Technical Note: Deterministic Linkage of Police Force Data and National Drug Treatment Monitoring Service Data

Rationale, Method, and Validation Checks

Ashley Mills (ajsm@kent.ac.uk)

2025-12-16

Table of Contents

1 Introduction

This document accompanies the quantitative component of the Police Drug Diversion (PDD) evaluation, which assesses whether diversion schemes operating in selected police force areas influence offending and drug treatment outcomes for drug-involved suspects. The broader study design—described in detail in (Stevens et al. 2023) uses observational cohorts of adults who came into contact with the police for qualifying drug-related offences between October 2021 and September 2022. Outcomes are derived from national administrative datasets, including the Police National Computer (PNC) for reoffending measures and the National Drug Treatment Monitoring System (NDTMS) for treatment-related indicators. The core analytic strategy involves comparing individuals eligible for diversion in PDD forces with those in carefully matched comparison forces, controlling for age, sex, ethnicity, prior offending, prior treatment, reason for police contact, and temporal and force-level confounders.

To enable this analysis, individual-level linkage is required between police force datasets (PFD), PNC, and NDTMS. The participating police forces supplied lists of eligible suspects to the Department of Health and Social Care (DHSC) for central linkage. PNC linkage was carried out by Ministry of Justice according to their own criteria, whereas NDTMS linkage was performed by the DHSC secondment team. The reliability of downstream analyses depends critically on the quality, transparency, and defensibility of the linkage process. The aim of the present document is therefore to describe the variables available for NDTMS linkage, outline the constraints imposed by the data structure, and justify the selection of a deterministic matching strategy. Given the absence of a viable gold standard and the empirical behaviour of the available identifiers, deterministic linkage represents the most robust and interpretable approach for matching PFD and NDTMS records within the parameters of the study.

2 Available variables

Linkage between PFD (Police Force Data) and NDTMS (National Drug Treatment Monitoring System) is necessary in order to assess the drug treatment outcomes. The two datasets have the following variables in common:

1. FirstNameInitial - The initial of the first name (PFD has full first name and must be converted)
2. LastNameInitial - The initial of the last name (PFD has full last name and must be converted)
3. DOB - Date of birth
4. Sex - Sex category: `M`,`F`, or `NA`
5. EthCoarse - Categorical ethnicity from: `WHITE`,`MIXED`,`ASIAN`,`BLACK`,`OTHER`, and `NA`
6. ForceName - Name of the force that the suspect was involved with for PFD and the name of the force that the NDTMS user's postcode sector belongs to for the NDTMS data.

Note that while ForceName represents the policing force of contact in PFD and the postcode-derived force area in NDTMS, both function as geographic anchors rather than institutional identifiers, and are therefore appropriate for spatial restriction rather than exact institutional equivalences.

3 Linkage validation

One of the fundamental issues linking between data sets X and Y on columns C is validating that matches are true matches and that non-matches are true negatives. Ideally we would like a linkage with the following properties:

Match Type	Ideal Value	Fellegi-Sunter Correlates
True Positive	High (close to 100%)	m
False Negative	Low (close to 0%)	$1 - m$
True Negative	High (close to 100%)	u
False Positive	Low (close to 0%)	$1 - u$

We can train a model using the Fellegi-Sunter method to estimate the m and u parameters for each matching variable and matching level within each variable and as such derive probability estimates for whether one record matches another. This process however relies on two key parameters that must be estimated: the prior match probability which represents the probability that two random records match and the posterior probability threshold which is assumed to represent a true match and is used as a filter to extract the final matching pairs. We will discuss these two in turn.

3.1 Prior match probability

The prior match probability is the probability that two records picked at random will match and serves to construct the initial estimates of m and u. This can be crucial in assuring a good model. For example: if we were matching records from year one of a university course with year two at the same university we would expect a higher prior match probability (correlated with the course attrition rate) than if we were matching records from year one and two at different universities.

In our case we have data collected from police about suspects who are either picked up for a drug offence or are known to be associated with drugs. And then we have NDTMS which is a database of drug treatment cases throughout the UK. Already then these datasets are likely to overlap significantly but the extent to which this is the case is unknown.

We can make the prior match probability even higher by restricting NDTMS to those people who entered into drug treatment AFTER contact with the police in our police data. Since one of the diversion mechanisms employed by some police forces is enrollment into drug treatment, the logical sequence implies a greater overlap with this restricted NDTMS set.

Furthermore we don’t actually care how many people in the whole of NDTMS match our police records, since the outcome we are analysing is intrinsically a post-contact one. It just makes for a good story to say that some certain % of all NDTMS records match and then the % subset which match with respect to the outcome we care about. It’s a nice-to-have but not essential for the analysis.

The prior match probability can be estimated by assigning a true-positive likelihood to a deterministic match rule and then randomly sampling the data pool, so this leaves us with the following question:

Q1: If a PFD record and an NDTMS record match on {FirstNameInitial,LastNameInitial,DOB,Sex,Ethnicity,ForceName}, then how likely is it that this is a true match?

If we had a gold standard sample of records that we knew were a true match between PFD and NDTMS we could then estimate the likelihood of a deterministic match on these variables representing a true match. Unfortunately we do not have such a mapping.

3.2 PFD data as a proxy for a gold-standard

What we do have is a subset of the PFD data that is uniquely identifiable by PNC number. So we can estimate how precise our matching variables are in Q1 *within* the PFD data. The table below examines how the specificity of matching changes as we increase the set of matching variables (ordered by their cardinality). The table below shows the number of 1:1 matches that the deterministic match rule creates.

Number of 1:1 deterministic matches between PFD and NDTMS records based on varying sets of matching variables. The table illustrates how the inclusion of different variables (FirstNameInitial, LastNameInitial, DOB, Sex, ForceName) impacts the precision of the linkage, showing the number of unique, unambiguous matches identified at each stage.

Columns	Unique Combinations	Number linked	% linked
FirstNameInitial	26	0	0.00
FirstNameInitial, LastNameInitial	613	0	0.00
FirstNameInitial, LastNameInitial, Sex	1133	55	0.16
FirstNameInitial, LastNameInitial, Sex, DOB	34738	34182	97.69
FirstNameInitial, LastNameInitial, Sex, DOB, ForceName	34942	34475	98.53
FirstNameInitial, LastNameInitial, Sex, DOB, ForceName, EthCoarse	35401	34064	97.35

As can be seen from the table `FirstNameInitial` and `LastNameInitial` are not discerning enough in combination to produce any 1:1 links. Adding in `Sex` creates a few matches but nothing noteworthy. It is only with the addition of `DOB` that the number of unique matches increases and then it is a very good match. Adding in the `ForceName` further discriminates individuals and is likely to be more powerful when matching with

NDTMS since NDTMS individuals can come from anywhere in the country, not just our selected police forces.

Adding in `Ethnicity` however reduces the precision of the deterministic match because some of the `PNCNumber` in the match calculation now map to more than one `Ethnicity`. This is likely because this variable is mostly police-reported and this process is subject to error. Examples of this from manual observation of the data: when Given that `Ethnicity` seems prone to error we have decided to exclude it from the match variables.

Initials and sex along with DOB appear to be very good within the PFD data at determining genuine matches. If we assume that this is also true between PFD and NDTMS then we can expect a good match from a deterministic rule.

We can also compute how many individuals appear in our data in more than one police force, to get an estimate of how localised individuals are within the capture window. Only 15 of 28542 individuals (~0.05%) occur in more than one force (and never more than two forces). While this is limited by being able to determine moves only between police forces we have evidence for, the very low number indicates that this cohort is likely to be fairly stationary within a window of 1 year.

3.3 Posterior threshold probability

The second issue concerns how we classify matches once a scoring system is in place. In probabilistic linkage frameworks, each candidate pair is assigned a posterior probability of being a true match, and a threshold is chosen to determine which pairs are accepted. The choice of threshold is crucial: set too low and many false positives are admitted, set too high and many true matches are lost. In practice, thresholds are usually calibrated against a gold standard (e.g. clerical review or auxiliary identifiers).

In our case, no such gold-standard mapping exists between PFD and NDTMS. This creates a fundamental problem: posterior probabilities cannot be validated. A match scored as 0.9 in a Fellegi–Sunter model may or may not be a true match, and without ground truth we cannot know. Introducing unverifiable probabilistic thresholds would therefore embed assumptions into the linkage without any empirical way to test or justify them. Given the policy-facing nature of the analysis, such unverifiable modelling would weaken rather than strengthen the evidence base.

4 Deterministic linkage

4.1 Justification

Given the limitations outlined above, deterministic linkage provides the only defensible approach for this analysis. The available matching variables in PFD and NDTMS are few in number, and crucially there is no gold-standard dataset against which a probabilistic linkage model could be calibrated. Without a validated estimate of the prior match probability or an empirically defensible posterior probability threshold, probabilistic scoring would inject unverifiable assumptions into the linkage process and could not be reliably interpreted.

By contrast, the deterministic rule based on {FirstNameInitial, LastNameInitial, DOB, Sex, ForceName} has several advantages.

First, its empirical properties can be directly assessed using the subset of PFD records with unique PNC identifiers. These tests show that initials and sex alone are insufficient, but the addition of date of birth yields a highly discriminating rule that produces a large number of true 1:1 matches and very few ambiguous links. Adding ForceName further improves the specificity of linkage by restricting matches to plausible local catchments. These findings indicate that, within the PFD data, the chosen variable set has high precision for identifying unique individuals.

Second, the underlying population structure makes this rule even more reliable when applied across PFD and NDTMS. Individuals in the PFD cohort are extremely unlikely to appear in more than one police force within the capture window (0.05% observed cross-force movement). NDTMS treatment entries are likewise geographically anchored. This spatial localisation increases the discriminative power of the deterministic rule and reduces the probability of spurious matches.

Third, we already expect a high degree of overlap between the two systems being linked. Police involvement for drug possession or related offences, and entry into structured drug treatment, are not independent processes: many police forces explicitly refer individuals into treatment services as part of diversion pathways, and individuals with entrenched drug-related problems frequently appear in both systems. This structural correlation raises the prior match probability and further increases the plausibility that a deterministic match on strong identifiers represents a genuine link rather than a chance alignment.

Ethnicity was excluded from the match criteria because incorporating it reduced precision in the PFD validation exercise—an effect consistent with known limitations of officer-defined ethnicity and the potential for misclassification. The deterministic rule therefore uses only variables that demonstrate both stability and discriminative strength.

Taken together, the evidence shows that deterministic linkage:

1. Provides empirically validated precision using internal gold-standard proxies (PNC-unique PFD records).
2. Avoids reliance on untestable probabilistic assumptions that cannot be validated without ground truth.
3. Reflects the real-world overlap between policing and treatment pathways, increasing the plausibility of true matches.
4. Minimises the risk of false positives, which is essential in policy-facing evaluation of treatment outcomes.

For these reasons, deterministic linkage represents the most transparent, defensible, and methodologically robust approach for linking PFD and NDTMS in this study.

4.2 Limitations

This approach will necessarily miss some true matches where identifiers are incomplete or inconsistently recorded; however, given the study aim and the absence of validation data, prioritising precision over recall is methodologically appropriate. Any remaining linkage error under this approach will preferentially bias results toward the null rather than generate spurious treatment effects.

4.3 Results

An extract of the complete NDTMS database was obtained from DHSC dated September 2025. Agency DAT codes were mapped to police force areas to provide geographical locations for each individual. NDTMS records were deterministically linked on {FirstNameInitial, LastNameInitial, DOB, Sex, ForceName} with the 44,535 individuals in the police force data, matching 17,579 unique individuals in NDTMS.

5 References

Stevens, Alex, Nadine Hendrie, Matthew Bacon, Steve Parrott, Mark Monaghan, Emma Williams, Dan Lewer et al. “Evaluating police drug diversion in England: protocol for a realist evaluation.” Health & justice 11, no. 1 (2023): 46.