Converting Optum EHR Oncology module into OMOP CDM

ETL logic and concepts mapping overview

INTRO

- The Optum® Enriched Oncology Data set is a group of tables that can supplement the Optum® de-identified Electronic Health Record dataset.
- It includes specific oncology concepts important for understanding the progression of the disease, which often not available in structured formats, particularly the tumor, node, and metastasis (TNM) values, stage information and biomarkers.
- It is obtained from patient records using NLP methods
- As of 2022, there are approximately 1.9 million patients with at least one solid tumor ICD-9 or ICD-10 diagnosis included in the data set.

Overall logic

source entry	target domain	target vocabulary
Histology	Condition	SNOMED, ICDO3
Topography	Condition	SNOMED, ICDO3
Laterality	Condition	SNOMED, ICDO3
Behavior (in situ, malignant or benign)	Condition	SNOMED, ICDO3
summary stage	Measurement	Cancer Modifier
metastasis location	Measurement	Cancer Modifier
TNM	Measurement	Cancer Modifier
tumor grade	Measurement	Cancer Modifier
characteristics: advanced, carcinomatosis, extensive, infiltrative, invasive, localized, etc	Measurement	Cancer Modifier
Biomarkers	Measurement	OMOP genomic, LOINC, SNOMED
evaluation system: Binet Stage, Durie/Salmon Stage, ECOG performance status, FIGO Stage, Gleason, Gleason score	Measurement	Cancer Modifier
Tumor size	Measurement	Cancer Modifier
treatment regimen	Episode*	HemOnc
Tumor progression	Episode*	Episode
treatment response	Observation	no mapping - not supported by the vocabulary model

Precoordinated into a single concept

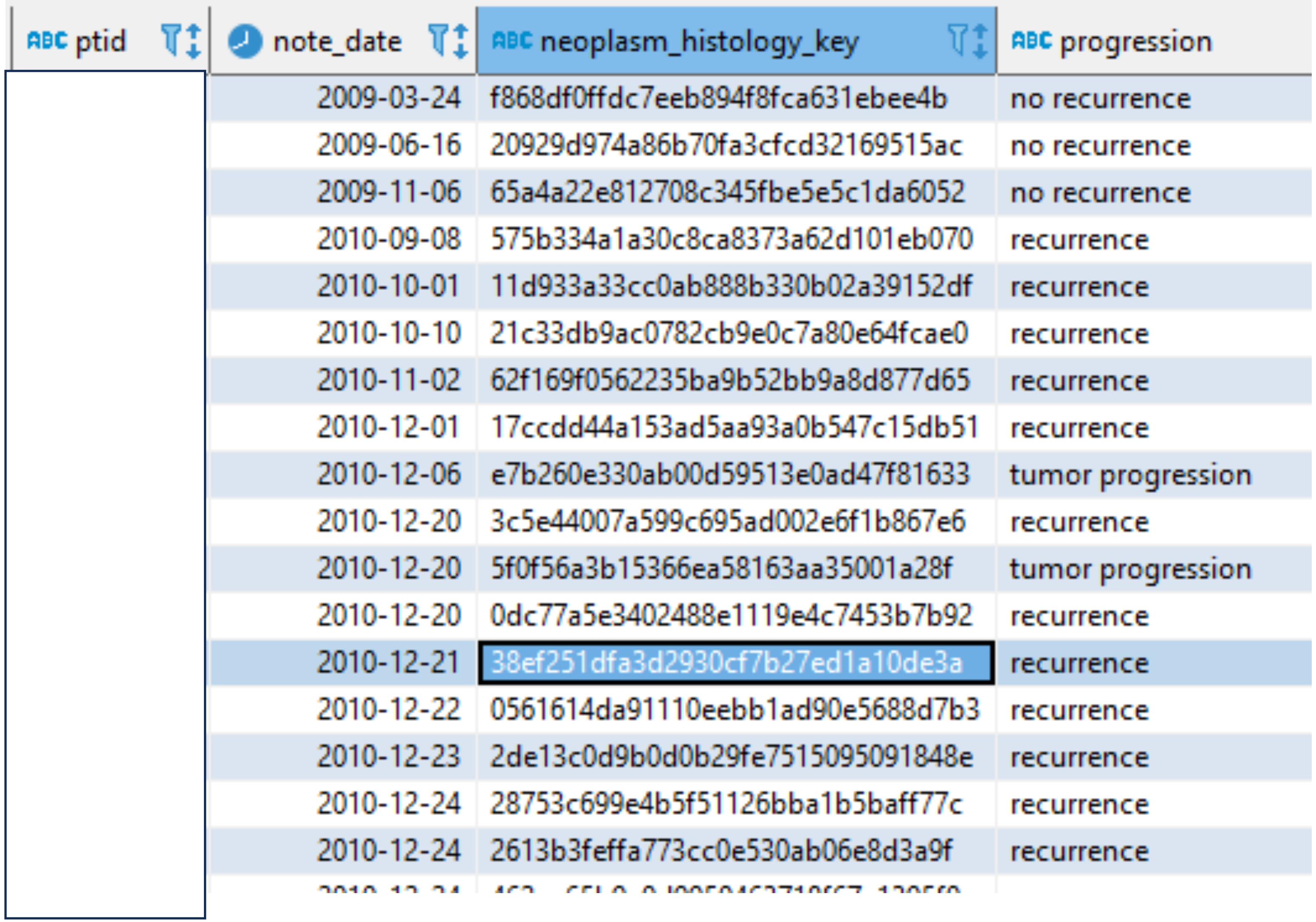
Expression level of immunostaining (“0”, “1+”, “2+”, “3+”, 100%, 90%) mapped to ‘positive’, ‘negative’, ‘equivocal’

Calculated the largest size and mapped to “*Largest Dimension of Tumor*”, *others to “Dimension of Tumor”*

*will be mapped in the next data refresh

Tumor progression to Episode

Source data example of a single patient



ptid	note_date	neoplasm_histology_key	progression
	2009-03-24	f868df0ffdc7eeb894f8fca631ebee4b	no recurrence
	2009-06-16	20929d974a86b70fa3cfcd32169515ac	no recurrence
	2009-11-06	65a4a22e812708c345fbe5e5c1da6052	no recurrence
	2010-09-08	575b334a1a30c8ca8373a62d101eb070	recurrence
	2010-10-01	11d933a33cc0ab888b330b02a39152df	recurrence
	2010-10-10	21c33db9ac0782cb9e0c7a80e64fcae0	recurrence
	2010-11-02	62f169f0562235ba9b52bb9a8d877d65	recurrence
	2010-12-01	17ccdd44a153ad5aa93a0b547c15db51	recurrence
	2010-12-06	e7b260e330ab00d59513e0ad47f81633	tumor progression
	2010-12-20	3c5e44007a599c695ad002e6f1b867e6	recurrence
	2010-12-20	5f0f56a3b15366ea58163aa35001a28f	tumor progression
	2010-12-20	0dc77a5e3402488e1119e4c7453b7b92	recurrence
	2010-12-21	38ef251dfa3d2930cf7b27ed1a10de3a	recurrence
	2010-12-22	0561614da91110eebb1ad90e5688d7b3	recurrence
	2010-12-23	2de13c0d9b0d0b29fe7515095091848e	recurrence
	2010-12-24	28753c699e4b5f51126bba1b5baff77c	recurrence
	2010-12-24	2613b3feffa773cc0e530ab06e8d3a9f	recurrence
	2010-12-24	[illegible]	

Remission Episode
2009-03-24 – 2010-09-08

Disease Recurrence Episode

Progression Episode

Disease Recurrence Episode

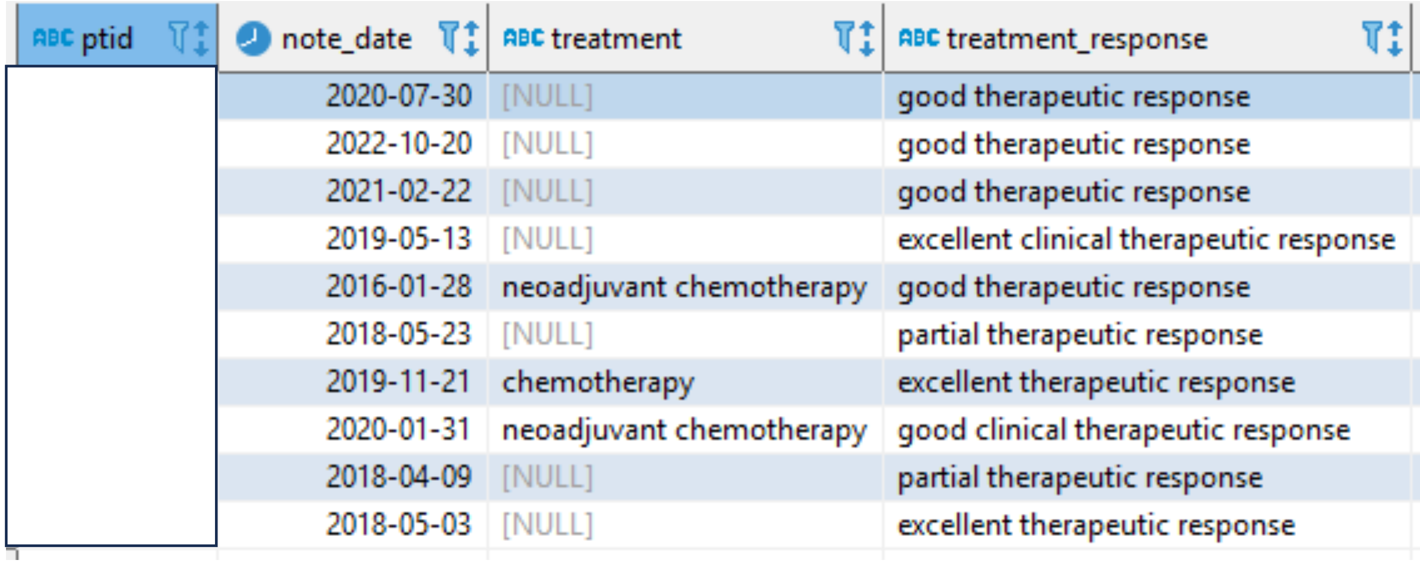
Data elements that can’t be mapped. Treatment response

Treatment response terms:

- good therapeutic response
- excellent therapeutic response
- complete therapeutic response
- partial therapeutic response
- complete pathologic therapeutic response
- very good partial response
- minimal residual disease response
- good clinical therapeutic response
- excellent clinical therapeutic response
- fair therapeutic response

Example of the data

Treatment response in different patients



ptid	note_date	treatment	treatment_response
	2020-07-30	[NULL]	good therapeutic response
	2022-10-20	[NULL]	good therapeutic response
	2021-02-22	[NULL]	good therapeutic response
	2019-05-13	[NULL]	excellent clinical therapeutic response
	2016-01-28	neoadjuvant chemotherapy	good therapeutic response
	2018-05-23	[NULL]	partial therapeutic response
	2019-11-21	chemotherapy	excellent therapeutic response
	2020-01-31	neoadjuvant chemotherapy	good clinical therapeutic response
	2018-04-09	[NULL]	partial therapeutic response
	2018-05-03	[NULL]	excellent therapeutic response

Data cleansing

Data entries were removed where exist:

- “in situ” and “invasive” at the same day.
- inconsistent numeric and narrative biomarkers results, for example

numeric result = “+1”, narrative result = “positive mutation” in ERBB2/HER2 measurement.*

- *More such rules to be applied:*
  - E.g. Positive and negative biomarker status in the same patient
- Event tables were deduped if at the same date there was the same information
  - condition_source_value in Conditions,
  - combination of measurement_source_value, value_as_number, value_source_value in Measurement.

*****

**A score of "1+" suggests that there is a low level of HER2 protein present in the cells. This low level is considered within the normal range, and so the cancer is unlikely to respond to therapies that target HER2. Therefore, a "1+" score is usually interpreted as a negative result for HER2 overexpression.*

Concept mapping

source_Name	target_concept_name
erb-b2 receptor tyrosine kinase 2 (ERBB2 or HER2/neu)	ERBB2 (erb-b2 receptor tyrosine kinase 2) gene variant measurement
estrogen receptor/progesterone receptor (ER/PGR)	ESR1 Protein Expression measurement
estrogen receptor/progesterone receptor (ER/PGR)	PGR (progesterone receptor) gene variant measurement
marker of proliferation Ki-67 (MKI67 or Ki-67)	MKI67 (marker of proliferation Ki-67) gene variant measurement
CD274 molecule (CD274 or PD-L1 or PDL1)	CD274 (CD274 molecule) gene variant measurement
adenocarcinoma	Malignant adenomatous neoplasm
carcinoma	Malignant epithelial neoplasm
squamous cell carcinoma	Squamous cell carcinoma
basal cell carcinoma	Neoplasm defined only by histology: Basal cell carcinoma, NOS
in situ ductal carcinoma	Intraductal carcinoma in situ of breast
lung non-small cell carcinoma	Non-small cell lung cancer
multiple myeloma	Multiple myeloma
malignant mammary neoplasm	Malignant tumor of breast
prostatic adenocarcinoma	Adenocarcinoma of prostate
lung adenocarcinoma	Adenocarcinoma of lung

Biomarkers were mapped mostly to the OMOP Genomic vocabulary, Generic Variation concept class.
77% distinct concepts are mapped to OMOP Genomic, 19% to SNOMED or LOINC,
4% are not mapped, but those have low frequency.

Conditions are mapped well with histology information included, but sometimes it's only histology (in yellow), so you need to define the topography and histology separately when phenotyping.

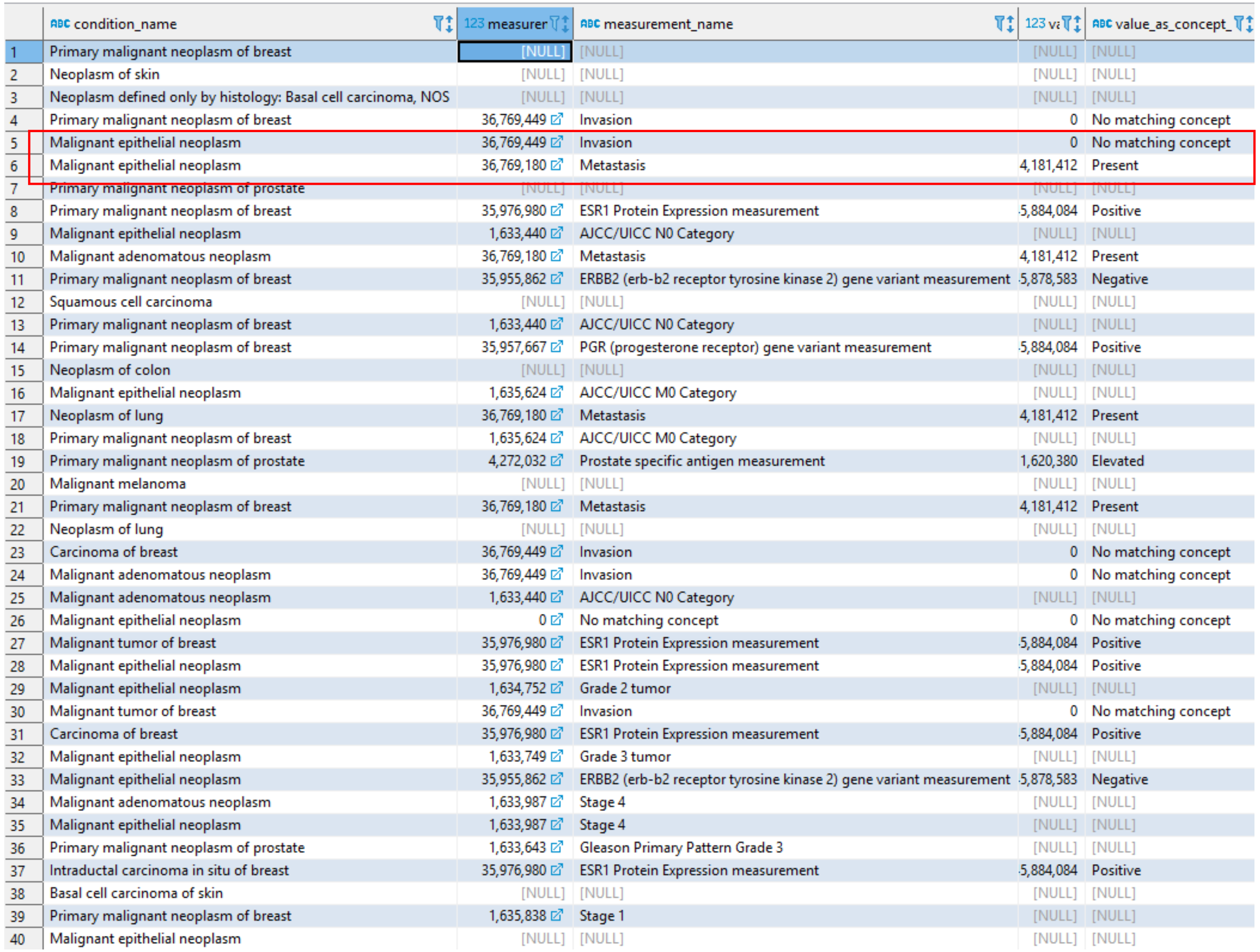
Cancer characteristics that can’t be mapped

source term	comment
locally advanced	
not metastatic	in theory can be mapped to metastasis+absent, but I afraid people will not use it, + our tools such as CD, doesn't look at values. Is there a concept for ‘non-metastatic’ – localized or something?
not invasive	
not in situ	
advanced	
localized	
carcinomatosis	there's such Condition, should be measurement
not malignant	
oligometastatic	
multicentric	
extensive	

Data evaluation

- 1) Conditions and measurements connected grouped
- 2) Create a Cancer cohort and evaluate the distribution of cancer modifiers
- 3) Look for impossible combination of events

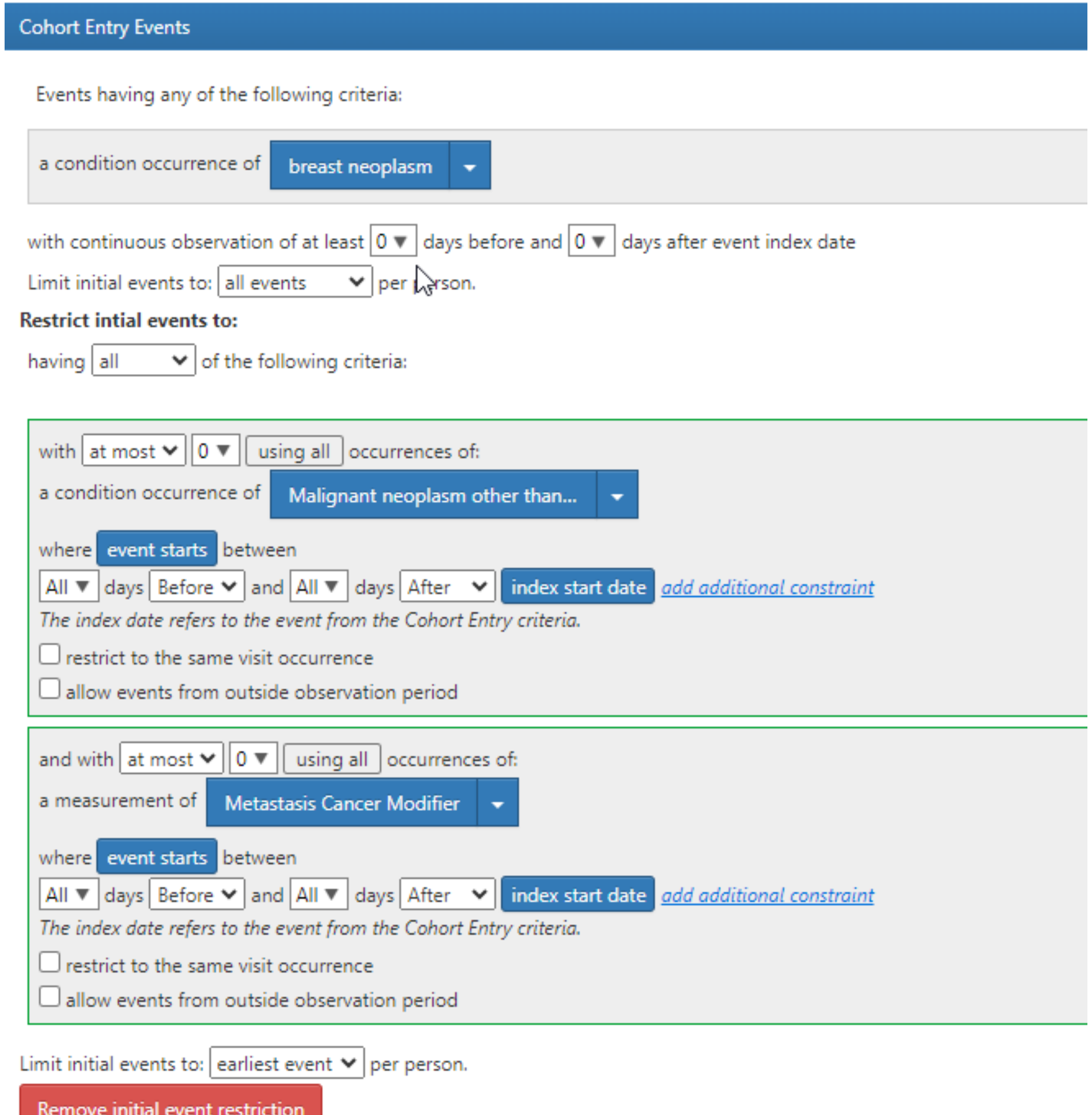
Top 40 condition-measurement combinations defined using MEASUREMENT modifiers



	condition_name	measurer	measurement_name	va	value_as_concept_
1	Primary malignant neoplasm of breast	[NULL]	[NULL]	[NULL]	[NULL]
2	Neoplasm of skin	[NULL]	[NULL]	[NULL]	[NULL]
3	Neoplasm defined only by histology: Basal cell carcinoma, NOS	[NULL]	[NULL]	[NULL]	[NULL]
4	Primary malignant neoplasm of breast	36,769,449	Invasion	0	No matching concept
5	Malignant epithelial neoplasm	36,769,449	Invasion	0	No matching concept
6	Malignant epithelial neoplasm	36,769,180	Metastasis	4,181,412	Present
7	Primary malignant neoplasm of prostate	[NULL]	[NULL]	[NULL]	[NULL]
8	Primary malignant neoplasm of breast	35,976,980	ESR1 Protein Expression measurement	5,884,084	Positive
9	Malignant epithelial neoplasm	1,633,440	AJCC/UICC N0 Category	[NULL]	[NULL]
10	Malignant adenomatous neoplasm	36,769,180	Metastasis	4,181,412	Present
11	Primary malignant neoplasm of breast	35,955,862	ERBB2 (erb-b2 receptor tyrosine kinase 2) gene variant measurement	5,878,583	Negative
12	Squamous cell carcinoma	[NULL]	[NULL]	[NULL]	[NULL]
13	Primary malignant neoplasm of breast	1,633,440	AJCC/UICC N0 Category	[NULL]	[NULL]
14	Primary malignant neoplasm of breast	35,957,667	PGR (progesterone receptor) gene variant measurement	5,884,084	Positive
15	Neoplasm of colon	[NULL]	[NULL]	[NULL]	[NULL]
16	Malignant epithelial neoplasm	1,635,624	AJCC/UICC M0 Category	[NULL]	[NULL]
17	Neoplasm of lung	36,769,180	Metastasis	4,181,412	Present
18	Primary malignant neoplasm of breast	1,635,624	AJCC/UICC M0 Category	[NULL]	[NULL]
19	Primary malignant neoplasm of prostate	4,272,032	Prostate specific antigen measurement	1,620,380	Elevated
20	Malignant melanoma	[NULL]	[NULL]	[NULL]	[NULL]
21	Primary malignant neoplasm of breast	36,769,180	Metastasis	4,181,412	Present
22	Neoplasm of lung	[NULL]	[NULL]	[NULL]	[NULL]
23	Carcinoma of breast	36,769,449	Invasion	0	No matching concept
24	Malignant adenomatous neoplasm	36,769,449	Invasion	0	No matching concept
25	Malignant adenomatous neoplasm	1,633,440	AJCC/UICC N0 Category	[NULL]	[NULL]
26	Malignant epithelial neoplasm	0	No matching concept	0	No matching concept
27	Malignant tumor of breast	35,976,980	ESR1 Protein Expression measurement	5,884,084	Positive
28	Malignant epithelial neoplasm	35,976,980	ESR1 Protein Expression measurement	5,884,084	Positive
29	Malignant epithelial neoplasm	1,634,752	Grade 2 tumor	[NULL]	[NULL]
30	Malignant tumor of breast	36,769,449	Invasion	0	No matching concept
31	Carcinoma of breast	35,976,980	ESR1 Protein Expression measurement	5,884,084	Positive
32	Malignant epithelial neoplasm	1,633,749	Grade 3 tumor	[NULL]	[NULL]
33	Malignant epithelial neoplasm	35,955,862	ERBB2 (erb-b2 receptor tyrosine kinase 2) gene variant measurement	5,878,583	Negative
34	Malignant adenomatous neoplasm	1,633,987	Stage 4	[NULL]	[NULL]
35	Malignant epithelial neoplasm	1,633,987	Stage 4	[NULL]	[NULL]
36	Primary malignant neoplasm of prostate	1,633,643	Gleason Primary Pattern Grade 3	[NULL]	[NULL]
37	Intraductal carcinoma in situ of breast	35,976,980	ESR1 Protein Expression measurement	5,884,084	Positive
38	Basal cell carcinoma of skin	[NULL]	[NULL]	[NULL]	[NULL]
39	Primary malignant neoplasm of breast	1,635,838	Stage 1	[NULL]	[NULL]
40	Malignant epithelial neoplasm	[NULL]	[NULL]	[NULL]	[NULL]

No topography

Cohort definition: Neoplasm of breast excluding other neoplasms, cancer modifiers as inclusion criteria



Cohort Entry Events

Events having any of the following criteria:

a condition occurrence of breast neoplasm

with continuous observation of at least 0 days before and 0 days after event index date

Limit initial events to: all events per person.

Restrict intial events to:

having all of the following criteria:

with at most 0 using all occurrences of:

a condition occurrence of Malignant neoplasm other than...

where event starts between

All days Before and All days After index start date add additional constraint

The index date refers to the event from the Cohort Entry criteria.

□ restrict to the same visit occurrence

□ allow events from outside observation period

and with at most 0 using all occurrences of:

a measurement of Metastasis Cancer Modifier

where event starts between

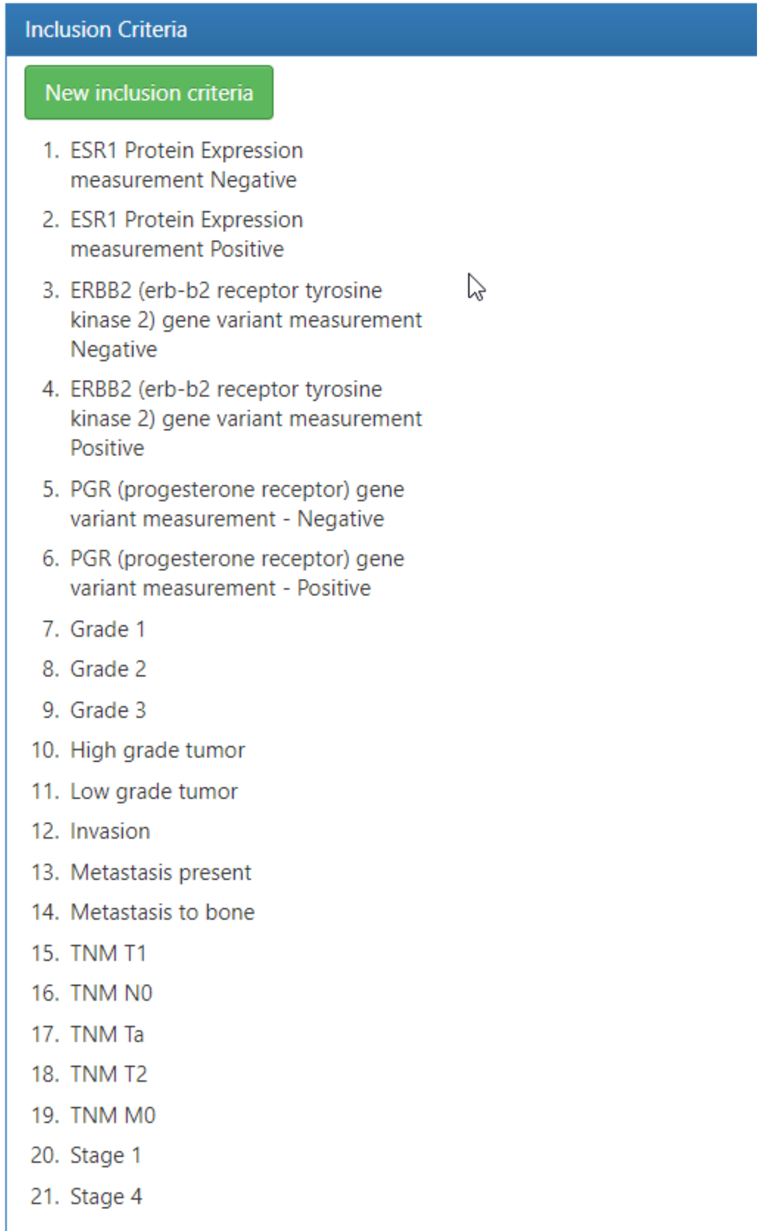
All days Before and All days After index start date add additional constraint

The index date refers to the event from the Cohort Entry criteria.

□ restrict to the same visit occurrence

□ allow events from outside observation period

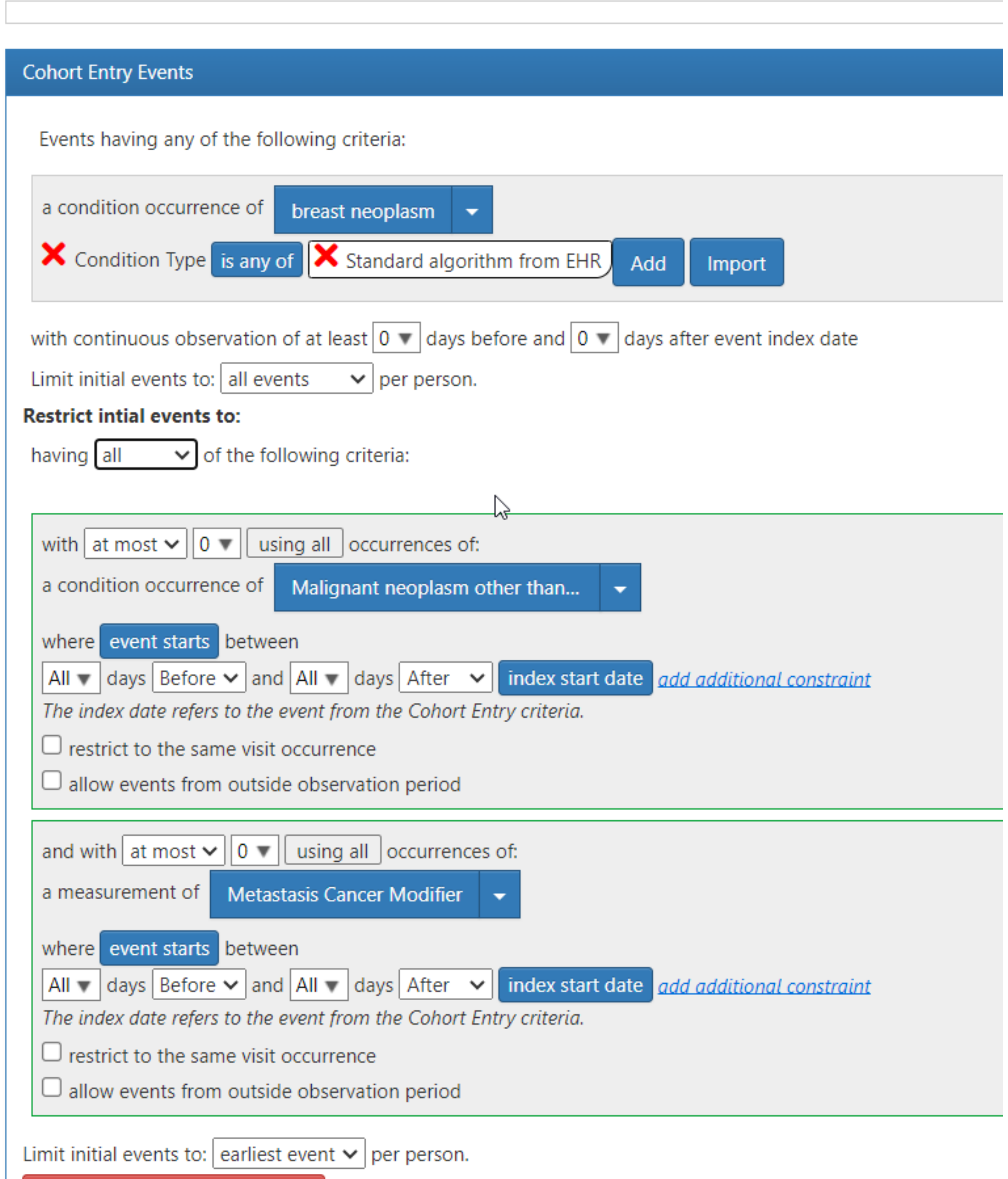
Limit initial events to: earliest event per person.



Inclusion Criteria

1. ESR1 Protein Expression measurement Negative
2. ESR1 Protein Expression measurement Positive
3. ERBB2 (erb-b2 receptor tyrosine kinase 2) gene variant measurement Negative
4. ERBB2 (erb-b2 receptor tyrosine kinase 2) gene variant measurement Positive
5. PGR (progesterone receptor) gene variant measurement - Negative
6. PGR (progesterone receptor) gene variant measurement - Positive
7. Grade 1
8. Grade 2
9. Grade 3
10. High grade tumor
11. Low grade tumor
12. Invasion
13. Metastasis present
14. Metastasis to bone
15. TNM T1
16. TNM N0
17. TNM Ta
18. TNM T2
19. TNM M0
20. Stage 1
21. Stage 4

Cohort definition: Neoplasm of breast excluding other neoplasms, cancer modifiers as inclusion criteria. Specific condition type



Cohort Entry Events

Events having any of the following criteria:

a condition occurrence of breast neoplasm

Condition Type is any of Standard algorithm from EHR

with continuous observation of at least 0 days before and 0 days after event index date

Limit initial events to: all events per person.

Restrict intial events to:

having all of the following criteria:

with at most 0 using all occurrences of:

a condition occurrence of Malignant neoplasm other than...

where event starts between

All days Before and All days After index start date *add additional constraint*

The index date refers to the event from the Cohort Entry criteria.

□ restrict to the same visit occurrence

□ allow events from outside observation period

and with at most 0 using all occurrences of:

a measurement of Metastasis Cancer Modifier

where event starts between

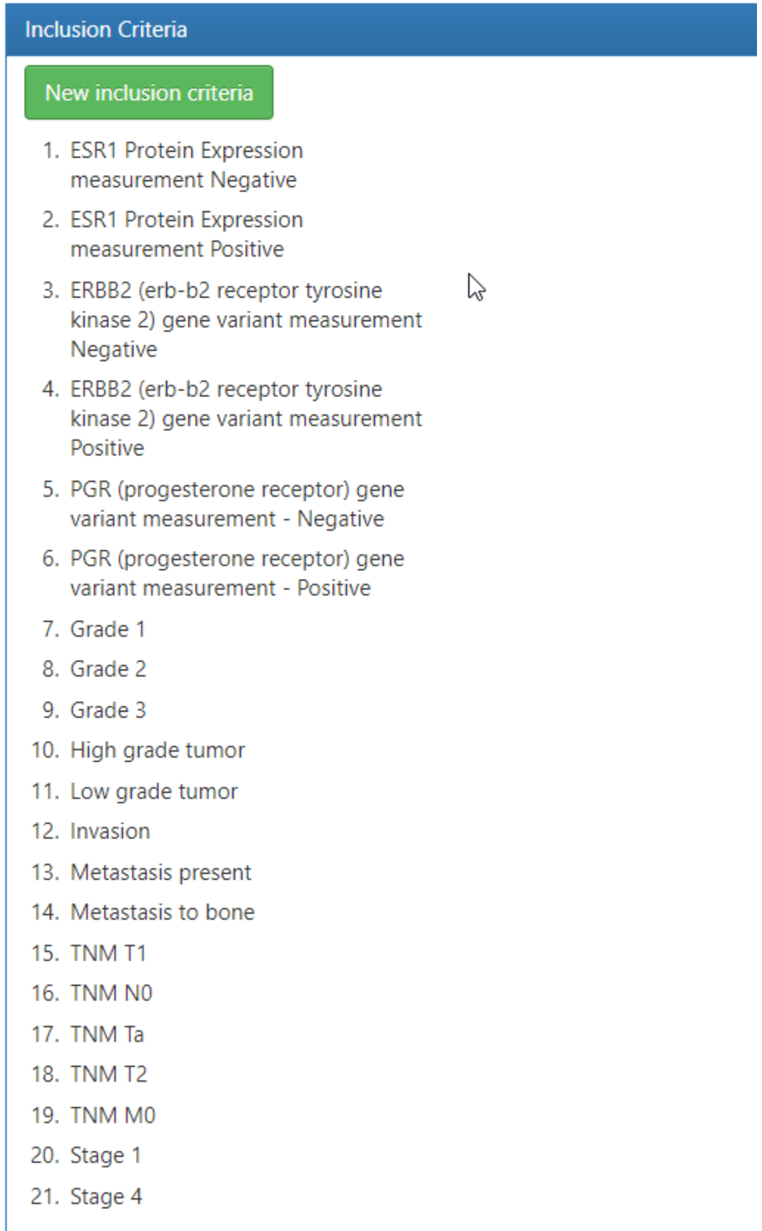
All days Before and All days After index start date *add additional constraint*

The index date refers to the event from the Cohort Entry criteria.

□ restrict to the same visit occurrence

□ allow events from outside observation period

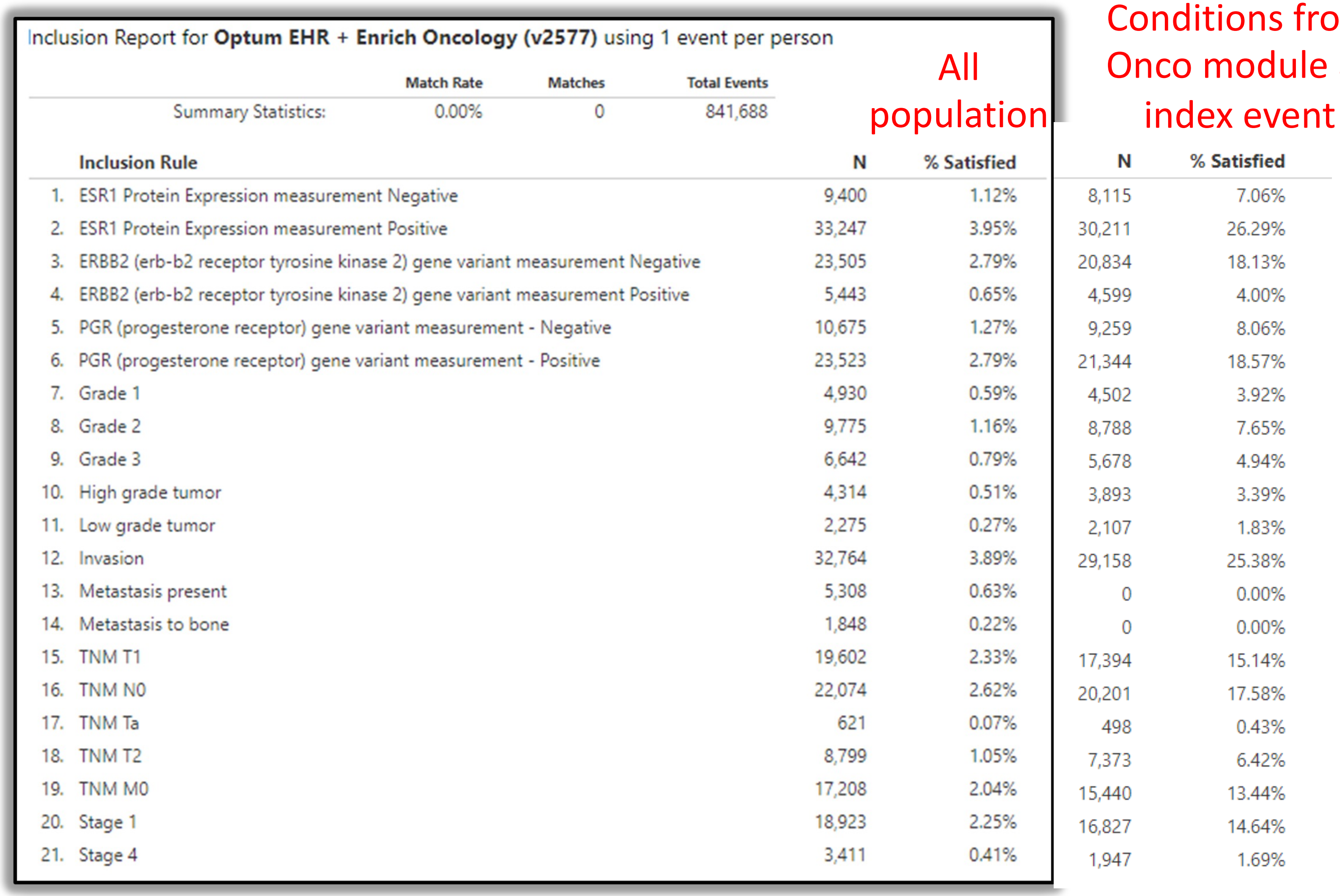
Limit initial events to: earliest event per person.



Inclusion Criteria

1. ESR1 Protein Expression measurement Negative
2. ESR1 Protein Expression measurement Positive
3. ERBB2 (erb-b2 receptor tyrosine kinase 2) gene variant measurement Negative
4. ERBB2 (erb-b2 receptor tyrosine kinase 2) gene variant measurement Positive
5. PGR (progesterone receptor) gene variant measurement - Negative
6. PGR (progesterone receptor) gene variant measurement - Positive
7. Grade 1
8. Grade 2
9. Grade 3
10. High grade tumor
11. Low grade tumor
12. Invasion
13. Metastasis present
14. Metastasis to bone
15. TNM T1
16. TNM N0
17. TNM Ta
18. TNM T2
19. TNM M0
20. Stage 1
21. Stage 4

Patients with cancer modifiers



Inclusion Report for **Optum EHR + Enrich Oncology (v2577)** using 1 event per person

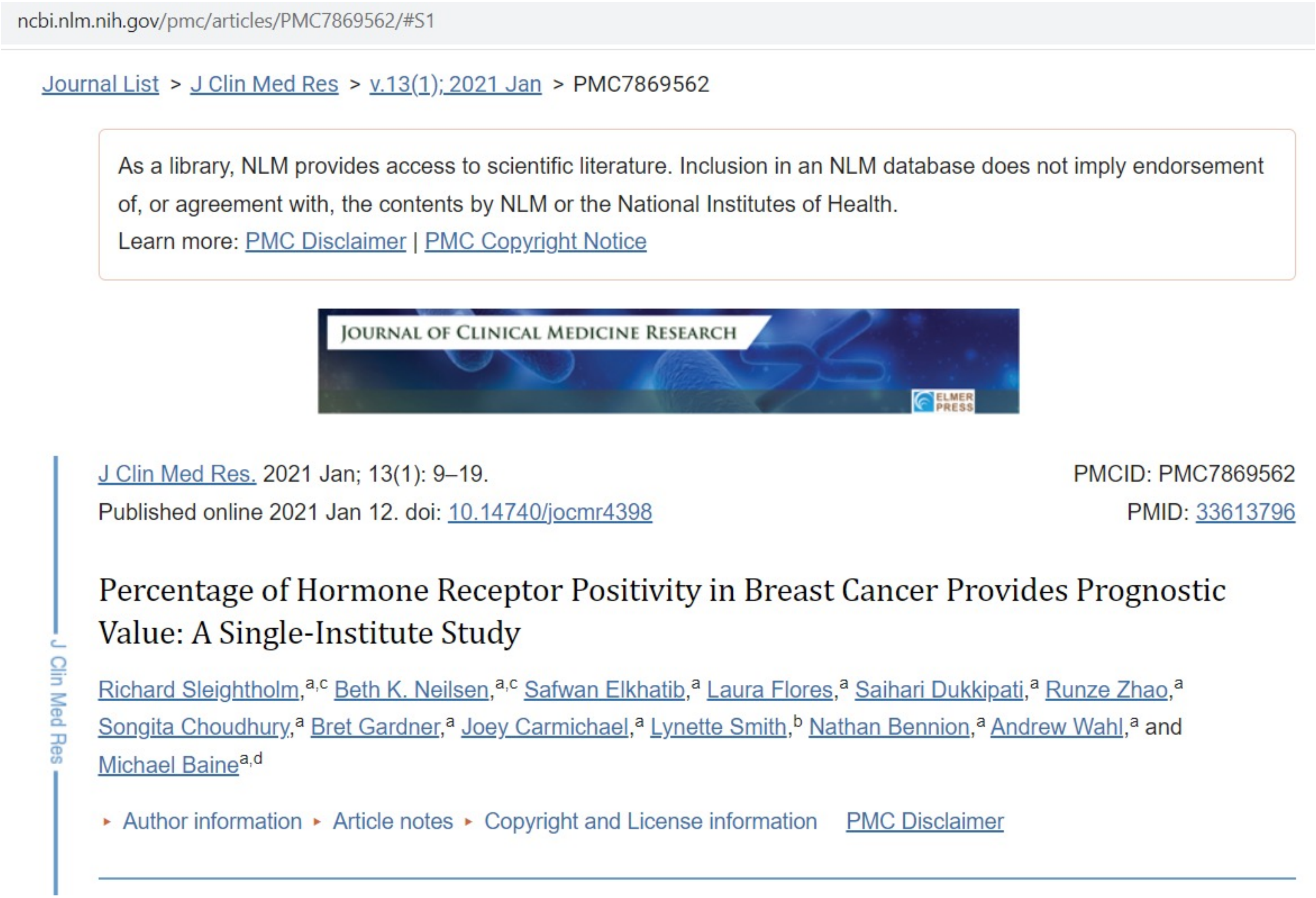
	Match Rate	Matches	Total Events
Summary Statistics:	0.00%	0	841,688

All population

Conditions from Onco module as index event

	Inclusion Rule	N	% Satisfied	N	% Satisfied
1.	ESR1 Protein Expression measurement Negative	9,400	1.12%	8,115	7.06%
2.	ESR1 Protein Expression measurement Positive	33,247	3.95%	30,211	26.29%
3.	ERBB2 (erb-b2 receptor tyrosine kinase 2) gene variant measurement Negative	23,505	2.79%	20,834	18.13%
4.	ERBB2 (erb-b2 receptor tyrosine kinase 2) gene variant measurement Positive	5,443	0.65%	4,599	4.00%
5.	PGR (progesterone receptor) gene variant measurement - Negative	10,675	1.27%	9,259	8.06%
6.	PGR (progesterone receptor) gene variant measurement - Positive	23,523	2.79%	21,344	18.57%
7.	Grade 1	4,930	0.59%	4,502	3.92%
8.	Grade 2	9,775	1.16%	8,788	7.65%
9.	Grade 3	6,642	0.79%	5,678	4.94%
10.	High grade tumor	4,314	0.51%	3,893	3.39%
11.	Low grade tumor	2,275	0.27%	2,107	1.83%
12.	Invasion	32,764	3.89%	29,158	25.38%
13.	Metastasis present	5,308	0.63%	0	0.00%
14.	Metastasis to bone	1,848	0.22%	0	0.00%
15.	TNM T1	19,602	2.33%	17,394	15.14%
16.	TNM N0	22,074	2.62%	20,201	17.58%
17.	TNM Ta	621	0.07%	498	0.43%
18.	TNM T2	8,799	1.05%	7,373	6.42%
19.	TNM M0	17,208	2.04%	15,440	13.44%
20.	Stage 1	18,923	2.25%	16,827	14.64%
21.	Stage 4	3,411	0.41%	1,947	1.69%

Compare with the article results



ncbi.nlm.nih.gov/pmc/articles/PMC7869562/#S1

Journal List > J Clin Med Res > v.13(1); 2021 Jan > PMC7869562

As a library, NLM provides access to scientific literature. Inclusion in an NLM database does not imply endorsement of, or agreement with, the contents by NLM or the National Institutes of Health.
Learn more: PMC Disclaimer | PMC Copyright Notice

JOURNAL OF CLINICAL MEDICINE RESEARCH

J Clin Med Res. 2021 Jan; 13(1): 9–19.
Published online 2021 Jan 12. doi: 10.14740/jocmr4398

PMCID: PMC7869562
PMID: 33613796

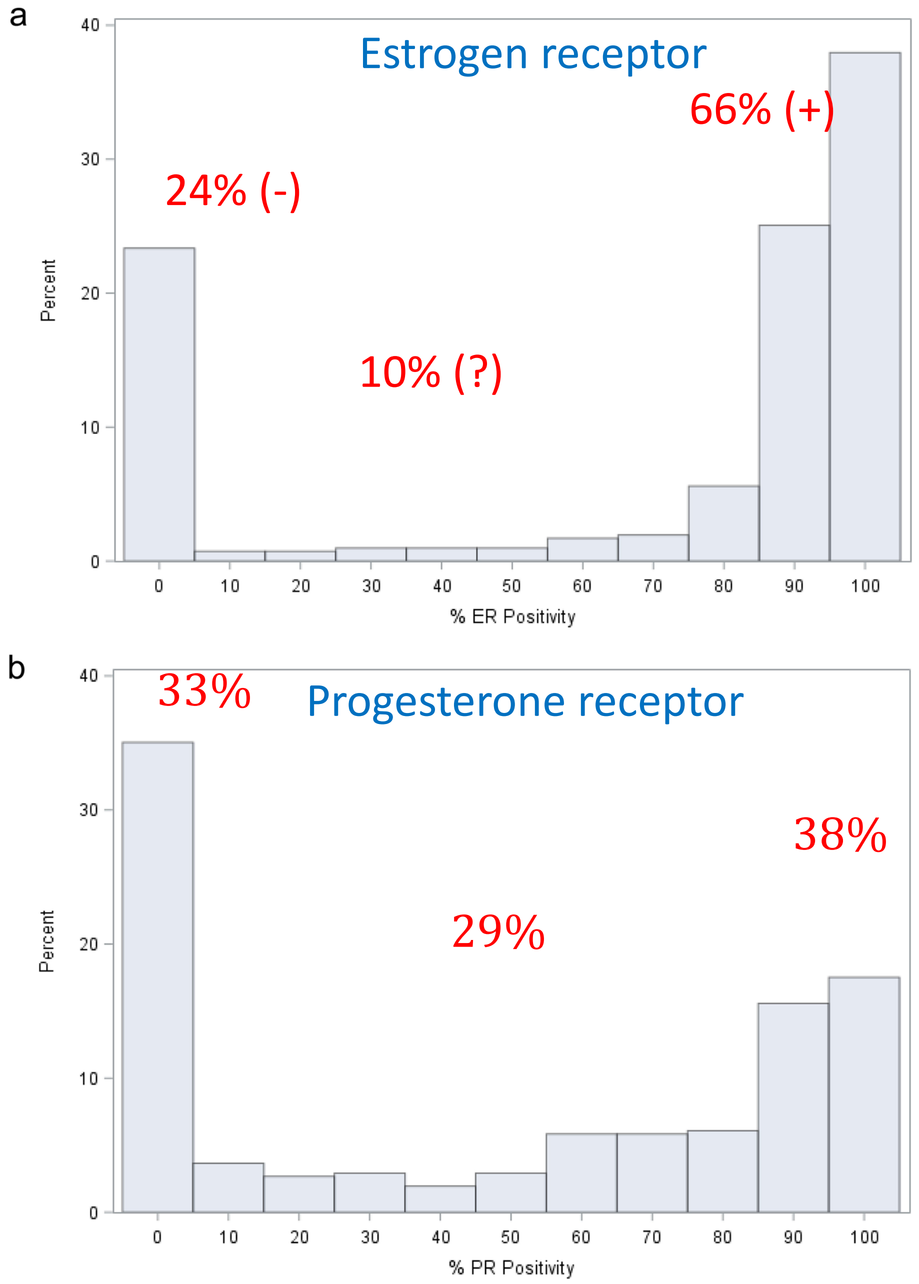
Percentage of Hormone Receptor Positivity in Breast Cancer Provides Prognostic Value: A Single-Institute Study

Richard Sleightholm,[a,c] Beth K. Neilsen,[a,c] Safwan Elkhatib,[a] Laura Flores,[a] Saihari Dukkipati,[a] Runze Zhao,[a] Songita Choudhury,[a] Bret Gardner,[a] Joey Carmichael,[a] Lynette Smith,[b] Nathan Bennion,[a] Andrew Wahl,[a] and Michael Baine[a,d]

▸ Author information ▸ Article notes ▸ Copyright and License information PMC Disclaimer

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7869562/

in our data: 22%(-) / 78(+)



in our data: 31% (-) / 69(+)

Future development

- **Tumor progression** will be mapped to the Episode table in the next iteration
- **Line of therapy** to be mapped to the HemOnc vocabulary with subsequent run and check of the ARTEMIS
- **Data cleansing algorithms** to be improved

Discussion

- Use cases – we can participate in a network study
- Data cleansing approaches
- Data validation algorithms
- Not mapped data elements