May 9, 2017

Proposed Formats for Data Extraction

Summary of Tables

To collect information on the study sample, as it evolves over time, including which patients have diabetes mellitus (DM), and handle exclusion of the pregnancy period, we propose the following format. We specify the variables we will collect for each visit and formats for providing those variables in later tables.

Table 1. Summary information for study sample and DM and pregnancy information. Each patient will have one row. If date is not available, "NULL" value will be recorded.

Patient ID, [string]	**The first visit*, [date]****	**Total number of visits*, [integer]**	**DM onset date*, [date]**	**Death date, [date]**	**Pregnancy #1 date***, [date]**	**Pregnancy #2 date, [date]**	**…**	**Pregnancy #21 date, [date]**
1	2012-01-15	2	NULL	NULL	NULL	NULL		NULL
2	2010-11-03	45	NULL	2015-12-31	NULL	NULL		NULL
3	2013-11-03	15	2014-12-03	NULL	NULL			NULL
4	2016-01-27	16	NULL	NULL	NULL	NULL		NULL
…								
N	2010-04-20	89	2010-12-03	NULL	2011-07-02	2014-09-02		NULL

* Determined from inpatient, outpatient and emergency department visits recorded on different dates. See the most recent version of Definitions_Appendix_A-[date].docx for details.

** Each site should confirm to us separately (**not** in this table) any date shifting they conduct relative to actual dates. We hope to obtain from each site (i) visits date identified to year and month [we do not need exact date]; (ii) to the extent (i) is not possible, a limited time shift (same number of days for all visits by each patient (no more than 30 days) plus a dummy variable for whether an encounter is before or after Medicaid expansion on 2014.01.01.

*** See Definitions_Part1 and Definitions_Appendix_A (Part A1) for definition of pregnancy. We use pregnancy dates to exclude pregnant women from the sample during the period around their pregnancy. If more than 10 pregnancies, fill in dates for first 10.

Table 2. Demographic variables. Each patient will have one row.

Description	Name	Source	Value	Format
Arbitrary person-level identifier used to link across tables.	PATID	PCORNET (DEMOGRAPHIC table)		SAS Char(x)
Date of birth*	BIRTH_DATE		use SAS Date format We would prefer to receive *birth year and month as separate variables. If underlying data has exact birthdate, code will extract only year and month*	Numeric
Gender	SEX		F=Female M=Male A = Ambiguous NI=No information UN=Unknown OT=Other *code will combine A, NI, UN, and OT*	
Race	RACE		01=American Indian or Alaska Native 02=Asian 03=Black or African American 04=Native Hawaiian or Other Pacific Islander 05=White 06=Multiple race 07=Refuse to answer NI=No information UN=Unknown OT=Other *code will combine 06, 07, NI, UN, OT*	SAS Char(2)
Hispanic	HISPANIC		Y=Yes N=No R=Refuse to answer NI=No information UN=Unknown OT=Other *code will combine R, NI, UN, OT*	SAS Char(2)

* If this date is date shifted, please advise us on your date shifting rules. We believe that the date shifting we agreed on is as follows:

All GPC sites except MCW: Provide year and month, but not exact date. This can be achieved by providing actual year and month for an "index" encounter, perhaps the first encounter for that patient in the data, and then providing other dates relative to that index encounter. But it would also work on our end to eliminate all "day of month" values, and give all dates as year and month.

MCW: All dates for each patient will be shifted by +/- 10 days (but not 0 days) relative to actual date. The shift will be consistent for each patient.

Capricorn sites: [*to come]

General examples for tables specified below.

Tables are for visualization purposes only. All numbers are made up values.

Example 1. Example of how dates should be partially redacted to provide year and month, but not exact date or time of service. An example shows records for a single patient, with multiple measures of diastolic blood pressure on the same day. When we collapse data to year and month, we also take an average of the within-day measures. Thus, in the example, the collapsed data has two entries for 2014-02 and two entries for 2015-01. We create two new variables to keep track of the order of multiple entries within each month came first.

The new "order" variables are:

MEASURE_DATE_orderNumber = 1 for first visit in a given month; 2 for second visit, and so on.

DAYS_from_First_YYMM: Number of days from first actual date of the first encounter. For this patient, the first encounter is Feb. 15, 2014, so this variable measures days from February 15, 2014 (hence the value is 00). The second encounter is Feb. 16, 2014, so this variable will be "01". This variable allows us to assess relative dates, while still preserving the overall limit on date identification to year-month format. This variable should be omitted for encounters in the first year-year

One could recreate MEASURE_DATE_orderNumber from DAYS_from_First_YYMM, but it seemed simpler to us to collect both up front.

Comment for Medical College of Wisconsin (MCW): All GPC sites except MCW will convert actual date to year and month format. MCW will start with date-shifted dates, where the shift is +/- 10 days, and then remove the shifted date. Thus, some records that MCW reports as being in, say, February 2014 will actually reflect visits in late January or early February 2014; some visits in early February will be reported as being in January; and some visits in late February will be reported as being in March.

Actual data format:

PATID	ENCOUNTERID	VITALID	MEASURE_DATE	MEASURE_TIME	DIASTOLIC
1	1	1	2014-02-15	12:01	75
1	2	20	2014-02-16	15:55	83
1	2	2	2014-02-16	16:00	92
1	2	34	2014-02-16	16:05	82
1	2	1	2014-02-16	16:10	96
1	2	308	2014-02-16	16:15	71
1	2	55	2014-02-16	16:20	71
1	2	13	2014-02-16	16:25	75
1	2	90	2014-02-16	16:30	73
1	2	92	2014-02-16	16:45	65
1	2	21	2014-02-16	17:00	65
1	2	13	2014-02-16	17:15	82
1	2	23	2014-02-16	17:30	79
1	67	400	2015-01-10	18:56	74
1	67	25	2015-01-10	21:00	70
1	67	29	2015-01-10	23:04	67
1	11	16	2015-01-20	13:40	77
1	11	44	2015-01-20	13:45	69
1	11	40	2015-01-20	13:46	69
1	11	56	2015-01-20	13:50	88
1	11	9	2015-01-20	14:00	72
1	11	1001	2015-01-20	14:15	68
1	11	100	2015-01-20	14:30	68
1	11	78	2015-01-20	14:45	74
1	11	32	2015-01-20	15:00	78

Converted data format:

PATID	ENCOUNTERID	VITALID	MEASURE_DATE	DAYS_from_FirstEncounter_Date	DIASTOLIC
1	1	1	2014-02	00	75
1	2	20	2014-02	01	77
1	67	400	2015-01	329	70
1	11	16	2015-01	339	72

Commented [BB1]: Alona to create new yyyy and mm variables

Example 2. Showing example on how multi-patient data should be reported

Commented [BB2]: Alona to fill in new column

Patient ID	ENCOUNTERID	LAB_RESULT_CM_ID	LAB_ORDER_DATE or SPECIMEN_DATE or RESULT_DATE*	DAYS_from_FirstEncounter_Date	LAB_NAME	LAB_LOINC	RESULT_NUM	RESULT_UNIT
1	1	1	2014-02	00	B1	XXX	XXX	XXX
1	2	16	2015-02	359	A1	XXX	XXX	XXX
1	2	2	2015-02	365	A2	XXX	XXX	XXX
1	3	3	2015-03	393	A1	XXX	XXX	XXX
2	25	4	2011-02	00	A1	XXX	XXX	XXX
2	255	5	2011-07	150	A1	XXX	XXX	XXX
2	30	6	2011-08	181	A1	XXX	XXX	XXX
3	27	7	2011-05	00	A1	XXX	XXX	XXX
3	27	8	2011-05	8	B1	XXX	XXX	XXX
3	17	9	2011-05	13	A1	XXX	XXX	XXX
3	37	10	2011-05	18	C1	XXX	XXX	XXX

*Will depend on site. Should be consistent within the site.

Table 3. Crosswalk for Patients, Encounters and Dates. Each patient will have multiple ENCOUNTERIDs/rows. We can determine the number of encounters for each patient by counting the number of distinct ENCOUNTERIDs for each patient. *Note: Date of first visit and selected other dates are also recorded in table 1.*

Description	Name	Source for Variables	Values	Format
Arbitrary person identifier used to link across tables.	PATID	PCORNET (ENCOUNTER table)		SAS Char(x)
Arbitrary encounter identifier. Used to link across tables, including the ENCOUNTER, DIAGNOSIS, and PROCEDURES tables.	ENCOUNTERID			SAS Char(x)
Encounter or admission date.	ADMIT_DATE		*See above for date blinding process*	depends on how dates are provided
	DAYS_from_FirstEncounter_Date			INT
Encounter type	ENC_TYPE		AV=Ambulatory Visit ED=Emergency Department EI=Emergency Department Admit to Inpatient Hospital Stay (permissible substitution) IP=Inpatient Hospital Stay IS=Non-Acute Institutional Stay OA=Other Ambulatory Visit NI=No information UN=Unknown OT=Other *code will combine EI,IS,OA,NI,UN,OT*	SAS Char(2)
Arbitrary facility code that identifies specific hospital or clinic.	FACILITYID		Y=Yes N=No R=Refuse to answer NI=No information UN=Unknown OT=Other *code will combine R, NI, UN, OT*	SAS Char(x)

Table 4. Prescription Medicines. Each patient may have multiple rows, one or more for each visit. Each encounter will have a separate encounter id. Some encounters may involve actual visits, others will just involve a phone prescription. If a patient receives multiple prescriptions in a single encounter, the patient will have a separate row for each prescription. These separate prescriptions will have the same ENCOUNTERID but different PRESCRIBINGID.

Description	Name	Source for Variables	Value	Format
Defined above	PATID	PCORNET (PRESCRIBING table)		SAS Char(x)
Defined above	ENCOUNTERID			SAS Char(x)
Arbitrary identifier for each unique prescription	PRESCRIBINGID			SAS Char(x)
RxNorm concept identifier (CUI) at highest available specificity.	RXNORM_CUI		If more than one option exists for mapping, the following ordered strategy may be adopted: 1)Semantic generic clinical drug 2)Semantic Branded clinical drug 3)Generic drug pack 4)Branded drug pack	Numeric(8)
Date prescription was ordered by provider.	RX_ORDER_DATE		*See above for date blinding process*	depends on how dates are provided
	DAYS_from_FirstEncounter_Date		*See Example 1 above*	INT
Provider code for provider who prescribed the medication (pseudo-identifier with crosswalk to real identifier)	RX_PROVIDERID			SAS Char(x)
days supply ordered, specified in the prescription.	RX_DAYS_SUPPLY			Numeric(8)
No. of refills ordered (not including original prescription).	RX_REFILLS		If no refills are ordered, value should be zero	Numeric(8)

Table 5. Vital Signs. Each patient may have multiple rows, one or more for each visit. Each visit will have a separate encounter id. If multiple vital signs are obtained in a single visit, the patient will have a separate row for each vital sign. These separate vital signs will have the same ENCOUNTERID but different VITALID.

Multiple measures in the same day. For vital signs with multiple measures on the same day (for example, heart rate or blood pressure, please compute the median for that day and report the median value, using the first VITALID for that day.

Note: We will not separately capture BMI, and instead compute this ourselves from height and weight.

Description	**Name**	**Source for Variables**	**Value**	**Format**
Defined above	PATID	PCORNET (VITAL table)		SAS Char(x)
Defined above	ENCOUNTERID			
Date of vitals measure.	MEASURE_DATE		*See above for date blinding process*	Depends on how dates are provided
	DAYS_from_FirstEncounter_Date		*See Example 1 above*	INT
Arbitrary identifier for each unique VITAL record.	VITALID			SAS Char(x)
Vital sign label e.g., WT for weight; SYSTOLIC for systolic blood pressure	Each of the following vitals should be reported in own column: HT(height in inch) WT (weight in pounds) SYSTOLIC (systolic blood pressure in mm Hg) DIASTOLIC (diastolic blood pressure in mm Hg) SMOKING(in categorical values)		*For the SMOKING variable:* Label as 1 (current smoker) if any current smoking code (01=Current every day smoker, 02=Current some day smoker, 05=Smoker, current status unknown, 07=Heavy tobacco smoker, 08=Light tobacco smoker) during this month. If no codes for this month, use most recent prior code. Label as 2 (former smoker) if 03=Former smoker recorded during the month, **and** no current smoker record during the month). If no codes for this month, use most recent prior code. [We can create a long-term quitter variable based on repeated "former smoker" observations]	

			Label as 3 (never smoker) if for **whole available period** patient has no smoking or former smoking codes 01,02,03, 05,07,08, and at least one 04 code (= Never smoker). Label as 4 (unknown) if for **whole available period** patient has no codes except for (06=Unknown if ever smoked, NI=No information, UN=Unknown, OT=Other)	

Table 6a. **Lab Results for labs included in PCORnet**. PCORnet allows for a very limited set of labs. Most of this limited set of labs are not relevant for this project. The only relevant labs available through PCORnet are AIC, LDL, and HGB. We will need to extract other labs from the underlying EHR (see Table 6b).

There will be separate rows for each lab value, for each encounter. Rows for the same patient and the same encounter will have the same ENCOUNTERID but different LAB_RESULT_CM_IDs.

We understand that different sites use different dates to record lab tests. Some have the date when the labs were ordered (in PCORnet this is LAB_ORDER_DATE), some have the date when the lab specimen was collected (SPECIMEN_DATE) and some have the date when lab results are available (which is not a PCORnet field). We only need one of these dates, but ask that each site tell us which date you are providing. In particular, if you have available only the date when results are available, which PCORnet field do you use to record this date?

Description	Name	Source	Value	Format
Defined above	PATID	PCORNET (LAB_RESULT_CM table)		SAS Char(x)
Defined above	ENCOUNTERID			
Date test was ordered.	LAB_ORDER_DATE			SAS Date (Numeric)
Arbitrary identifier for each unique LAB_RESULT_CM record.	LAB_RESULT_CM_ID			SAS Char(x)
Date specimen was collected.	SPECIMEN_DATE			SAS Date (Numeric)
	DAYS_from_FirstEncounter_Date		*See Example 1 above*	INT
Standardized/converted result for quantitative results. Should be null for qualitative results	RESULT_NUM			SAS Char(8)
Converted/standardized units for the result.	RESULT_UNIT		SAS Char(11)	
HbA1c	LAB_NAME		A1C=Hemoglobin A1c	
LDL cholesterol	LAB_NAME		LDL=Low-density lipoprotein	
CREATININE	LAB_NAME		Creatinine	
CK	LAB_NAME		Creatine kinase total	
CK_MB	LAB_NAME		Creatine kinase MB	

CK_MBI	LAB_NAME		Creatine kinase MB/creatine kinase total	
TROP_I	LAB_NAME		Troponin I cardiac	
TROP_T_QL	LAB_NAME		Troponin T cardiac (qualitative)	
TROPT_QN	LAB_NAME		Troponin T cardiac (qualitative)	
HGB	LAB_NAME		Hemoglobin	

Table 6b. Lab results not included in PCORnet. PCORnet allows for a very limited set of labs (see Table 6a). We therefore need to extract other labs from the underlying EHR (this Table 6b). There will be separate rows for each lab value, for each encounter. Rows for the same patient and the same encounter will have the same ENCOUNTERID but different LAB_RESULT_CM_IDs.

We understand that different sites use different dates to record lab tests. Some have the date when the labs were ordered (which PCORnet calls LAB_ORDER_DATE), some have the date when the lab specimen was collected (which PCORnet calls SPECIMEN_DATE) and some have the date when lab results are available (which is not a PCORnet field). CAPriCORN uses date fields that match the PCORnet definitions, but we understand that the GPC sites call this date START_DATE, but the meaning of this variable differs across sites. We ask that each site tell us which date you are providing.[1]

Cleaning steps: Lab values can be recorded with errors in raw data. For labs extracted from PCORnet (Table 6a), PCORnet has established filters to exclude at least some wrongly recorded data. For the additional labs in this table, we will provide filters, based on those used in the CAPriCORN network.

Description	Name	Source	Value	Format
Defined above	PATID	Local i2b2		SAS Char(x)
Defined above	ENCOUNTERID			
Date test was ordered (for CAPriCORN sites)	LAB_ORDER_DATE			SAS Date (Numeric)
Date specimen was collected (for CAPriCORN sites)	SPECIMEN_DATE			SAS Date (Numeric)
	DAYS_from_FirstEncounter_Date		*See Example 1 above*	INT
General lab date (for GPC sites)	START_DATE			
Standardized/converted result for quantitative results. This variable should be null for qualitative results	RESULT_NUM			SAS Char(8)

[1] Notes: We decided not to ask for OGTT (oral glucose tolerance test) due to the wide variety of tests, because this test is not very commonly used, and the test protocols vary as to both glucose dose and measurement period;

Converted/standardized units for the result.	RESULT_UNIT		SAS Char(11)	
Logical Observation Identifiers, Names, and Codes (LOINC)	LOINC			
Total cholesterol	LAB_NAME		*Numeric value is expected* *List of LOINC codes will be decided during in-person meeting*	
HDL cholesterol	LAB_NAME		*List of LOINC codes will be decided during in-person meeting*	
Triglycerides	LAB_NAME		*Categorical value (non-null result is available) is expected.* *List of LOINC codes will be decided during in-person meeting*	
BNP	LAB_NAME		*Numeric value is expected* *List of LOINC codes will be decided during in-person meeting*	
EMG	LAB_NAME		An electromyogram *Categorical value (non-null result is available) is expected.* *List of LOINC codes will be decided during in-person meeting*	
Random Glucose	LAB_NAME		*Numeric value is expected* *List of LOINC codes will be decided during in-person meeting*	
Fasting Glucose	LAB_NAME		*Numeric value is expected* *List of LOINC codes will be decided during in-person meeting*	
Urine microalbumin	LAB_NAME		*Categorical value (non-null result is available) is expected.* *List of LOINC codes will be decided during in-person meeting*	

Table 7. Non-urgent visits. Specialized table, for non-urgent visits (routine, check-up and well-being visits). Each patient will have multiple rows with multiple ENCOUNTERIDs and corresponding PROCEDURESID and DIAGNOSISID. The set of relevant visits is specified through encounter code (ICD-9 or -10) or procedure code (CPT code). See Definitions_Appendix_B, Part 1, Tables AppB-1 and AppB-2). PATID, ENCOUNTERID, PX_DATE and ADMIT_DATE should match in both tables during collection of data. Only ADMIT_DATE will be reported as date of visit. PX_DATE is listed in this table for educational reasons.

Description	Name	Source	Value	Format
Defined above	PATID	PCORNET (PROCEDURE table for CPT codes)		SAS Char(x)
Defined above	ENCOUNTERID			
	ENC_TYPE			
Defined above (Table 3)	ADMIT_DATE		*See Example 1 above*	
	ADMIT_DATE_orderNumber		=1 for first encounter on ADMIT_DATE; =2 for second encounter, etc.	
Arbitrary identifier for each unique procedure	PROCEDURESID			SAS Char(x)
Procedure code.	PX		*See Definitions_Appendix_B1 for details*	
Procedure code type.	PX_TYPE		09=ICD-9-CM 10=ICD-10-PCS 11=ICD-11-PCS CH=CPT or HCPCS LC=LOINC ND=NDC (National Drug Code) RE=Revenue NI=No information UN=Unknown OT=Other *Code will pick up ICD-9-CM and ICD-10-CM codes only, and will combine NI, UN, OT*	
Date the procedure was performed.	PX_DATE		SAS Date	
	DAYS_from_FirstEncounter_Date		*See Example 1 above*	INT
Arbitrary identifier for each unique diagnosis.	DIAGNOSISID	PCORNET (DIAGNOSIS table for ICD codes)		SAS Char(x)

Diagnosis code.	DX		*See Definitions_Appendix_B, Part B1 for details*	
Diagnosis code type	DX_TYPE		09=ICD-9-CM 10=ICD-10-CM 11=ICD-11-CM SM=SNOMED CT NI=No information UN=Unknown OT=Other *Code will pick up ICD-9-CM and ICD-10-CM codes only and will combine NI, UN, OT*	

Table 8. Immunization. Each patient will have multiple rows with multiple ENCOUNTERIDs and corresponding PROCEDURESID.

Description	Name	Source	Value	Format
Defined above.	PATID	PCORNET (PROCEDURE table for CPT codes)		SAS Char(x)
Defined above	ENCOUNTERID			
Defined above	PROCEDURESID			SAS Char(x)
Procedure code.	PX		See Definitions_Appendix_B for exact codes	
Procedure code type.	PX_TYPE		09=ICD-9-CM 10=ICD-10-PCS 11=ICD-11-PCS CH=CPT or HCPCS LC=LOINC ND=NDC RE=Revenue NI=No information UN=Unknown OT=Other	
Date the procedure was performed.	PX_DATE		SAS Date	
	DAYS_from_FirstEncounter_Date		*See Example 1 above*	INT
Defined above.	DIAGNOSISID			SAS Char(x)

Commented [BB3]: Table is not completed. We need to decide whether immunization records are complete enough to be worth capturing. Bernie thinks no, too many patients get vaccines from pharmacies, flu vaccines from employers, etc.

Defined above. .	ADMIT_DATE	PCORNET (DIAGNOSIS table for ICD codes)		SAS Date (Numeric)
Diagnosis code.	DX		*See Definitions_Appendix_B, Part B2 for details*	
Diagnosis code type	DX_TYPE		09=ICD-9-CM 10=ICD-10-CM 11=ICD-11-CM SM=SNOMED CT NI=No information UN=Unknown OT=Other *Code will pick up ICD-9-CM and ICD-10-CM codes only*	

Table 9. Health outcomes.

Elements in this part will be provided after discussion with all sites.

Currently see three examples of outcomes in a separate file

Table 10. Diagnoses

Description	Name	Source	Value	Format
Arbitrary person-level identifier used to link across tables.	PATID	PCORNET (DIAGNOSIS table for ICD codes)		SAS Char(x)
Arbitrary encounter-level identifier. The ENCOUNTERID should be present if the prescribing activity is directly associated with an encounter	ENCOUNTERID			

Arbitrary identifier for each unique record	DIAGNOSISID			SAS Char(18)
Principal discharge diagnosis flag. Relevant only on IP and IS encounters. For ED, AV, and OA encounter types, mark as X=Unable to Classify. (Billing systems do not require a primary diagnosis for ambulatory visits (eg, professional services).)	PDX		P=Principal S=Secondary X=Unable to Classify NI=No information UN=Unknown OT=Other	
Diagnosis code.	DX			
Encounter type.	ENC_TYPE		AV=Ambulatory Visit ED=Emergency Department EI=Emergency Department Admit to Inpatient Hospital Stay (permissible substitution) IP=Inpatient Hospital Stay IS=Non-Acute Institutional Stay OA=Other Ambulatory Visit NI=No information UN=Unknown OT=Other	
Admission date	ADMIT_DATE		*Dates are specific to each site rules.*	SAS Date (Numeric)
	DAYS_from_FirstEncounter_Date		*See Example 1 above*	INT