the medication
- CodeValue - the amount prescribed
- Codeunits - the instruction (e.g. "take 1 in the morning")

## Assumptions and comments

- Patients start taking their medication the day after the date of the prescription. Justification:
  - we only know when the opioid was prescribed, we don't know when it was collected, but it could very well be next day
  - if the Rx is e.g. 4 times a day then it is unlikely that the patient would take full dose on the day of prescription
  - it is easier programmatically to assume the prescription is taken in full days rather than assigning partial amounts to the first and last days
- If a patient has the same clinical code more than once on a single day we assume it is a duplicate rather than multiple prescriptions on the same day
- The prescribed medicine can be viewed as an upper bound of what was actually taken because although it was prescribed we don't know if it was
  - collected from a pharmacy, or
  - taken correctly. This is different to the secondary care data which is actual administrations of opioids rather than just the prescription.

## Cleaning method

- Duplicate lines (where the PseudonymisedID, EntryDate and ReadCode columns are the same) almost always have the same CodeValue, but for the few cases where they differ we assume the higher amount.
- Duplicate lines (where the PseudonymisedID, EntryDate and ReadCode columns are the same) usually have one row with a valid Codeunits column, and one with a null value. In those cases we take the valid value and disregard the null value. Where there are different Codeunits, they usually correspond to the same thing (e.g. **ONE TABLET TWICE A DAY** and **ONE IN THE MORNING, ONE IN THE PM**). For the few case where they don't we take the widest range, e.g. if the duplicate instructions where **ONE A DAY**, and **2-4 PER DAY**, then we would treat this as **1-4 PER DAY**.

## Process overview

This is the general overview of the cleaning process. There are also sections below with specific cleaning operations for different opioid forms e.g. tablets, patches, oral solution.

For each prescription we establish the following:

1. **DATE**: The date of the prescription
2. **NAME**: The opioid name
3. **SIZE**: One of
  - **MG\_PER\_TABLET** - mgs per tablet
  - **MG\_PER\_ML** - the strength of oral solutions
  - **MCG\_PER\_HOUR** - the strength of patches
4. **TOTAL\_PRESCRIBED\_MG**: Total mgs prescribed
5. **MG\_PER\_DAY**: Mgs per day consumed
6. **MME**: The morphine milligram equivalent

**DATE** is taken straight from the **EntryDate** column.

---

**NAME** and **SIZE** are mapped directly from the clinical code (Read v2 and EMIS codes). Each code corresponds uniquely to the **NAME** (e.g. morphine / codeine / fentanyl), and the **SIZE** of the opioid (e.g. 30mg tablets, 5mg/ml syrup, 12mcg/hour patch)

Every clinical code in the data that corresponds to an opioid has been mapped to the active ingredient and size. For example

Code	Description	Type	Ingredient	Size
dia2.	CO-CODAMOL 8mg/500mg tablets	Read v2	codeine	8mg/tablet
OXM/2665NEMIS	Oxycodone Hydrochloride M/R tablets 10 mg	EMIS	oxycodone	10mg/tablet
o42S.	MATRIFEN 12micrograms/hr patches	Read v2	fentanyl	12mcg/hour
dj1l.	ORAMORPH 10mg/5mL liquid 100mL	Read v2	morphine	2mg/ml

The exception is the code **djks.** which seems to correspond to two different formulations of Oxycodone in the data - 10mg/ml and 1mg/ml. The instructions given to the patient confirm this, e.g. for 10mg/ml there are instructions like "take 15mg (1.5ml)", and for 1mg/ml there are instructions like "take 10mg (10ml)". Therefore we use the ReadCode AND the description to determine which formulation was actually prescribed.

**MME** is taken from a predefined lookup as follows (only opioids and forms that exist in the data are listed below):

Opioid	Route	MME
codeine	oral	0.15
morphine	oral	1.0
oxycodone	oral	1.5
fentanyl	oral	130.0
fentanyl	topical	100.0
tramadol	oral	0.1
co-codamol	oral	0.15
buprenorphine	topical	75.0
buprenorphine	sublingual	30.0
dihydrocodeine	oral	0.25
hydromorphone	oral	4.0
meptazinol	oral	0.02
pethidine	oral	0.1
tapentadol	oral	0.4
pentazocine	oral	0.37
methadone	oral dose <= 20 mg/day	4
methadone	oral dose <= 40 mg/day	8
methadone	oral dose <= 60 mg/day	10
methadone	oral dose > 60 mg/day	12

**TOTAL\_PRESCRIBED\_MG** depends on the form:

- for tablets, we multiply the **SIZE** (in mg/tablet) by the CodeValue (as the number of tablets prescribed)

- for oral solutions, we multiple the **SIZE** (in mg/ml) by the CodeValue (as the amount of mls prescribed)
- for patches, we multiple the **SIZE** (in mcg/hr), by 24 (to get mcg/day), by 3 or 7 (depending on the patch wearing length), and by the CodeValue (as the number of patches prescribed)
  - sometimes the CodeValue for patches is "2 packs", which we take to mean 10 patches as this only occurs for Fencino prescriptions which come in 5-packs.
- for sachets, we multiple the **SIZE** (in mg/sachet) by the CodeValue (as the number of sachets prescribed)

**MG\_PER\_DAY** is derived from the **Codeunits** column, which is the instruction given to the patient. The R package **doseminer** is used to convert these into a **LOWEST\_DAILY\_DOSE** and a **HIGHEST\_DAILY\_DOSE**. E.g. "Take 1-2, 3 times a day" would have a **LOWEST\_DAILY\_DOSE** of 3 and a **HIGHEST\_DAILY\_DOSE** of 6. Some instructions are parsed incorrectly by doseminer and so are manually changed.

The **LOWEST\_DAILY\_DOSE** and **HIGHEST\_DAILY\_DOSE** are created as:

- a number of tablets/sachets, for tablets/sachets
- a number of mls for oral solutions
- a number of patches for patches

Once the above are determined, we have enough info to calculate how much opioid (and MME) is consumed daily, and when prescriptions end due to the supply being exhausted. For each prescription:

1. We calculate a **LOWEST\_DAILY\_MG** and a **HIGHEST\_DAILY\_MG**
  - **LOWEST\_DAILY\_MG** = **LOWEST\_DAILY\_DOSE** x **SIZE** (for tablets, sachets, and oral solutions)
  - **LOWEST\_DAILY\_MG** = **LOWEST\_DAILY\_DOSE** x **SIZE** x 24 (for patches as the **SIZE** is mcg/hour)
  - **NB** If the opioid is optional (e.g. the instruction includes something like **as required**) then the **LOWEST\_DAILY\_MG** is 0
2. We look for the next prescription of the same drug
  - For a prescription that is followed by other similar prescriptions:
    - We calculate the **DAILY\_DOSE\_MG** required to exhaust the supply exactly at the point of the next prescription
    - If the **DAILY\_DOSE\_MG** is within the range [**LOWEST\_DAILY\_MG**, **HIGHEST\_DAILY\_MG**], then we assume that is the actual daily dose
    - If the **DAILY\_DOSE\_MG** is above the **HIGHEST\_DAILY\_MG**, then we assume the patient took the **HIGHEST\_DAILY\_MG**
    - If the **DAILY\_DOSE\_MG** is below the **LOWEST\_DAILY\_MG**, then we assume the patient took the medication at an average of the **LOWEST\_DAILY\_MG** and the **HIGHEST\_DAILY\_MG**, until it ran out, then had a gap until their next prescription
  - For the last prescription of an item
    - We assume they took the medication as instructed and take the **DAILY\_DOSE\_MG** as the average of the **LOWEST\_DAILY\_MG** and the **HIGHEST\_DAILY\_MG**
  - For both, if the instruction is unclear or missing and we can't calculate a **LOWEST\_DAILY\_MG** or a **HIGHEST\_DAILY\_MG**, then we
    - use the most recent instruction for the same medication for this patient, or
    - if that doesn't exist we assume they exhausted their supply exactly in time for their next prescription, or
    - if it's the last prescription, or the exhaust dose is smaller than a minimum amount (e.g. less than 1 tablet a day), then we assume a 28 day prescription and use a **DAILY\_DOSE\_MG** to exhaust the medication in 28 days

## Category

In case it is useful to analysts, each entry in the final table contains a **Category** which tries to describe the underlying prescription from which the daily dose is calculated. The categories are:

Category	Description
Next Rx within range	This is not the last prescription of the particular medication. The next prescription of this item is within the expected date range assuming the patient takes the prescription correctly. The daily dose is set to exhaust the supply before the next Rx.
Next Rx too far in future	This is not the last prescription of the particular medication. Even if the patient took the lowest prescribed dose, they will run out before the next prescription. The daily dose is the average of the lowest and highest daily dose.
Next Rx too soon	This is not the last prescription of the particular medication. Even if the patient took the highest prescribed dose, they would still have medicine at the time of their next prescription. The daily dose is set as the highest daily dose, but the ExhaustDose is set to exhaust the supply (overdose) before the next Rx.
Next Rx exists but unknown length of current Rx	This is not the last prescription of the particular medication. There is not enough information to calculate the length. We use the previous instruction for daily dose. If that doesn't exist we assume a daily dose to exhaust the supply. If that daily dose would be too small (e.g. < 1 tablet a day), then we assume a 28 day Rx, or 1 dose per day, whichever is smaller.
Last Rx - known length	This is the last prescription of the particular medication. There is enough information to calculate the length. Daily dose is the average of the lowest and highest daily doses.
Last Rx - unknown length	This is the last prescription of the particular medication. There is not enough information to calculate the length. Daily dose is based on a 28 day Rx, or 1 dose per day, whichever is smaller.

## Exhaust Dose

In the situation where a patient receives a subsequent prescription, before they would have used up the previous prescription, the assumption above is that they used the highest prescribed dose. An alternative would have been to assume that they took more than the highest prescribed dose, and exhausted their supply by the time of the next prescription. In these situations we use the `DoseToExhaustInMg` and `ExhaustMME` columns to show how much they could have consumed based on what they had in their possession. This may be useful for sensitivity analyses, or as an alternative analysis using different assumptions.

## Injections

Less than 1% of the prescriptions are for injections, and most of these are likely palliative care - which is not of interest in this analysis. Injections are therefore ignored from the main output file. Instead a file `/dbfs/mnt/analysis-outputs/primary-care-injections.csv` is created with the following columns:

- `PseudonymisedID` - the unique patient id
- `FirstInjectionPrescription` - the first date of an injection prescription
- `LastInjectionPrescription` - the last date of an injection prescription
- `NumberOfPrescriptions` - the total number of injection prescriptions
- `DifferenceInDaysBetweenFirstAndLast` - number of days between first and last injection prescription

This should be enough to confirm that most are palliative care and can be ignored - but if there are a substantial number of records that may not be palliative care, then this can be revisited.

## Patches

Patches are typically administered for a certain number of days. We use the instruction (every 4 days) to determine the duration. However, for patches where the instruction is unclear, e.g. "User schedule", then we assume that fentanyl patches are worn for 3 days, and buprenorphine patches are worn for 7 days.