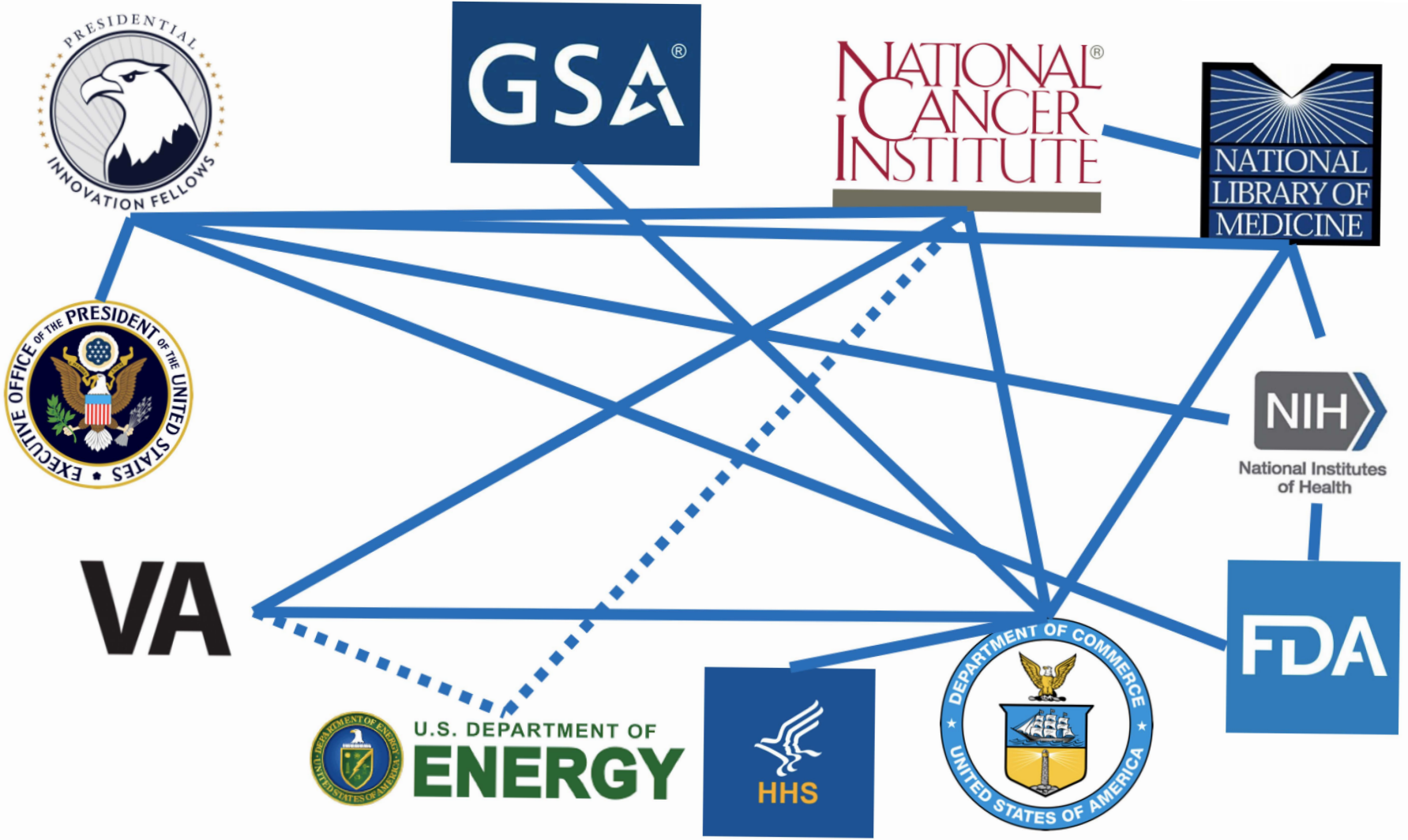
Natural Language Processing (NLP) at the Agency Interface

MLAI R&D

Gil Alterovitz, PhD, FACMI and
Justin Koufopoulos
Presidential Innovation Fellows

Use Case: Make finding clinical trials easy for patients

Approach: Reducing clinical trial recruitment burden by making data more machine-readable and leveraging NLP


PRESIDENTIAL
INNOVATION FELLOWS
GSA
NATIONAL CANCER INSTITUTE
NATIONAL LIBRARY OF MEDICINE
EXECUTIVE OFFICE OF THE PRESIDENT OF THE UNITED STATES
NIH
National Institutes of Health
VA
FDA
U.S. DEPARTMENT OF ENERGY
HHS
DEPARTMENT OF COMMERCE
UNITED STATES OF AMERICA

Move from one-way communication to two-way, interactive and iterative conversation

NIH and FDA Request for Public Comment on Draft Clinical Trial Protocol Template for Phase 2 and 3 IND/IDE Studies

Notice Number: NOT-OD-16-043

Key Dates

Release Date: March 17, 2016
Response Date: April 17, 2016

Related Announcements

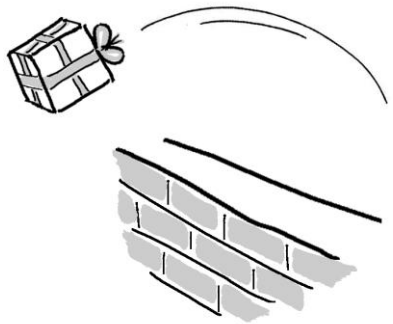
NOT-OD-17-064

Issued by

National Institutes of Health (NIH)
U.S. Food and Drug Administration (FDA)

Purpose

Background



Request for Information (RFI)

Listening sessions

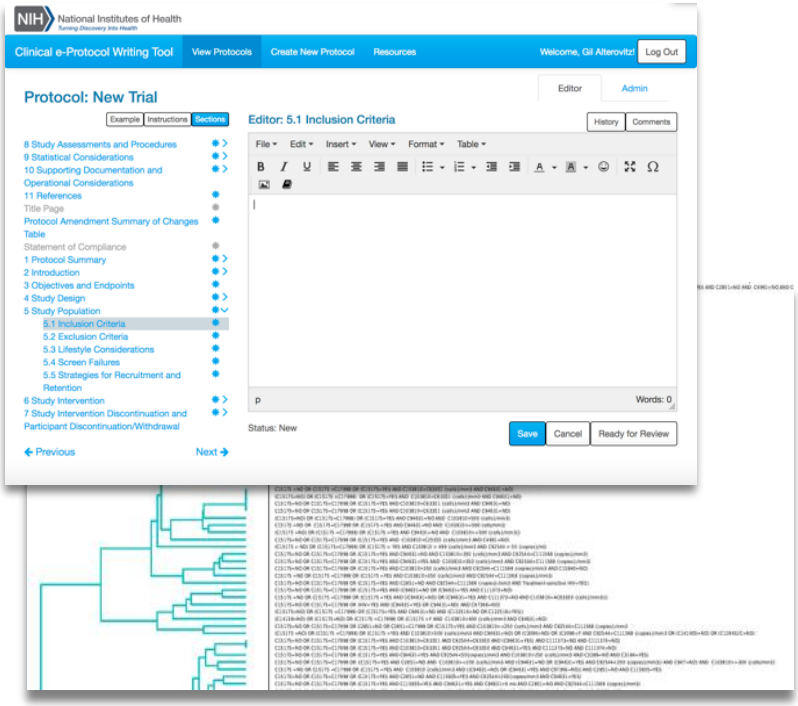
Consider how policies, particularly in science and technology can be piloted quickly and efficiently

Roundtables

Iterative feedback

Collaborative Pilot Program

Consider how policies, particularly in science and technology can be piloted quickly and efficiently



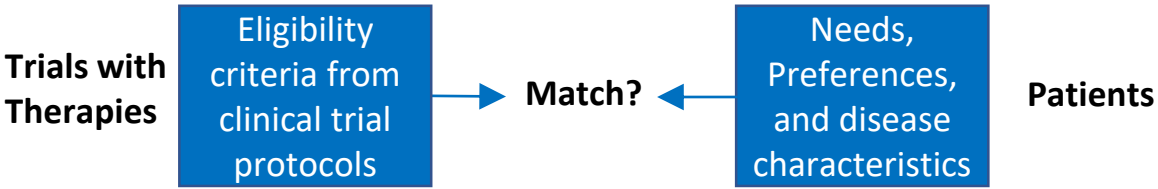
Using AI to incentivize good practice in clinical trials

Bringing together government (e.g. VA, GSA, NIH, FDA), Academia, Industry

Structuring Criteria Using Terminology and Boolean Logic

- **Data Set 1: Eligibility Criteria**
  - Source: Clinical Trials Reporting Program) CTRP trials from 342 trials open to accrual Oct. 2017
  - Type: NCI Enterprise Vocabulary Services (EVS) coded data, Boolean expressions
  - HIV, Hemoglobin, Platelets, and White blood cell count

- **Data Set 2: Participants Data**
  - Source: 100 participants based on contact center phone calls
  - Type: EVS coded data (Boolean)
  - Cancer Site (Primary), Cancer Type, Stage, Cell Type, Grade, State/Institution/City, ZIP Code, Type of Trial, Treatment History



Need clarity and uniformity to facilitate NLP/AI

- Eligibility language is not structured
- Eligibility language is complex: text can be confusing for humans (and machines)
- Eligibility language is not standardized

Inclusion Criteria:

- For pre-surgical patients
- Suspected diagnosis of resectable non-small cell lung cancer; cancers with a histology of "adenosquamous" are considered a type of adenocarcinoma and thus a "nonsquamous" histology; patients with squamous cell carcinoma are eligible only if the registering site has EA5142 Institutional Review Board (IRB) approved
- Suspected clinical stage of IIIA, II (IIA or IIB) or large IB (defined as size >= 4cm); Note: IB tumors < 4cm are NOT eligible; stage IB cancer based on pleural invasion is not eligible unless the tumor size is >= 4cm
- For post-surgical patients
- Completely resected non-small cell lung cancer with negative margins (R0); patients with squamous cell carcinoma are eligible only if the registering site has EA5142 IRB approved
- Pathologic stage IIIA, II (IIA or IIB) or large IB (defined as size >= 4 cm); Note: IB tumors < 4cm are NOT eligible; stage IB cancer based on pleural invasion is not eligible unless the tumor size is >= 4cm
- Eastern Cooperative Oncology Group (ECOG) performance status 0-1
- No patients who have received neoadjuvant therapy (chemo- or radio-therapy) for this lung cancer
- No locally advanced or metastatic cancer requiring systemic therapy within 5 years prior to registration; no secondary primary lung cancer diagnosed concurrently or within 2 year prior to registration
- No prior treatment with agents targeting EGFR mutation, ALK rearrangement, and PD-1/PD-L1/CTLA-4
- No patients known to be pregnant or lactating
- Patients who have had local genotyping are eligible, regardless of the local result
- No patients with recurrence of lung cancer after prior resection
- Note: Post-surgical patients should proceed to registration immediately following preregistration
- Completely resected NSCLC with negative margins (R0); cancers with a histology of "adenosquamous" are considered a type of adenocarcinoma and thus a "nonsquamous" histology; patients with squamous cell carcinoma are eligible only if the registering site has EA5142 IRB approved

Structuring Eligibility – Example: Platelets with NCI Thesaurus Terms

- **Encode With Logic**
  - Platelet count greater than or equal to 100,000 platelets per cubic millimeter
  - Platelet ≥ 100,000/mm^3
  - NCIt Code: C51951 ≥ 100,000 UCUM Code: /mm^3)
- **Criteria Fill-in Template**
  - Platelet count greater than or equal to _100,000__ platelets per cubic millimeter ___without_transfusion___.

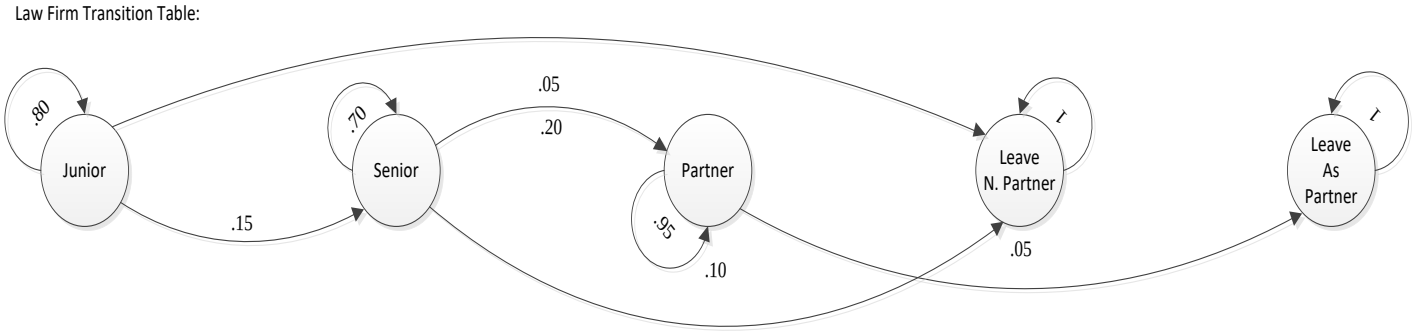
Potential Uses for AI/NLP in Patient/Trial Matching

1. **Matching patients to trials:** Match patients to trials based on eligibility criteria/patient information (and vice versa).
2. **Standardized protocol templates:** Take text (and/or structured eligibility criteria) and cluster to find templates that can be used to in future to standardize protocols when they are initially written for a trial.
3. **Curator assistant:** Use NLP to produce suggested structured ontological concepts and logic for curators to consider when doing manual curation (e.g. for Clinical Trials Reporting Program)
4. **Patient data:** Use NLP to extract concepts form patient data that could be used for matching to trials.
5. **Chatbots**: Use speech recognition, voice synthesis, and NLP for protocol text/structured eligibility criteria to engaging patients looking for trials.

How can we incentivize standardization via NLP?

- **Challenges:**
  - Field is evolving (new treatments)
  - There is no incentive for writers of protocols to structured language
  - Doing structuring after protocols are submitted is hard (may not know original meaning of the writer). Hard to scale curation for all criteria
  - NLP/AI not perfect for predicting structure
- **Potential Solution:**
  - Use AI/NLP to present templates with data-based consensus options (based on text and frequency of already used in trials database so far) at the Point-of-need (i.e. when originally writing the trial).
  - Incentivize via 1) templates to make writing protocols faster and 2) consensus scoring to encourage standardization.
  - Create database for standardized phrases/criteria and variants (which can be more easily structured)
- **Potential Outcome:**
  - This gradually leads to standardization in writing protocols for most criteria, thereby making curation/structuring much more scalable (since only have to do a few standard cases as well as much fewer special cases).

Incentivization for Standardization Framework: Absorption Markov Chain via Consensus Score



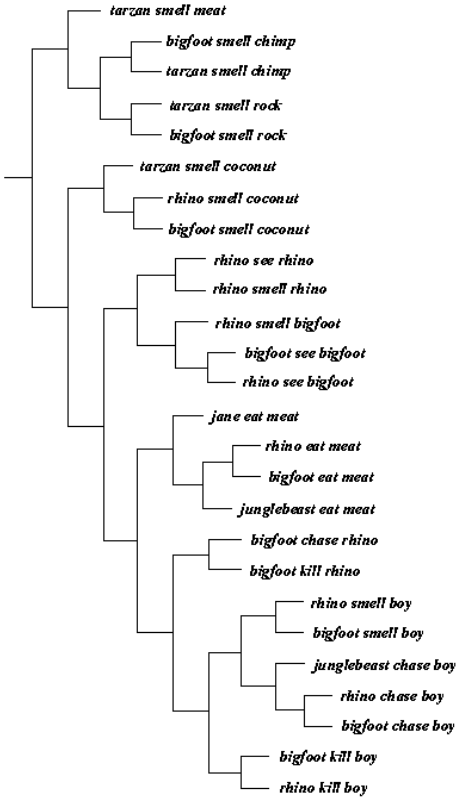
Absorption Markov Chain:
After n->infinite transitions on transition matrix w will tend to reach an absorption state.

$$P = \begin{bmatrix} .80 & .15 & 0 & .05 & 0 \\ 0 & .70 & .20 & .10 & 0 \\ 0 & 0 & .95 & 0 & .05 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Law Firm Example:
Initial state: Junior
Absorption states: Leave without getting partner. Leave as partner.

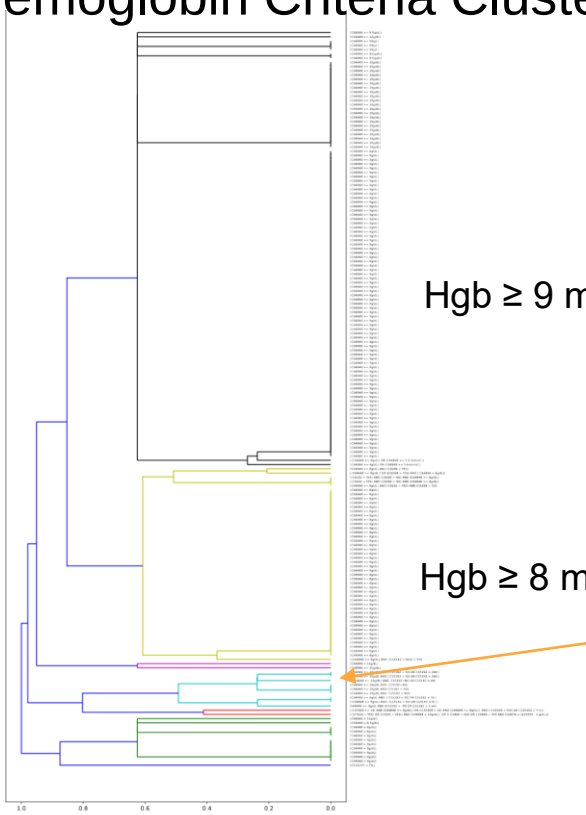
NIH Data Science Hackathon Goal:
Toward Structured Protocol Authoring Standards and Templates

- AI for Data-based Protocol Standardization
- Load in criteria text: e.g. all 342 Cancer Therapy Evaluation Program (CTEP) trials.
- Cluster based on similarity (e.g. semantics/entity/logical expression) to find patterns.
- Extract popular templates with options for curators to select from (with frequency of popularity).



Unsupervised Learning: Clustering Logic Statements

Hemoglobin Criteria Clusters



Hgb ≥ 9 mg/dL

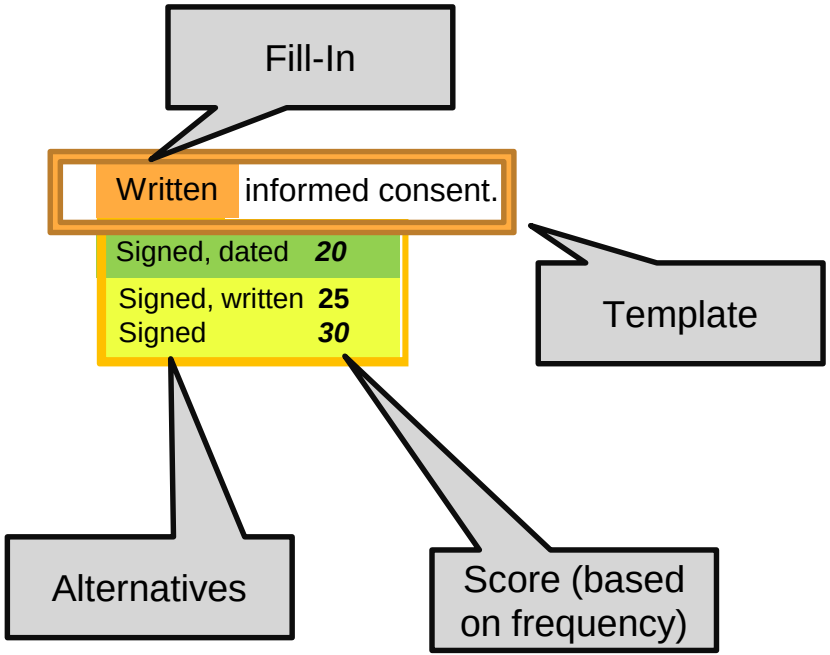
Hgb ≥ 8 mg/dL

Additional constraints:
No transfusion in past 28 days

Hemoglobin Criteria Frequencies

Cutoff	Proportion (%)
9 g/dl	49.71%
8 g/dl	29.71%
10 g/dl	17.14%
12 g/dl	0.57%
.009 g/dl (9 mg/dl) **(TYPO?)**	1.14%
11 g/dl	1.14%
8.5 g/dl	0.57%

Standardizing via Alternatives and Consensus Score: Change in Phrasing/Semantics



Find Template in Source Text:

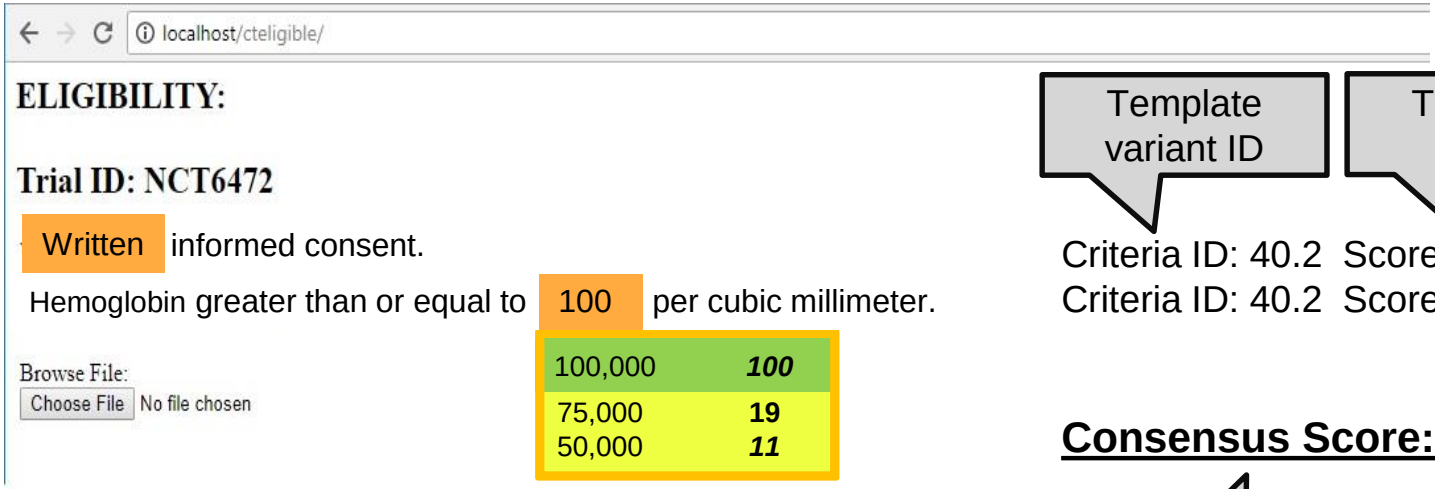
Signed, dated informed consent.
Signed, written informed consent.
Written informed consent.
…

Template:
_____ informed consent

Output Fill-ins/Score:

Signed, dated / 20
Signed, written / 25
Written / 100

Standardizing via Alternatives and Consensus Score: Change in Phrasing/Semantics

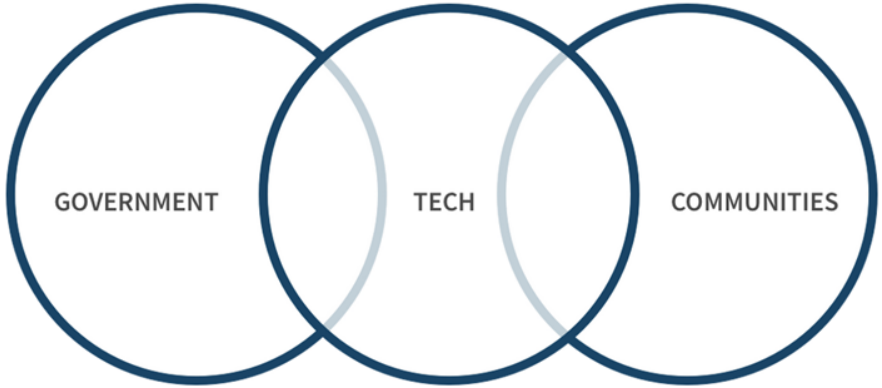


Iterative approach

- Collect relevant datasets and code, then funnel through a 12-week (starting Sept.), user-centered agile sprint process to facilitate AI/NLP for an experimental therapy ecosystem.
- Demo final tools and showcase data, code on code.gov and other .gov websites.

What is The Opportunity Project?

The Opportunity Project is a process for engaging government, communities, and the technology industry to create digital tools that address our greatest challenges as a nation. This process helps to empower people with technology, make government data more accessible and user-friendly, and facilitate cross-sector collaboration to build new digital solutions.



Summary

- Created datasets so can compare different patient-clinical trial matching approaches “apple to apple.”
- Clustered results based on Boolean logic and free text with NLP
- Found fill-in templates via leveraging Boolean logic for released criteria data and for free text by criteria
- Designed framework for incentivizing standardization in criteria using absorption Markov chain approach via consensus score for protocols
- Created proof of concept web interface for entering criteria and getting data-based suggestions
- Resulted in rectifying issues/suggested changes to: business processes, database API/code, and underlying protocol data.
- Now preparing to launch opportunity project to iterate and extend current work.

Acknowledgements

- Sheila Prindiville
- Gisele Sarosy
- Samantha Finstad
- Shahin Assefnia
- David Loose
- Neesha Desai
- Sam Isa
- Nina Bianchi
- Peter Garrett
- Nelvis Castro
- Mary Anne Bright
- Candace Deaton Maynard
- Tony Kerlavage
- David Patton
- Brent Coffey
- Anna Steinfeld
- Jerry Sheehan
- Joshua Di Frances
- Paul Fearn
- Justin Koufopoulos
- Drew Zachary
- Kristen Honey
- And many more…

Thanks!

Contact:
gil.alterovitz@pif.gov and justin.koufopoulos@pif.gov

"Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Networking and Information Technology Research and Development Program."

The Networking and Information Technology Research and Development (NITRD) Program

Mailing Address: NCO/NITRD, 2415 Eisenhower Avenue, Alexandria, VA 22314

Physical Address: 490 L'Enfant Plaza SW, Suite 8001, Washington, DC 20024, USA Tel: 202-459-9674, Fax: 202-459-9673, Email: nco@nitrd.gov, Website: https://www.nitrd.gov