Appendix

A. Supplementary material

Appendix 1 Clause summaries with examples from the eMERGE phenotype algorithms

Clause Summary	Example clause**	Unique Clause Count
Calculate event measurement <with reference range / threshold> <with temporal restriction>	Calculate median height for subject whose <age>=18 years> old after removing <height measures <36 inches or >90 inches> [121_ExtremeObesity]	14
Count of event <with relevant event linkage> <with reference range / threshold] [with temporal restriction>	<Anyone ≥2 years of age> with at least one positive <inpatient or outpatient> C. diff antigen or toxin test [70_Cdiff] >= 2 visits with [a primary care provider] over a <minimum of a 3 year continuous period of enrollment> [149_CaMRSA]	301
Demographics extraction	Age>=40 [216_BPH]	24
Event definition	Patient-specific index date for controls is the date of earliest qualifying hospital admission if qualified by the hospital admission criteria [70_Cdiff]	17
Relevant unstructured document identification or NLP tool configuration	Choosing the Right NLP System for Your Reports from the two NLP systems provided. [188_CAAD]	17
Temporal associations from multiple events <with relevant event linkage> <within temporal window>	Having at least one eye exam within last two years if living or the last two years prior to date of death where <provider specialty is Ophthalmology or Optometry>. [172_AMD] Has all SBP<135 and DBP<90 one month AFTER BP meds [68_ResHTN]	68
Term search or event extraction from document/document section/parsed document	At least one non-negated mention from Table A, Column 1 (Disorder related mentions) and Table A, Column 2 (Anatomical site related mentions) either in the SAME or	48

	adjacent sentences in a 'section of interest [109_VTE]	
Other (i.e., algorithm description, categorize phenotype, cohort identification with non-EHR data)	Likely super obese (Case Type #2): Not Case Type 1 and >=75% of BMI measures >=50 kg/m^2 [121_ExtremeObesity]	25
Total		514*

*Some clauses may have multiple clause summaries; **Example clauses are from eMERGE phenotypes from PheKB; [] refers to a phenotype; __refers to the main part of clause summary; <> refers to the quantifier in clause summary

Appendix 2 2019 Portability-Customization Tasks Review/Comment

This webpage will collect your local experiences and opinions on identified customization tasks at clause level. The customization task is defined as what work needs to be done locally for the implementation site before a computable phenotype (e.g., a SQL query, SAS/R implementation, KNIME application, Atlas query) can be run against your local data repository.

In each of the questions below, we will present a clause extracted from an eMERGE phenotype's pseudocode, followed by a list of customization tasks. Each clause is presented with a clause ID in square brackets at the beginning.

Please only check the customization tasks you <u>disagree</u> with (meaning, you don't believe it is needed) and comment on why you disagree in the "Other comments" text input box. Please also describe other customization tasks in your local implementation in the "Other comments". **If you have no or very few clause level comments, you can also comment directly in the manuscript draft instead of submitting this survey.**

Please use one machine for one participant. This way, your responses will be saved after clicking "Next" button and you will be able to return to it later until you submit the comments.

***1. Your name and affiliation**

2. [587_CKD_2] IV. 3. Calculate eGFR from SCr by CKD-EPI formula.
[CKD-EPI formula is provided and SCr (serum creatinine) has been defined and extracted from previous rules]

- ☐ Implement the CKD-EPI formula in SQL
- ☐ Other comments _____

3. [70_Cdiff_1] 1.1 Anyone ≥2 years of age with at least one positive inpatient or outpatient C. diff antigen or toxin test (includes positive tests for strain A, strain B, or strains A and B).

- ☐ Specify an operational definition of inpatient or outpatient

- ☐ Find relevant data and identify "C. Diff antigen or toxin test" labs
- ☐ Explore data to find what are positive tests
- ☐ Other comments _____

4. [70_Cdiff_19] 3) has been exposed to a class 2 (moderate risk for C. diff) or class 3 (high risk for C. diff) antibiotic, and

[Class 2/3 antibiotic drug names are provided in a table]

- ☐ Map the provided medication names to RxCUI codes that are used in the EHR
- ☐ Other comments _____

5. [70_Cdiff_24] Patient-specific index date for controls is the date of earliest qualifying hospital admission if qualified by the hospital admission criteria

- ☐ Specify an operational definition of "hospital admission" for implementation
- ☐ Other comments _____

6. [729_PAD_3] Selection of clinical notes and rename the selected notes.

[Clinical notes have been defined and identified from previous rules]

- ☐ Retrieve note meta information for renaming notes
- ☐ Other comments _____

7. [729_PAD_4] d. Provide an index date file [to the software] that contains a tab separated patient id and an index date that will act as a cutoff point for note processing.

[Clinical notes have been defined and identified from previous rules]

- ☐ Retrieve patient id and index date of identified related notes and generate the file
- ☐ Other comments _____

8. [729_PAD_5] NLP-PAD algorithm will only process the information on selected note sections, note types and service groups.

[note sections, note types and service groups names are provided in a table]

- ☐ Explore unstructured data to identify "note sections" programmatically

- Specify institution specific "note sections", "note types" and "service groups" corresponding to what the developing site has provided
- Customize and debug the NLP-PAD algorithm software
- Other comments _____

9. [149_CaMRSA_2] 2: caMRSA Case Criteria:

2a: Gold standard MRSA definition:

Bacteria culture-confirmed MRSA.

*****

"Bacteria culture-confirmed MRSA" is defined as:

a. Find the following or similar keywords in the text results of a bacterial culture lab test (methicillin OR oxacillin are the only 2 antibiotics for which to search):

i. MRSA: this was only screenings at NU so was not used at NU, but at other sites may be different

ii. Methicillin-resistant Staphylococcus aureus

iii. "STAPHYLOCOCCUS AUREUS" AND "OXACILLIN RESISTANT"

AND

b. “present” or “positive” or other affirmative mention (unqualified by negation, uncertainty, or historical reference) for cases

c. “absent” or “negative” or other negative mention for controls

- Identify "bacterial culture lab test" report
- Explore unstructured data to find if the keywords provided are also used locally
- Search MRSA Keywords from unstructured data
- Implement affirmative and negative mention of the proposed/customized keywords
- Other comments _____

10. [149_CaMRSA_4] 2b: Silver standard MRSA definition:

SSTI diagnosis within +/- 7 days of that diagnosis

*****

SSTI for cases is defined as:

i. Must have the following or similar keywords in the text results of a bacterial culture lab

test, such as skin, wound, boil, abscess, but also recognizing that anatomic sites (e.g. foot/hand/leg/buttock, etc.) are also typically classified as an SSTI. Thus, keywords we used are:

SKIN / WOUND / BOIL / ABSCESS / FOOT / HAND / LEG / BUTTOCK / BREAST / CARBUNCLE / FURUNCLE / FINGER / TOE / CELLULITIS / IMPETIGO / SUBCUTANEOUS / HAIR / HYDRADENITIS / NOSE / NASAL

ii. else, if cannot search this text, must have 1 of the ICD-9 codes from Table 1 below, within 1 week of the MRSA positive culture

Table 1: ICD-9-CM code: 611, 680.*, 681.*, 682.*, 684.*, 686.9, 704.8, 705.83

- Identify "bacterial culture lab test" report
- Explore unstructured data to find if the keywords provided are also used locally
- Search relevant keywords from unstructured lab test report
- Compile a machine readable input file containing complete ICD codes and keywords
- Map the provided ICD9 codes to ICD10 codes considering the EHR contains ICD10 coded diagnosis
- Other comments _____

11. [149_CaMRSA_7] Control: who receive routine primary care within your study site defined as >= 2 visits with a primary care provider over a minimum of a 3 year continuous period of enrollment.

*****

Continuous enrollment is defined at Group Health as a period of enrollment in a Group Health-administered insurance or integrated health care plan for at least 5 years, allowing interruptions or gaps in coverage of up to three months. [Sites may implement continuous enrollment in whatever manner they deem to be consistent with continuous enrollment]

- Specify an operational definition of "primary care provider" for implementation
- Potentially incomplete/unavailable provider information
- Find "health care plan" information
- Unavailable "health care plan" information
- Define "Continuous enrollment" for implementation if no "health care plan" data is available
- Identify where to find visits linking provider information from the EHR

□ Other comments _____

12. [149_CaMRSA_9] Control: No prior h/o SSTI

*****

i. Exclude if any of the above list of keywords for cases is found in the text result of a lab test

ii. AND use the ICD-9 codes in Table 1 as exclusions if they occurred at any time

[Keywords and ICD-9 codes are provided in the text]

□ Find "bacterial culture lab test" report

□ Explore unstructured data to find if the keywords provided are also used locally

□ Search keywords from unstructured data

□ Compile a machine readable input file containing complete ICD codes and keywords

□ Map the provided ICD9 codes to ICD10 codes considering the EHR contains ICD10 coded diagnosis

□ Populate NLP search results of lab reports for SQL query if unstructured data is not stored in the database

□ Other comments _____

13. [149_CaMRSA_10] Control: No prior h/o MRSA infection

*****

"History of MRSA infection" is defined as: use ICD-9 codes 041.12, 482.42, 482.41, 038.12, V02.54, V12.04, V09.0. Search anywhere there are diagnosis codes, and also in the problem list. If you have text in your problem list, please also search for MRSA that way.

[Keywords are provided in the text]

□ Find "problem list"

□ Explore unstructured data to find if the keywords provided are also used locally

□ Search keywords from problem list

□ Compile a machine readable input file containing complete ICD codes and keywords

□ Map the provided ICD9 codes to ICD10 codes considering the EHR contains ICD10 coded diagnosis

- Populate the search results of the problem list for SQL query if the problem list is not stored in the database
- Other comments _____

14. [188_CAAD_pseudoCode_1] 6.1. In the period preceding the date of the last known contact with the patient (e.g., an outpatient or inpatient encounter) the patient must have evidence of contact that satisfies at least one of the following three criteria:
6.1.1. Contact on at least two (2) occasions ≥365 days apart in the preceding 5 years (e.g., 6/23/2011 and 9/30/2014), OR

- Specify an operational definition of "outpatient or inpatient encounter" for implementation
- Define "known contact" for implementation
- Define "preceding 5 years" for implementation
- SQL implementation of "contact having ≥365 days apart"
- Other comments _____

15. [188_CAAD_pseudoCode_2] 6.1.2. Contact in at least three (3) different calendar quarters in the preceding 5 years (e.g., 1/1/2014 and 3/15/2014 and 6/30/2014), OR

- Define "known contact" for implementation
- Define "preceding 5 years" for implementation
- Implement "contact in different calendar quarters" in SQL
- Other comments _____

16. [188_CAAD_pseudoCode_9] 8.1. Have evidence from a carotid imaging study of ≤15% carotid artery stenosis bilaterally (i.e., there is no evidence of stenosis >15% in either carotid artery),

- Find "carotid imaging study" report
- Extract percentage information of carotid artery stenosis from unstructured reports
- Other comments _____

17. [188_CAAD_NLP_1] Choosing the Right NLP System for Your Reports
Two NLP systems are provided. One system assumes that each report to be processed

contains relevant information (carotid artery stenosis) from an imaging study of carotid artery stenosis.

The other system assumes that incoming reports may or may not be relevant imaging studies and therefore attempts to determine, as the first step in processing each report, whether or not the report is relevant.

- Find "carotid imaging study" report
- Check if the report contains "carotid artery stenosis" information
- Customize and debug the NLP system for extracting carotid artery stenosis information
- Other comments _____

18. [216_BPH_4] Population include: Exclude those with Prostate & Bladder Cancer using ICD-9, Tumor Registry (if available) and keywords:

ICD-9: (1 or more of any on separate days):

OR Tumor Registry: Primary site = any of the following (eg. C619 Prostate gland)

OR Keyword: (1 or more of any of the following keywords in Problem List): 'prostate cancer', 'malignant tumor of prostate', 'bladder CA', 'bladder cancer'.

- Find "problem list"
- Find "tumor registry" data
- Map the provided ICD9 codes to ICD10 codes considering the EHR contains ICD10 coded diagnosis
- Search the keywords from unstructured "problem list"
- Populate the search results of the problem list for SQL query if the problem list is not stored in the database
- Other comments _____

19. [216_BPH_8] Control: 3 or more outpatient visits in any 2-year period for age ≥40

- Specify an operational definition of "outpatient visits" for implementation
- Other comments _____

20. [236_Appendicitis_4] Case 4. Positive result of post-surgical biopsy report (SNOMED-CT CUI rules, see Figure 2, Table 4) YES->CASE, NO->EXCLUDE

[Table 4 is a list of SNOMED-CT CUIs and description for different groups]

*****

For a positive pathology report, note must have 1 CUI from inflammation group (Group 1), AND 1 or more from any of the other groups (Groups 2-6).

- Find relevant "pathology report"
- Compile a machine readable input file containing complete relevant SNOMED-CT CUIs
- Search CUIs from an NLP tool parsed report
- Other comments _____

21. [236_Appendicitis_5] Case 5. Systemic antibiotics (>2days treatment, starting on encounter of diagnosis(step 1), see Table 3) YES->CASE, NO->6

[Table 3 provides a list of medication names]

- Specify an operational definition of drug treatment duration for implementation
- Map the provided medication names to RxCUI codes that are used in the EHR
- Specify an operational definition of "on encounter of diagnosis" for implementation
- Compile a machine readable input file containing complete medication names
- Other comments _____

22. [170_StatinMACE_9] Revascularization while on statin: • Statin prescribed prior to the revascularization event in medical records at least 180 days

[Medication names provided]

- Compile a machine readable input file containing complete medication names
- Map the provided medication names to RxCUI codes that are used in the EHR
- Other comments

23. [172_AMD_1] 1a. -Select all subjects from the PMRP (Personalized Medicine Research Project) cohort who have:

- **Consented**
- **Did not withdraw from the study**
- **Include subjects with contact_for_research = 'N'**

- **Include subjects where questionnaires have been scanned**

□ Specify a Localized definition of “PMRP (Personalized Medicine Research Project) cohort” and its corresponding covariates

□ Find “PMRP subject” relevant information from the local cohort

□ Other comments _____

24. [172_AMD_12] 12a. Select from subjects (see step 11a above) having at least one eye exam within the last two years if living or the last two years prior to date of death if deceased, using CPT codes: 92002, 92004, 92012, 92014, 92018, 92019, or 99201, 99202, 99203, 99204, 99205, 99211, 99212, 99213, 99214, 99215, 99241, 99242, 99243, 99244, 99245 where provider specialty is Ophthalmology or Optometry.

□ Identify where to find “provider specialty” information linking procedure from the EHR

□ Specify an operational definition of “provider specialty is Ophthalmology or Optometry” for implementation

□ Potentially incomplete or unavailable information for “department” information

□ Compile a machine readable input file containing complete procedure codes

□ Other comments _____

25. [120_DM/HtnCKD_11] Case: B10. Type 2 diabetes?

[Use existing eMERGE T2DM algorithm that is developed by Northwestern University]

□ Implement another existing phenotype

□ Other comments _____

26. [120_DM/HtnCKD_23] Exclusion criteria: We did not allow patients that had serum creatinine tests on consecutive days or within the same day as we assumed these were ‘inpatients’. For patients that had duplicate serum creatinine tests meaning they had two tests at the same date/time but different values we kept the test with the maximum value in the algorithm.

□ Find lab code for "serum creatinine"

□ Implement "consecutive days" in SQL

□ Other comments _____

27. [125_CardiorespiratoryFitness_2] 2. In the remaining medical records, flag any record indicating cardiac stress test by using combinations of relevant codes detected simultaneously on the same date as listed below. Then set the first detect date as the test date for that patient.

S1: ECG (any of the CPT-4 codes: 93015; 93016; 93017; 93018) + Echo (CPT-4 code 93350); OR Echo code 93351 only;

S2: ECG (any of the CPT-4 codes: 93015; 93016; 93017; 93018) + Cardiac nuclear test (CPT-4 codes: A9500);

S3: ECG (any of the CPT-4 codes: 93015; 93016; 93017; 93018) + Oxygen uptake (any of the CPT-4 codes: 94680; 94681);

S4: ECG (any of the CPT-4 codes: 93015; 93016; 93017; 93018) only.

S5: If the patient is flagged as S1 and S3 simultaneously, set that patient as S5.

- Compile a machine readable input file containing complete CPT codes
- Other comments _____

28. [224_GERD_2] Case Inclusion: - Individual's medical record includes two or more prescriptions for GERD -related medications (see Table 1)

[Drug generic and brand names are provided in a table]

- Compile a machine readable input file containing complete drug names
- Map the provided medication names to RxCUI codes that are used in the EHR
- Other comments _____

29. [224_GERD_8] Control: - No history of relevant medications (see Table1)

[Drug generic and brand names are provided in a table]

- Compile a machine readable input file containing complete drug names
- Map the provided medication names to get RxCUI codes that are used in the EHR
- Find out if "no history of" means searching both structured and unstructured data (for example Past Medical History of clinical notes) should for medication use
- Search drug names / codes from relevant unstructured data if it is required
- Other comments _____

30. [99_OcularHtn_12] 12a: -Select subjects having two or more ocular hypertension diagnoses (see 3a above) within ophthalmology/optometry (see 10a above), where the earliest and most recent diagnosis are at least fourteen days apart.

[3a] Diagnoses of Ocular Hypertension

'ICD 9' = '365.04'

'HICDA 2' = '375.3'

'HICDA 1' = '378.909'

[10a] ocular hypertension diagnoses codes given by ophthalmologist/optometrist within ophthalmology/optometry department(s)

- Compile a machine readable input file containing complete diagnosis codes
- Map the provided ICD9 codes to ICD10 codes considering the EHR contains ICD10 coded diagnosis
- Identify where to find diagnosis linking department and provider information from the EHR
- Specify an operational definition of "ophthalmologist/optometrist within ophthalmology/optometry" for implementation
- Potentially incomplete/unavailable department/provider information
- Other comments _____

31. [99_OcularHtn_13] 13a: Select subjects having two or more ocular hypertension diagnoses (see 3a above) within ophthalmology/optometry (see 10a above), where the earliest and most recent diagnosis are at least fourteen days apart, and subjects age at earliest ocular hypertension diagnosis is 40 years or older.

[3a] Diagnoses of Ocular Hypertension

'ICD 9' = '365.04'

'HICDA 2'= '375.3'

'HICDA 1' = '378.909'

[10a] ocular hypertension diagnoses codes given by ophthalmologist/optometrist within ophthalmology/optometry department(s)

- Compile a machine readable input file containing complete diagnosis codes

- Map the provided ICD9 codes to ICD10 codes considering the EHR contains ICD10 coded diagnosis
- Identify where to find diagnosis linking department and provider information from the EHR
- Specify an operational definition of "ophthalmologist/optometrist within ophthalmology/optometry" for implementation
- Potentially incomplete/unavailable department/provider information
- Other comments _____

32. [147_HeartFailure_8] Free Text Ejection Fraction Results - Priority Metric

(Optional – Use only if numeric EF measurements are not available)

EF Result Categories	Free Text Variations
Preserved	**normal, supernormal, low-normal, moderate**
Reduced	**abnormal, reduced, low, severe, decreased**

- Explore unstructured data to find if the free text variations provided are also used locally
- Extract EF measurements from relevant reports
- Other comments _____

33. [147_HeartFailure_11] Step 4: identifying HF control:

- **EF ≥ 50% if measured or patient does not have echocardiographic measurements**

- Find relevant documents for "echocardiographic measurements"
- Extract EF measurements from relevant documents if unstructured data is used
- Other comments _____

34. [92_Diverticulosis_B_3] At least 1 colonscopy?

- Check if "colonscopy" refers to procedure codes or relevant report
- Find relevant report of colonoscopy if "colonscopy" refers to report
- Check the existence of the relevant report if "colonscopy" refers to report
- Other comments _____

35. [109_VTE_1] Step 1: NLP-based execution

In the algorithm there needs to be one of the following two conditions to be considered a case:

1) At least one non-negated mention from Table A, Column 1 (Disorder related mentions) and Table A, Column 2 (Anatomical site related mentions) either in the SAME or adjacent sentences in a 'section of interest' OR

2) At least one non-negated term from the following list in a 'section of interest' from the Table B, Column 1 (Explicit "Stand-alone" mentions) and/or Table C, Column 1 (Other mentions) without any of the exclusion terms (Table B, Column 2).

[Keywords are provided from Table A, B and C]

- Find relevant notes
- Specify "section of interest" from locally used/identified sections
- Search non-negated keywords from same or adjacent sentences from selected section
- Other comments _____

36. [162_Autism_4] Case: DSM-V Symptom Criteria

4a. Each of DSM-V criteria set A (A1,A2,A3) + At least 2 from criteria set B (B1,B2,B3,B4)

4b. Both criteria set C and D

[162_Autism glossary of terms.pdf provides dictionary]

- Find relevant notes
- Search symptom keywords from the relevant notes
- Compile a machine readable input file containing complete keywords
- Identify where from clinical notes to search "DSM-IV Symptom criteria"
- Other comments _____

37. [161_EarlyChildhoodObesity_7] Case: 7. More than 50% of BMI Measurements > 76th percentile YES->8 NO->EXCLUDE

[URL of WHO/CDC BMI-for-age percentiles is provided]

[Assuming the BMI measurements have been extracted from previous steps]

- Understand "76th percentile" and "WHO/CDC BMI-for-age percentiles"
- Download and import WHO/CDC BMI-for-age percentile file for SQL query use

□ Other comments _____

38. [97_AAA_20] EXCLUDES: EXLUDE TYPE 3:TYPE 3a:

IS not a Case Type 1

Is not a Case Type 2

Has not had an encounter in the last 5 years

[Case Type 1 and Case Type 2 have been identified and implemented from the previous rules]

□ Specify an operational definition "encounter" for implementation

□ Other comments _____

39. [112_Zoster_3] Case Incl: I-3 Has ≥5 years of continuous enrollment (or encounter) history since age 35.

Note: Implementations of "continuous enrollment" may vary by institution. Continuous enrollment is defined at Group Health as a period of enrollment in a Group Health-administered insurance or integrated health care plan for at least 5 years, allowing interruptions or gaps in coverage of up to three months. Gaps of this size are allowed because they typically represent administrative data inconsistencies rather than actual interruptions in access to care.

Sites may implement continuous enrollment in whatever manner they deem to be consistent with continuous enrollment, including use of more granular rules.

□ Specify an operational definition of "encounter" for implementation

□ Find "health care plan" information

□ Unavailable "health care plan" information

□ Define "continuous enrollment" for implementation if no "health care plan" data is available

□ Other comments _____

40. [9_Cataracts_Main_2] 2a -Select subjects who have had at least 1 "Inclusion" Cataract surgery;

- **Use the Marshfield Clinic Charges file (contains CPT codes and charges).**

- **Select the following CPT codes '66982', '66983', '66984', '66985', '66986','66830', '66840', '66850', '66852', '66920', '66930', '66940'.**
- **Exclude traumatic, congenital and juvenile cataract surgery codes.**
- **Exclude reversed and reversal records and include only production records.**
- **The provider must be a clinical provider.**

[Relevant ICD9 codes and CPT codes are provided]

- Identify where to find "provider" information linking procedure data from the EHR
- Incomplete/Unavailable provider information
- Specify an operational definition of "clinical provider" for implementation
- Compile a machine readable input file containing complete ICD and CPT codes
- Identify local data corresponding to "Marshfield Clinic Charges file"
- Specify what are "reversed and reversal records" and "production records" in EHR data
- Other comments _____

41. [9_Cataracts_Main_8] 5b Use NLP to search for general “Inclusion” cataract terms including one or more of the following CUIs:

[List of CUIs are provided in text]

Or meeting any of the following rules:

[MedLEE terms finding/descriptor/Region/Bodyloc rules are provided in text]

[Documents have been identified from previous algorithm rules]

- Install MedLEE
- Use MedLEE to parse relevant documents
- Search CUIs from MedLEE parsed parsed documents
- Specify what are "reversed and reversal records" and "production records" in EHR data
- Compile a machine readable input file containing complete keywords
- Other comments _____

42. [9_Cataracts_Main_10] 5d Locate ophthalmology form documents

- **Using feedback from a domain expert, identify the records in your EHR that contain ophthalmology information, specifically concerning cataracts. This is highly dependent on each institution's EHR and data collection strategies."**

☐ Discuss with domain experts and explore data to find documents containing "ophthalmology concerning cataracts" information

☐ Search unstructured documents for finding relevant notes

☐ Other comments _____

43. [9_Cataracts_Main_16] 9a Refer to ID 5 for information on how cataract was found (or not) using NLP and ICR techniques

[Documents have been identified from previous algorithm rules]

☐ Read another rule for understanding NLP and ICR techniques to search cataract

☐ Other comments _____

44. [9_Cataracts_SubytpeNLP_1] 1a This process took place in steps 5 & 9 of the "Pseudo code for the Cataract Phenotype". Identify all documents that have the term "Cataract" embedded in the text of electronic documents. This is a filtering mechanism to reduce the number of documents that will require NLP and ICR processing.

[Documents have been identified from previous algorithm rules]

☐ Search "cataract" from relevant notes

☐ Other comments _____

45. [9_Cataracts_SubytpeNLP_3] 3a Determine which eye the cataract was found. (left eye and right eye)

[list of CUI, MedLEE Descriptors of Region and Bodyloc are provided for searching cataract mention in the same sentence]

[Documents have been identified from previous algorithm rules]

☐ Use MedLEE to parse relevant notes

☐ Compile a machine readable input file containing complete CUIs and descriptors information

☐ Search CUI and related descriptors from MedLEE parsed notes

☐ Other comments _____

46. [10_Dementia_1] We are looking for people with a minimum of 5 visits with the DX of interest OR a dementia drug fill.

[DX ICD9 codes and drug names are provided in the text]

- Identify where to find "visit" information linking diagnosis or drug
- Map the provided ICD9 codes to ICD10 codes considering the EHR contains ICD10 coded diagnosis
- Compile a machine readable input file containing complete ICD codes and drug names
- Other comments _____

47. [16_PAD_1] Ankle-brachial index (ABI) for ascertaining PAD.

[The background information of how ABI is measured, check the pseudocode for details]

- Find data of ABI
- Extract ABI from relevant reports if unstructured data is used
- Other comments _____

48. [16_PAD_8] Mutivariable logistic regression model to ascertain PAD.

This model uses the following billing code variables to predict presence of PAD in these patients. A description of the billing codes is provided in the table below.

An integer score based on the beta-coefficients was created and used to predict presence of PAD.

Variable	DF	Estimate	StdErr	WaldChiSq	ProbChiSq	Score
Intercept	1	-1.58	0.04	1421.53	0.0000	-6
CPT4Px73725	1	0.66	0.16	16.33	0.0001	3
…	…	…	…	…	…	…
ICD9Px8848	1	0.65	0.13	23.94	0.0000	3
…	…	…	…	…	…	…

- Implement the logistic regression model in SQL
- Map the provided ICD9 codes to ICD10 codes considering the EHR contains ICD10 coded diagnosis

- Compile a machine readable input file containing complete ICD and CPT codes
- Other comments _____

49. [8_QRS_4] ECG Impression must not contain evidence of heart disease concepts. Method: Natural Language Processing (NLP) on ECG impression. Will exclude all but negated terms (e.g., exclude those with possible, probable, or asserted bundle branch blocks). Should also exclude normalization negations like "LBBB no longer present." ***[List of disease UMLS CUIs are provided in the text]***

- Find EKG report for extracting "ECG impression"
- Use NLP tool to parse EKG report
- Search non-negated UMLS CUIs from the NLP tool parsed reports
- Compile a machine readable input file containing complete disease UMLS CUIs
- Other comments _____

50. [8_QRS_5] ECGs was not recorded during presence of sodium channel blocking drugs [List of medication names are provided in text]. Method: Taken from last clinic note or problem list before the ECG, can be simplified to "anytime before" the ECG

- Map the provided medication names to RxCUI codes that are used in the EHR
- Find medication list from problem list or clinic note
- Search drug names from clinical note or problem list
- Compile a machine readable input file containing complete medication names
- Populate the search results for SQL query if clinical note or problem list is not stored in the database
- Other comments _____

51. [8_QRS_7] Notes contain no evidence of heart disease concepts before ECG time or within one month following Method:

- **NLP for notes, Problem Lists at or near ECG time, ignoring Family Medical History and Allergy sections (using section tagger)**
- **ICD9 and CPT codes at or near ECG time describing heart disease**

- **Labs:**
    - **Positive cardiac enzymes (CPK-MB > 8, Troponin > 0.05)**
    - **BNP > 100**

[List of disease UMLS CUIs are provided in the text]

- Use/Customize section tagger to extract FMH and allergy section from notes
- Search non-negated UMLS CUIs from the NLP tool parsed notes
- Compile a machine readable input file containing complete UMLS CUIs, ICD codes and CPT codes
- Map the provided medication names to RxCUI codes that are used in the EHR
- Identify lab code for BNP, CPK-MB, Troponin
- Explore data to check lab units for identified labs
- Populate the search results from notes and problem lists for SQL query if notes and problem lists are not stored in the database
- Other comments _____

52. [8_QRS_8] Must have at least a problem list and/or note containing non-empty (can say "none") medication list and past medical history before or immediately after the time of the ECG.

Method: Note section tagging to detect non-empty past medical history and medication sections. [using section tagger]

- Identify problem list
- Explore notes to identify "medication list, past medical history" programmatically
- Specify operational definition of "non-empty (can say "none") medication list" for implementation
- Use/Customize section tagger to extract non-empty relevant sections
- Other comments _____

53. [8_QRS_9] Heart disease defined

Presence of ICD9 before ECG (page 4) [List of ICD9 codes are provided in text]

Presence of CPT codes representing cardiac surgery (page 5) [List of ICD9 codes are provided in text]

Presence of concepts in free-text clinical notes, search for either:

1. concepts (for use with natural language processing systems, page 10) [List of UMLS CUI are provided in text] *****Use this list if UMLS concept-identification tools (e.g., MetaMap, KnowledgeMap, MedLEE) are used to process the text documents and ECG impressions beforehand.

- or -

2. strings (for text searching with SQL or other systems, page 21) - must be "probable" or "asserted" and not in a "Family Medical History" section) indicative of:

[List of string/keywords are provided]***(assuming non-negated, not in the presence of a family history section)**

Note: Use this list instead of the UMLS CUIs if concept identification is not performed

- Find relevant clinical notes
- Search non-negated heart disease strings from clinical notes
- Use/Customize MetaMap/KnoweledgeMap/MedLEE for parsing clinical notes
- Search non-negated heart disease UMLS CUI codes from NLP tool parsed notes
- Compile a machine readable input file containing complete ICD codes, CPT codes, UMLS CUIs and keywords
- Map the provided medication names to RxCUI codes that are used in the EHR
- Explore notes to identify "Family Medical History" programmatically
- Populate the search results from clinical notes for SQL query if the clinical notes are not stored in the database
- Other comments _____

54. [17_RBC_7] General Selection Process for RBC Indices and ESR Samples V. Patients taking medications affecting ESR and RBC indices: We used the medications listed in Appendix B2 to screen our study participants for the use of medications that can potentially affect RBC indices (mainly taken for autoimmune/connective tissue disorders and epilepsy). NLP was implemented to that regard. Samples collected from patients taking these medications were excluded if they fell in a 4-month period centered upon the date at which such use was detected in the EMR.

[Medication names are provided from the table]

- Map the provided medication names to RxCUI codes that are used in the EHR
- Search medication names from relevant list if unstructured medication list is used
- Compile a machine readable input file containing complete medication names
- Populate the search results for SQL query if unstructured medication list is not stored in the database
- Other comments _____

55. [19_WBC_1] Lab Results with the following components (where available):

1. LEUKOCYTE COUNT (K/uL)

2. NEUTROPHILS (% of LC)

3. BANDS (% of LC)

4. LYMPHOCYTES (% of LC)

5. MONOCYTES (% of LC)

6. EOSINOPHILS (% of LC)

7. BASOPHILS (% of LC)

8. PLATELET COUNT (K/uL)

9. Mean Platelet Volume (MPV)

- Find lab code for required labs
- Explore lab data to check if the lab units are same with as described, otherwise convert lab units for lab data extraction
- Other comments _____

56. [68_ResHTN_2] Case step 1: CASE Type 2: Has two outpatient (if possible) measurements of SBP > 140 or DBP > 90 at least one month after meeting medication criteria while still on 3 simultaneous* med classes

***Simultaneous is defined as evidence that they are taking the medications concurrently. Such evidence could be presence of the medications in the same medication list (e.g., problem list, clinic note, or discharge summary) or via medication refill data [medication class and medication names are provided in text]**

- Specify an operational definition of outpatient for implementation
- Identify where to find "visit" information linking lab from the EHR

- Find lab code for "SBP" and "DBP"
- Explore lab data to check if the lab units are same with as described, otherwise convert lab units for lab data extraction
- Compile a machine readable input file containing complete keywords
- Map the provided medication names to RxCUI codes that are used in the EHR if structured medication list is used
- Find medication list from problem list, clinic note, or discharge summary
- Search medication names from identified medication list
- Populate the medication search results for SQL query if the unstructured medication list is not stored in the database
- Implement "multiple simultaneous med classes" in SQL
- Other comments _____

57. [68_ResHTN_6] Case step 2: 2.3 Exclude if GFR < 30 ml/min before the time of meeting the CASE 1 or 2 definitions or within 6 months after meeting the medication definition. ***eGFR is calculated using MDRD formula which serum creatinine, age and race are required

- Implement eGFR formula in SQL
- Find lab code for "serum creatinine"
- Explore lab data to check if the lab units are same with as described, otherwise convert lab units for lab data extraction
- Other comments _____

58. [68_ResHTN_9] Controls – Case 1: has 1 med from med classes below (and never has more than 1 simultaneous med class, although the med class can change)
***Simultaneous is defined as evidence that they are taking the medications concurrently. Such evidence could be presence of the medications in the same medication list (e.g., problem list, clinic note, or discharge summary) or via medication refill data.**
[medication class and medication names are provided in text]

- Find medication list from problem list, clinic note, or discharge summary
- Search medication names from identified medication list

- Compile a machine readable input file containing complete medication names and classes
- Map the provided medication names to RxCUI codes that are used in the EHR if structured medication list is used
- Populate the medication search results for SQL query if the unstructured medication list is not stored in the database
- Implement "more than 1 simultaneous med class" in SQL
- Other comments _____

59. [68_ResHTN_10] Controls – Case 1: Has all SBP<135 and DBP < 90 one month AFTER BP meds (require at least 1 BP measurement)

- Find lab code for "SBP" and "DBP"
- Explore lab data to check if the lab units are same with as described, otherwise convert lab units for lab data extraction
- Find unstructured medication list
- Compile a machine readable input file containing complete medication names and classes
- Search medication names from identified unstructured medication list
- Map the provided medication names to RxCUI codes that are used in the EHR if structured medication list is used
- Populate the medication search results for SQL query if the unstructured medication list is not stored in the database
- Other comments _____

60. [68_ResHTN_11] Has no outpatient (if possible) measurement of SBP > 140 or DBP > 90

- Specify an operational definition of outpatient for implementation
- Identify where to find visit information linking lab from the EHR
- Find lab code for "SBP" and "DBP"
- Explore lab data to check if the lab units are same with as described, otherwise convert lab units for lab data extraction
- Other comments _____

61. [14_Hypothyroidism_3] Case Inclusion: Require at least 2 instances of either medication or lab (a combination is acceptable) with at least 3 months between the first and last instance of medication or lab

[Medication names are provided in the text]

*******Case lab names/values**

Hypothyroidism: TSH >5 or FT4 <0.5

Anti-thyroglobulin antibodies: H-TGA, ThyrAB, AThyg- positive

Anti-thyroperoxidase: H-TPO, TPO, AThyP - positive

Anti-thyroid antibodies: ThyAb – positive

- Compile a machine readable input file containing complete medication names
- Map the provided medication names to RxCUI codes that are used in the EHR
- Find lab codes of all required labs
- Explore lab data to check if the lab units are same with as described, otherwise convert lab units for lab data extraction
- Explore lab data to check how positive labs are presented
- Other comments _____

62. [14_Hypothyroidism_8] Time-dependent case exclusion: Recent pregnancy TSH/FT4 (any pregnancy billing code or lab test if all Case Definition codes, labs, or medications fall within 6 months before pregnancy to one 1 year after pregnancy)

[Pregnancy ICD9 list is provided in text; Pregnancy lab name is provided]

- Compile a machine readable input file containing complete ICD codes
- Map the provided ICD9 codes to ICD10 codes considering the EHR contains ICD10 coded diagnosis
- Find lab codes of TSH/FT4
- Explore lab data to check if the lab units are same with as described, otherwise convert lab units for lab data extraction
- Other comments _____

63. [14_Hypothyroidism_9] Time-dependent case excl: Recent contrast exposure (all abnormal lab or medication references occurring within 6 weeks following a contrast study)

*****Contrast exposure case exclusion (cannot be a case if all abnormal lab or medication references occur within 6 weeks of a contrast study)**

Multiple methods could be used for this. One proposal would be to use NLP to identify radiology reports with "intravenous contrast" (e.g., optiray, radiocontrast, iodine, omnipaque, visipaque, hypaque, ioversol, diatrizoate, iodixanol, isovue, iopamidol, conray, iothalamate, renografin, sinografin, cystografin, conray, iodipamide) keywords that are not negated. Another method would be to identify radiology tests that could have contrast (such as CT scans) and exclude these.

MRIs, gastrograffin, and barium contrast, however, are not exclusions."

- Find out what "abnormal lab or medication" refer to
- Compile a machine readable input file containing complete keywords, medication names
- Map the provided medication names to RxCUI codes that are used in the EHR
- Find lab codes of all required labs
- Explore lab data to check if the lab units are same with as described, otherwise convert lab units for lab data extraction
- Explore lab data to check how abnormal labs are presented
- Explore radiology reports to find if the "intravenous contrast" keywords provided are also used locally
- Search non-negated "intravenous contrast" keywords from radiology reports
- Populate NLP search results for SQL query if unstructured data is not stored in the database
- If radiology tests are used, need to define relevant lab tests
- Other comments _____

64. [14_Hypothyroidism_10] Control: No billing codes for hypothyroidism, no evidence of thyroid replacement meds

[Relevant ICD9 codes and Thyroid-altering medications names are provided in the text]

- Compile a machine readable input file containing complete ICD codes and medication names

- Map the provided ICD9 codes to ICD10 codes considering the EHR contains ICD10 coded diagnosis
- Map the provided medication names to RxCUI codes that are used in the EHR
- Specify an operational definition of “billing codes for hypothyroidism” for implementation
- Other comments _____

65. [14_Hypothyroidism_12] Control: Must contain at least two Past Medical History sections and Medication lists (could substitute two non-acute clinic visits or requirement for annual physical)

- Specify institution specific section headers for "Past Medical History" and "Medication lists"
- Search clinical notes to identify required sections
- Specify an operational definition of "non-acute clinic visits" or "requirement for annual physical" for implementation
- Other comments _____

66. [14_Hypothyroidism_13] Control exclusion: Any cause of hypo- or hyper-thyroidism

- Understand "any cause"
- Specify an operational definition of "Any cause of hypo- or hyper-thyroidism" for implementation
- Other comments _____

67. [15_Lipids_7] Genotyped pts

- Specify a Localized definition of "genotyped" for implementation
- Other comments _____

68. [15_Lipids_14] Earliest Dx date

[Dx refers to type 1 or type 2 diabetes mellitus and cancer; related ICD9 codes are provided in tables]

- Compile a machine readable input file containing complete ICD codes
- Map the provided ICD9 codes to ICD10 codes considering the EHR contains ICD10 coded diagnosis

□ Other comments _____

69. [15_Lipids_16] >= LDL-C measure before exclusion date?

[The earliest exclusion date has been extracted from its previous rule]

□ Understand if "exclusion date" refers to "earliest exclusion date"

□ Other comments _____

70. [11_DiabeticRetinopathy_4] 2. Confirmed diagnosis/problem mention of diabetic retinopathy: OR, the patient has any of the codes or terms listed in a problem list. Depending on how problem lists are stored in the medical record, this may require NLP to determine the problems.

[List of keywords terms are provided in text]

□ Find "problem list"

□ Compile a machine readable input file containing complete keywords

□ Search codes or terms from "problem list"

□ Populate the problem list search results for SQL query if the problem list is not stored in the database

□ Other comments _____

71. [11_DiabeticRetinopathy_6] 4. Eye exam within the past 2 years

The patient is shown to have had at least one eye exam within the past two years of the reference date. If a patient is deceased, it will be within two years prior to date of death. Eye exam is defined as either:

***Any encounter with a provider in an Optometry or Ophthalmology department**

****If appointment data are available, identify appointments in an Optometry or Ophthalmology department, or an encounter with a provider with a specialty of Optometry or Ophthalmology.**

Or

***Any of the following CPT codes for Evaluation & Management (E&M) that were submitted by a provider in an Optometry or Ophthalmology department. [List of CPT codes are provided in text]"**

- Find "death" data
- Specify an operational definition of "encounter" for implementation
- Identify where to find "department" information linking "appointment data" from the EHR
- Incomplete or unavailable "department" information
- Identify where to find "department" information linking procedure (E&M) from the EHR
- Compile a machine readable input file containing complete CPT codes
- Other comments _____

72. [602_FH_3] All patients above 18 years of age and with a lipid profile in EHR will be the parent sample set for identifying cases and controls.
Index date is defined as a date with the highest LDL-C levels in the EHR calculated using the Friedewald equation* or measured directly.
***[LDL-C] = [Total cholesterol] - [HDL-C] - ([TG]/5) for mg/dL**
[LDL-C] = [Total cholesterol] - [HDL-C] - ([TG]/2.2) for mmol/L
[Lipid profile refers to LDL-C, TG]

- Find lab codes of lipid profile
- Explore lab data to check if the lab units are same with as described, otherwise convert lab units for lab data extraction
- Implement the equation for index date in SQL
- Other comments _____

73. [514_CRC_4] 5.2.2. During the period spanning 365 days before through 365 days after the date of a qualifying CRC diagnosis code, has at least one procedure code indicating a surgical procedure to treat CRC (Table 5.2.B).
[CPT code and ICD-9Proc codes are provided from the table]

- Compile a machine readable input file containing complete ICD and CPT codes
- Map the provided ICD9 codes to ICD10 codes considering the EHR contains ICD10 coded diagnosis
- Other comments _____

74. [514_CRC_11] 6.1.3. Has no surgical pathology reports containing the terms “colon,” “cecal,” or “cecum,” ever.

- Find "surgical pathology report"
- Search the provided terms from surgical pathology reports
- Other comments _____

75. [514_CRC_14] 6.2.3. The negative lab tests span a period of time ≥1,826 days(i.e., the difference between the earliest and latest negative lab test is at least 1,826 days or five years).

[Relevant lab LOINC codes are provided in a table]

- Compile a machine readable input file containing complete LOINC codes
- Explore lab data to check how negative labs are presented
- Other comments _____

76. [584_Migraine_4] Control: No ICD codes related to disease of nervous system (320 - 359.99; G00 - G99) or brain tumor (191.xx, 225.xx; D33.2, C71.9), preferably match for age, race and gender with cases.

- Compile a machine readable input file containing complete ICD codes
- Map the provided ICD9 codes to ICD10 codes considering the EHR contains ICD10 coded diagnosis
- Specify an operational definition of "match for age, race and gender" for SQL implementation
- Other comments _____

77. [582_ChronicRhinosinusitis_11] Flowchart: CRS Dx @>=2 CRS specialty visits **with otolaryngologist [annotation is from published paper]**

[ICD9CM and ICD10CM codes are provided in excel]

- Identify where to find "specialty" information linking diagnosis from the EHR
- Compile a machine readable input file containing complete ICD codes
- Other comments _____

78. [528_CIN_6] Serum creatinine (SCr) values within two years prior to DCA, excluding the day of the administration of contrast, will be used as a baseline. If the patient has more than one SCr value prior to DCA the value closest to DCA will be used as a baseline. Post-procedure SCr will be populated for up to 7 days following DCA starting from the day after contrast administration. We will use the highest SCr during these 7 days. The LOINC codes (logical observation identifiers names and codes) for creatinine are listed in table 5.

- Compile a machine readable input file containing complete LOINC codes
- Explore lab data to check if the lab units are same with as described, otherwise convert lab units for lab data extraction
- Implement "closest or highest value" in SQL
- Other comments _____

79. [585_RheumatoidArthritis_1] 2.1. Feature Dictionary

The table below lists the features for the RA phenotype algorithm and their associated beta coefficients (weights). The weights are used to derive a predicted probability of RA or no RA for each subject. The list of codes for each feature is listed in the Appendices (Section 3).

Feature_ID Nomenclature: [Phenotype] + '_COD_' + [Feature Type Abbreviation] , where 'COD' means coded variable.

Features:

DX_RA: Coded mentions of a Rheumatoid Arthritis diagnosis per subject

DX_Lupus: Coded mentions of a Lupus diagnosis per subject

DX_Psoriatic: Coded mentions of a Psoriatic arthritis diagnosis per subject

LAB_RF: A lab test for Rheumatoid Factor.

patient_dxenct: Total number of encounters (visits), per subject, with a coded diagnosis (any diagnosis not limited to RA), limit to one occurrence per date, an inpatient stay that spans multiple days will be counted as one date.

[Check pseudocode for detailed feature table]

[ICD-9 and ICD10 codes for DX features, LONIC codes for LAB feature are provided in excel file]

- Identify where to find "encounters (visit)" information linking diagnosis from the EHR

- Implement multiple features in SQL
- Other comments _____

80. [588_NALFD_6] NAFLD Case 3: 3) Diagnostic: Evidence of hepatic steatosis, in clinical notes: a. Method - Natural Language: Text positive for diagnostic phrases (Table 3)
[Table 3 lists NASH terms]

- Find relevant notes
- Compile a machine readable input file containing complete NASH terms
- Explore unstructured clinical notes to find if the NASH terms provided are also used locally
- Search non-negated NASH terms from clinical notes
- Other comments _____

81. [772_ArrhythmiasEKGintervals_2] "Normal" ECG: Recorded in an outpatient setting. Method: If algorithm does not exist for classifying inpatient vs. outpatient: • Greater than >24 hours from hospitalization window based on admission/discharge date. • Presence of sinus rhythm
*******Sinus rhythm: Include only EKGs for which the words "sinus rhythm" appears within the EKG interpretation AND a PR-interval is defined"**

- Specify an operational definition of "outpatient setting" for implementation
- Find EKG report
- Extract "sinus rhythm" information from EKG reports
- Understand/Identify "the EKG interpretation AND a PR-interval is defined"
- Populate "sinus rhythm" search results for SQL query if EKG report is not stored in the database
- Other comments _____

82. [772_ArrhythmiasEKGintervals_5] ECGs was not recorded during presence of potentially EKG altering drug
Method: Taken from last clinic note or problem list before the ECG, can be simplified to "any time before" the ECG (see list below)
[Medication names are provided from text and excel file]

- Find relevant note and problem list
- Compile a machine readable input file containing complete medication names
- Search medication names from relevant sections of clinical notes
- Populate medication search results for SQL query if relevant note or problem list is not stored in the database
- Other comments _____

83. [772_ArrhythmiasEKGintervals_7] Notes contain no evidence of heart disease before ECG time or within one month following as defined below

Method: • ICD9/10 and CPT codes (see list) ever before or within one month following the date of the ECG describing heart disease:

• Labs: •• BNP > 100* (LOINC: 42637-9); •• nT-BNP > 300* (LOINC: 71425-3); •• Trop-I > 0.05 (LOINC: 42757-5); •• Trop-T > 0.05 (LOINC: 48425-3)

[ICD9/ICD10 codes/code ranges and CPT codes are provided in text and excel file]

- Compile a machine readable input file containing complete ICD and CPT codes
- Explore lab data to check if the lab units are same with as described, otherwise convert lab units for lab data extraction
- Other comments _____

84. [884_OvUtCa_4] Subjects meeting the above case definition must be classified according to each of the following binary indicators:

A. Ever satisfied criterion for ovarian cancer per column G in Table 1.

1.Age at date of earliest known ovarian cancer.

B. Ever satisfied criterion for uterine cancer per column G in Table 1.

1. Age at date of earliest known uterine cancer.

C. Ever satisfied criterion for fallopian cancer per column G in Table 1.

1. Age at date of earliest known fallopian cancer.

D. Ever satisfied criterion for peritoneal cancer per column G in Table 1.

1. Age at date of earliest known peritoneal cancer.

E. Ever satisfied criterion for endometrial cancer per column G in Table 1.

1. Age at date of earliest known endometrial cancer.

[Diagnosis codes of ICD9 and ICD10 are provided in Table 1]

- Compile a machine readable input file containing complete ICD codes
- Implement multiple disease classification in SQL
- Other comments _____

85. [755_Autoimmunity_3] Control Cohort: Condition B. No instances of any positive serologies as defined by institutional and assay recommendations from a list of serologies (Table 3). Codes are listed in fileAIDalgorithm_V1_coding_control.csv under variable names starting with "Serology".

[Serology tests names are in table 3. Relevant LOINC codes are provided in the txt file]

- Explore lab data to check if the lab units are same with as described, otherwise convert lab units for lab data extraction
- Explore lab data to check how positive labs are presented
- Other comments _____

86. [519_MetforminResponse_1] Step 1: Defining a Diabetes Free Cohort

Exclude patients if any of the following criteria if occurred prior to the start date.

- **Random glucose > 200 mg/dL**
- **Fasting Glucose ≥ 126 mg/dL**
- **Two instances of diabetes diagnoses codes at least 60 days apart (Appendix A)**
- **Exposure to a diabetes medication (see Medication Data Dictionary)**

*******Implementation Note: Mayo Clinic had a completely electronic prescription database as of 1/1/2004. To enhance our ability to get metformin initiators as opposed to prevalent users we excluded all patients with prior evidence of diabetes (Type 1 and 2) and/or diabetic medication prior to 2004. The start date will need to be customized at each site but note that metformin was approved by the FDA on 03/03/1995.**

[ICD9 and ICD10 codes are provided in a table, medication ingredient RxCUI are provided in a table]

- Find lab codes of all required labs

- Explore lab data to check if the lab units are same with as described, otherwise convert lab units for lab data extraction
- Compile a machine readable input file containing complete ICD codes and RxCUI codes
- Map the provided ingredient RxCUI to non-ingredient RxCUI codes that are also used in the EHR
- Find out when is the metformin usage date
- Other comments _____

87. [93_ColonPolyp_3] controls must have not had any polyps in any of their colonoscopies.

- Find relevant pathology report
- Explore unstructured data to find relevant keywords
- Search keywords from related pathology reports
- Other comments _____

Appendix 3 2019 Time Effort Estimation and OMOP Assistance for Phenotyping Survey

Background

The purpose of the survey is to elicit your opinion on identified customization tasks, gain your empirical knowledge of estimating time effort for implementing the tasks, and collect your comments on how OMOP assists phenotype implementation.

Please use one machine for each participant for answering the survey. This way, you will be able to save your answers and return to it later until you submit the survey.

1. What EHR data do you use for phenotyping?

□ Please select all that apply

□ Enterprise Clinical Data Warehouse

□ Clinical data in OMOP Common Data Model

□ Clinical data in I2B2

□ Clinical data in PCORnet

□ Other (please specify) _____

2. What is your role?

_____

3. How many years of experience in developing or implementing electronic phenotypes?

O <= 1 year

O 2 to 5 years

O > 5 years

Time effort estimation for finishing customization task

Please use your best knowledge to estimate how long on average it takes to complete each task, assuming you work on each task continuously and full-time.

*4. Compile machine readable input file containing complete codes/string names/string keywords from the pseudocode

O NA (Not a customization task)

O 1 to 60 minutes

O 1 to 8 hours

O 1-5 days (8 working hours/day)

O 1-2 weeks (5 working days/week)

O 3-4 weeks (5 working days/week)

O 1-2 months

O >=3 months

O Unable to estimate, please explain why

_____

*5. Map provided ICD-9 codes to ICD-10 codes

O NA (Not a customization task)

O 1 to 60 minutes

O 1 to 8 hours

O 1-5 days (8 working hours/day)

O 1-2 weeks (5 working days/week)

O 3-4 weeks (5 working days/week)

O 1-2 months

O >=3 months

O Unable to estimate, please explain why

_____

*6. Map provided ingredient RxCUI to non-ingredient RxCUI

O NA (Not a customization task)

O 1 to 60 minutes

O 1 to 8 hours

O 1-5 days (8 working hours/day)

O 1-2 weeks (5 working days/week)

O 3-4 weeks (5 working days/week)

O 1-2 months

O >=3 months

O Unable to estimate, please explain why

_____

*7. Map provided medication names to RxCUI

O NA (Not a customization task)

O 1 to 60 minutes

O 1 to 8 hours

O 1-5 days (8 working hours/day)

O 1-2 weeks (5 working days/week)

O 3-4 weeks (5 working days/week)

O 1-2 months

O >=3 months

O Unable to estimate, please explain why

_____

*8. Map provided lab names to local codes (e.g., LONIC)

O NA (Not a customization task)

O 1 to 60 minutes

O 1 to 8 hours

O 1-5 days (8 working hours/day)

O 1-2 weeks (5 working days/week)

O 3-4 weeks (5 working days/week)

O 1-2 months

O >=3 months

O Unable to estimate, please explain why

_____

*9. Specify an operational definition of a specific EHR data element for implementation, the data element objectively exists in the EHR, but needs local knowledge for defining it; for example primary care provider, department information linking diagnosis

O NA (Not a customization task)
O 1 to 60 minutes
O 1 to 8 hours
O 1-5 days (8 working hours/day)
O 1-2 weeks (5 working days/week)
O 3-4 weeks (5 working days/week)
O 1-2 months
O >=3 months
O Unable to estimate, please explain why

_____

*10. Specify an operational definition of an event for implementation, the event is defined by using EHR data or other research data, while the event itself is not part of EHR data element; for example continuous enrollment

O NA (Not a customization task)
O 1 to 60 minutes
O 1 to 8 hours
O 1-5 days (8 working hours/day)
O 1-2 weeks (5 working days/week)
O 3-4 weeks (5 working days/week)
O 1-2 months
O >=3 months
O Unable to estimate, please explain why

_____

*11. Find data source for a group of data or one data type, for example pathology report, problem list, lab test report

O NA (Not a customization task)

O 1 to 60 minutes

O 1 to 8 hours

O 1-5 days (8 working hours/day)

O 1-2 weeks (5 working days/week)

O 3-4 weeks (5 working days/week)

O 1-2 months

O >=3 months

O Unable to estimate, please explain why

_____

*12. Explore how a specific data element is presented from structured data, for example lab reference range

O NA (Not a customization task)

O 1 to 60 minutes

O 1 to 8 hours

O 1-5 days (8 working hours/day)

O 1-2 weeks (5 working days/week)

O 3-4 weeks (5 working days/week)

O 1-2 months

O >=3 months

O Unable to estimate, please explain why

_____

*13. Explore how a specific data element is presented from unstructured data, for example keywords for some disease

O NA (Not a customization task)

O 1 to 60 minutes

O 1 to 8 hours

O 1-5 days (8 working hours/day)

O 1-2 weeks (5 working days/week)

O 3-4 weeks (5 working days/week)

O 1-2 months

O >=3 months

O Unable to estimate, please explain why

_____

*14. Understand unstructured data whereas programming and domain knowledge are generally required, this knowledge is not specific to a phenotype; for example identifying relevant section header, or relevant note type.

O NA (Not a customization task)

O 1 to 60 minutes

O 1 to 8 hours

O 1-5 days (8 working hours/day)

O 1-2 weeks (5 working days/week)

O 3-4 weeks (5 working days/week)

O 1-2 months

O >=3 months

O Unable to estimate, please explain why

_____

*15. Pre-process unstructured data for checking the existence of specific note, report, or section

O NA (Not a customization task)

O 1 to 60 minutes

O 1 to 8 hours

O 1-5 days (8 working hours/day)

O 1-2 weeks (5 working days/week)

O 3-4 weeks (5 working days/week)

O 1-2 months

O >=3 months

O Unable to estimate, please explain why

_____

*16. Search unstructured notes or reports with regular expression keywords matching (RegEx)

O NA (Not a customization task)

O 1 to 60 minutes

O 1 to 8 hours

O 1-5 days (8 working hours/day)

O 1-2 weeks (5 working days/week)

O 3-4 weeks (5 working days/week)

O 1-2 months

O >=3 months

O Unable to estimate, please explain why

_____

*17. Search unstructured notes or reports with regular expressions keywords matching with negation (NegEx) or uncertainty, or historical reference detection

O NA (Not a customization task)

O 1 to 60 minutes

O 1 to 8 hours

O 1-5 days (8 working hours/day)

O 1-2 weeks (5 working days/week)

O 3-4 weeks (5 working days/week)

O 1-2 months

O >=3 months

O Unable to estimate, please explain why

_____

*18. Extract clinical information from unstructured notes or reports using advanced NLP implementation (specifically entity and relation extraction and linking information, e.g., extract heart rate from ECG report)

O NA (Not a customization task)
O 1 to 60 minutes
O 1 to 8 hours
O 1-5 days (8 working hours/day)
O 1-2 weeks (5 working days/week)
O 3-4 weeks (5 working days/week)
O 1-2 months
O >=3 months
O Unable to estimate, please explain why

_____

*19. Install and customize an NLP tool for parsing clinical documents (e.g., install cTAKES to parse all clinical notes)

O NA (Not a customization task)
O 1 to 60 minutes
O 1 to 8 hours
O 1-5 days (8 working hours/day)
O 1-2 weeks (5 working days/week)
O 3-4 weeks (5 working days/week)
O 1-2 months
O >=3 months
O Unable to estimate, please explain why

_____

*20. Extract clinical information from an NLP tool parsed clinical documents (e.g., search heart disease concepts (UMLS CUIs) from MedLEE parsed ECG report).

O NA (Not a customization task)

O 1 to 60 minutes

O 1 to 8 hours

O 1-5 days (8 working hours/day)

O 1-2 weeks (5 working days/week)

O 3-4 weeks (5 working days/week)

O 1-2 months

O >=3 months

O Unable to estimate, please explain why

_____

*21. Populate search results from unstructured clinical documents for SQL query use if relevant unstructured data is not stored in the database

O NA (Not a customization task)

O 1 to 60 minutes

O 1 to 8 hours

O 1-5 days (8 working hours/day)

O 1-2 weeks (5 working days/week)

O 3-4 weeks (5 working days/week)

O 1-2 months

O >=3 months

O Unable to estimate, please explain why

_____

*22. Understand phenotype algorithm pseudocode clause

O NA (Not a customization task)

O 1 to 60 minutes

O 1 to 8 hours

O 1-5 days (8 working hours/day)

O 1-2 weeks (5 working days/week)

O 3-4 weeks (5 working days/week)

O 1-2 months

O >=3 months

O Unable to estimate, please explain why

_____

*23. Implement complicated logic in SQL (e.g., implement extrapolating height at serum creatinine measurement time from its pre- and post- height measurement based on a formula)

O NA (Not a customization task)

O 1 to 60 minutes

O 1 to 8 hours

O 1-5 days (8 working hours/day)

O 1-2 weeks (5 working days/week)

O 3-4 weeks (5 working days/week)

O 1-2 months

O >=3 months

O Unable to estimate, please explain why

_____

*24. Check if required data is incomplete or unavailable

O NA (Not a customization task)

O 1 to 60 minutes

O 1 to 8 hours

O 1-5 days (8 working hours/day)

O 1-2 weeks (5 working days/week)

O 3-4 weeks (5 working days/week)

O 1-2 months

O >=3 months

O Unable to estimate, please explain why

_____

*25. Implement another existing phenotype as one rule of a phenotype

O NA (Not a customization task)

O 1 to 60 minutes

O 1 to 8 hours

O 1-5 days (8 working hours/day)

O 1-2 weeks (5 working days/week)

O 3-4 weeks (5 working days/week)

O 1-2 months

O >=3 months

O Unable to estimate, please explain why

_____

*26. In your opinion, what are other tasks that are not identified above? Please elaborate.

O NA (Not a customization task)

O 1 to 60 minutes

O 1 to 8 hours

O 1-5 days (8 working hours/day)

O 1-2 weeks (5 working days/week)

O 3-4 weeks (5 working days/week)

O 1-2 months

O >=3 months

O Unable to estimate, please explain why

_____

Feedback on how OMOP assists phenotype implementation

*27. What are your experiences with OMOP?

Please select all that apply

☐ Have NOT used OMOP at all

☐ Understand OMOP CDM

- Understand OMOP vocabulary
- Have experiences in ETL
- Have populated structured data in OMOP
- Have populated unstructured data in OMOP

*28. Which customization tasks would be facilitated by using the OMOP common data model and sharing OMOP-based SQL implementation? Please select all that apply.

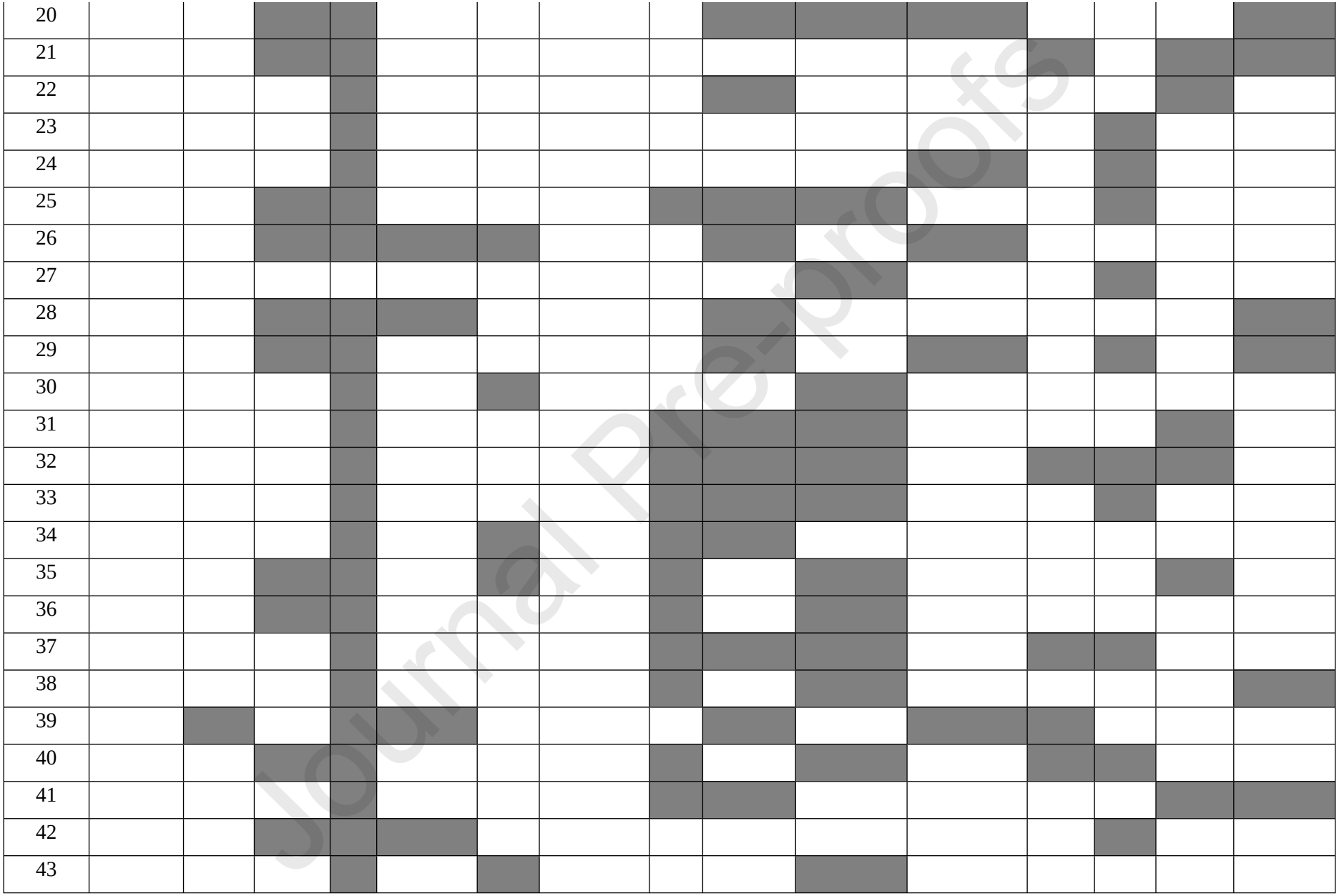
- Compile machine readable input file containing complete codes/string names/string keywords from the pseudocode
- Map provided ICD-9 codes to ICD-10 codes
- Map provided ingredient RxCUI to non-ingredient RxCUI
- Map provided medication names to RxCUI
- Map provided lab names to LONIC codes
- Specify an operational definition of a specific EHR data element for implementation, the data element objectively exists in the EHR, but needs local knowledge for defining it; for example primary care provider, department information linking diagnosis
- Specify an operational definition of an event for implementation, the event is defined by using EHR data or other research data, while the event itself is not part of EHR data element; for example continuous enrollment
- Find data source for a group of data or one data type, for example pathology report, problem list, lab test report
- Explore how a specific data element is presented from structured data, for example lab reference range
- Explore how a specific data element is presented from unstructured data, for example keywords for some disease
- Understand unstructured data whereas programming and domain knowledge are generally required, this knowledge is not specific to a phenotype; for example identifying relevant section header, or relevant note type
- Pre-process unstructured data for checking the existence of specific note, report, or section
- Search unstructured notes or reports with regular expression keywords matching (RegEx)

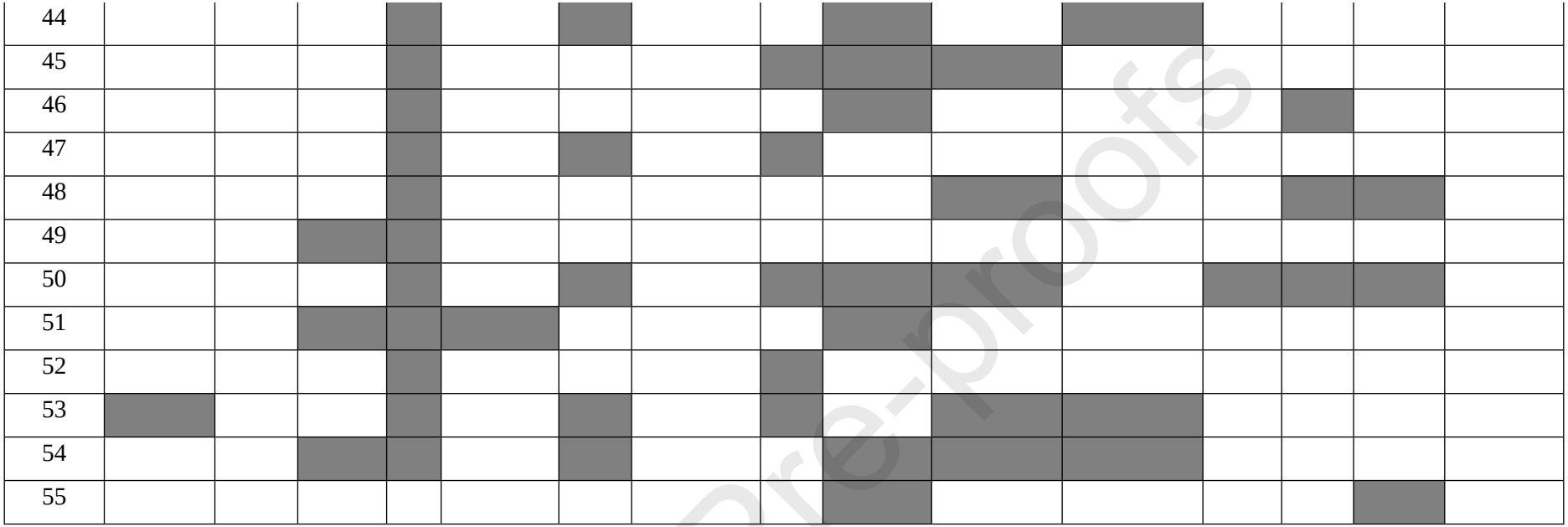
- Search unstructured notes or reports with regular expressions keywords matching with negation (NegEx) or uncertainty, or historical reference detection
- Extract clinical information from unstructured notes or reports using advanced NLP implementation (specifically entity and relation extraction and linking information)
- Install and customize an NLP tool for parsing clinical documents
- Extract clinical information from an NLP tool parsed clinical documents
- Populate search results from unstructured clinical documents for SQL query use if relevant unstructured data is not stored in the database
- Understand phenotype algorithm pseudocode clause
- Implement complicated logic in SQL
- Check if required data is incomplete or unavailable
- Implement another existing phenotype as one rule of a phenotype
- Other tasks or any comments (please specify)

_____

Appendix 4 Different data types used in the phenotype algorithms with source citation provided

Phenotype	Allergy	Death	Demo	Dx	Encounter	Visit	Family Hx	Lab	Procedure	Rx / Medical device	Provider / Specialty / Department	Problem list	Note	Report	non-EHR*
1				■	■	■	■	■		■	■				
2			■	■				■	■						
3			■	■		■		■	■	■			■		■
4													■		
5			■	■						■				■	■
6			■	■		■		■			■	■		■	■
7				■	■			■	■					■	■
8			■	■		■			■	■					■
9			■	■	■				■	■	■			■	■
10			■	■		■		■	■	■		■			
11				■				■	■	■		■			
12				■						■					
13			■	■					■		■				■
14			■	■		■		■	■	■			■		
15				■					■						
16			■	■		■				■					
17			■	■		■				■					
18			■	■		■				■					
19			■	■					■	■	■				■





*e.g., cancer registry, cohort meta information

1: 18_T2DM; Type 2 Diabetes Mellitus phenotype algorithm. https://phekb.org/phenotype/type-2-diabetes-mellitus (accessed 2019 February 7).

2: 587_CKD; Chronic Kidney Disease phenotype algorithm. https://phekb.org/phenotype/chronic-kidney-disease (accessed 2019 February 7).

3: 70_Cdiff; Clostridium Difficile Colitis phenotype algorithm. https://phekb.org/phenotype/clostridium-difficile-colitis (accessed 2019 February 7).

4: 729_PAD; Peripheral artery disease (PAD) -NLP phenotype algorithm. https://phekb.org/phenotype/pad-nlp-2017 (accessed 2019 February 7).

5: 926_pneumonia; Pneumonia phenotype algorithm. https://phekb.org/phenotype/pneumonia-vumc-emerge-v51 (accessed 2019 February 7).

6: 149_CaMRSA; Community associated MRSA (Methicillin-resistant Staphylococcus aureus) phenotype algorithm. https://phekb.org/phenotype/camrsa (accessed 2019 February 7).

7: 188_CAAD; CAAD (Carotid Artery Atherosclerosis Disease) phenotype algorithm. https://phekb.org/phenotype/caad-carotid-artery-atherosclerosis-disease (accessed 2019 February 7).

8: 216_BPH; Benign Prostatic Hyperplasia (BPH) phenotype algorithm. https://phekb.org/phenotype/benign-prostatic-hyperplasia-bph (accessed 2019 February 7).

9: 236_Appendicitis; Appendicitis phenotype algorithm. https://phekb.org/phenotype/appendicitis (accessed 2019 February 7).

10: 121_ExtremeObesity; Extreme Obesity phenotype algorithm. https://phekb.org/phenotype/extreme-obesity (accessed 2019 February 7).

11: 170_StatinMACE; Statins and MACE phenotype algorithm. https://phekb.org/phenotype/statins-and-mace (accessed 2019 February 7).

12: 146_Asthma; Asthma phenotype algorithm. (accessed 2019 February 7).

13: 172_AMD; phenotype algorithm. https://phekb.org/phenotype/asthma (accessed 2019 February 7).

14: 120_DM/HtnCKD; Diabetes/Hypertension-Associated Chronic Kidney Disease phenotype algorithm. https://phekb.org/phenotype/diabeteshypertension-associated-chronic-kidney-disease (accessed 2019 February 7).

15: 125_CardiorespiratoryFitness; Cardiorespiratory Fitness phenotype algorithm. https://phekb.org/phenotype/cardiorespiratory-fitness-algorithm-emerge-mayo-network-phenotype (accessed 2019 February 7).

16: 179_ADHD; ADHD phenotype algorithm. https://phekb.org/phenotype/adhd-phenotype-algorithm (accessed 2019 February 7).

17: 184_AtopicDermatitis; Atopic Dermatitis phenotype algorithm. https://phekb.org/phenotype/atopic-dermatitis-algorithm (accessed 2019 February 7).

18: 224_GERD; Gastroesophageal Reflux Disease (GERD) phenotype algorithm. https://phekb.org/phenotype/gastroesophageal-reflux-disease-gerd-phenotype-algorithm (accessed 2019 February 7).

19: 100_Glaucoma; Glaucoma phenotype algorithm. https://phekb.org/phenotype/glaucoma (accessed 2019 February 7).

20: 99_OcularHtn; Ocular Hypertension phenotype algorithm. (accessed 2019 February 7).

21: 147_HeartFailure; phenotype algorithm. https://phekb.org/phenotype/ocular-hypertension (accessed 2019 February 7).

22: 92_Diverticulosis; Diverticulosis and Diverticulitis phenotype algorithm. https://phekb.org/phenotype/diverticulosis-and-diverticulitis (accessed 2019 February 7).

23: 109_VTE; Venous Thromboembolism (VTE) phenotype algorithm. https://phekb.org/phenotype/venous-thromboembolism-vte (accessed 2019 February 7).

24: 162_Autism; Autism phenotype algorithm. https://phekb.org/phenotype/autism (accessed 2019 February 7).

25: 161_EarlyChildhoodObesity; Severe Early Childhood Obesity phenotype algorithm. https://phekb.org/phenotype/severe-early-childhood-obesity (accessed 2019 February 7).

26: 97_AAA; Abdominal Aortic Aneurysm (AAA) phenotype algorithm. https://phekb.org/phenotype/abdominal-aortic-aneurysm-aaa (accessed 2019 February 7).

27: 90_ACEIcough; ACE Inhibitor (ACE-I) induced cough phenotype algorithm. https://phekb.org/phenotype/ace-inhibitor-ace-i-induced-cough (accessed 2019 February 7).

28: 112_Zoster; Herpes Zoster phenotype algorithm. https://phekb.org/phenotype/herpes-zoster (accessed 2019 February 7).

29: 9_Cataracts; Cataracts phenotype algorithm. https://phekb.org/phenotype/cataracts (accessed 2019 February 7).

30: 10_Dementia; Dementia phenotype algorithm. https://phekb.org/phenotype/dementia (accessed 2019 February 7).

31: 16_PAD; Peripheral Arterial Disease phenotype algorithm. https://phekb.org/phenotype/peripheral-arterial-disease-2012 (accessed 2019 February 7).

32: 8_QRS; Cardiac Conduction (QRS) phenotype algorithm. https://phekb.org/phenotype/cardiac-conduction-qrs (accessed 2019 February 7).

33: 17_RBC; Red Blood Cell Indices phenotype algorithm. https://phekb.org/phenotype/red-blood-cell-indices (accessed 2019 February 7).

34: 19_WBC; White Blood Cell Indices phenotype algorithm. https://phekb.org/phenotype/white-blood-cell-indices (accessed 2019 February 7).

35: 68_ResHTN; Resistant hypertension phenotype algorithm. https://phekb.org/phenotype/resistant-hypertension (accessed 2019 February 7).

36: 13_Height; Height phenotype algorithm. https://phekb.org/phenotype/height (accessed 2019 February 7).

37: 14_Hypothyroidism; Hypothyroidism phenotype algorithm. https://phekb.org/phenotype/hypothyroidism (accessed 2019 February 7).

38: 15_Lipids; Lipids phenotype algorithm. https://phekb.org/phenotype/lipids (accessed 2019 February 7).

39: 11_DiabeticRetinopathy; Diabetic Retinopathy phenotype algorithm. https://phekb.org/phenotype/diabetic-retinopathy (accessed 2019 February 7).

40: 602_FH; Electronic Health Record-based Phenotyping Algorithm for Familial Hypercholesterolemia. https://phekb.org/phenotype/electronic-health-record-based-phenotyping-algorithm-familial-hypercholesterolemia (accessed 2019 February 7).

41: 514_CRC; Colorectal Cancer (CRC) phenotype algorithm. https://phekb.org/phenotype/colorectal-cancer-crc (accessed 2019 February 7).

42: 584_Migraine; Migraine phenotype algorithm. https://phekb.org/phenotype/migraine (accessed 2019 February 7).

43: 487_Epilepsy; Epilepsy/Antiepileptic drug response algorithm. https://phekb.org/phenotype/epilepsyantiepileptic-drug-response-algorithm (accessed 2019 February 7).

44: 582_ChronicRhinosinusitis; CRS (Chronic Rhinosinusitis) phenotype algorithm. https://phekb.org/phenotype/crs-chronic-rhinosinusitis (accessed 2019 February 7).

45: 528_CIN; Contrast Induced Nephropathy phenotype algorithm. https://phekb.org/phenotype/contrast-induced-nephropathy (accessed 2019 February 7).

46: 656_HearingLoss; Hearing Loss phenotype algorithm. https://phekb.org/phenotype/hearing-loss (accessed 2019 February 7).

47: 585_RheumatoidArthritis; Rheumatoid Arthritis (RA) phenotype algorithm. https://phekb.org/phenotype/rheumatoid-arthritis-ra (accessed 2019 February 7).

48: 588_NALFD; Non-alcoholic fatty liver disease (NALFD) & Alcoholic Fatty Liver Disease (ALD) phenotype algorithm. https://phekb.org/phenotype/non-alcoholic-fatty-liver-disease-nalfd-alcoholic-fatty-liver-disease-ald (accessed 2019 February 7).

49: 915_IntellectualDisability; Intellectual Disability phenotype algorithm. https://phekb.org/phenotype/intellectual-disability (accessed 2019 February 7).

50: 772_ArrhythmiasEKGintervals; (Arrhythmias) EKG Intervals phenotype algorithm. https://phekb.org/phenotype/arrhythmias-ekg-intervals-phenotypegenotype-correlations (accessed 2019 February 7).

51: 884_OvUtCa; Ovarian/Uterine Cancer (OvUtCa) phenotype algorithm. https://phekb.org/phenotype/ovarianuterine-cancer-ovutca (accessed 2019 February 7).

52: 755_Autoimmunity; Autoimmune Disease phenotype algorithm. https://phekb.org/phenotype/autoimmune-disease-phenotype (accessed 2019 February 7).

53: 519_MetforminResponse; https://phekb.org/phenotype/metformin-response phenotype algorithm. https://phekb.org/phenotype/metformin-response (accessed 2019 February 7).

54: 1105_Anxiety; Anxiety phenotype algorithm. https://phekb.org/phenotype/anxiety-algorithm (accessed 2019 February 7).

55: 93_ColonPolyp; Colon Polyps phenotype algorithm. https://phekb.org/phenotype/colon-polyps (accessed 2019 February 7).