Clintek 11
Data Driven Analysis

Presented By: Yves Greatti
February 3rd 2023

Problem Statements

- **Given the Audit table, tell us what information you can gather about the study?**
- **Pick 2 itemOIDS. Use your statistics and ML knowledge to clean the data for these two itemOIDS and remove anomalous points you think exists in the data.**

Missing Data and Initial Pre-Processing

- Clintek11 is a study of a drug effect
- 35 unique patients in Audit and Query table
- Most of the missing data in Audit and Query is related to the columns ‘Unnamed: 11’ and ‘Unnamed: 12’ (>99%). We dropped these columns.
- 79,234 rows in Audit
- 5,795 rows in Query
- Prescribed Dose is 60

FormOID Description

FormOID	Description
AE	Adverse Events
CM	Concomitant Meds
DA	Disposition
DS	Drug Accountability
EX	Exposure
IE	Inclusion/Exclusion
MH	Medical History
SU	Substance Use
VS	Vital Sign

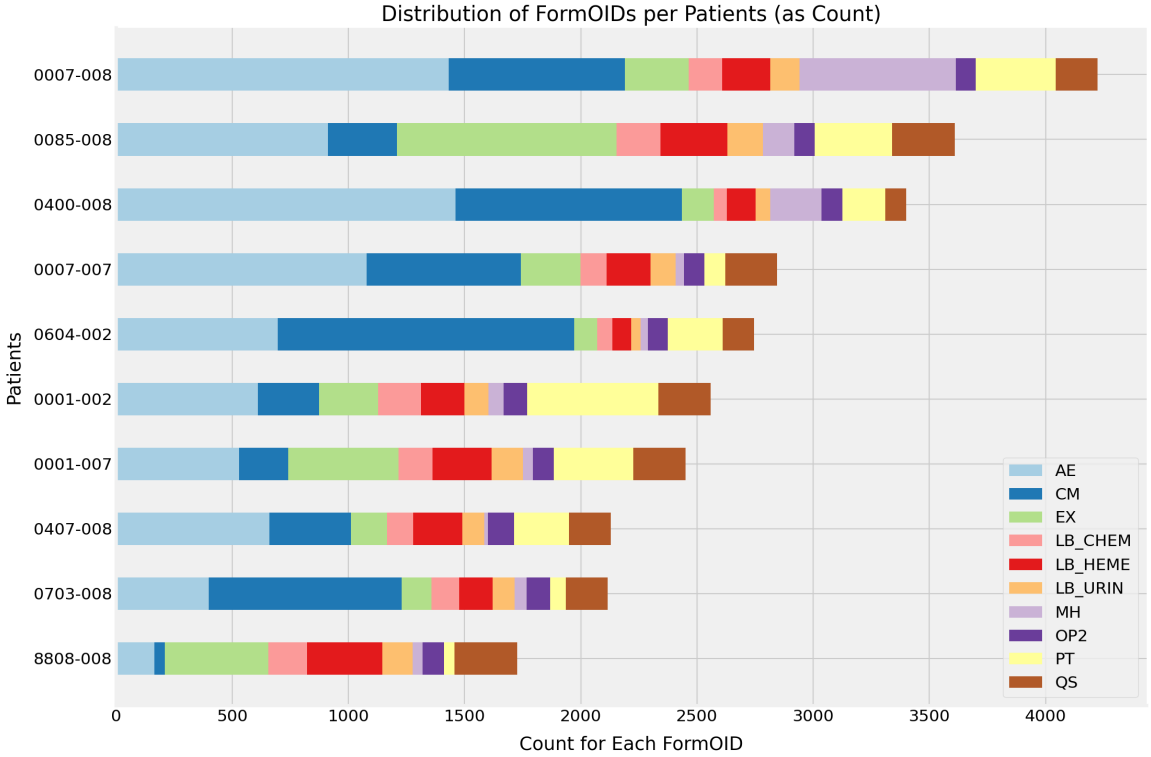
Top 10 Characteristics By Data Points (in %)

Attributes	Audit	Query	Audit ∩ Query
SubjectID	0007-008, 0085-008, 0400-008, 0007-007, 0001-002, 0001-007, 0604-002, 0703-008, 0407-008, 8808-008	0085-008, 0001-007, 0007-008, 0085-007, 8800-007, 0085-002, 0001-002, 0007-007, 0407-008, 0084-008	0001-002, 0001-007, 0007-007, 0007-008, 0085-008, 0407-008
FormOID	AE, CM, EX, PT, QS, LB_HEME, MH, LB_CHEM, OP2, LB_URIN	LB_HEME, EX, AE, PT, LB_CHEM, CM, LB_CHEM2, LB_URIN, LB_UM, LB_COAG	AE, CM, EX, LB_CHEM, LB_HEME, LB_URIN, PT
ItemOID	CM.CMEVNO, CM.CMENDAT, CM.CMONGO, CM.CMSTDAT, CM.CMINDC, CM.CMINDSP, CM.CMDSTXT, CM.CMTRT, CM.CMDOSFRQ, CM.CMDOSU	AE.AETERM, EX.EXACTION, PT.PTPRDAT, EX.EXPDOS, EX.EXREGCH, CM.CMEVNO, LB_UM.LBUMDAT', PT.PTTRT, CM.CMTRT', AE.AEENDAT	CM.CMEVNO, CM.CMTRT

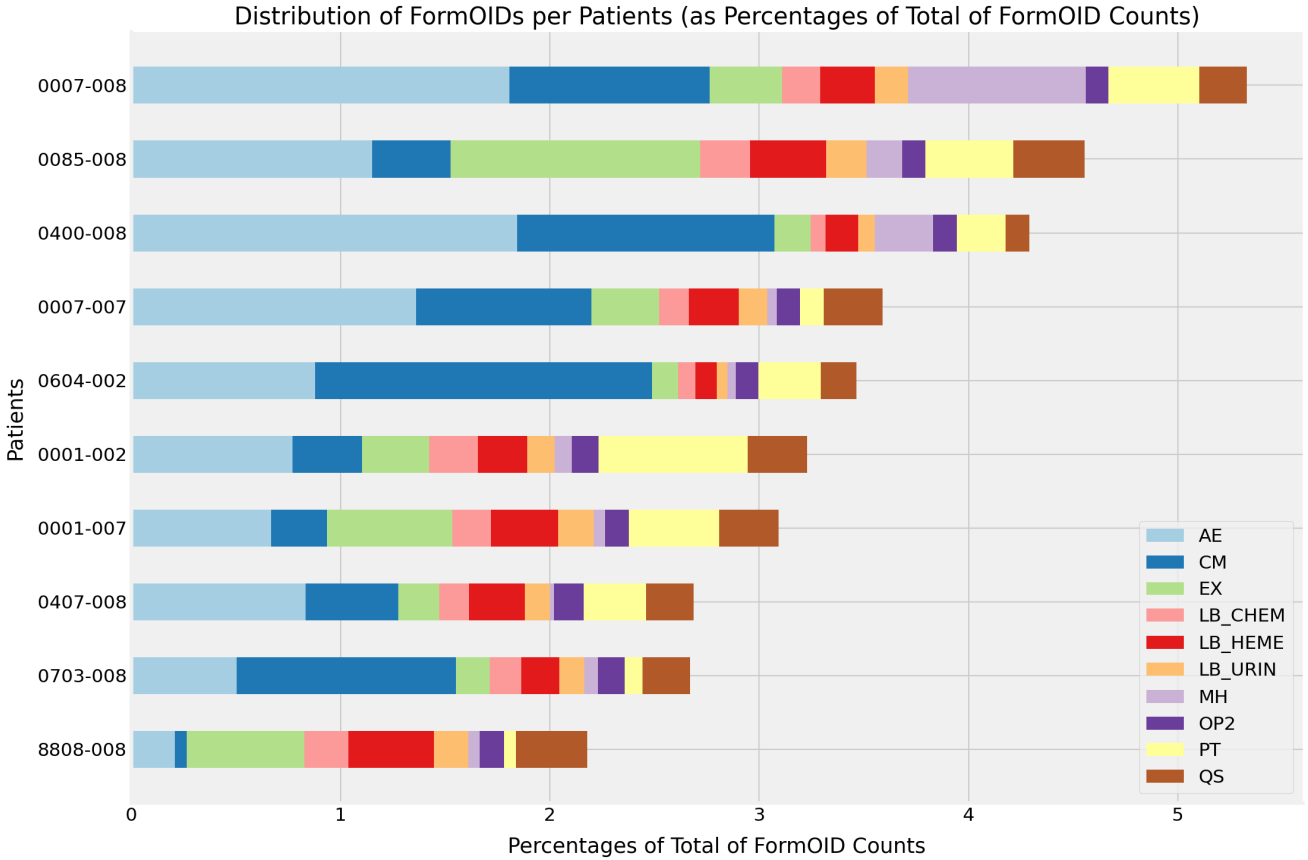
Distribution of FormOIDS Per Top 10 Patients

- **All 35 unique patients in Audit and Query tables cannot not be plotted easily**
- **Top 10 patients have the most data points in AE and CM**
- **Patient 007-008 and 0085-008 have the most data points**
- **007-007 has almost the double number of data points compared to 8808-008**

Distribution of FormOIDS per Patients in audit table

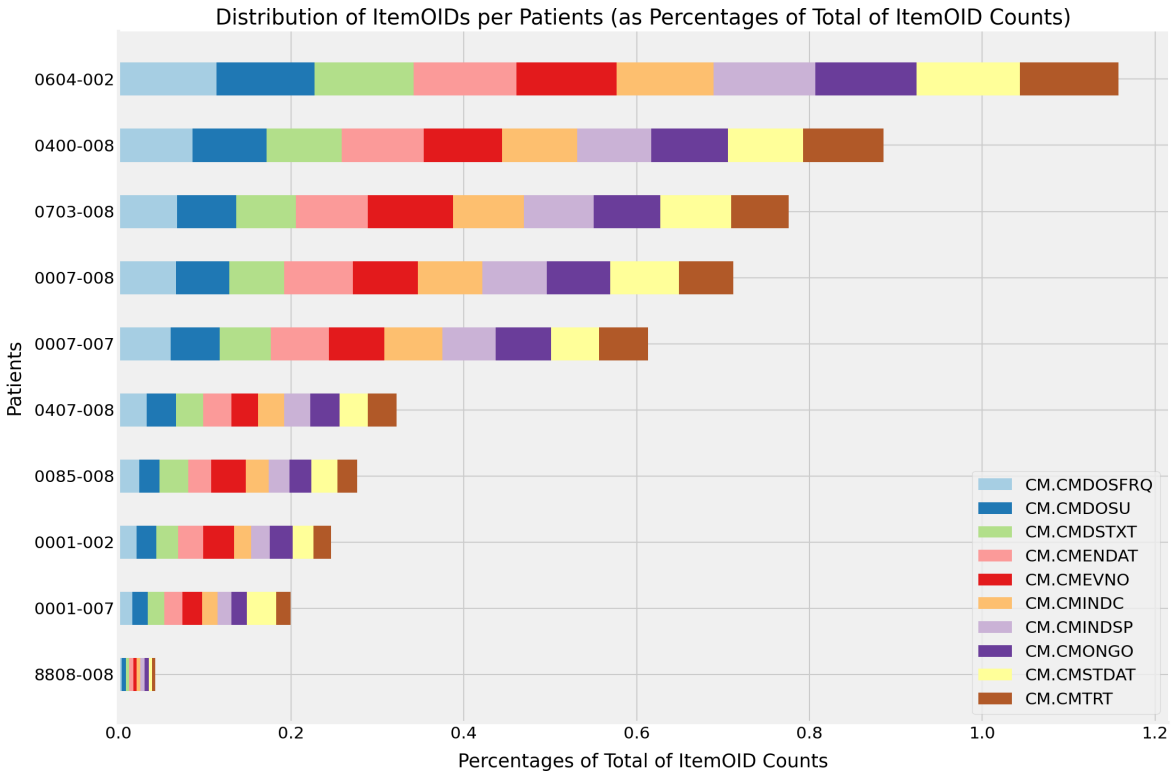


Distribution of FormOIDs per Patients in audit table (%)



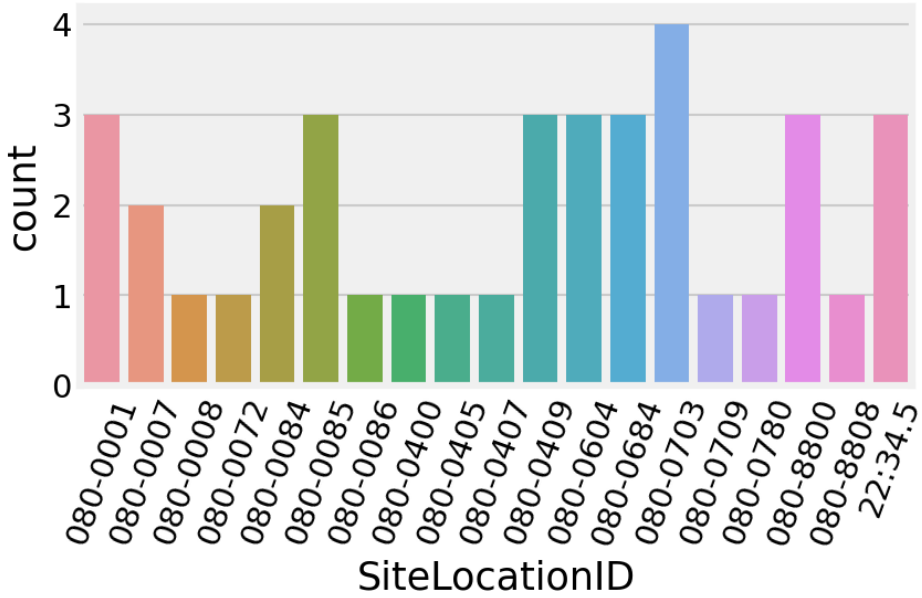
Distribution of ItemOIDs per Patient in audit table (%)

CM ItemOiDS are prevalent among the top 10 patients



Site Locations

- **Forms for a patient were mostly collected at only one site but some sites were visited by more than one patient.**



Urea patient Event Flow

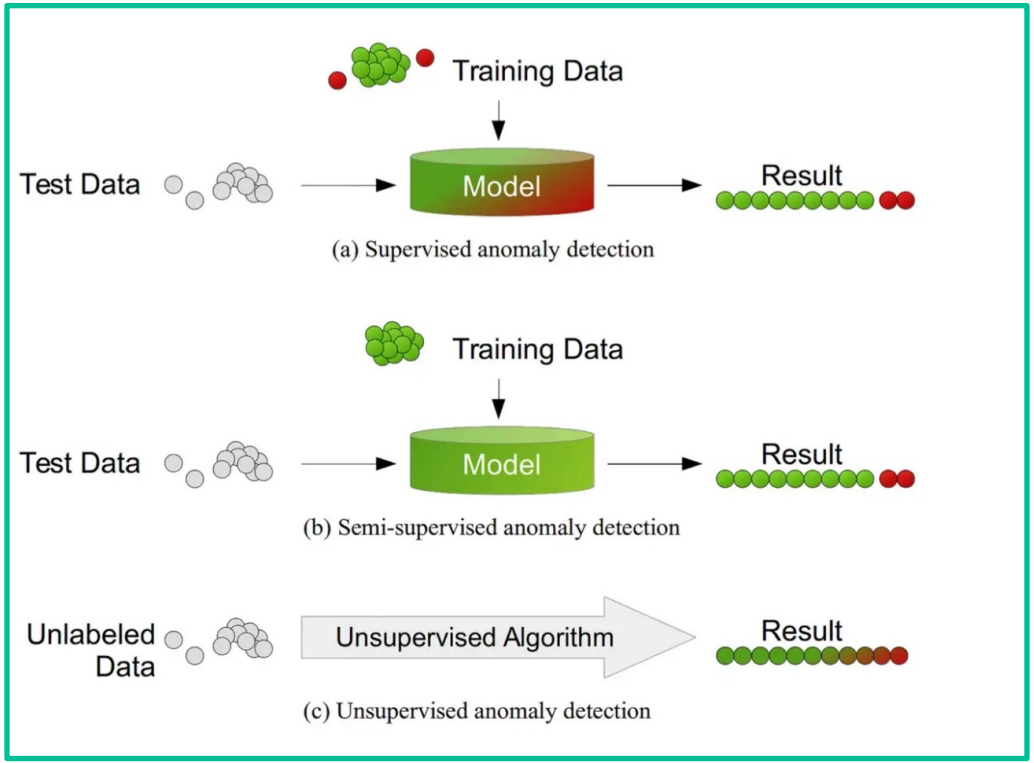
	OpenDate	CloseDate	ItemValue	QueryValue	QueryResponse
491	2015-11-10	2015-11-13	5	The lab parameter is blank. If it was not perf...	updated
5502	2016-02-23	2016-04-21	5.4	Was serum chemistry sample collected? is Yes. ...	NaN
4040	2016-03-01	2016-04-21	6.1	Data is not recorded. If not done, Please upda...	Updated
3867	2016-03-29	2016-03-30	5.4	5.4 on 5/29/15 per source. Please update - Th...	updated
3868	2016-03-29	2016-03-30	6.1	6.1 per source. Please complete. This is the...	-
3011	2017-01-19	2017-01-24	6.1	Please assess if CS or NCS from the drop down ...	Updated
3016	2017-01-19	2017-01-24	5.4	Please assess if CS or NCS from the drop down ...	Updated

	EntryDate	ModifiedDate	ItemValue	ActionCategory
72678	2015-06-18	2018-08-03	NaN	EnteredEmpty
77043	2015-06-18	2018-08-03	NaN	EnteredEmpty
59114	2015-08-14	2018-08-03	NaN	EnteredEmpty
49932	2015-09-17	2018-08-03	NaN	EnteredWithMissingCode
49552	2015-09-18	2018-08-03	NaN	EnteredEmpty
40367	2015-11-10	2018-08-03	5	EnteredWithChangeCode
16286	2016-03-29	2018-08-03	5.4	EnteredWithChangeCode
16287	2016-03-29	2018-08-03	6.1	EnteredWithChangeCode

Statistics of Number of Days Between Open and Answer Dates

	Statistics
Count	**4942**
Mean +/- SD	**41 +/- 63**
Min - Max	**0 - 514**
25%	**1**
50%	**12**
75%	**55**

Outlier Detection Methodologies



ItemOIDS Outlier Selection

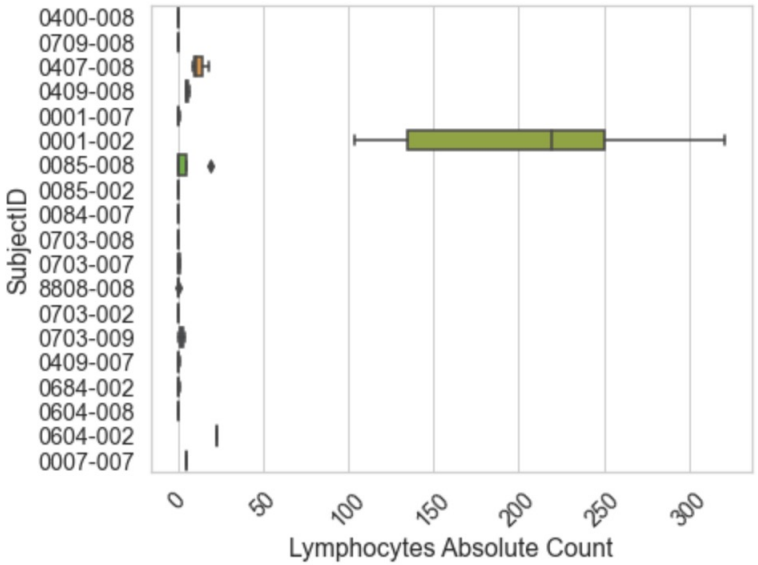
ItemOID	Overall	perc
CM.CMEVNO	1004	1.27
CM.CMENDAT	942	1.19
CM.CMONGO	890	1.12
CM.CMSTDAT	887	1.12
CM.CMINDC	880	1.11
CM.CMINDSP	874	1.10
CM.CMDSTXT	865	1.09
CM.CMTRT	838	1.06
CM.CMDOSFRQ	833	1.05
CM.CMDOSU	830	1.05

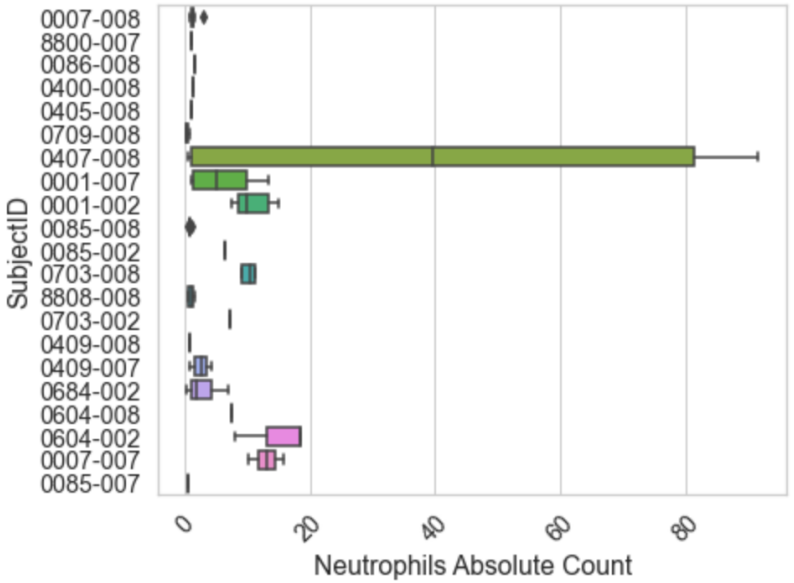
Top 10 Audit ItemOID by Count

ItemOID	Overall	perc
EX.EXPDOS	95	6.10
LB_HEME.LYMPHABS	66	4.24
CM.CMEVNO	63	4.05
LB_HEME.NEUTABS	61	3.92
LB_HEME.RBC	49	3.15
EX.EXDOSE	44	2.83
LB_HEME.BANDNEUT	43	2.76
LB_HEME.MONOABS	40	2.57
LB_HEME.NEUT	37	2.38
LB_HEME.LYMPH	36	2.31

Top 10 Query ItemOID by Count

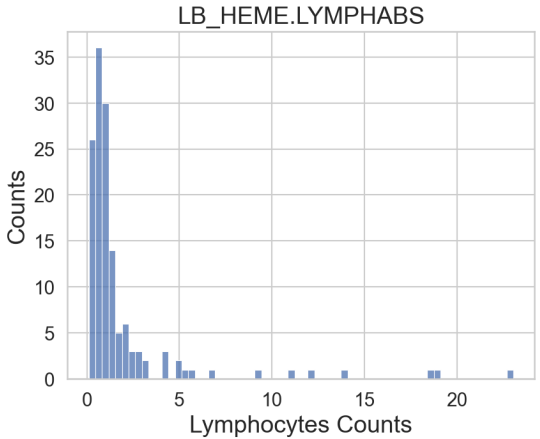
Distribution of Lymphocyte and Neutrophil Count

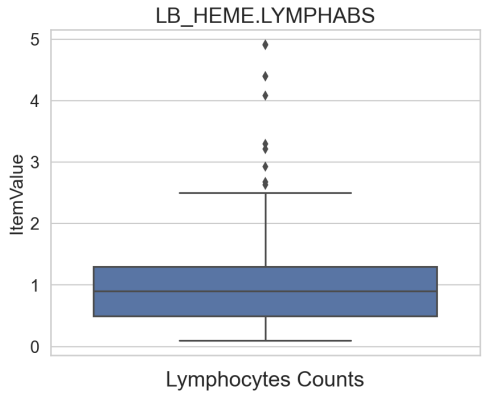




Lymphocyte (abs) Count Outlier Detection

Count	149
Mean ± SD	13.26 ± 49.80
Min	0.10
25%	0.60
50%	1.00
75%	1.90
Max	321.31





Lymphocyte Count Outlier Detection

```
MAD = Median(|x_i - Median(x)|)

clf = MADClassifier ()
clf.fit(vals)
preds = clf.predict(vals)

4.08, 4.4 , 4.4 , 4.9 , 4.91, 5.4 , 5.8 , 6.9 , 9.2 , 11.2 , 12,
14. , 18.7 , 18.81, 23.06, 52. , 103.41, 114.38, 141.09, 207.39,
231.99, 234.17, 299.83, 321.31
```

```
clf = ECODClassifier()
clf.fit(vals)
preds = clf.predict(vals)

0.1 , 0.11, 0.2 , 0.2 , 0.2 , 0.2 , 0.2 , 114.38, 141.09,
207.39, 231.99, 234.17, 299.83, 321.31
```

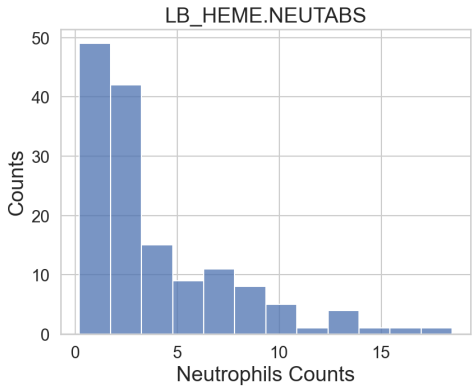
	LB_HEME.LYMPHABS
count	49.000000
mean	0.831837
std	1.037583
min	0.200000
25%	0.400000
50%	0.590000
75%	0.800000
max	4.910000

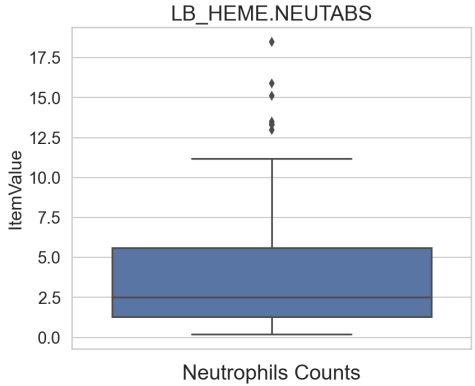
Query LB_HEME.LYMPHABS Statistics

Ecod: unsupervised outlier detection using empirical cumulative distribution functions

Neutrophil (abs) Count Outlier Detection

Count	147
Mean ± SD	3.86 ± 3.63
Min	0.20
25%	1.28
50%	2.50
75%	5.60
Max	18.48





Neutrophil Count Outlier Detection

- **MADClassifier**

12.99, 13.3 , 13.38, 13.5 , 15.11, 15.9 , 18.48, 20.3 , 20.55, 78.1 , 80.9 , 83.6 , 91.7

- **ECODClassifier**

0.2 , 0.21, 0.22, 0.32, 0.43, 0.44, 0.47, 0.47, 15.9 , 18.48, 20.3 , 20.55, 78.1 , 80.9 , 83.6 , 91.7

	LB_HEME.NEUTABS
count	59.000000
mean	4.915763
std	5.280215
min	0.210000
25%	0.940000
50%	1.500000
75%	8.900000
max	18.480000

Query LB_HEME.NEUTABS Statistics

Key TakeAways

- The data is more or less clean depending the FormOID
- Median time for query action to be resolved is 12 days
- AE and CM FormOIDS are the prevalent data in audit

Future Work

- More Data Analysis (Audit, Query tables)
- Statistic significance of the outlier detection (C.I.)
- Supervised outlier detection

Questions