Mental-functioning Classification Efficacy using an NER on the Gold Standard Corpus for Activity information

Guy Divita[1], Maryanne Sacco[1], Kathleen Coale[1], Rafael Jimenez Silva[1], Rebecca Parks[1], Elizabeth Rasch[1]

[1] Rehabilitation Medicine Department, National Institutes of Health Clinical Center, Bethesda, MD

Abstract

Objective

We introduce a Named Entity Recognition (NER) application that categorizes mental-functioning mentions. Mental-functioning information extracted and characterized from clinical documentation is salient for coordination of care, health outcomes evaluation, disability determination and health research. We concentrate on Communication and Cognition (ComCog) and Interpersonal Interactions and Relationships (IPIR) categories being the most representative mental-functioning activities.

Methods and Materials

We prior developed an ontology with a clinical rehabilitation informatics perspective and derived rule-based NER applications to explore and define mental-functioning. This paper reports on the task of classifying mentions that have already been tagged as mental-functioning mentions by an upstream NER. The Gold Standard Corpus for Activity Information (GoSCAI) is an available clinical corpus that have been annotated for human functioning information based on the WHO's International Classification of Functioning, Disability and Health. This paper presents the mental-functioning categorization NER, leveraging terminology and knowledge derived from the ontology, benchmarked on annotated clinical records from the GoSCAI.

Results

The NER's ComCog weighted average F-Score across the 20 categories is 0.717. The NER's IPIR weighted average F-Score across the 8 categories is 0.726. The inter-rater reliability and support are also reported.

Discussion

The number of training mentions for each category was not evenly distributed, with many categories having insufficient training support and majority classes skewing the over-all performance. The IRR indicate finding mental-functioning mentions is cognitively challenging. Classifying mental-functioning is even more difficult.

Conclusions

The rule-based NER performs admirably despite the uneven training coverage for these cognitively challenging categorization tasks.

1 Introduction

Mental-functioning information extracted and characterized from clinical documentation is a salient source of data for coordination of health and social care, as well as health outcomes evaluation, disability determination and health research and is information that clinicians express interest in having [1].

While social determinants of health (SDOH) combined with traditional health condition observations provide strong outcome predictive power [2], we argue that mental-functioning observations provides an even stronger predictive mental health outcomes signal, but to date, little has been done to capture and/or combine mental-functioning with SDOH along with health condition data for more accurate outcomes prediction, diagnosis and condition classification. Accessing mental functioning information such as functional communication and cognition and interpersonal skills is of great use when applied to determine needs and predictions of social functioning, type and need of social support, and mental health care needs and outcomes. While our work focuses on mental functioning mention extraction in individual clinical records, this work could/should be extended to be combined with structured medical record data to provide population or epidemiological level outcomes analysis.

M*ental-functioning* activity information is challenging to identify in health records, due to over emphasis of documentation at the mental body-function level. Accurately characterizing a patient's current and past *mental-functioning* has been elusive because it has been inadequately described within standard terminologies and ontologies and often confused with the related topic of mental-function.[1 3] The Ecological Mental-functioning Ontology (EMFO)[4-6] was developed and now distributed as a mechanism to explore and define *mental-functioning* from a clinical rehabilitation informatics perspective. The aim of this paper is to present the refinement and benchmarking of a NER that classifies *mental-functioning*, leveraging terminology and knowledge derived from the EMFO, benchmarked on annotated clinical records from the Gold Standard Corpus for Activity Information (GoSCAI)[7]. The GoSCAI is a corpus of de-identified clinical notes that have been annotated for human functioning information based on the framework of the World Health Organization (WHO) International Classification of Functioning, Disability and Health (ICF)[8]. The corpus, annotations, and machine learned predictive models are now disseminated by permission [7]. The rule based NER, the subject of this paper, is disseminated separately via an open source license [4], as the rule based NER and associated terminology products contain no personal health information (PHI) or personal individual information (PII) and are not subject to the additional restrictions put in place to insure patient and provider security.
This NER was first developed for a Social Security Administration (SSA) use case [9], tuned for a second population, is an evolution of v3NLP Framework software originally developed for the VA.

2 Prior Work

This work is built in conjunction as part of a larger body of work that includes

- An Ecological Model of Mental Functioning Ontology developed to scope out what mental-functioning is from a rehabilitation informatics perspective [4]
- Mental Functioning Annotation guidelines, specifically for ComCog and IPIR [10, 11]
- Manual annotations created on an SSA claimant corpus used for SSA use cases followed by manual annotations created on the GoSCAI corpus [7]
- Large Language Models developed to identify mental-functioning containing sentences within clinical notes first trained on SSA claimant clinical records then trained and benchmarked on the GoSCAI corpus (unpublished, manuscript in preparation)

- A rule based NER to identify and classify mental-functioning sentences, built to be employed to also classify those sentences identified by the Large Language Models [4]

The rule-based NLP platform employed for this work, referred to as java-nlp-Framework [12], was adapted from the V3NLP Framework [13] and Sophia [14] which were used for symptom extraction [15] and finding mentions of sexual trauma in clinical notes of veterans [16]. The framework employed is built upon Apache's Unstructured Information Management Architecture [17] (UIMA) NLP platform, has a pedigree from Unified Medical Language System (UMLS) concept extraction in biomedical literature, MetaMap [18] as the java code base was derived from MMTx [19]. More recently, Java-nlp-Framework has been crafted to find body-function information within clinical texts in support of efforts by the SSA to identify such information [12].

2.1 Ecological Mental-functioning Ontology Background

The Ecological Mental-functioning Ontology (EMFO) describes classes and relationships having to do with the domain of *mental-functioning*. *Mental-functioning* is an individual's behaviors, activities and participation in daily life. These behaviors can be observed by others, such as clinicians, and documented in health records, and could also be reported by the patient themselves. This ontology was created initially to identify what is and is not *mental-functioning* in a larger effort to find *mental-functioning* mentions within clinical records for evidence to potentially aid social security disability adjudicators in making informed decisions from medical evidence.
It is important to point out that *mental-functioning* is different but related to mental-functions. Mental health domain experts have stressed the importance to distinguish and highlight mentions within clinical text that have to do with *mental-functioning* at the activity level and not include mentions solely at the body/mental-function level. Mental-functions are classified as body functions, a person's intrinsic physiological capability, and are not evidence of activities nor participation. Documenting that a patient has the capability to do calculations (evidence of mental-function), is different from documentation based on observation of calculating the tip for a restaurant bill (evidence of *mental-functioning*).
Standard medical terminologies, including Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) [20] and Medical Subject Headings (MeSH) [21], are distillations of language used within clinical notes. Leveraging existing medical terminologies to identify language of *mental-functioning* in clinical records was limited, as terminologies include much coverage around mental-functions and less coverage of activity functioning, further exacerbated by definitions muddling the distinctions between the two.

Initially, the ontology was based on a basic structure defined in the International Classification of Functioning, Disability and Health (ICF)[8] that explicitly separates functions of the body from functions at the activities and participation level, or what we call functioning. Overall, the ontology does recognize the ICF structure broadly with a few deviations and additions reflecting an evolved understanding of the field since the ICF's introduction in 2001.

Mental-functioning is related to mental-functions, which is related to body function, which is related

to body structures, and so on. All of this exists within contextual and environmental factors that affect functioning in daily life activities. As such, the EMFO includes classes and relationships that have to do with body function, body structures, context and environment along with feedback mechanisms that round out the description of the theoretical model. The Ecological part of this ontology's name is an acknowledgement of the contextual, environmental, and feedback components incorporated within the ontology.

2.2 NER Background

The java-nlp-Framework contains a traditional NLP pipeline based on the UIMA platform. This platform is structured as a pipeline composed of individual modules or software-annotators that perform work on input documents that decompose the text of each input file into constituent parts including tokens and sentences. The employed pipeline contains annotators that do an admirable job of identifying semi-templated components including clinically significant sections, section headings, slots and their values, questions and answers, and checkboxes.

The pipeline described in this paper includes a dictionary-lookup software-annotator that provides the bulk of the computational effort that the rest of the pipeline relies upon (see Figure 1).

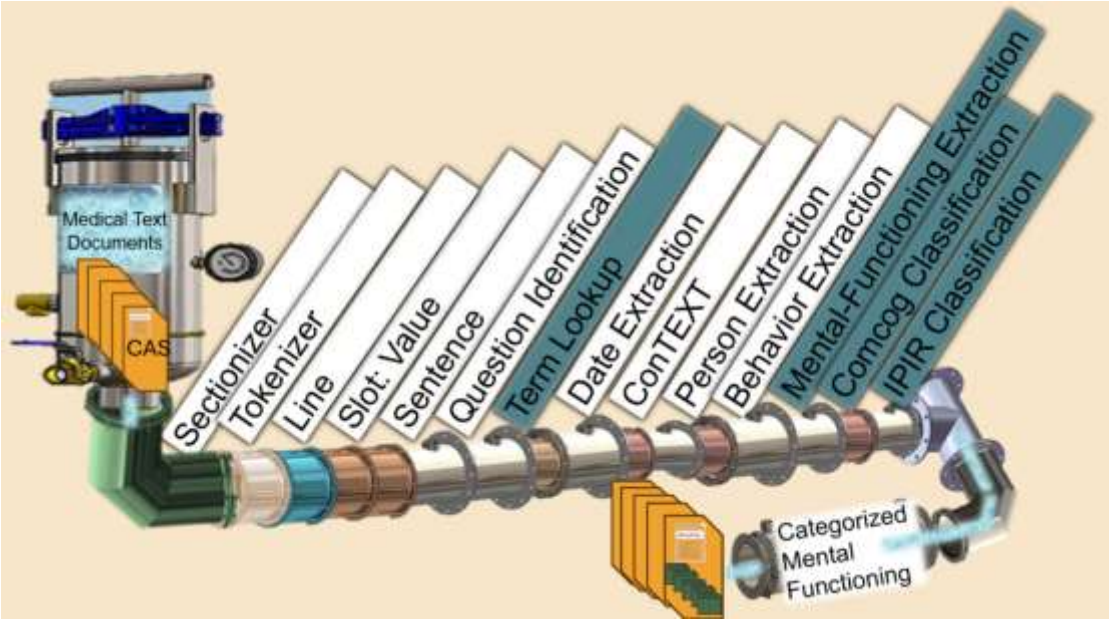

Figure 1: *Mental-functioning* NER Pipeline

This dictionary-lookup has three significant qualities worth mentioning. Dictionary entries can be multi-word keys, allowing terms like *diabetes mellitus* to be chunked together into a single unit. The dictionaries are crafted and curated for specific purposes including finding general English terms, dictionaries to support finding disease names, symptoms, proper names and the like. Each dictionary entry includes meta information that is carried along with matches, most importantly, potential semantic categories by which the entry can be classified. For example, the entry for *listening* has the category "Communication" associated with it. It is the categories associated with matched terms that downstream software-annotators utilize.

Dictionaries have been crafted and deployed in the pipeline to chunk terms that are known to not be relevant to downstream tasks, for example, a dictionary of people's names, a dictionary of place names, a role dictionary, dictionaries for behavior and emotion. Non-dictionary terms are also found using software-annotators that are regular-expression-based to find date, time and number expressions.

The NER framework takes formatted input documents from text, and General Architecture of Text Engineering [22] (GATE) types and output UIMA flavored xml, and GATE formats. For this paper, GATE formatted manual annotations were the input and both GATE and UIMA formats were output.

2.3 Manual Annotations on a Clinical Corpus

The GoSCAI corpus includes 484 de-identified clinical notes rich in activity mentions from note types that include occupational and vocational therapy (OT/VOC), physical therapy (PT), recreation therapy (RT), speech and language pathology (SLP), social work (SW), or miscellaneous (MISC) that includes psychiatry, neurology and physiatry [7].

Guidelines [10 11], first created for the SSA use case [9], were used and in some cases refined for situations that were not encountered in the SSA use case to codify how the samples get annotated. Physical and occupational therapist domain experts manually annotated according to the guidelines. The guidelines used for the ComCog and IPIR annotation tasks are published on Zenodo [10 11].

In all, 484 clinical records were manually annotated for both ComCog and IPIR including 386 used for training, and 98 for testing. As an aside, the same records were also manually annotated for the other activity related mentions including mobility and self-care and domestic life (SCDL). Each record was singly annotated for this training set, as an acceptable Inter-Rater Reliability (IRR) had already been established on the SSA use case. We did establish an IRR on this set by having the annotators doubly annotate the test set. See Table 3b for the IRR results for the ComCog sample, and Table 4b for the IRR results for the IPIR set.

3 Methods

The work that is described in this paper represents the task of further classifying sentences that previously had been identified with *mental-functioning* content from ComCog and IPIR categories via leveraging terminology derived from the EMFO. This work relies on prior annotations done in an upstream process that used a modified spaCy tokenizer/sentence splitter[23] and relied on manual *mental-functioning* annotations on those sentences. The sections that follow are descriptions and specifics of the downstream annotators that identified and created the *mental-functioning* annotations.

3.1.1 Mental-functioning NER Pipeline and Annotator

The *mental-functioning* NER pipeline (see Figure 1) includes software-annotators that create evidence of behavior, emotion and support (financial, social, institutional) from the terms in these

semantic categories. This is followed in the pipeline by a software-annotator that creates *mental-functioning* evidence and *mental-functioning* mention mark-ups. The Mental-Functioning Ontology Annotator identifies and creates *mental-functioning* evidence mark-ups from terms that have semantic categories that generalize up to *mental-functioning* within the EMFO.

The basic algorithm within the annotator is over-generative, and filters are applied to weed out spurious mentions. There are ad-hoc filter annotators in the pipeline to filter out documents that include forms, where a templated text contains what looks like *mental-functioning* mentions, but are not, since the mentions are within templated text instructions to providers on how to fill out the form. Similarly, mentions that can be attributed to the author of the document, not the client or patient, are filtered out. These mentions have evidence that includes the use of pronouns such as *we* and *our* along with use of pronouns such as *you*, *your*, *his*, *her*, *she*, and *him*. These mentions are also tagged with *provider* attribution as a feature at the term and sentence level.

The Mental-Functioning Ontology software-annotator creates *ComCog_yes* and *IPIR_yes* mark-ups for sentences that contain such content for sentences that have not otherwise been filtered out. This NER can take input from previously machine-annotated documents with *ComCog_yes*, and *IPIR_yes* that were either manually annotated from machine segmented sentences or from the Bidirectional Encoder Representations from Transformer (BERT) [24] based models trained on the manual annotations. The NER can utilize these upstream broad classifications in addition to the evidence gathered to this point in the pipeline.

3.1.2 Communication & Cognition (ComCog) Software-annotator

The ComCog software-annotator looks for terms that the dictionary has categorized with a ComCog category within the span of the *ComCog_yes* sentence. When found, a category evidence markup is created, and a category markup is created for the span of the *ComCog_yes*. The ComCog Sub-Categorization annotator further sets 20 attributes to each *ComCog_yes* annotation. For each *ComCog_yes* sentence span, all the EMFO generated evidence that generalizes to ComCog activities from the Ontology, along with behavior and support evidence that cover that span, are gathered and if there is any of these found, further processed. There are 16 ComCog attributes taken from 3 chapters in the ICF activities and participation component having to do with learning and applying knowledge, general tasks and demands and communication, and 4 attributes specifically requested by the SSA use case which included adaptation, memory, pacing and persistence (see Table 1) [8]. The ComCog annotator creates evidence mark-ups for each term it deems ComCog evidence, and the sentence span with the evidence in it (that has not been filtered out by non-relevant counter evidence) is used to create one or more of the ComCog sub-category mark ups.

Table 1: *ComCog categories*

ComCog Categories ICF Codes and Code Name	ICF Code Description
d110-d129 Purposeful Sensory Experiences	Includes watching and listening and other basic senses intentionally to experience stimuli
d130-d159 Basic Learning	Acquiring language, rehearsing, acquiring information, learning to read, write or calculate
d160 Focusing Attention	Intentionally focusing on specific stimuli, such as by filtering out distracting noises
d163 Thinking	Formulating and manipulating ideas, concepts, and images
d166 Reading	Performing activities involved in the comprehension and interpretation of written language for the purpose of obtaining general knowledge or specific information
d170 Writing	Using or producing symbols or language to convey information
d172 Calculating	Performing computations by applying mathematical principles to solve problems that are described in words and producing the results
d175 Solving Problems	Finding solutions to questions or situations by identifying and analyzing issues, developing options and solutions, evaluating potential effects of solutions, and executing a chosen solution
d177 Making Decisions	Making a choice among options, implementing the choice, and evaluating the effects of the choice
d179 Applying Knowledge Other	Represent information about managing money
d210-d220 Undertaking Tasks	General aspects of carrying out single or multiple tasks
d230 Carrying Out Daily Routine	Carrying out simple or complex and coordinated actions in order to plan, manage and complete the requirements of day-to-day procedures or duties
d240 Handling Stress	Carrying out actions to manage and control the psychological demands required to carry out tasks demanding significant responsibilities and involving stress, distraction, or crises
d310-d329 Receiving Comm	Receiving spoken, nonverbal, written or sign language messages
d330-d349 Producing Communication	Speaking, non-speech vocal expression, singing, or producing non-verbal, written, or sign language
d350-d369 Conversation	Conversations, discussion, or using communication devices
SSA Variables	**Variable description**
Applied Memory	Explicit descriptions of memory ability or inability as applied to activities of communication and cognition
Adaptation	Adaptation requires evidence of modifying an activity/task to allow improved performance at that task or activity
Pacing	The speed at which one works while sustaining an activity (e.g., slow, quick, how often one needs breaks to get a task completed
Persistence	The ability to stick with a task over time (including work tasks) and sustaining an activity over a period of time for a cognitive reason

3.1.3 Non-Relevant Evidence

There is a filter to throw out mentions that are not IPIR, ComCog, or *mental-functioning* related due to some known context. For example, if the sentence has the word *invoice*, or *billing* in it, this is not an EMFO mention. There were 1389 terms marked as *NotEMFO* which include *case number*, *swelling* and *please*. These non-relevant evidence terms were garnered and added from frequency distributions of terms appearing in false-positive mentions and never appearing in true-positive or false-negative mentions.

3.1.4 ComCog and IPIR Dictionary Creation and Curation

Vocabularies from the UMLS [25] with potential of having relevant *mental-functioning* information were reviewed. Beyond the ICF [8], these included the MeSH [21], SNOMED-CT [15], Medical Dictionary for Regulatory Activities(MedDRA) [26], and Thesaurus of Psychological Index Terms (PSY) [27]. Seed concepts from each of the EMFO's 4 quadrants (i.e., input, throughput, output, and feedback) were identified to traverse through the UMLS as a whole. The UMLS was systematically traversed through descending hierarchical and non-relationships as they existed via the content within MRREL and MRHIER [28] files. Paths were followed based on UMLS concept-level relationships, rather than the more exact or accurate method of only following relationships asserted within and among the same source at the atom level. This approach was an effort to broaden our ability to extract terms at the expense of retrieving some non-relevant terms during development. Manual culling steps were taken to remove the more egregious fallacious terms afterwards. Some seed term areas were fruitful, picking up much content. Other seed terms were less fruitful, picking up terminology that is still covered only by the ICF.

3.1.4.1 VerbNet and Additional Coverage

In the initial pilot annotation tasks, the domain expert annotators were highlighting observable behaviors within the sample clinical corpus, but these functioning mentions had no coverage within any UMLS sources. They were, not surprisingly, mostly verb and adverbial phrases. VerbNet [29] is a catalog of English Verb classes, by thematic role which covers almost all known English verbs. Within VerbNet, the annotators identified 162 of the 322 classes related to *mental-functioning* and then subsequently classified each class back as either ComCog or IPIR. The annotators subsequently sub-classified the VerbNet classes with ComCog and IPIR categories, resulting in an additional dictionary used within the NER. This addition greatly increased coverage.

3.1.4.2 Terminology from the Manual Annotations

The dictionaries were augmented and tuned by utilizing terms within the manual annotations which had no coverage in the initial dictionaries, (i.e., no UMLS coverage) using only terms discovered in the training sets, in particular, discovered from a review of the NER's false-negatives. These annotations gave us the opportunity to further curate the dictionaries by culling terms which spuriously identified *mental-functioning* from a review of the high-frequency false-positives in the training set.

3.1.5 Interpersonal Interactions and Relationships (IPIR)

The IPIR Sub-Categorization annotator further sets 8 attributes to each *IPIR_yes*. For each *IPIR_yes* sentence span, all the EMFO generated evidence that generalizes to IPIR Activities from the Ontology, along with IPIR Participation evidence, Behavior and Support evidence that cover that span are gathered, and if there is any of these found, further processing is done (see table 2) [30].

Table 2: *IPIR Categories*

IPIR Categories ICF Code and Code Names	**ICF Code Description**
d710-d729 General Interpersonal Interactions	Interacting, maintaining and managing interactions with people in a contextually and socially appropriate manner, such as by showing consideration and esteem when appropriate, or responding to the feelings of others as well as, as by regulating emotions and impulses, controlling verbal and physical aggression, acting independently in social interactions, and acting in accordance with social rules and conventions
d730 Relating with Strangers	Engaging in temporary contacts and links with strangers for specific purposes, when asking for directions or other information, or making a purchase
d740 Formal Relationships	Creating and maintaining specific relationships in formal settings
d7400 Relating with Persons in Authority	Creating and maintaining formal relations with people in positions of power of a higher rank or prestige relative to one's own position
d750 Informal Social Relationships	Entering into relationships with others, such as casual relationships with people living in the same community or residence, or with co-workers, students, playmates, people with similar backgrounds or professions.
d760 Family Relationships	Creating and maintaining kinship relationships, such as those with members of the nuclear family, extended family, foster and adopted family and step-relationships, more distant relationships such as second cousins, or legal guardians.
d770 Intimate Relationships	Creating and maintaining close or romantic relationships between individuals, such as husband and wife, lovers or sexual partners
d779 IPIR, other	General relationships where no particular relationship can be specified such as “others”, “other people”, “crowds”, “someone”, etc.

The IPIR annotator creates evidence mark-ups for each term it deems IPIR evidence, and the sentence span with the evidence in it (that has not been filtered out by non-relevant counter evidence) is used to create one or more of the IPIR sub-category mark-ups.

3.1.6 IPIR Heuristics

The pipeline includes software-annotators for each of the IPIR categories. Each software-annotator

works roughly the same, where the terms of each sentence of each section are iterated through accumulating evidence to be acted upon for each sentence and each section. Below are specifics about what evidence each software-annotator is looking for and triggered on.

ICF d710-d729 General Interpersonal Interaction mark-ups are created when there is emotion evidence present. Interaction mentions are also created when the sentence span included person evidence and behavior evidence.

ICF d730 Relating with Stranger mark-ups are created when there is evidence of stranger relationship evidence. These kinds of terms come from a dictionary of 100 terms like *shopper*, *public_places*, and *visitor*. The bulk of these terms were garnered from examples seen in the data.

ICF d740 and d7400 Formal Relationship mark-ups are created when terms categorized as non-authority or sometimes-authority are present. These kinds of terms come from a person-role and title dictionary that was augmented with non-authority, sometimes-authority and authority-position semantic categories, garnered from the Bureau of Labor Statistics and other public sources [31-33]. Those terms categorized as authority-position had d7400 mentions made.

ICF d750 Informal Social Relationship mark-ups are created when there is informal relationship evidence found in the span of each *IPIR_yes* sentence. If there is no such evidence, but there is the presence of pronouns, an informal social relationship markup was made. This works only because we already know this is an *IPIR_yes* sentence from an upstream software-annotator, so cases where the pronouns relate to possessions rather than to other people would have already been filtered out. The UMLS terms categorized with family history and person roles were the impetus for the informal relationships dictionary employed here.

ICF d760 Family Relationship mark-ups are created when there is family history evidence found in the span of the *IPIR_yes* sentence. The underlying terminology used here are the terms in the UMLS categorized with their Family-History semantic type.

ICF d770 Intimate Relationship mark-ups are created when there is intimate relationship evidence found in the span of the *IPIR_yes* sentence. A small dictionary of 130 terms was created mostly from the guideline examples, observations, introspection and use of thesauri to garner adjectives and nouns that indicate an intimacy. Such terms include *significant other*, *old lady*, and *inseparable*.

ICF d779 Particular interpersonal relationships, other specified and unspecified mark-ups are also created when there is "other" relationship evidence found in the span of *IPIR_yes* sentence. There were only 30 or so terms identified with this tag, which included *others*, *nobody*, and *withdrawn*. The guidelines and training data were the source for this terminology.

4 Results

The annotations and efficacy were done at the sentence level: See table 3a for the ComCog test

efficacy. Table 3b shows the inter-rater reliability for each category. See table 4a for the IPIR test efficacy, 4b for the IRR. Peach cell backgrounds highlight categories that were seen in the IRR and adequate training examples.

Table 3: *ComCog Category Efficacy*

	Table 3a: GoSCAI ComCog Test Sample Category Results				**Table 3b: GoSCAI ComCog IRR Sample Category Results**	
ComCog Categories	**F-Score**	**Recall**	**Precision**	**%Support**	**F-Score**	**% Support**
d110-d129	0.762	0.667	0.889	0.57	0.500	1.02
d130-d159	0.426	0.417	0.435	2.24	0.225	1.72
d160	0.773	0.680	0.895	1.72	0.710	1.41
d163	0.535	0.472	0.616	8.82	0.598	9.62
d166	0.462	0.750	0.333	0.39	0.300	0.27
d170	0.000	0.000	0.000	0.13	0.500	0.12
d172	0.500	0.500	0.500	0.03	0.250	0.16
d175	0.133	0.083	0.333	0.37	0.350	0.86
d177	0.617	0.673	0.569	10.33	0.553	9.73
d179	0.286	0.250	0.333	0.10	0.250	0.23
d210 d220	0.492	0.565	0.435	8.08	0.451	10.71
d230	0.200	0.167	0.250	0.54	0.174	1.45
d240	0.764	0.778	0.750	3.20	0.470	3.13
d310-d329	0.628	0.670	0.591	6.59	0.625	6.22
d330-d349	0.870	0.919	0.827	49.56	0.778	41.71
d350-d369	0.396	0.528	0.317	2.00	0.428	2.50
Applied Memory	0.577	0.508	0.667	3.20	0.381	5.36
Adaptation	0.400	0.437	0.368	0.76	0.172	01.76
Pacing	0.000	0.000	0.000	0.03	0.00	0.04
Persistence	0.452	0.583	0.368	1.32	0.077	1.99
Micro Avg	0.463	0.482	0.474		0.647	
Macro Avg	0.628	0.660	0.607		0.389	
Weighted Avg	0.717					
ComCog Extraction						
ComCog_yes	Not the subject of this paper				.80	

Table 4*: GoSCAI IPIR Categorization Efficacy*

	Table 4a: GoSCAI IPIR Test Sample Category Results				Table 4b: GoSCAI IPIR IRR Sample Category Results	
IPIR Categories	**F-Score**	**Precision**	**Recall**	**% Support**	**F-Score**	**% Support**
d710-d729 General Interpersonal Interactions	0.690	0.681	0.698	22.20	0.310	22.89
d730 Relating with Strangers	0.667	0.500	1.00	00.00	0.000	00.34
d740 Formal Relationships	0.696	0.692	0.700	16.47	0.289	17.32
d7400 Relating with Persons in Authority	0.111	0.067	0.333	00.26	0.250	00.30
d750 Informal Social Relationships	0.702	0.664	0.744	14.98	0.430	10.75
d760 Family Relationships	0.808	0.781	0.838	27.90	0.461	28.28
d770 Intimate Relationships	0.859	0.867	0.850	13.63	0.405	11.80
d779 IPIR, other	0.416	0.301	0.672	04.56	0.117	08.32
Micro Avg	0.712				0.394	
Macro Avg	0.619	0.569	0.729		0.283	
Weighted Avg	0.726	0.664	0.750			
IPIR Extraction						
	Not the subject of this paper				.699	

5 Error Analysis

Among the common failures were semi-templated formats, mostly in the form of slot: value constructions where the sentence segmentor was inconsistent with either chunking too much as one sentence, inadvertently scooping up whole sections as a sentence to be classified or splitting up the slot or question part as one sentence, and the value or answer part in a second sentence. This created situations where annotators marked sentences consisting of "Yes", because it was the answer to a relevant question posed as the slot in the prior sentence.

Similarly, relevant questions with no or negated answers as the next sentence also were the sources of failure, where the NER would identify the question as relevant when it was not because it had no answer. There were a number of templated recreation therapy notes in this corpus that turned out to be a source of inconsistency among the domain experts for the ComCog classification task. These were so prevalent that we reviewed several of the common templates from the training set, had the annotators adjudicate, and based on their updated decisions had all the instances of those templates across the entire corpus updated.

There were multiple categories that had little or no representation either in the initial inter-annotator agreement set, training set, and/or the test set. This raised issues when categorized mentions showed up in the test set, but the category did not show up in the training set. We did not have the resources to re-sample to augment the corpus with examples from the non-majority class categories to improve the models.

As the task to classify mentions already known to be a ComCog or IPIR mention, an answer was required. There were cases where none of the model annotators produced an answer. In such cases, the *d779 IPIR, other category* was chosen for IPIR mentions that could not otherwise be categorized and the majority class *d330-d349* chosen for ComCog mentions that it could not otherwise be categorized. This strategy worked well for ComCog, but not for IPIR. An informal review of the *d779* false positives indicated almost 80% were due to not being otherwise classified. Much of the false negative failures of all the IPIR models are hidden within the *d779* statistic.

6 Discussion

The ComCog Categorization micro average F-Score of 0.463 in table 3A reveals that there were many categories where the application performed predictably poorly. This can be explained by not enough training as seen via the percent of support or proportion of the training examples devoted to each category compounded with the IRR performance seen in table 3b, indicating that the training is likely to be inconsistent. The macro average F-Score, 0.628, tells the story that the good performance within the 5 majority classes pull the score up, and the weighted average F-Score of 0.717, proportionally weights the F-Score by the training exposure in each class.

The IPIR categorization micro average F-Score of 0.712 in table 4a displays a robustness of using a dictionary-based approach despite having mediocre to poor IRR across the board. Three of the 8 categories that had disproportionately small support and/or IRR for the IPIR categorization task, explaining why the macro average F-Score was lower at 0.619 due to not having an overwhelming majority class, but where taking into account the weighting by support, the weighted average F-Score was decent 0.726.

The larger task was to identify ComCog and IPIR mentions in clinical text, which is being done outside the scope of this paper. Given already identified mentions, this task reports on the efficacy to identify them. The choices made for inclusion in the corpus centered around ComCog and IPIR mentions without delving into sampling around the categories. The annotator training centered around mutually identifying ComCog and IPIR at an acceptable rate of .7 average F-Score seen in table 3b and 4b before proceeding to further classifying. Due to resource constraints, we did not initially look at or further train the annotators until each category had an acceptable IRR. The IRR results indicate that the annotators were able to identify mental-functioning mentions, but the task of further classifying them is more problematic. That being said, for IPIR, the dictionary, rule-based approach performed admirably above the IRR levels, and for the ComCog classifiers, acceptably for most of the categories.

7 Future Work

Going forward, the rule-based system could be further tuned by augmenting with new training data to include at least a minimum number of mentions within each category to train from, and a feedback task to review output from the NER so as to alter mentions that were suggested but got missed to weed out annotator inconsistencies that crop up.

8 Conclusions

Mental-functioning extracted and characterized from clinical documentation is a valuable source of

functional data for health and social care purposes such as disability determination, health research, health outcomes evaluation, and it is information that clinicians express interest in having. However, mental-functioning activity information can be challenging to identify in health records due to the over emphasis of documentation at the mental body function level.

The EMFO was developed to aid in accurately finding these mentions through semantic information that distinguishes mental-function from evidence of *mental-functioning*. A rule-based NER was developed from an existing codebase, leveraging the EMFO to extract and classify *mental-functioning* mentions. This NER was first developed and benchmarked using OCR'd SSA claimant records and then re-tuned and distributed on the (available) GoSCAI corpus. The weighted average ComCog Subclassification F-Score on this set is 0.717. The weighted average IPIR subclassification F-Score was 0.726. The benchmarked NER is one part in a series of tasks to provide the informatics community with resources related to functioning.

9 Acknowledgements

9.1.1 Author contributions

GD oversaw the development of the EMFO and NER. BD oversees the deliverables and provided analytics. MS developed the EMMF and the EMFO, is one of the domain experts, and an annotator of this corpus. KC developed the EMFO, is one of the domain experts and an annotator. RJS did the sampling, set up the annotation tasks, provided statistics, and was a domain expert and an annotator of this corpus. RP is a domain expert and annotator of this corpus. ER is the head of this branch, is a domain expert, and provided the framing for this task.

9.1.2 Supplementary material

9.1.3 Funding

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH). The contributions of the NIH author(s) were made as part of their official duties as NIH federal employees, are in compliance with agency policy requirements, and are considered Works of the United States Government. However, the findings and conclusions presented in this paper are those of the author(s) and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services.

9.1.4 Conflicts of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

9.1.5 Software Availability

The NER and EMFO ontology are available on GitHub at: https://github.com/CC-RMD-EpiBio/EcologicalMentalFunctioningOntology

9.1.6 Data Availability

The GoSCAI corpus description is available from zenodo at: https://zenodo.org/records/15528545. The dataset is restricted to those researchers who request and receive permission to use the data

and agree to the NIH Clinical Center's terms of use.

10 References

1. Goldman HH, Porcino J, Divita G, Goldman HH, Porcino J, Divita G, Zirikly A, Desmet B, Sacco M, Marfeo E, McDonough C, Rasch E, Chan L. Informatics research on mental health functioning: decision support for the social security administration disability program. Psychiatric Services 2023;74(1):56-62.
2. Chen M, Tan X, Padman R. Social determinants of health in electronic health records and their impact on analysis and risk prediction: a systematic review. Journal of the American Medical Informatics Association 2020;27(11):1764-73
3. Zirikly A, Desmet B, Newman-Griffis D, et al. Information extraction framework for disability determination using a mental functioning use-case. JMIR Medical Informatics 2022;**10**(3):e32245.
4. Divita G, et al. The Ecological Mental Functioning Ontology. 2024. https://github.com/CC-RMD-EpiBio/EcologicalMentalFunctioningOntology/.
5. Sacco MJ, Divita G, Rasch E. Development of an ontology to characterize mental functioning. Disability and rehabilitation 2024;**46**(16):3739-48.
6. Divita G, Sacco M, Coale K, Parks R, Rasch E. Building an ecological mental functioning ontology: An Informatics Perspective. Proceedings of the International Conference on Biomedical Ontologies 2024; Enschede, The Netherlands.
7. A Gold Standard Corpus for Activity Information (GoSCAI). Zenodo 2025 doi: 10.5281.
8. WHO. *International Classification of Functioning, Disability and Health*. Geneva: World Health Organization. 2001
9. Divita G, Desmet B, Sacco MJ, Coale K, Silva RJ, Parks R, Rasch E. Classifying Mental-functioning using a Named Entity Recognition Tool: Based on the Ecological Mental-functioning Model Ontology. Rehabilitation Medicine Department, National Institutes of Health Clinical Center, Bethesda, MD, 2025. https://github.com/CC-RMD-EpiBio/EcologicalMentalFunctioningOntology/blob/main/publications/OntologyNERPaper.pdf
10. NIH CC RMD Epidemiology & Biostatistics Section. Annotation Guideline for Free-Text Activity Functioning Information: Communication & Cognition. Secondary Annotation Guideline for Free-Text Activity Functioning Information: Communication & Cognition 2025a. https://zenodo.org/records/13910167.
11. NIH CC RMD Epidemiology & Biostatistics Section. Annotation Guideline for Free-Text Activity Functioning Information: Interpersonal Interactions & Relationships. Secondary Annotation Guideline for Free-Text Activity Functioning Information: Interpersonal Interactions & Relationships 2025b. https://zenodo.org/records/13774684.
12. Divita G, Coale K, Maldonado JC, Silva RJ, Rasch E. Extracting body function information using rule-based methods: Highlighting structure and formatting challenges in clinical text. Frontiers in Digital Health 2022;**4**:914171.
13. Divita G, Carter ME, Tran L-T, et al. v3NLP framework: tools to build applications for extracting concepts from clinical text. eGEMs 2016;**4**(3):1228.
14. Divita G, Zeng QT, Gundlapalli AV, Duvall S, Nebeker J, Samore MH. Sophia: a expedient UMLS concept extraction annotator. AMIA Annual Symposium Proceedings; 2014, p. 467.

15. Divita G, Luo G, Tran L-TT, Workman TE, Gundlapalli AV, Samore MH. General symptom extraction from VA electronic medical notes. MEDINFO 2017: Precision Healthcare through Informatics: IOS Press, 2017:356-60.
16. Divita G, Brignone E, Carter ME, et al. Extracting sexual trauma mentions from electronic medical notes using natural language processing. MEDINFO 2017: Precision Healthcare through Informatics: IOS Press, 2017:351-55.
17. Ferrucci D, Lally A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. Natural Language Engineering 2004;**10**(3-4):327-48.
18. Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association 2010;**17**(3):229-36.
19. Divita G, Tse T, & Roth L. Failure analysis of MetaMap transfer (MMTx). MEDINFO 2004; 2004. IOS Press.
20. Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. Studies in health technology and informatics 2006;**121**:279.
21. Lipscomb CE. Medical subject headings (MeSH). Bulletin of the Medical Library Association 2000;88(3):265.
22. Cunningham H. GATE, a general architecture for text engineering. Computers and the Humanities 2002;36:223-54.
23. Vasiliev Y. *Natural language processing with Python and spaCy: A practical introduction*: No Starch Press, 2020.
24. Koroteev MV. BERT: a review of applications in natural language processing and understanding. arXiv preprint arXiv:2103.11943 2021.
25. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research 2004;**32**(suppl_1):D267-D70.
26. Jaasu NM, Kamaraj R, Seetharaman R. MedDRA (medical dictionary for regulatory activities). Research journal of pharmacy and technology 2018;**11**(10):4751-54.
27. Gallagher LA. Thesaurus of psychological index terms. (No Title) 2005.
28. NLM. UMLS Reference Manual [Internet]. Secondary UMLS Reference Manual [Internet]. https://www.ncbi.nlm.nih.gov/books/NBK9685/.
29. Palmer M, Bonial C, Hwang JD. 17 VerbNet: Capturing English Verb Behavior, Meaning, and Usage. The Oxford handbook of cognitive science 2016:315.
30. Organization WH. *International Classification of Functioning, Disability, and Health: Children & Youth Version: ICF-CY*: World Health Organization, 2007.
31. Job titles for social workers. Secondary Job titles for social workers. https://www.wcupa.edu/education-socialwork/gradSocialWork/documents/socialworkjobtitles.pdf.
32. U.S. Bureau of Labor Statistics 2010 Census Occupational Titles and Code List. Secondary U.S. Bureau of Labor Statistics 2010 Census Occupational Titles and Code List 2010. https://www.bls.gov/cps/cenocc2010.htm.
33. ONGIG Job Titles: The Difinitive Guide. Secondary ONGIG Job Titles: The Difinitive Guide 2025. https://www.ongig.com/job-titles#/#hierarchy.