Supplementary material: Data Linkage process

Retrospective data linkage will examine contacts with the health and justice systems prior to and post referral for the whole data set held by PALM from 2000-2014 of approximately 4000 clients addressing aims one-three.

Data collection
Data for the complete PALM client database (TED) from the year 2000 will be linked to administrative data from a series of health and criminal states government data sets from two states in Australia. An external independent data linkage organisation will be used for data linkage of personal identifiers within the data sets specified currently in the Master Link Key and external data set personal identifiers provided by data custodians to create a study unique identifier or Project-Specific Person Identification Number or (PPN or 'linkage key'). De-identified data for the 5 years prior to first record in PALM starting in 2000 where available, to most recently available will then be requested from data custodians against the PPN. De-identified data attached to the PPN will then be provided to the research team (figure 2).

Insert Figure S1. Flow diagram of data linkage process here

Record linkage brings together information that relates to the same individual from different data sources. In this way it is possible to construct chronological sequences of health events for individuals. Combined, these individual 'stories' create a larger story about the health and life pathways of these young people in two Australian states.

In brief, two separate datasets will be generated in the process:

1. for linkage containing only a limited set of personal identifiers created and held by the data linkage organisation only; and
2. for analysis by the research team, with name and full address removed ('de-identified') but including the approved health and justice information, age, gender and postcode.

This process ensures that:

- Data linkage organisation staff performing the linkage use demographic variables but do not have access to the administrative information about the individuals;
- Data custodians only have access to data within their data collections; and
- Researchers receive data which contains no identifying variables, or variables which provide a link back to the data linkage organisation's MLK or external custodian data sets.

Step 1: Seek approval to conduct research

- Data linkage organisation (DL org) to request and feasibility assessment.
- Data custodian support.
- Research team seeks ethics approvals

Step 2: Data linkage organisation to create PPN for study and request identifiers from other external custodians

- External custodian 1 (Ted Noffs) provides an encrypted source record number (SRN) and personal identifiers from TED to DL org.
- DL org assigns PPN* to each SRN in TED.
- Two external data custodians and MLK custodians requested to send personal identifiers in date range specified with encrypted ID from their data set.

External Custodian 2: Criminal justice in state 1

External Custodian 3: Criminal justice in state 2

MLK Custodians: Health data sets in both states

Step 3: DL org to conduct identifier matching and request data variables be sent to research team

- DL org uses probabilistic matching and assigns the DL org PPN from TED for records belonging to same person to all other data custodians matched identifiers
- DL org sends PPN to data custodians with encrypted ID from their data set requesting data variables be extracted for each PPN and sent to research team.

Step 4: Data sets sent to research team

- Custodians send data variables requested against PPNs to research team.

Step 5: Study data base and research

- Research team receives and merges de-identified data files on secure server, cleans and analyses to answer research questions.