**E-MTM Data Validation Work Plan**

**Goal**: Automate the process of validating Enhanced Medication Therapy Management (E-MTM) encounter data using contents in supporting documents provided by Medicare Part D E-MTM participants (i.e. sponsors).

**Overview of work flow:**

% encounter records

* with drug id
* identify location of service
* with valid SNOMED CT code
* etc…

Simple counting work that has been done by PHSR, no automation is needed

internal validation:  
(simple counting)

raw encounter data

primary task

find evidence of encounter in document

Scope of DMG’s automation work

external validation

detect encounter details in document content but not in encounter data

secondary task

supporting document

**Outline of Tasks:**

1. Find evidence of encounter in document
   1. Identify columns in encounter data that can be validated (ids, date of encounter, description of encounter, drug RxCUI code, etc.) using supporting document information
   2. Associate supporting document with encounters using RECORD and CNTRCT\_ID columns in encounter data, and identify rows in encounter data (encounters) that can be validated
   3. Extract contents of supporting document PDFs
      1. Extract structured textual content (e.g. tables with directly usable rows/columns - e.g. encounter date)
      2. Extract unstructured textual content (e.g. paragraphs)
      3. Extract text from images (e.g. screenshot)
      4. Validate content extraction - compare machine and human extraction results.
   4. Structuralize extracted contents as needed. This often involves identifying and preserving tables/mapping relationships between textual elements, e.g. ‘Date of Encounter’ should be associated with a date entry ‘08/26/2019’.
      1. Validate content structuralization - compare machine and human structuralization results
   5. Compare cleaned up, structuralized content data against encounter data
      1. Design a rule to identify ‘matches’ - e.g. ‘Discussed with patient’ in encounter data and ‘Talked to patient about drug use’ from supporting document content should be considered good matches. This involves the design of a semantic distance function to measure proximity of semantics expressed by different terms.
      2. Flag ‘bad matches’ - i.e. flag encounters for which our algorithm shows low likelihood of finding needed evidence in supporting documents, thus need subsequent manual curation. This requires defining a numerical threshold for the likelihood of match to flag the bad matches.
   6. Validate the flagging of mismatches
      1. Determine the minimum required sample size to train a NLP model (e.g. identifying certain encounters from a paragraph)
      2. Use a sample of manually checked encounters (with or without evidence from documents), compute performance of the matching algorithm, and determine optimal numerical threshold above.
   7. Flag sponsor compliance status (full compliance = evidence found in documents for all encounters, partial compliance = % of encounters with evidence found, etc.)
   8. Flag certain non-compliance types of interest
      1. Recommendation to change medication - does document suggest which new medication should be used?
      2. Other non-compliance types? - check with PHSR.
2. Detect encounter details in document content but not in encounter data
   1. Supervised method
      1. Check with PHSR about specific additional encounter details of interest.
      2. Identify a set of keywords, and search for presence in documents
      3. Design a scoring function to make recommendation for manual checking for encounter details
      4. Validate the recommendation algorithm - compare machine and human content extraction results
   2. Unsupervised method
      1. Identify topics embedded in text in documents
      2. In consult with PHSR, design a recommendation algorithm to flag certain topics for manual checking for encounter details
      3. Validate the recommendation algorithm - compare machine and human content extraction results.