Genentech
A Member of the Roche Group

Cervical Cancer Screening

Help prevent cervical cancer by identifying at-risk populations

$100,000 · 40 teams · a year ago

Training Data

0 files

Preview

Data Introduction

In this competition, you are to predict whether a patient receives regular cervical cancer screening (pap smear), given the medical records of over 3 million women in the United States.

A patient is defined as a screener if the patient has had a pap smear in the last 5 years. This patient must be between the age of 25 and 65, and has not been diagnosed of cervical cancer or had hysterectomies.

Please note that because of the data source, this data provided in the competition may have some sampling bias, therefore may not reflect the prevalence of cervical cancer screening of the general population.

This competition has 11 relational data tables. Relationships between the datasets are captured in the following schema:

File descriptions

patient_train.csv, patient_test.csv

Patient Table			
Field Name	Col Nbr	Data Type	Description
PATIENT_ID	1	NUMBER(22)	Patient ID
PATIENT_AGE_GROUP	2	VARCHAR(5)	Patient Age in 3 year increments (18-20, 21-23, 24-26, etc.)
PATIENT_GENDER	3	VARCHAR(1)	Gender
PATIENT_STATE	4	VARCHAR(2)	State
ETHNICITY	5	VARCHAR(20)	Patient Ethnicity, where available: Caucasian, African American, Hispanic, All Other
HOUSEHOLD_INCOME	6	VARCHAR(14)	Patient Range of Household Income, where available: Less than or equal to $49,999, $50-99,999 , $100k + and Unknown)
EDUCATION_LEVEL	7	VARCHAR(20)	Patient highest education attained, where available: High School or Less, Some College, "Associate degree and above" and Unknown
IS_SCREENER	8	VARCHAR(1)	Whether or not a user receives regular pap smear

patient_activity_head.csv.gz (md5=c8781bdef8e818dff645b84f22974d50)

Patient Activity Table			
Field Name	Col Nbr	Data Type	Description
PATIENT_ID	1	NUMBER(22)	Patient ID
ACTIVITY_TYPE	2	VARCHAR(1)	Values: A = All Claims, R = Retail Claims Only For example, A includes all claims in the database, whereas R includes only claims from a retail pharmacy
ACTIVITY_YEAR	3	VARCHAR(4)	Year of database activity. A record will exist in this table if there is activity in the Rx database for the given patient and year combination, based on the given activity type value.
ACTIVITY_MONTH	4	VARCHAR(1)	Month of database activity. A record will exist in this table if there is activity in the Rx database for the given patient and month combination, based on the given activity type value. Note this field is associated to the YR field above.

prescription_head.csv.gz (md5=a0b4d8830dbaebc2437970ceeca07962)

Rx Table			
Field Name	Col Nbr	Data Type	Description
CLAIM_ID	1	VARCHAR(22)	Claim ID
PATIENT_ID	2	NUMBER(22)	Patient ID
DRUG_ID	3	VARCHAR(13)	Drug ID
PRACTITIONER_ID	4	NUMBER(22)	Practitioner ID (DS_WRITER_GID)
REFILL_CODE	5	VARCHAR(2)	0 = New Rx, else value = Refill Number
DAYS_SUPPLY	6	NUMBER(3)	Days Supply
RX_FILL_DATE	7	DATE	RX Fill Date
RX_NUMBER	8	VARCHAR(32)	Encrypted Rx Script Number
PAYMENT_TYPE	9	VARCHAR(60)	Payment Type for the Rx claim: Cash Commercial Medicare Medicaid Assistance etc. Unknown NULL

drugs.csv

Drug Table			
Field Name	Col Nbr	Data Type	Description
DRUG_ID	1	VARCHAR(13)	Drug ID (Primary Key)
NDC11	2	VARCHAR(11)	NDC11
DRUG_NAME	3	VARCHAR(60)	Drug Name
BGI	4	VARCHAR(1)	Brand Generic Indicator
BB_USC_CODE	5	VARCHAR(5)	Blue Book USC Code
BB_USC_NAME	6	VARCHAR(60)	Blue Book USC Name
DRUG_GENERIC_NAME	7	VARCHAR(60)	Drug Generic Name
DRUG_STRENGTH	8	VARCHAR(10)	Drug Strength
DRUG_FORM	9	VARCHAR(40)	Drug Form
PACKAGE_SIZE	10	NUMERIC(11,3)	Quantity Per Package
PACKAGE_DESCRIPTION	11	VARCHAR(20)	Description of Package
MANUFACTURER	12	VARCHAR(15)	Manufacturer Name
NDC_START_DATE	13	DATE	Date NDC Entered Market

physicians.csv

Physician Table (Optional)			
Field Name	Col Nbr	Data Type	Description
PHYSICIAN_ID	1	NUMBER(22)	PRC_REL_GID
PRACTITIONER_ID	2	NUMBER(22)	DS Writer GID (PRIMARY Key)
STATE	3	VARCHAR(2)	State
SPECIALTY_CODE	4	VARCHAR(3)	Specialty Code
SPECIALTY_DESCRIPTION	5	VARCHAR(75)	Specialty Description
CBSA	6	VARCHAR(30)	CBSA in which the physician is located

diagnosis_head.csv.gz (md5=23226f0041c4a3516604f2ede5cedfc7)

Diagnosis Table (Optional)			
Field Name	Col Nbr	Data Type	Description
PATIENT_ID	1	NUMBER(22)	Patient ID
CLAIM_ID	2	VARCHAR(22)	Claim ID
CLAIM_TYPE	3	VARCHAR(4)	Claim Type: HCFA/UB92 or MX/HX
DIAGNOSIS_DATE	4	VARCHAR(6)	Year and Month from the Service_From_Date
DIAGNOSIS_CODE	5	NUMBER	Diagnosis Code
PRIMARY_PRACTITIONER_ID	6	NUMBER(22)	Practitioner ID (DS_WRITER_GID)
PRIMARY_PHYSICIAN_ROLE	7	VARCHAR(4)	Physician Role Code: ORD MX Ordering Practitioner PRV MX Providing Practitioner RFR MX Referring Practitioner RND MX Rendering Practitioner ATG HX Attending Physician OPR HX Operating Physician UNK Role code Unknown (HX or MX Physician)

diagnosis_code.csv

Diagnosis Code Table (Optional)			
Field Name	Col Nbr	Data Type	Description
DIAGNOSIS_CODE	1	VARCHAR(30)	Diagnosis Code
DIAGNOSIS_DESCRIPTION	2	VARCHAR(255)	Diagnosis Description

procedure_head.csv.gz (md5=be63438561638cdb5b313b9904012a2d)

Procedure Table (Optional)			
Field Name	Col Nbr	Data Type	Description
PATIENT_ID	1	NUMBER	Patient ID
CLAIM_ID	2	NUMBER	Claim ID
CLAIM_LINE_ITEM	3	NUMBER10,0	AKA: SRVC_LINE_SEQ. CLAIM_GID and CLAIM_LINE_ITEM form a composite key. There are one to many line items for each CLAIM_ID
CLAIM_TYPE	4	VARCHAR22	Claim Type: HCFA/UB92 or MX/HX
PROCEDURE_CODE	5	VARCHAR248	Procedure Code
PROCEDURE_DATE	6	VARCHAR26	Year and Month from the Service_From_Date
PLACE_OF_SERVICE	7	VARCHAR260	Place of Service
PLAN_TYPE	8	VARCHAR216	Plan Type
PRIMARY_PRACTITIONER_ID	9	NUMBER18,0	Practitioner ID (DS_WRITER_GID)
UNITS_ADMINISTERED	10	NUMBER15,0	Units Administered
CHARGE_AMOUNT	11	NUMBER	Amount Charged
PRIMARY_PHYSICIAN_ROLE	12	VARCHAR210	Physician Role Code: ORD MX Ordering Practitioner PRV MX Providing Practitioner RFR MX Referring Practitioner RND MX Rendering Practitioner ATG HX Attending Physician OPR HX Operating Physician UNK Role code Unknown (HX or MX Physician)
ATTENDING_PRACTITIONER_ID	13	NUMBER18,0	Practitioner ID (DS_WRITER_GID)
REFERRING_PRACTITIONER_ID	14	NUMBER18,0	Practitioner ID (DS_WRITER_GID)
RENDERING_PRACTITIONER_ID	15	NUMBER18,0	Practitioner ID (DS_WRITER_GID)
ORDERING_PRACTITIONER_ID	16	NUMBER18,0	Practitioner ID (DS_WRITER_GID)
OPERATING_PRACTITIONER_ID	17	NUMBER18,0	Practitioner ID (DS_WRITER_GID)

procedure_code.csv

Procedure Code Table (Optional)			
Field Name	Col Nbr	Data Type	Description
PROCEDURE_CODE	1	VARCHAR(48)	Procedure Code
PROCEDURE_DESCRIPTION	2	VARCHAR(100)	Procedure Description

surgical_head.csv.gz (md5=162290192f0bb174ef4206d95f38531d)

Surgical Code Table (Optional)			
Field Name	Col Nbr	Data Type	Description
SURGICAL_CODE	1	VARCHAR(48)	Surgical Procedure Code
SURGICAL_DESCRIPTION	2	VARCHAR(100)	Surgical Procedure Description

surgical_code.csv

Surgical Table (Optional)			
Field Name	Col Nbr	Data Type	Description
PATIENT_ID	1	NUMBER(22)	Patient ID
CLAIM_ID	2	VARCHAR(22)	Claim ID
PROCEDURE_TYPE_CODE	3	VARCHAR(4)	HXPR or 0001 = Principal Procedure Code HX01 or 0002 = Other Procedure Code 1 HX02 or 0003 = Other procedure Code 2 HX03 or 0004 = Other Procedure Code 3 HX04 or 0005 = Other Procedure Code 4 HX05 or 0006 = Other Procedure Code 5
CLAIM_TYPE	4	VARCHAR(4)	Claim Type: UB92 or HX
SURGICAL_CODE	5	VARCHAR(48)	Surgical_code
SURGICAL_PROCEDURE_DATE	6	VARCHAR(6)	Year and Month from the Service_From_Date
PLACE_OF_SERVICE	7	VARCHAR(60)	Place of Service
PLAN_TYPE	8	VARCHAR(16)	Plan Type
PRACTITIONER_ID	9	NUMBER(22)	Practitioner ID (DS_WRITER_GID)
PRIMARY_PHYSICIAN_ROLE	12	VARCHAR(10)	Physician Role Code: ORD MX Ordering Practitioner PRV MX Providing Practitioner RFR MX Referring Practitioner RND MX Rendering Practitioner ATG HX Attending Physician OPR HX Operating Physician UNK Role code Unknown (HX or MX Physician)

sample_submission.csv

a sample submission file in the correct format.

Claim_Id's

The claim ID is a way of identifying the information received on a claim form (CMS 1450 and CMS1500) submitted by a physician. These claim forms can contain multiple services, so the claim ID ties them all back together. A claim may have multiple diagnoses associated with it. A unique record in the diagnosis table would be a patient_id/Claim_id/Srvc_date/diagnosis code. On the procedure and surgical tables, claims may have multiple lines. A unique record is a patient_id/Claim_id/srvc_line/procedure code.

Within the data, that is a key field used to appropriately link the diagnosis to the procedure, but the contestants SHOULD NOT try to link the Rx (prescription) claim ID to the diagnosis or procedure claim ID. The Rx claim ID is an entirely different claim submission from the pharmacy, so it has a separate claim ID.

The above data dictionary can be downloaded here.