**LLT Fuzzy Matching Instructions:**

**NOTE: To maximize reproducibility and facilitate downstream automation, please read and follow these instructions. Please do not modify the working fuzzy mapping files further (other than using sort on columns). In particular, do not transform all text into CAPS, or change the name of column headers.**

**Background:** We are harmonizing the lowest level (LLT) of medical history (MH) terms using the MedDRA vs21 standard (as we did previously with Adverse Event (AE) terms. At this point we have cleaned and formatted the original MH data and completed exact matching on the MH terms. Out of a total of 28720 unique MH rows, all but 7115 were successfully matched exactly. Fuzzy matching (with fuzzywuzzy) has been run on these remaining 7115. Fuzzywuzzy outputs the top 5 matches using any of it’s built in fuzzy matching strategies. For MH terms we have available three sources of terms available for attempting to harmonized MH LLTs to the MedDRA standard. These included (MHTERM), available for all rows, LLT\_NAME, and MHMODIFY. Fuzzy matching was thus applied to each (whenever available). Therefore in the fuzzy matching files there are 5 columns giving the candidate fuzzy matches for each of these three term sources.

Manual Fuzzy mapping instructions:

1. Go through each row in your assigned mapping file and manually enter values (as described below) to indicate the ‘best’ LLT match for each.
2. During this process, only ever enter/modify values in the following columns:

\*FZMatch\_Choice\_ID\_MHTERM (yellow) – If the best match is available among the 5 (MHTERM) fuzzy matches enter the integer (1-5) for the corresponding column where that best match term is located. *i.e. FZMatch\_1\_MHTERM = 1, FZMatch\_2\_MHTERM = 2, FZMatch\_3\_MHTERM = 3, FZMatch\_4\_MHTERM = 4, FZMatch\_5\_MHTERM = 5 (Leave cell blank otherwise)*

\* FZMatch\_Copied\_Term (orange) – If an integer was entered in FZMatch\_Choice\_ID\_MHTERM leave this cell blank. Otherwise COPY AND PASTE the exact, best matching text MedDRA standard term/phrase into this cell. This may either be copied from one of the other two sets of fuzzy matching results (i.e. MHMODIFY, or LLT\_NAME), or taken directly from the MedDRA LLT standard term file, if there is a better match not identified with fuzzy matching. Please make sure only terms where currency= Y are taken from the LLT standard file.

\*LLT\_map\_code (grey) – This is where the ‘mapping quality is recorded. We will be using our global term match quality codes below:

* 0 = Exact Match - DL1\_FT1 and DL1\_FT2 both match (i.e. MHTERM = LLT\_NAME column)
* 1 = Exact Match - DL1\_FT1 match (i.e. MHTERM match)
* 2 = Exact Match - DL1\_FT2 match (i.e. LLT\_NAME match)
* 3 = Exact Match - DL1\_FT3 match (i.e. MHMODIFY match)
* 4 = Fuzzy Matching (High Confidence)
* 5 = Fuzzy Matching (Medium Confidence)
* Blank = Unmapped/NA

During this fuzzy matching process this cell should be filled in for every row! Only the values 4, 5, or 6 should be entered. Enter 4 if there is a relatively clear best term match that should not require any further follow up. Enter 5 if the ‘best’ chosen match is not clearly ideal, and there is reasonable uncertainty that this match is best. Lastly enter 6 if no ‘best’ match could reasonably be identified from the available information. In this case, no values will be entered for cells of the two columns described above. Cells tagged with a 6 demand follow up by an expert with domain knowledge. Please remember to check all three sets of fuzzy match information (when available). The alternative terms can be useful in more clearly identifying a ‘best’ match.

Dictionary of Columns in Fuzzy mapping files:

Effort\_Split – Temporary indices entered to split up the fuzzy matching task into separate files.

ROW\_INDEX – The original row index of the entire dataset. We can always resort on this value to get all entries into their original order.

MHTERM – The Medical History term, available for every entry in the data.

FZMatch\_1\_MHTERM – Top scoring fuzzywuzzy match to the MHTERM.

FZMatch\_2\_MHTERM – 2nd highest scoring fuzzywuzzy match to the MHTERM.

FZMatch\_3\_MHTERM – 3rd highest scoring fuzzywuzzy match to the MHTERM.

FZMatch\_4\_MHTERM – 4th highest scoring fuzzywuzzy match to the MHTERM.

FZMatch\_5\_MHTERM – 5th highest scoring fuzzywuzzy match to the MHTERM.

FZMatch\_Choice\_ID\_MHTERM – see above

FZMatch\_Copied\_Term – see above

LLT\_map\_code – see above

MHMODIFY – The alternative, ‘modified’ medical history term, available for some entries in the data.

FZMatch\_1\_ MHMODIFY – Top scoring fuzzywuzzy match to the MHMODIFY.

FZMatch\_2\_ MHMODIFY – 2nd highest scoring fuzzywuzzy match to the MHMODIFY.

FZMatch\_3\_ MHMODIFY – 3rd highest scoring fuzzywuzzy match to the MHMODIFY.

FZMatch\_4\_ MHMODIFY – 4th highest scoring fuzzywuzzy match to the MHMODIFY.

FZMatch\_5\_ MHMODIFY – 5th highest scoring fuzzywuzzy match to the MHMODIFY.

LLT\_NAME – Another alternative, medical history term, available for some entries of the data. Presumably this term conforms to the MedDRA standard, but we cannot verify which version.

FZMatch\_1\_ LLT\_NAME – Top scoring fuzzywuzzy match to the LLT\_NAME.

FZMatch\_2\_ LLT\_NAME – 2nd highest scoring fuzzywuzzy match to the LLT\_NAME.

FZMatch\_3\_ LLT\_NAME – 3rd highest scoring fuzzywuzzy match to the LLT\_NAME.

FZMatch\_4\_ LLT\_NAME – 4th highest scoring fuzzywuzzy match to the LLT\_NAME.

FZMatch\_5\_ LLT\_NAME – 5th highest scoring fuzzywuzzy match to the LLT\_NAME.

PT\_NAME – can ignore here

HLT\_NAME – can ignore here

HLGT\_NAME – can ignore here

SOC\_NAME – can ignore here

T\_LLT – can ignore here

T\_LLT\_CODE – can ignore here