We clustered trials in the following four files from the ../nci_data/dataset1-trials directory, based on each trial's curated inclusion criteria, expressed in part through UMLS codes, as stored in the "Boolean" column:
- "Hemoglobin_CTEP Trials_072018"
- "Platelets_CTEP Trials_072018"
- "WBC_CTEP Trials_072018"
- "HIV_CTEPTrials_072018"

First, we excluded trials with missing criteria in the Boolean column:
- "Hemoglobin_CTEP Trials_072018"
    Excluding 173 of 347 rows
    After exclusion, 174 rows remain
- "Platelets_CTEP Trials_072018"
    Excluding 148 of 342 rows
    After exclusion, 194 rows remain
- "WBC_CTEP Trials_072018"
    Excluding 276 of 342 rows
    After exclusion, 66 rows remain
- "HIV_CTEPTrials_072018"
    Excluding 123 of 342 rows
    After exclusion, 219 rows remain

Next, we parsed the boolean expression in a crude way to obtain information needed for the feature extraction below. We primarily focused on extracting (1) triples that represent individual criteria and (2) operators ("AND" and "OR"). As an example of a criterion triple, the criterion "C64848 >= 8g/dL" maps to the triple ('C64848', '>=', '8g/dL').

The parsing described above results in a sequence of operators and a sequence of triples for each trial. We ignore nesting of criteria disjunctions and conjunctions. While this is a significant simplification, our approach seems to work well despite the simplification, perhaps due to the pairwise relations captured by the features described below. In some cases, we encountered triples that were incomplete, perhaps due to manual annotation error. We filled in this missing data using a placeholder value.

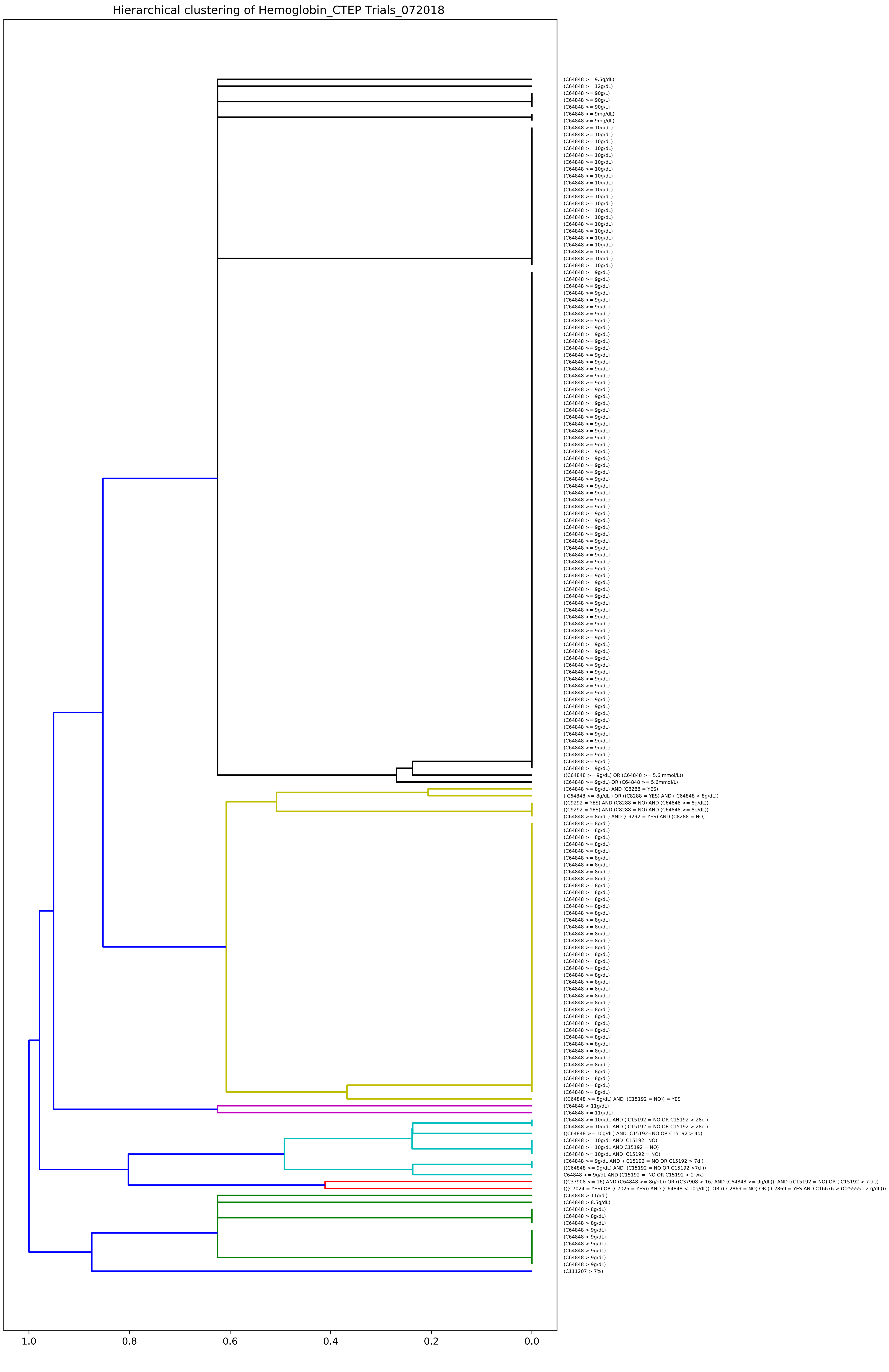
We define the following features based on the sequence of operators and the sequence of triples for each trial. Each triple contains three elements, the left, center, and right elements. The counts below are taken over the triples or operators in each trial's list. Most commonly, each element or triple only occurs 0 or 1 times. However, in some cases, elements or entire triples are repeated, e.g., the same UMLS code occurs in several triples, or a triple is repeated in two clauses. The features are as follows:
- The count of each triple element (left, center, or right).
- The count of each pair of triple elements.

- The count of each triple.
- The count of each pair of triples.
- The count of each operator.
- The count of each pair of operators.

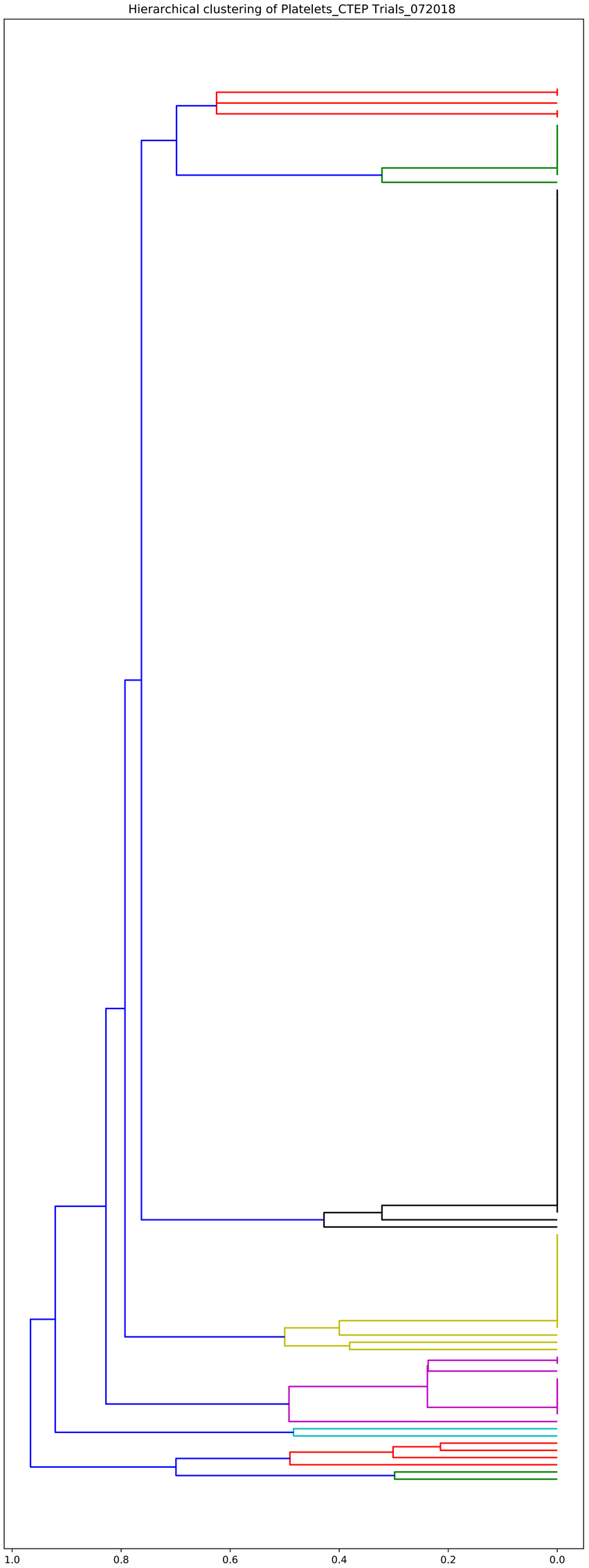
For example, "(C64848 >= 9g/dL) OR (C64848 >= 5.6mmol/L)' maps to the following non-zero features:
  - The count of each triple element (left, center, or right).
    'l_count_C64848': 2.0,
    'c_count_>=': 2.0,
    'r_count_5.6mmol/L': 1.0,
    'r_count_9g/dL': 1.0,
  - The count of each pair of triple elements.
    'lc_count_(C64848, >=)': 2.0,
    'lr_count_(C64848, 5.6mmol/L)': 1.0,
    'lr_count_(C64848, 9g/dL)': 1.0,
    'cr_count_(>=, 5.6mmol/L)': 1.0,
    'cr_count_(>=, 9g/dL)': 1.0,
  - The count of each triple.
    "triple_count_('C64848', '>=', '5.6mmol/L')": 1.0,
    "triple_count_('C64848', '>=', '9g/dL')": 1.0,
  - The count of each pair of triples.
    "triple_pair_count_('C64848', '>=', '5.6mmol/L')_('C64848', '>=', '5.6mmol/L')": 1.0,
    "triple_pair_count_('C64848', '>=', '5.6mmol/L')_('C64848', '>=', '9g/dL')": 1.0,
    "triple_pair_count_('C64848', '>=', '9g/dL')_('C64848', '>=', '5.6mmol/L')": 1.0,
    "triple_pair_count_('C64848', '>=', '9g/dL')_('C64848', '>=', '9g/dL')": 1.0
  - The count of each operator.
    'operator_count_OR': 1.0,
  - The count of each pair of operators.
    'operator_pair_count_OR_OR': 1.0,

Based on these features, we carry out hierarchical clustering using complete linkage and cosine similarity. I tried a few variants and this combination seemed to give the best results. It makes sense that cosine similarity works well for these sparse count features.

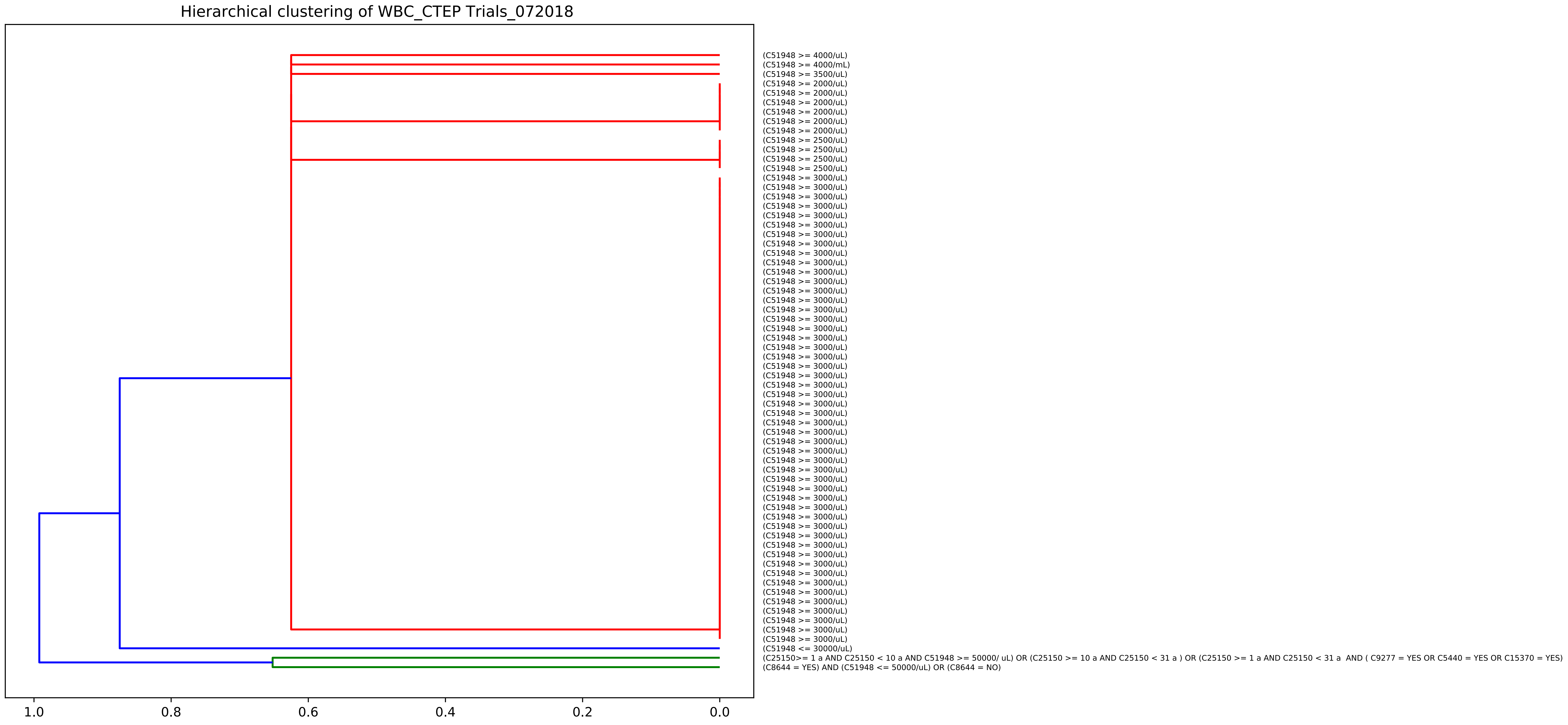
We plot dendrograms representing each clustering in *.clustering.pdf. We report features (alongside the original data) in *.features.csv. We report the linkage matrix in *.linkage_matrix.csv.

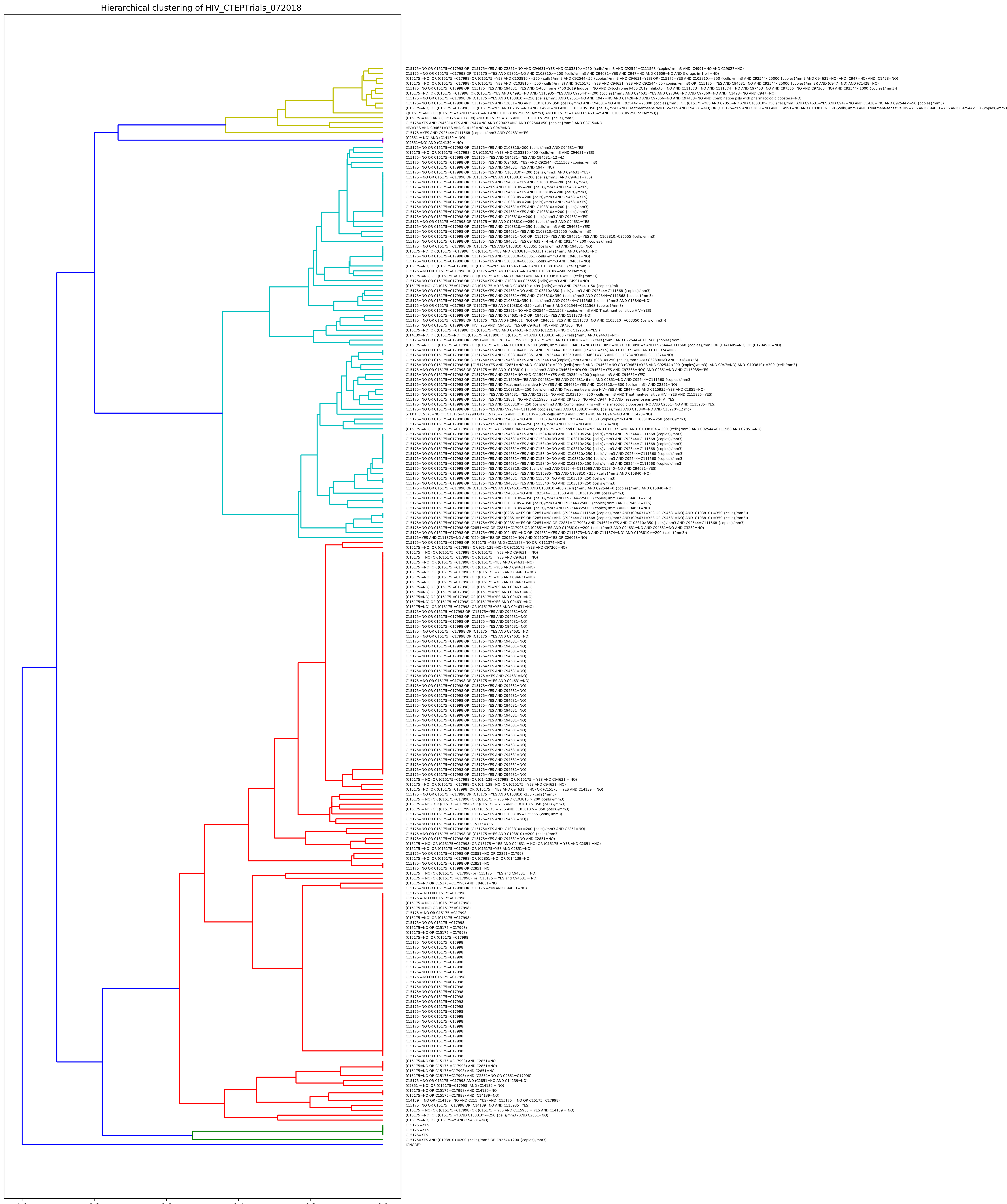

Hierarchical clustering of Hemoglobin_CTEP Trials_072018
(C64848 >= 9.5g/dL)
(C64848 >= 12g/dL)
(C64848 >= 90g/L)
(C64848 >= 90g/L)
(C64848 >= 90g/L)
(C64848 >= 9mg/dL)
(C64848 >= 9mg/dL)
(C64848 >= 10g/dL)
(C64848 >= 10g/dL)
(C64848 >= 10g/dL)
(C64848 >= 10g/dL)
(C64848 >= 10g/dL)
(C64848 >= 10g/dL)
(C64848 >= 10g/dL)
(C64848 >= 10g/dL)
(C64848 >= 10g/dL)
(C64848 >= 10g/dL)
(C64848 >= 10g/dL)
(C64848 >= 10g/dL)
(C64848 >= 10g/dL)
(C64848 >= 10g/dL)
(C64848 >= 10g/dL)
(C64848 >= 10g/dL)
(C64848 >= 10g/dL)
(C64848 >= 10g/dL)
(C64848 >= 10g/dL)
(C64848 >= 10g/dL)
(C64848 >= 10g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
(C64848 >= 9g/dL)
((C64848 >= 9g/dL) OR (C64848 >= 5.6 mmol/L))
(C64848 >= 9g/dL) OR (C64848 >= 5.6mmol/L)
(C64848 >= 8g/dL) AND (C8288 = YES)
(C64848 >= 8g/dL) OR ((C8288 = YES) AND (C64848 < 8g/dL))
((C9292 = YES) AND (C8288 = NO) AND (C64848 >= 8g/dL))
((C9292 = YES) AND (C8288 = NO) AND (C64848 >= 8g/dL))
(C64848 >= 8g/dL) AND (C9292 = YES) AND (C8288 = NO)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
(C64848 >= 8g/dL)
((C64848 >= 8g/dL) AND (C15192 = NO)) = YES
(C64848 < 11g/dL)
(C64848 >= 11g/dL)
(C64848 >= 10g/dL AND (C15192 = NO OR C15192 > 28d)
(C64848 >= 10g/dL AND (C15192 = NO OR C15192 > 28d)
((C64848 >= 10g/dL) AND C15192=NO OR C15192 > 4d)
(C64848 >= 10g/dL AND C15192=NO)
(C64848 >= 10g/dL AND C15192 = NO)
(C64848 >= 10g/dL AND C15192 = NO)
(C64848 >= 9g/dL AND (C15192 = NO OR C15192 > 7d)
((C64848 >= 9g/dL) AND (C15192 = NO OR C15192 >7d))
C64848 >= 9g/dL AND (C15192 = NO OR C15192 > 2 wk)
((C37908 <= 16) AND (C64848 >= 8g/dL)) OR ((C37908 > 16) AND (C64848 >= 9g/dL)) AND ((C15192 = NO) OR (C15192 > 7 d))
(((C7024 = YES) OR (C7025 = YES)) AND (C64848 < 10g/dL)) OR ((C2869 = NO) OR (C2869 = YES AND C16676 > (C25555 - 2 g/dL)))
(C64848 > 11g/dl)
(C64848 > 8.5g/dL)
(C64848 > 8g/dL)
(C64848 > 8g/dL)
(C64848 > 8g/dL)
(C64848 > 9g/dL)
(C64848 > 9g/dL)
(C64848 > 9g/dL)
(C64848 > 9g/dL)
(C64848 > 9g/dL)
(C64848 > 9g/dL)
(C111207 > 7%)
1.0
0.8
0.6
0.4
0.2
0.0

Hierarchical clustering of Platelets_CTEP Trials_072018



C51951 >= 70000 / uL
C51951 >= 70000 / uL
C51951 >= 30000 / uL
C51951 >= 150000 / uL
C51951 >= 150000 / uL
C51951 >= 50000 / uL
C51951 >= 50000 / uL
C51951 >= 50000 / uL
C51951 >= 50000 / uL
C51951 >= 50000 / uL
C51951 >= 50000 / uL
C51951 >= 50000 / uL
C51951 >= 50000 / uL
C51951 >= 50000 / uL AND C51951 < 28d
C51951 >= 100000 / uL
C51951 >= 100000 / uL
C51951 >= 100000 / uL


Hierarchical clustering of WBC_CTEP Trials_072018
(C51948 >= 4000/uL)
(C51948 >= 4000/mL)
(C51948 >= 3500/uL)
(C51948 >= 2000/uL)
(C51948 >= 2000/uL)
(C51948 >= 2000/uL)
(C51948 >= 2000/uL)
(C51948 >= 2000/uL)
(C51948 >= 2000/uL)
(C51948 >= 2500/uL)
(C51948 >= 2500/uL)
(C51948 >= 2500/uL)
(C51948 >= 2500/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 >= 3000/uL)
(C51948 <= 30000/uL)
(C25150>= 1 a AND C25150 < 10 a AND C51948 >= 50000/ uL) OR (C25150 >= 10 a AND C25150 < 31 a) OR (C25150 >= 1 a AND C25150 < 31 a AND (C9277 = YES OR C5440 = YES OR C15370 = YES)
(C8644 = YES) AND (C51948 <= 50000/uL) OR (C8644 = NO)
1.0
0.8
0.6
0.4
0.2
0.0


Hierarchical clustering of HIV_CTEPTrials_072018
1.0
0.8
0.6
0.4
0.2
0.0