ResDAC
RESEARCH DATA ASSISTANCE CENTER

Introduction to DE-SynPUF

04/09/2013

Presented by Elizabeth Hair, PhD, NORC at University of Chicago

Moderated by Erin Mann, ResDAC

About ResDAC

- Centers for Medicare and Medicaid (CMS) contractor
- Offer free assistance to researchers interested in using Medicare and Medicaid data for research
- Provide a range of services related to CMS data
  - Assistance Desk
  - Workshops and Outreach

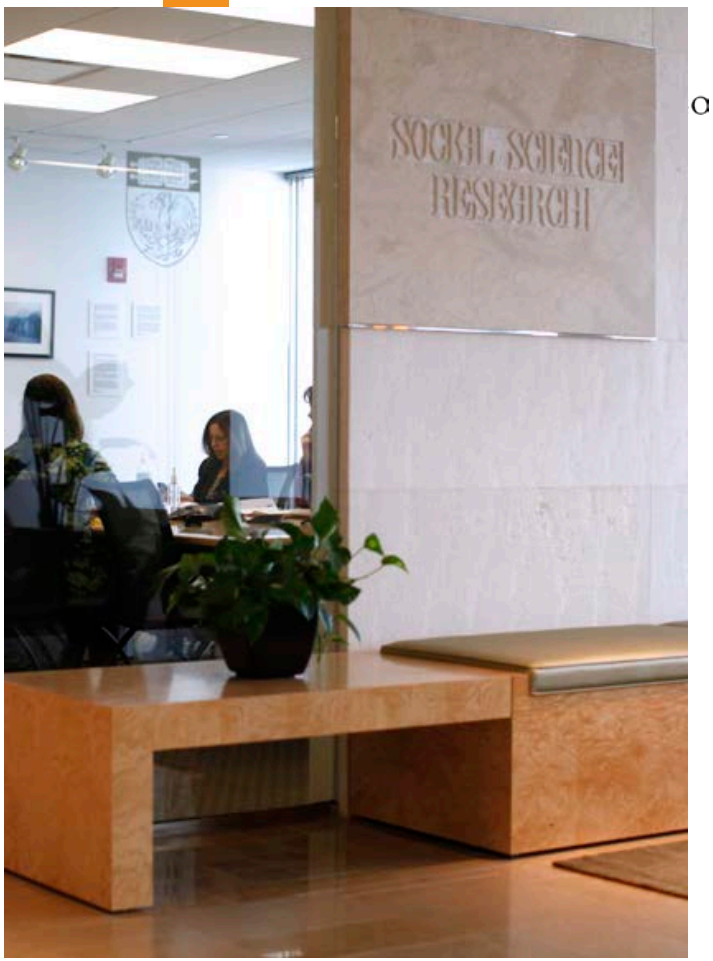
Webinar Series on CMS Data

- 02/28 – Introduction to CMS Data
- 03/19 – Non-Identifiable Data
- 04/04 –Cost Reports
- **04/09 - DE-SynPUFs**
- 04/10 –Limited Data Sets
- 04/25 –Research Identifiable Data
- 05/02 – Utilization Data

View past webinars and register for upcoming webinars at the ResDAC website (www.resdac.org)

Overview of Data Entrepreneurs' Synthetic PUF (DE-SynPUF) for CMS Medicare Claims Data

The CMS Research Data Assistance Center (ResDAC)
and
NORC at the University of Chicago



Presentation for ResDAC

April 9, 2013 at 12PM CT/ 1PM ET

NORC at the UNIVERSITY of CHICAGO

DE-SynPUF Team

- Technical Team
  - Avi Singh
  - Josh Borton
  - Amanda Tzy-Chyi Yu
  - Al Crego
- Re-identification Team
  - Fritz Scheuren
  - Susan Hinkins
  - Patrick Baier
- Project Management

  CMS
  - Chris Haffer

  IMPAQ
  - Erkan Erdem
  - Slava Katz

  NORC
  - Mike Davern
  - Elizabeth Hair
  - Margrethe Montgomery

Agenda

NORC at the UNIVERSITY of CHICAGO

- Background
- DE-SynPUF Development
- Comparison of DE-SynPUF and Real Data
- Re-identification and Certification
- Documentation for DE-SynPUF
- Next steps

Background

NORC at the UNIVERSITY of CHICAGO

DE-SynPUF Development

Purpose of Data Entrepreneurs' SynPUF (DE-SynPUF)

NORC at the UNIVERSITY of CHICAGO

- New type of 'synthetic' file useful for data entrepreneurs for software and application development and training purposes
- Preserve detailed data structure of key variables at beneficiary and claim levels
  - Data is fully 'synthetic' for disclosure safety
  - Little or no analytic utility due to lack of preservation of interdependence between variables
- Created file that were certified and released as a Public Use File (PUF) in Feb 2013.

Guiding Principle for DE-SynPUF

NORC at the UNIVERSITY of CHICAGO

- Essential that the confidentiality of the Medicare beneficiaries are protected
- The DE-SynPUF is based on a 5% sample of Medicare beneficiaries and includes beneficiary summary data, inpatient, outpatient, carrier, and PDE claims data
- Same structure, metadata and size that allows entrepreneurs to build tools that will work on DE-SynPUF as well as the real data
- Very limited analytic utility to ensure the file is safe from potential re-identification threats

Summary of requests from DE

- **Structure:** Roundtable participants stressed to us their preference for a file that mirrored the "real" CMS data even if it was of low analytic utility.
- **Geography**: County would be most useful. State is not helpful; zip code would be useful, but is not necessary.
- **Time**: Year is not enough; day is needed.
- **Linking**: The ability to link across files is important.
- **Longitudinal data**: Three years of data would be better than one year of data.
- **Data Documentation**: Metadata on types and structure of variables would be helpful.
- **Provider information:** Include provider and institution IDs

DE-SynPUF Contents

- 5% sample of enrolled Medicare beneficiaries in 2008
- 3 years of claims (2008, 2009, 2010)
  - Inpatient
  - Outpatient
  - Carrier
  - PDE – Prescription Drugs
- Detailed contents of each table in Appendix A

DE-SynPUF Contents (cont.)

NORC at the UNIVERSITY of CHICAGO

- **DE-SynPUF Subsample Files:**
  - 20 separate subsamples.
  - In each subsample, there are 8 CSV files that contain the raw data for that subsample.
  - Each subsample contains all the beneficiary data and claims data for the subsample of beneficiaries.
  - Users can work with anywhere from 1 to all 20 subsamples.

DE-SynPUF Contents (cont.)

NORC at the UNIVERSITY of CHICAGO

Table 1. File Names of the Eight CSV Files Pertaining to Five File Types in each DE-SynPUF Subsample

File type	CSV File name	Number of Years of Data
Beneficiary Summary DE-SynPUF	DE1_0_2008_Beneficiary_Summary_File_Sample_#	1
	DE1_0_2009_Beneficiary_Summary_File_Sample_#	1
	DE1_0_2010_Beneficiary_Summary_File_Sample_#	1
Inpatient Claims DE-SynPUF	DE1_0_2008_to_2010_Inpatient_Claims_Sample_#	3
Outpatient Claims DE-SynPUF	DE1_0_2008_to_2010_Outpatient_Claims_Sample_#	3
Prescription Drug Events (PDE) DE-SynPUF	DE1_0_2008_to_2010_Prescription_Drug_Events_Sample_#	3
Carrier Claims DE-SynPUF	DE1_0_2008_to_2010_Carrier_Claims_Sample_#A	3
	DE1_0_2008_to_2010_Carrier_Claims_Sample_#B	3

NOTE: The "#" symbol takes on the values from 1 – 20 and is the subsample number (e.g., subsample 1 the 2008 Beneficiary Summary DE-SynPUF is called "DE1_0_2008_Beneficiary_Summary_File_Sample_1")

SAS READIN Files

NORC at the UNIVERSITY of CHICAGO

- The provided SAS READIN programs allow users to specify which subsamples to read in.
- The SAS READIN programs read in CSV data files and transform them into SAS data sets.
- There are five SAS READIN programs: one for each file type.
- Document: Instructions for the SAS READIN Files
- Users are advised to carefully read the instructions included in each of the five SAS READIN program before running any one of them

DE-SynPUF Description

- The data structure is very similar to the CMS limited data sets, albeit with a smaller number of variables
- Programs and procedure designed using the SynPUF are fully functional when applied to CMS limited data sets

DE-SynPUF Description (cont.)

NORC at the UNIVERSITY of CHICAGO

- The variable names in DE-SynPUF were kept the same as those in the real Medicare data unless the data values were altered to protect provider or beneficiary privacy. In those rare cases when the data values were significantly altered, we added the prefix “SP_” to the original variable name.

Some comparison of DE-SynPUF and Real Data

Comparison of Actual and DE-SynPUF Estimates – Demography

NORC at the UNIVERSITY of CHICAGO

Gender	DE-SynPUF (%)	2008 5% (%)
Male	44	45
Female	56	55

Race	DE-SynPUF (%)	2008 5% (%)
White	83	83
Black	11	10
Other	4	4
Hispanic	2	2

Comparison of Actual and DE-SynPUF Estimates – Year of Birth

NORC at the UNIVERSITY of CHICAGO

Year of Birth	DE-SynPUF (%)	2008 5% (%)
post 1973	5	5
1964-1973	8	8
1954-1963	13	12
1944-1953	16	15
1939-1943	19	19
1934-1938	24	24
1929-1933	7	7
1924-1928	5	5
1919-1923	2	3
pre 1919	1	1

Comparison of Actual and DE-SynPUF Estimates – Claims

	DE-SynPUF	2008	DE-SynPUF	2009	DE-SynPUF	2010
	Percent[a] (%)	Percent (%)	Percent (%)	Percent (%)	Percent (%)	Percent (%)
Inpatient	14	16	16	15	11	15
Outpatient	51	50	63	50	49	50
Carriers	73	70	80	70	76	70
PDE	63	53	79	56	74	57

Note: [a] Percent of beneficiaries with at least one claim in a certain claim type

Comparison of Actual and DE-SynPUF Estimates – Reimbursement

NORC at the UNIVERSITY of CHICAGO

	DE-SynPUF	2008	DE-SynPUF	2009	DE-SynPUF	2010
	Mean	Mean	Mean	Mean	Mean	Mean
Total Inpatient	2,550	2,850	2,500	3,050	1,450	3,050
Total Outpatient	850	1,150	1,050	1,250	600	1,300
Total Carriers	1,550	2,100	1,750	2,250	1,100	2,350
Total PDE	1,950	3,150	1,750	3,300	1,200	3,350

NORC at the UNIVERSITY of CHICAGO

Documentation for DE-SynPUF

DE-SynPUF Documentation

- Methodology report (for CMS only)
- For CMS website and public distribution
  - User's Manual
    - Which includes a basic data utility analysis
  - Codebook of Variables
  - FAQ
  - SAS read-in programs
- NORC is funding an experimental metadata manager tool that could be used to help disseminate and distribute data products like DE-SynPUF
  - All information loaded into the experimental metadata manager are in the public domain

Implications

- What do these data files mean for Data Entrepreneurs' ability to use CMS data?
  - Created files with the elements they asked for including the ability to create programs that can be run on the "real" data
  - Can be used to develop codes for application development with realistic data (in terms of structure and complexity)

Questions?

Thank You!

NORC at the UNIVERSITY of CHICAGO

insight for informed decisions™

Appendix A: DE-SynPUF Tables

NORC at the UNIVERSITY of CHICAGO

Beneficiary Table – part 1

NORC at the UNIVERSITY of CHICAGO

#	Variable names	Labels
1	DESYNPUF_ID	DESYNPUF: Beneficiary Code
2	BENE_BIRTH_DT	DESYNPUF: Date of birth
3	BENE_DEATH_DT	DESYNPUF: Date of death
4	BENE_SEX_IDENT_CD	DESYNPUF: Sex
5	BENE_RACE_CD	DESYNPUF: Beneficiary Race Code
6	BENE_ESRD_IND	DESYNPUF: End stage renal disease Indicator
7	SP_STATE_CODE	DESYNPUF: State Code
8	BENE_COUNTY_CD	DESYNPUF: County Code
9	BENE_HI_CVRAGE_TOT_MONS	DESYNPUF: Total number of months of part A coverage for the beneficiary.
10	BENE_SMI_CVRAGE_TOT_MONS	DESYNPUF: Total number of months of part B coverage for the beneficiary.
11	BENE_HMO_CVRAGE_TOT_MONS	DESYNPUF: Total number of months of HMO coverage for the beneficiary.
12	PLAN_CVRG_MOS_NUM	DESYNPUF: Total number of months of part D plan coverage for the beneficiary.
13	SP_ALZHDMTA	DESYNPUF: Chronic Condition: Alzheimer or related disorders or senile
14	SP_CHF	DESYNPUF: Chronic Condition: Heart Failure
15	SP_CHRNKIDN	DESYNPUF: Chronic Condition: Chronic Kidney Disease
16	SP_CNCR	DESYNPUF: Chronic Condition: Cancer
17	SP_COPD	DESYNPUF: Chronic Condition: Chronic Obstructive Pulmonary Disease
18	SP_DEPRESSN	DESYNPUF: Chronic Condition: Depression
19	SP_DIABETES	DESYNPUF: Chronic Condition: Diabetes

Beneficiary Table – part 2

#	Variable names	Labels
20	SP_ISCHMCHT	DESYNPUF: Chronic Condition: Ischemic Heart Disease
21	SP_OSTEOPRS	DESYNPUF: Chronic Condition: Osteoporosis
22	SP_RA_OA	DESYNPUF: Chronic Condition: rheumatoid arthritis and osteoarthritis (RA/OA)
23	SP_STRKETIA	DESYNPUF: Chronic Condition: Stroke/transient Ischemic Attack
24	MEDREIMB_IP	DESYNPUF: Inpatient annual Medicare reimbursement amount
25	BENRES_IP	DESYNPUF: Inpatient annual beneficiary responsibility amount
26	PPPYMT_IP	DESYNPUF: Inpatient annual primary payer reimbursement amount
27	MEDREIMB_OP	DESYNPUF: Outpatient Institutional annual Medicare reimbursement amount
28	BENRES_OP	DESYNPUF: Outpatient Institutional annual beneficiary responsibility amount
29	PPPYMT_OP	DESYNPUF: Outpatient Institutional annual primary payer reimbursement amount
30	MEDREIMB_CAR	DESYNPUF: Carrier annual Medicare reimbursement amount
31	BENRES_CAR	DESYNPUF: Carrier annual beneficiary responsibility amount
32	PPPYMT_CAR	DESYNPUF: Carrier annual primary payer reimbursement amount

Inpatient Claims Table

NORC at the UNIVERSITY of CHICAGO

#	Variable names	Labels
1	*DESYNPUF_ID*	DESYNPUF: Beneficiary Code
2	*CLM_ID*	DESYNPUF: Claim ID
3	*SEGMENT*	DESYNPUF: Claim Line Segment
4	*CLM_FROM_DT*	DESYNPUF: Claims start date
5	*CLM_THRU_DT*	DESYNPUF: Claims end date
6	*PRVDR_NUM*	DESYNPUF: Provider Institution
7	*CLM_PMT_AMT*	DESYNPUF: Claim Payment Amount
8	*NCH_PRMRY_PYR_CLM_PD_AMT*	DESYNPUF: NCH Primary Payer Claim Paid Amount
9	*AT_PHYSN_NPI*	DESYNPUF: Attending Physician – National Provider Identifier Number
10	*OP_PHYSN_NPI*	DESYNPUF: Operating Physician – National Provider Identifier Number
11	*OT_PHYSN_NPI*	DESYNPUF: Other Physician – National Provider Identifier Number
12	*CLM_ADMSN_DT*	DESYNPUF: Inpatient admission date
13	*ADMTNG_ICD9_DGNS_CD*	DESYNPUF: Claim Admitting Diagnosis Code
14	*CLM_PASS_THRU_PER_DIEM_AMT*	DESYNPUF: Claim Pass Thru Per Diem Amount
15	*NCH_BENE_IP_DDCTBL_AMT*	DESYNPUF: NCH Beneficiary Inpatient Deductible Amount
16	*NCH_BENE_PTA_COINSRNC_LBLTY_AM(T)*	DESYNPUF: NCH Beneficiary Part A Coinsurance Liability Amount
17	*NCH_BENE_BLOOD_DDCTBL_LBLTY_AM(T)*	DESYNPUF: NCH Beneficiary Blood Deductible Liability Amount
18	*CLM_UTLZTN_DAY_CNT*	DESYNPUF: Claim Utilization Day Count
19	*NCH_BENE_DSCHRG_DT*	DESYNPUF: Inpatient discharged date
20	*CLM_DRG_CD*	DESYNPUF: Claim Diagnosis Related Group Code
21-30	*ICD9_DGNS_CD_1 – ICD9_DGNS_CD_10*	DESYNPUF: Claim Diagnosis Code 1 – Claim Diagnosis Code 10
31-36	*ICD9_PRCDR_CD_1 – ICD9_PRCDR_CD_6*	DESYNPUF: Claim Procedure Code 1 – Claim Procedure Code 6
37-81	*HCPCS_CD_1 – HCPCS_CD_45*	DESYNPUF: Revenue Center HCFA Common Procedure Coding System 1 – Revenue Center HCFA Common Procedure Coding System 45

Outpatient Claims Table

NORC at the UNIVERSITY of CHICAGO

#	Variable names	Labels
1	*DESYNPUF_ID*	DESYNPUF: Beneficiary Code
2	*CLM_ID*	DESYNPUF: Claim ID
3	*SEGMENT*	DESYNPUF: Claim Line Segment
4	*CLM_FROM_DT*	DESYNPUF: Claims start date
5	*CLM_THRU_DT*	DESYNPUF: Claims end date
6	*PRVDR_NUM*	DESYNPUF: Provider Institution
7	*CLM_PMT_AMT*	DESYNPUF: Claim Payment Amount
8	*NCH_PRMRY_PYR_CLM_PD_AMT*	DESYNPUF: NCH Primary Payer Claim Paid Amount
9	*AT_PHYSN_NPI*	DESYNPUF: Attending Physician – National Provider Identifier Number
10	*OP_PHYSN_NPI*	DESYNPUF: Operating Physician – National Provider Identifier Number
11	*OT_PHYSN_NPI*	DESYNPUF: Other Physician – National Provider Identifier Number
12	*NCH_BENE_BLOOD_DDCTBL_LBLTY_AM*	DESYNPUF: NCH Beneficiary Blood Deductible Liability Amount
13-22	*ICD9_DGNS_CD_1 – ICD9_DGNS_CD_10*	DESYNPUF: Claim Diagnosis Code 1 – Claim Diagnosis Code 10
23-28	*ICD9_PRCDR_CD_1 – ICD9_PRCDR_CD_6*	DESYNPUF: Claim Procedure Code 1 – Claim Procedure Code 6
29	*NCH_BENE_PTB_DDCTBL_AMT*	DESYNPUF: NCH Beneficiary Part B Deductible Amount
30	*NCH_BENE_PTB_COINSRNC_AMT*	DESYNPUF: NCH Beneficiary Part B Coinsurance Amount
31	*ADMTNG_ICD9_DGNS_CD*	DESYNPUF: Claim Admitting Diagnosis Code
32-76	*HCPCS_CD_1 – HCPCS_CD_45*	DESYNPUF: Revenue Center HCFA Common Procedure Coding System 1 – Revenue Center HCFA Common Procedure Coding System 45

Carrier Claims Table

NORC at the UNIVERSITY of CHICAGO

#	Variable names	Labels
1	*DESYNPUF_ID*	DESYNPUF: Beneficiary Code
2	*CLM_ID*	DESYNPUF: Claim ID
3	*CLM_FROM_DT*	DESYNPUF: Claims start date
4	*CLM_THRU_DT*	DESYNPUF: Claims end date
5-12	*ICD9_DGNS_CD_1 – ICD9_DGNS_CD_8*	DESYNPUF: Claim Diagnosis Code 1 – Claim Diagnosis Code 8
13-25	*PRF_PHYSN_NPI_1 – PRF_PHYSN_NPI_13*	DESYNPUF: Provider Physician – National Provider Identifier Number
26-38	*TAX_NUM_1 – TAX_NUM_13*	DESYNPUF: Provider Institution Tax Number
39-51	*HCPCS_CD_1 – HCPCS_CD_13*	DESYNPUF: Line HCFA Common Procedure Coding System 1 – Line HCFA Common Procedure Coding System 13
52-64	*LINE_NCH_PMT_AMT_1– LINE_NCH_PMT_AMT_13*	DESYNPUF: Line NCH Payment Amount 1 – Line NCH Payment Amount 13
65-77	*LINE_BENE_PTB_DDCTBL_AMT_1 – LINE_BENE_PTB_DDCTBL_AMT_13*	DESYNPUF: Line Beneficiary Part B Deductible Amount 1 – Line Beneficiary Part B Deductible Amount 13
78-90	*LINE_BENE_PRMRY_PYR_PD_AMT_1 – LINE_BENE_PRMRY_PYR_PD_AMT_13*	DESYNPUF: Line Beneficiary Primary Payer Paid Amount 1 – Line Beneficiary Primary Payer Paid Amount 13
91-103	*LINE_COINSRNC_AMT_1 – LINE_COINSRNC_AMT_13*	DESYNPUF: Line Coinsurance Amount 1 – Line Coinsurance Amount 13
104-116	*LINE_ALOWD_CHRG_AMT_1 – LINE_ALOWD_CHRG_AMT_13*	DESYNPUF: Line Allowed Charge Amount 1 – Line Allowed Charge Amount 13
117-129	*LINE_PRCSG_IND_CD_1 – LINE_PRCSG_IND_CD_13*	DESYNPUF: Line Processing Indicator Code 1 – Line Processing Indicator Code13
130-142	*LINE_ICD9_DGNS_CD_1 – LINE_ICD9_DGNS_CD_13*	DESYNPUF: Line Diagnosis Code 1 – Line Diagnosis Code 13

Prescription Drug Claims Table

#	Variable names	Labels
1	*DESYNPUF_ID*	DESYNPUF: Beneficiary Code
2	*PDE_ID*	DESYNPUF: CCW Part D Event Number
3	*SRVC_DT*	DESYNPUF: RX Service Date
4	*PROD_SRVC_ID*	DESYNPUF: Product Service ID
5	*QTY_DSPNSD_NUM*	DESYNPUF: Quantity Dispensed
6	*DAYS_SUPLY_NUM*	DESYNPUF: Days Supply
7	*PTNT_PAY_AMT*	DESYNPUF: Patient Pay Amount
8	*TOT_RX_CST_AMT*	DESYNPUF: Gross Drug Cost