New OSHPD Linkage Products Release

The Office of Statewide Health Planning and Development (OSHPD) is happy to announce the release of new linked data products. The first two linked 2018 and 2019 Patient Discharge Data (PDD) to the California Comprehensive Master Death Files (CCMDF) (PDD_CCMDF_2018, 2019) and the other three linked 2017-2019 Emergency Department (ED) Data to the 2017 California Comprehensive Death Files (CCDF) (ED_CCDF_2017) and the 2018-2019 CCMDF (ED_CCMDF_2018, 2019). These products are produced using a probabilistic linkage model developed jointly by ChoiceMaker, LLC and OSHPD utilizing machine learning to mimic human intuition on record matching.
A brief comparison of the newly released linkage data products with the previously released products is listed in thetable below.

OSHPD Linkage Data Products								
Dataset Name	**New Release**		**Previously Released**					
	PDD_CCMDF_YYYY	**ED_CCDF_2017** **ED_CCMDF_YYYY**	**PDD_CCDF_YYYY**	**EDD_CCDF_YYYY**	**PDD_VSD_YYYY**	**PDD_Death_YYYY**	**EDD_Death_YYYY**	**ASD_Death_YYYY**
Years Available	2018-2019	2017-2019	2014-2017	2014-2016	1990-2013	2005-2013	2005-2013	2005-2013
Data Sources	PDD and CCMDF	EDD and CCDF/CCMDF	PDD and CCDF	EDD and CCDF	PDD and DSMF	PDD and DSMF	EDD and DSMF	ASD and DSMF
Model	Probabilistic Linkage with Machine Learning				Probabilistic Linkage	Deterministic Linkage		
Notes:	• The "YYYY" in the dataset name indicates the year of the source files linked • PDD: Patient Discharge Data; EDD: Emergency Department Data; ASD: Ambulatory Surgery Data • CCDF: the California Comprehensive Death File, used for 2014-2017 • CCMDF: the California Comprehensive Master Death File, used for 2018 and 2019 • DMSF: Death Statistical Master File							

There are some unique features in the released linkage data products produced by probabilistic approach with machine learning:

1. Record matching:
   1) The death records within one calendar year are linked to PDD and ED records within the same calendar year,not across all years;
   2) The death records are linked to all hospitalizations or all ED encounters of the same patient, not just the finalhospitalization or encounter of the same patient within the calendar year;

2. Variables within the datasets:
   1) Naming convention:
      - ‘IP_’ indicates that the variables originated in the PDD;
      - ‘ED_’ indicates that the variables originated in the ED;
      - ‘VD_’ indicates that the variables originated in the CCDF.

2) For all variables from the CCDF, the field sequence number is listed in the variable labels;
3) The unique identifiers:
    - For the death records is the state file number (VD_fileno);
    - For PDD and ED is created as a concatenation of data_id and pat_id as PDDid and EDid, respectively;
4) Two unique ChoiceMaker generated variables, the ChoiceMaker Probability (CM_probability) andChoiceMaker Decision (CM_decision), are provided in the linkage files:
    - CM_probability is the conditional probability or match probability computed by the Maximum Entropy model of a linked pair with a value between 0 and 1;
    - CM_decision indicates the decision made by ChoiceMaker according to the computed CM_probability on whether a linked pair is considered a match ("M", CM_probability >=0.95) or a hold ("H", 0.2 =< CM_probability <0.95). Pairs with CM_probability <0.2 are considered differ and are not included in the data product;
5) Two calculated variables DysAdmtDth and DysDschDth indicate the days from admission to death and from discharge to death, respectively;
6) The diagnosis codes and procedure codes in PDD and ED were updated from ICD-9-CM to ICD-10-CM and ICD-10-CM/PCS on 10/01/2015.