Lead Disease Matching for the SEC POC

NCI SEC POC Team

10/21/2020

Lead Disease Matching Algorithm

Let D_{ALL} be the set of all trials with the following attributes:

Attribute	Value
current_trial_status	active
primary_purpose.primary_purpose_code	treatment or screening
sites.recruitment_status	ACTIVE
record_verification_date_gte	two years prior to today's date

Let P be the set of NCIt codes expressed for a given potential study participant.

Let D_{API} be the set of trials returned by the CTAPI matching the diseases in P. Note that $D_{API} \subseteq D_{ALL}$.

Let MT be a relation that maps a C code into a set of maintypes. The domain of MT is the set of all C codes. The range of MT is the set of maintypes. Note that MT may yield an empty set or a set of one or more maintypes.

Let c be a NCIt C code. Traverse the NCIt digraph towards the root from c. Let l be the minimum distance between c and any maintype encountered during this traversal. The set returned by MT is the set of maintypes that are exactly distance l in the NCIt digraph from c.

Let

$$P_{mt} = \cup_{c \in P} MT(c)$$

That is, P_{mt} is the union of the set of maintypes computed from the set of potential study participant data.

Now consider a clinical trial $T \in D_{API}$. Let TD_{TL} be the set of c codes for the diseases associated with T with the following attributes :

Attribute	Value
inclusion_indicator	TRIAL
lead_disease_indicator	YES

Let

$$T_{mt} = \cup_{d \in TD_{TL}} MT(d)$$

.

That is T_{mt} is the union of the set of maintypes computed from the set of TRIAL level lead diseases.

A lead disease match is obtained for a given trial if

$$T_{mt} \cap P_{mt} \neq \emptyset$$

That is, there is at least one common maintype in the set of maintypes for the trial and the potential study participant data.