Annotation guidelines:
Risk factors for Heart Disease in Diabetic Patients

Amber Stubbs, Ozlem Uzuner, Vishesh Kumar, Stanley Shaw
April 1, 2014

1. Overview

This annotation task will create a set of medical documents that track the progression of heart disease in diabetic patients. Multiple records will be annotated for each patient, which will allow a general timeline to be created from the set. This project uses tags and attributes used to indicate the presence and progression of disease (diabetes, heart disease), associated risk factors (hypertension, hyperlipidemia, smoking status, obesity status, and family history), and the time they were present in the patient's medical history.

2. Tags

Each of the diseases and risk factors associated with this task has its own set of indicators that we will use to identify whether or not the disease or risk factor is present for that patient, and when it is present. This section discusses each tag that will be used in this annotation.

Every tag (except for those related to smoking status and family history) has an "indicator" attribute and a "time" attribute. The indicator attributes are described below; the time attribute is described in Section 3.

For each indicator, only text that matches the description of the indicator should be annotated as evidence that the risk factor is present. If an indicator is negated ("*no history of diabetes*"), that text should not be annotated.

Please note that medications now have their own tag—see section 2.8 for details.

2.1 Diabetes

indicator	description
mention	a diagnosis of Type 1 or Type 2 diabetes, or a mention of a preexisting diagnosis
High A1c	An A1c test value of over 6.5
High glucose	2 fasting blood glucose measurements of over 126

Table 1: indicators and descriptions for the Diabetes tag

Phrases that should be annotated for Diabetes:
- **Mention**: *patient has h/o DMII, diabetic ketoacidosis*
- **Med-Sulfonylureas**: *Medications on admission: Glyburide, ...*
- **A1c**: *7/18: A1c: 7.3*
- **Glucose**: Example: *(8:00am)glu: 145 [. . .] (8:00pm)glu: 139*

Phrases that should **not** be annotated for Diabetes:
- *Sister with h/o DMII*
- *pre-diabetic*
- *Will consider starting Glyburide at next visit*
- *A1c: 6.2*
- *Glucose after eating: 200*

2.2 CAD

indicator	description
mention	a diagnosis of CAD, or a mention of a history of CAD
event	– MI, STEMI, NSTEMI – revascularization procedures (bypass surgery, CABG, percutaneous) – cardiac arrest – ischemic cardiomyopathy
test result	– exercise or pharmacologic stress test showing ischemia – abnormal cardiac catheterization showing coronary stenoses (narrowing)
symptom	chest pain consistent with angina

Table 2: indicators and descriptions for the Diabetes tag

Phrases that should be annotated for CAD:
- **mention**: *PMH: significant for <u>CAD</u>*
- **med-Thienopyridines** *Discharge medications: continue on <u>Prasugrel</u>;*
- **med-Nitrates** *increase <u>Nitrostat</u>*
- **event**: *s/p <u>STEMI in 2004</u>; <u>CABG in 1999</u>*
- **test result**: *<u>dolbutamine stress test revealing ischemia</u>; <u>cath. of LAD revealed 50% lesion</u>*
- **symptom**: *patient being treated for <u>stable angina</u>*

Phrases that should **not** be annotated for CAD:
- *Ruled out for MI*
- *Atrial fib.*
- *Reason for admission: chest pain*
- *Performed stress test; neg. for ischemia*

2.3 Hyperlipidemia/Hypercholesterolemia

indicator	description
mention	a diagnosis of Hyperlipidemia or Hypercholesterolemia or a mention of the patient already having that condition
high cholesterol	total cholesterol of over 240

high LDL	LDL measurement of over 100 mg/dL

Table 3: Indicators and descriptions of Hyperlipidemia/Hypercholesterolemia

Phrases that should be annotated for Hyperlipidemia:
- **mention**: *I agree with his risk factor modification including control of his diabetes, hypertension, and <u>hypercholesterolemia</u>.*
- **me-Statin**: *Current medications: <u>Lipitor</u> [...]*
- **high cholesterol**: *The result of your latest chol. test is <u>250</u>*
- **high LDL**: *latest <u>LDL: 135</u>*

Phrases that should **not** be annotated for Hyperlipidemia:
- *Allergies: Lipitor*
- *At risk for high chol.*
- *Will start Lipitor if test results not improved on next visit*
- *LDL 90 mg/dL*
- *Chol. good @ 190*

2.4 Hypertension

indicator	description
mention	a diagnosis of Hypertension or a mention of a pre-existing condition
high blood pressure	BP measurement of over 140/90 mm/hg (if either value is high, the patient has hypertension)

Table 4: indicators and descriptions for the Hypertension tag

Phrases that should be annotated for Hypertension:
- **mention**: *PMH: <u>HTN</u>*
- **med-Beta blockers**: *His current medications include <u>Toprol</u>, and Glucophage*
- **high blood pressure:** *at admit, <u>bp 140/100</u>*

Phrases that should **not** be annotated for Hypertension:
- *At risk for HTN*
- *bp 120/80*
- *Allergies: Toprol*

2.5 Obesity

indicator	description
mention	a description of the patient as being obese
BMI	BMI over 30
waist circumference	Waist circumference measurement of:

	• men: 40 inches or more
	• women: 35 inches or more

Table 5: indicators and descriptions for the Obesity tag

Phrases that should be annotated for Obesity:
- **mention**: *57y/o <u>obese</u> white male*
- **BMI**: *spoke to patient about lowering <u>BMI (31.4 last August</u>)*
- **waist circumference**: *<u>42in waist</u> meas.*

Phrases that should **not** be annotated for Obesity:
- weight: 340
  - weight alone is not an indicator of obesity
- lost 150lbs over past year
  - weight/weight loss is not an indicator of obesity

2.6 <mark>Family History of premature CAD</mark>

The FAMILY_HIST tag has only an *indicator* attribute that has the values "present" and "not present". This tag should only be marked as present if the patient is has a **first-degree relative** (parents, siblings, or children) who was diagnosed **prematurely** (younger than 55 for male relatives, younger than 65 for female relatives) with **CAD**. Family histories of diabetes or any of the other risk factors should **not** be annotated.

Phrases that should be annotated for Family History of CAD:
- *Family/social History: <u>Father diagnosed w/ CAD at 49</u>*
- *<u>Fam.hist. significant for premature CAD</u>*

Phrases that should **not** be annotated for Family History of CAD:
- *Father w/ CAD, died at 82yrs*
  - unknown when diagnosis occurred
- *No known relatives with CAD*
  - *does not confirm CAD*
- *Both grandfathers prem. CAD*
  - not first-degree relatives

2.7 Smoker

The SMOKER tag does not have *indicator* or *time* attributes. Instead, it has a *status* attribute that indicates whether the person is **currently** a smoker (CURRENT), used to smoke but **quit over a year ago** (PAST), **smoked at some point** but it is unclear if they are still smoking or have quit (EVER), has **never** smoked (NEVER), or if their smoking is **not mentioned** (UNKNOWN).

For the purposes of this research, a person is considered to be a current smoker if they have smoked at all **within the past year.** A patient is given "past" status if they quit smoking more than a year ago,

"ever" status if it is unclear whether they have quit and "never" if the text states that they have never been a smoker. If no information is provided about their smoking status, they are to be labeled as "unknown".

Phrases that should be annotated for Smoker:
- **CURRENT**: *Patient says trying to quit <u>1pack/day habit</u>; <u>quit 6mos ago</u>*
- **PAST**: *Social history: patient used to smoke, <u>quit 10 yrs ago</u>*
- **EVER**: *<u>quit smoking</u>; <u>remote history of tobacco dependence</u>*
  - <u>Unless we know how "remote" the history is, it has to be "ever"</u>
- **NEVER**: *<u>denies tobacco</u>, alcohol, illicits*
- **UNKNOWN**: (no mention of smoking/tobacco use)

Phrases that should not be annotated for Smoker:
- *recreational marijuana*
- *chewing tobacco*

2.8 <mark>Medications</mark>

Due to overlaps between medications prescribed for different diseases/risk factors, MEDICATION is now its own tag.   The list below contains all the different categories of medications that we are interested in for this task.

Please remember that you should annotated every **category** of medication that the patient is on—do not only annotate one CAD medication if the person is taking, for example, both aspirin and a beta blocker —annotate both.

While you do not need to indicate the risk factor/disease related to each medication, the correlations are provided here for reference:

Diabetes: Metformin, insulin, sulfonylureas, thiazolidinediones, GLP-1 agonists, Meglitinides, DPP-4 inhibitors, Amylin, anti-diabetes medications, combinations including these
CAD: Aspirin, Thienopyridines, beta blockers, ACE inhibitors, nitrates, calcium-channel blockers, combinations including these
Hyperlipidemia: statins, fibrates, niacins, ezetimibes, combinations including these
Hypertension: beta-blockers, ACE inhibitors ARBs, Thiazide diuretics, calcium-channel blockers, combinations including these
Obesity: orlistat (xenical) or Lorqess (Lorcaserin)

The medication tag has two "type" attributes for identifying what category of drug(s) a medication falls into. For combination drugs, indicate both categories. For drugs that are only of one type, the "type2" attribute can be left blank.

Phrases that should be annotated for medications:
- **ACE inhibitor:** *Current medications: <u>Lisinopril</u>*
- **Thiazolidinediones & metformin:** *increase <u>Avandamet</u>*
- *<u>etc</u>*

Phrases that should not be annotated for medications:
- *Consider starting Humalog if next A1c measurement too high*
- *Allergic to Simvastatin*

3. The *Time* attribute

Every tag (except for SMOKER and FAMILY_HIST) has a time attribute that is used to show when the indicator for each medical problem is known to have existed. These reflect when the indicator occurred/was active in relation to the date the medical record was written. This document creation time (DCT) most often refers to an entire day, usually the admission date for a hospital stay, the day of an outpatient visit, or the date a report is written describing the events of a previous hospital stay or doctor's visit. Whichever of these is used, all indicators should be annotated in relation to the DCT. The possible values for the time attribute are:
- continuing
- before DCT
- during DCT
- after DCT
- not mentioned

The rest of this section discusses when to use each possible attribute value.

3.1 The DCT

All the files in this dataset start with the line "Record date: " followed by the calendar date that the record was written. This date is the Document Creation Time (DCT), and it is the date against with all the risk factors and indicators in the document are evaluated. Because the DCT is always a calendar date, you can think about it as a "container" for the events of the day—we aren't concerned with **when** in the container the events occurred, just whether they occurred inside it or overlap with it either before, after, or both.

In most records, the DCT is the date of the patient's visit to the doctor, or possibly the date of one day of a multi-stay visit. In either of those cases, simply evaluate the risk factors in the record in relation to the date of that record. In other words, if the record describes one day in a multi-day visit, don't try to extrapolate the events to the end of the patient's stay.

In some cases, the records are summary letters sent from a specialist to a PCP. In those cases, use the date of the letter as the DCT, rather than the date of the visit being described.

3.2 "Continuing" times

The "continuing" value for the time attribute is used to show that whatever the annotated indicator is, it was present before the DCT, during day of the DCT, and continued after the day of the DCT. In general, this attribute value will only be used for indicators that represent a **state or condition that is unlikely to change**, such as an existing diagnosis (and corresponding **mention**) of diabetes, CAD, hypertension, etc. This attribute may also be used if a medication is mentioned in the context of "**continue/increase** XXX med. following discharge".

Correct examples of "continuing" times:
- **CAD, mention**: _h/o CAD_
- **Diabetes, mention**: _diabetes diagnosed 2004_
- **Hyptertension, mention**: _PMH significant for HTN_
- **Hyperlipidemia, mention**: _current problems: Hyperlipidemia, GERD, Arthritis._
- **Obesity, mention**: _Obese white female, 45yo_
  - (note that this doesn't mean the person is obese their whole life, just in the time surrounding that DCT**)**
- **Medication**: _Discharge instructions: continue on Glucophage_
- **CAD, symptom**: _patient being treated for ongoing stable angina._
- **Medication:** _continue full-dose aspirin and plavix_

Incorrect "continuing" times:
- **Diabetes, mention**_: This visit confirms diabetes_
- **Obese, mention**_: weight still around 260_
- **CAD/HTN, mention:** _at risk for CAD/HTN_
- any **measurement** such as blood pressure, cholesterol, LDL, blood glucose, A1c, etc
- any **CAD-related event or test**

Note that "continuing" can be used for medications, events, etc on the Incorrect list only if "continuing" is being used to indicate that those things are not mentioned in the text (see Section 4.1 for more information).

3.2.1 "Continuing" medications

Previous versions of this guideline suggested that medications should not, in general, be considered continuing unless specifically stated. However, this assessment was incorrect. A more correct interpretation (provided by our MD consultants) is that **if a medication is mentioned in a list, whether it be "medications on admission" or "medications", or just in a list with no title, the assumption has to be that the patient was on them before and during the DCT, and unless a medication is mentioned specifically as being discontinued, will be on them afterwards as well.**

This assumption applies to write-ups of inpatient and outpatient visits, as well as to summary letters written from a specialist to a PCP. In the case of the summary letters, the DCT is still **the date that the document was written**, _not_ the date of the visit (which might not even be stated in the document itself).

3.3 "Before DCT" times

This attribute value is used to indicate that the indicator can only be stated to be present prior to the date of the record. This will most often apply to:
- lab values from previous tests
- events or symptoms that occurred prior to admission

Correct "before admit" annotations:

- **Hyperlipidemia, high LDL**: *lab values from previous visit: LDL: 135, [...]*
- **CAD, event**: *s/p CABG 2004*

Incorrect "before admit" annotations:
- **CAD, test result:** *An exercise Myoview performed in 01/1999 revealed no evidence of inducible ischemia*
  - If the test did reveal ischemia, then this would be a valid "before admit" annotation

3.4 "During DCT" times

The "during DCT" value for the *time* attribute indicates that a risk factor indicator occurred the day of the date on the record. This will generally apply to:
- tests or surgeries performed during the day the document was written
- medications prescribed during that day
- symptoms occurring that day (that are not indicated as being ongoing problems)

Note that if a patient is diagnosed with a condition such as CAD, diabetes, etc. for the first time during a visit, those mentions should be annotated as both "during DCT" and "after DCT" (see next section), as those conditions will be ongoing from that time.

Correct "during DCT" annotations:
- **Hypercholestorolimia, high cholestorol**: *today's lab values: Chol. 247 ...*
- **Hypertension, medication**: *started on Lisinopril*
- **CAD, symptom**: *patient developed chest pain during stay*
- **Diabetes, mention**: *patient confirmed as diabetic, will be trained in managing blood sugar*

Incorrect "during DCT" annotations:
- **Events** or **test results** that are from a previous visit
- Events, tests, medications, etc, that are being considered or scheduled, but whose results are not in the record

3.5 "After DCT" times

The "after DCT" value for the *time* attribute indicates that the risk factor indicator applies to the days after the date of the record. This will mostly be used for:
- newly prescribed medications (in conjunction with a "during DCT" annotation)
- ongoing diagnoses that were discovered during the day

Correct "after discharge" annotations:
- **Hyptertension, medication**: *Discharge medications: Diovan [...]*
- **Diabetes, mention**: *patient confirmed as diabetic, will be trained in managing blood sugar*
- **CAD, medication**: *patient continuing aspirin regimen begun this visit...*

Incorrect "after discharge" annotations:

- **Any medication**: that is only being suggested, *recommend talking to GP about starting (medication); will consider starting (medication) next visit*
- **Any mention or indicator** that is mentioned as a possibility, that is being recommended, or that is suggested, but that is not definitely performed

4. Annotation Procedure

The procedure for annotating explicit and implicit events is that every risk factor/indicator/time combination mentioned in the document should be addressed in the annotations.

4.2 Completed annotations

When an annotation is completed, all of the risk factors with multiple indicators (Diabetes, Hypertension, Hyperlipidemia, Obese) should, at minimum, have tag(s) for each indicator that is present in the text.

Note that it's fine to annotate multiple mentions of the same indicator if you want to, as long as all the indicators mentioned in the text have *at least one* tag.

5. Annotation software

Annotation will be done using MAE (Multi-purpose Annotation Environment). The newest test version is available at your account on the UAlbany server, packaged with the test annotation files and these instructions. On most systems you can run the .jar file by double-clicking, loading the DTD, then loading the file you want to annotate. Best practice is to run the software from the command line (see the included user guide).

6. A note on HIPAA/PHI

While the files we are sharing with you have been de-identified and changed to protect the identities of the patients and doctors, we ask that you not share these files or send them to anyone. If you know someone who wants access to the files, please direct them to Amber Stubbs (astubbs@albany.edu). If you find information in the files that seems to be original PHI, please contact Amber immediately.