1. Download All Human BioSamples with esearch

2. Format the esearch output file into

<ACC>-<AttrName>|<AttrVal> format (formatFile.pl)

3. Grep the entries with disease attribute values from the formatted file.

grep ‘disease’ humanAccAttr.output | awk -F”|” ‘{print $1”\t”$2}’ | awk -F”-“ ‘{print $1”\t”$2}’ > biosample\_w\_diseases.txt

4. Change the entered disease values to uppercase to normalize the casing difference. This file has all the biosample unique disease attribute values. There are total of 327 unique disease terms.

awk -F"\t" '{print toupper($3)}' biosample\_w\_disease.txt|grep -v '^-' |sort -u > biosample\_all\_disease\_term\_uniq.txt

5. Then run metaMapLite without ID with biosample\_all\_disease\_term\_uniq.txt

6. Generate a mapping file between entered term with preferred term from the metamap result file. There are total of 393 mapped preferred terms. (All disease terms)

awk -F"|" '{print $4"\t"$5"\t"$6"\t"$7}' biosample\_diseases.mmi |awk -F'"' '{print $1"\t"$2"\t"$4}' > biosample\_diseases\_mapping\_w\_prefer\_term\_sem\_code.txt

7. Grep only the sementic codes that are related with disease (biosample\_diseases.mmi). There are total of 229 terms.

grep 'acab\|anab\|comd\|cgab\|dsyn\|emod\|inpo\|mobd\|neop\|patf\|sosy' biosample\_diseases\_mapping\_w\_prefer\_term\_sem\_code.txt > biosample\_diseases\_mapping\_w\_prefer\_term\_sem\_code\_filtered.txt

8. Use the file from step 3 (biosample\_w\_diseases.txt) to map the entered disease terms to the preferred terms (only 1000 biosample entries with a disease are used for visualization).

esearch\_biosample\_w\_disease\_1000\_mapped.txt (in metamap folder)

Before Normalized = 40 terms

After Normalized = 23 terms