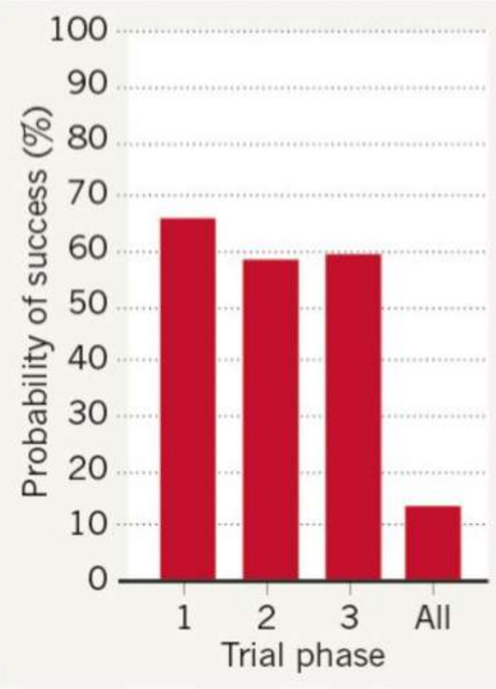
EFFICIENT AND GENERALIZABLE MODEL FOR CLINICAL TRIAL COHORT SELECTION

Cohort selection is labor intensive and time consuming

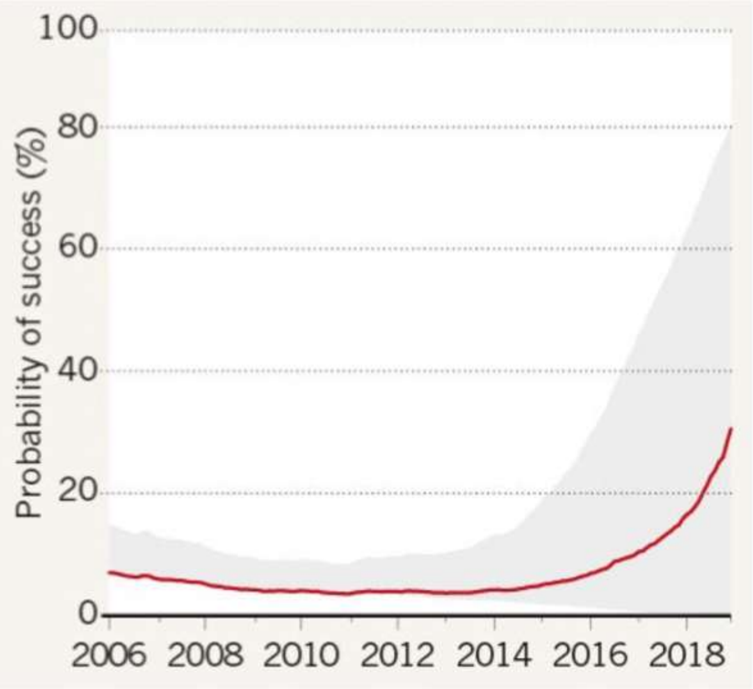
Clinical trials are the most risky and expensive part of the drug development process. More than 85% of the drugs fail during clinical trials and when a drug candidate fails in stage 3, the loss could be over $1B. While poor trial design is one reason they could fail, insufficient volunteer patient participation is another major reason. Patients could drop out during a trial for various reasons, causing the statistical power to drop below the minimum.

Furthermore, improper cohort selection and/or lower recruitment of subjects contribute to failure of meeting the pre-defined primary and secondary endpoints, resulting in rejections or suggestions to further modify the trial design by the FDA regulatory team. In other cases, a trial team could spend years and not enroll enough participants.

In a National Cancer Institute's 2016 study, between 2000 and 2011, 18% of trials found fewer than half the number of patients they were seeking after three or more years of search (Bennette et al., 2016).



Sources: ref 1



Sources: go.nature.com/2INT6SV

A 2019 study shows that from 2000 to 2015, only 13.8% of drug candidates successfully pass through all three stages of clinical trials.

Typical way researchers find clinical trials is by searching structured data, e.g. diagnosis codes, dates of exposure. They then have to manually review patients' chart files to match all the trials' inclusion and exclusion criteria. These searches are typically done by the IT department and then researchers have to review such large datasets generated and have to validate to find a small set of patients they can enroll. This tedious and time-consuming process is one of the key contributors to delayed study recruitment efforts.

One reason for this scattered inefficient approach is the limited utility of structured data, while up to 80-90% of healthcare information remains buried in the unstructured data, such as, free-from clinician notes and patient reported outcomes.

Recent studies reveal that 92% of inclusion and exclusion criteria benefit from including unstructured data in patient searches.
Leveraging AI on such unstructured data significantly improves precision and recall vs using structured data alone and can generate results rapidly. Clinicians' notes are one important unstructured data that can provide actionable insights. The labor intensiveness of traditional cohort selection cannot be underestimated. Extracting relevant information from free text clinical notes requires exhaustive reading by highly trained medical personnel, while only a tiny percent of the patients investigated wound up as suitable cohorts. In a recent study in Mass General Hospital on the benefit of radiotherapy for older women, the team managed to enroll only 636 people in 5 years, out of the roughly 40,000 patients known to exist in the US each year.

This is why, with the recent rapid advancement in AI, automating the text mining task using Natural Language Processing (NLP) became an exciting possibility, attracting many Pharma companies and academic researchers. The interest in this area can be illustrated with the 2018 challenge by National NLP Clinical Challenges (N2C2) on cohort selection, which had 47 research teams competing. This challenge has been active every one to two years since 2006, with sponsorship from Harvard Medical School, Blavatnik Institute Biomedical Informatics, and George Mason University, School of Engineering.

The intent is to boost innovation of informatics and AI technology in the field of healthcare. The sponsoring organizations contribute much of the preparatory work by, perhaps most importantly, taking up manually annotations of clinical documents. Their effort yielded many important lessons on the pros and cons of various approaches, which we leverage as a starting point for formulating our product vision.

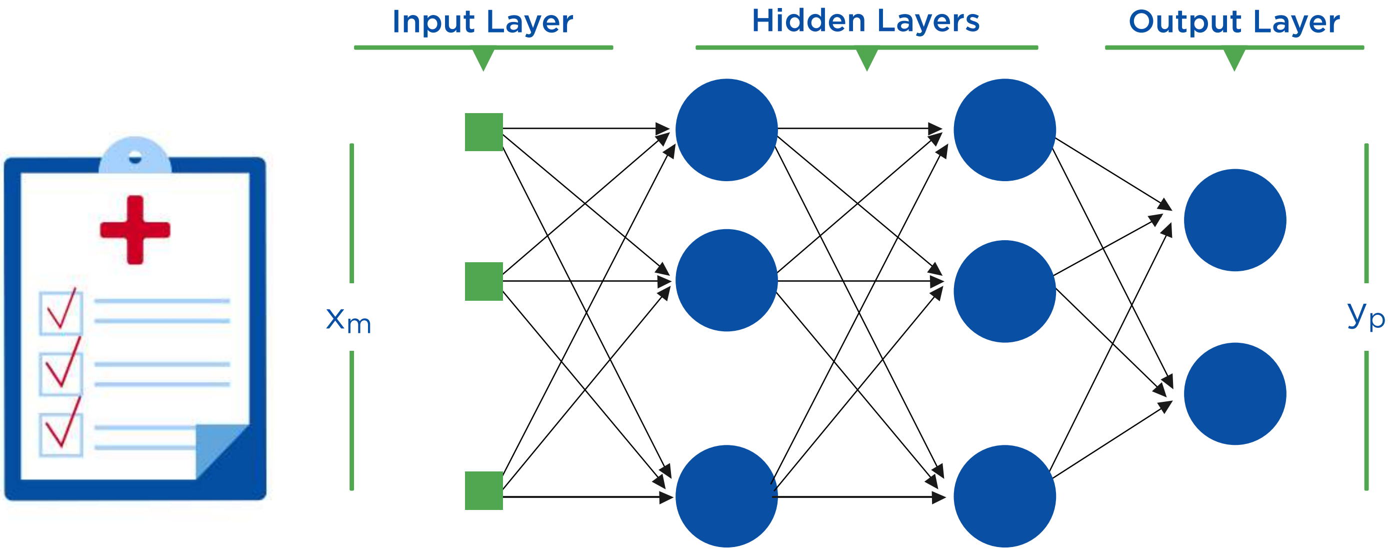
IT IS A DIFFICULT NLP PROBLEM

It cannot be overemphasized how much challenge remains in applying NLP to healthcare data. Medicine is a highly specialized field, often containing implicit information that requires background knowledge and context to decipher. It is also particularly rich in linguistic diversity; a heart attack can be written as myocardial infarction, myocardial infarct or simply MI. A successful NLP program must incorporate a comprehensive set of synonyms, abbreviations, and concept hierarchy.

However, if there is one area of severe obstacle in realizing the power of NLP, it is in the lack of annotated clinical text. If the prohibitive cost which puts many neural NLP and computer vision projects out of reach of investigators who are not Google, this particular domain exacerbates the acute shortage by many orders of magnitude, due to the qualification requirement of the annotators.

This shortcoming is made abundantly clear during the N2C2 challenge, which only supplied a paltry set of clinical notes for 288 patients, to target a study with 12 inclusion and exclusion criteria. Given the low sample count, it was relatively easy for the contestants to manually gather the necessary vocabulary (some with hardcoded character patterns) and rules to maximize their prediction accuracy. The use of simple classifiers per criterion directly exposes the artificial nature of pre-annotated data, a scenario which is clearly unattainable for the data in the wild.

Furthermore, given the limited dataset, there is no reason why a binary checklist cannot reach 100% accuracy per patient, after sufficient iterations in vocabulary expansion. The top contestant score of 0.91 in F1, although impressive in general, actually serves to highlight the inherent weakness of a classification approach. The tendency for hardcoding makes their methods not at all generalizable for a massive search over a large population data. The annotation effort per criterion also makes the prospect of matching the over 400,000 studies posted on clinicaltrials.gov unimaginable.



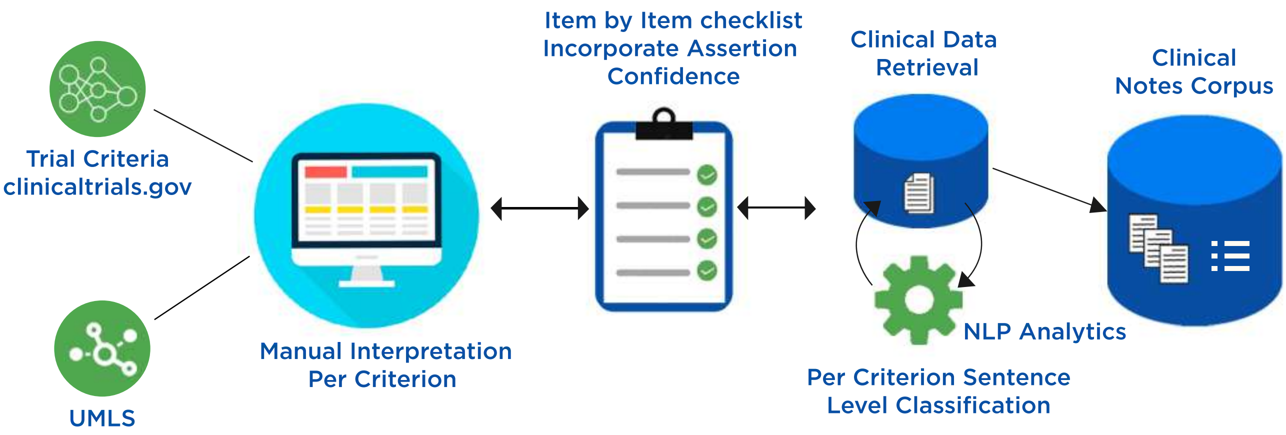
Conventional AI Approach

Requires manual labelling of clinical notes corpus

The classifier model is only reusable if another trial criterion is identically stated

Intensive annotation and not generalizable

Implossible to achieve 100% accuracy



Our sentence level matching can reach 100% accuracy given a comprehensive collection of vocabularies through our exploration tool

Towards creating a commercially viable product, the overall capability involves many of the well-known NLP techniques as illustrated below with common examples:

- **NER** - Named Entity Recognition. Whether a phrase is a disease, treatment, drug, etc.?
- **RE** - Relation Extraction. Is the drug improving the disease or causing side effects?
- **Negation** - For instance, a patient refused a vasectomy.
- **Assertion** - present, absent, possible, conditional, hypothetical, not associated
- **Date extraction** - e.g., 1/24/67: Normal iron studies.
- **Number extraction** - e.g., Digoxin 0.25 MG PO QD 5.

However, the NER portion needs to go way beyond determining the category of the phase.

A typical trial criterion may look like this:
ABDOMINAL SURGERY - History of intra-abdominal surgery, small or large intestine resection, or small bowel obstruction.

We would devise multiple linguistic approaches to identify patients who meet this criterion. Two examples are given below.

BY PROCEDURE

Requires the complete collections of procedures indicative of a surgical or intervention procedure in the abdominal region. A brief google search can already generate a quick list of over 10 such procedures (Adrenalectomy, Appendectomy, Bariatric surgery, Laparoscopic Gastric Bypass, ERCP, Pancreatography, Cholecystectomy, Esophageal surgery, Esophagectomy, etc). Using our iterative method described later, we could assemble an extensive and ever-growing list, suited for a procedural description.

BY ANATOMICAL SECTION

A different physician may choose different languages and specify the anatomical region instead. Any of the long list of body regions (Adrenal gland, common bile duct, biliary tract, Liver, Spleen, Bowel, Pancreas, Colon, Anorectal, Ileal, Kidney, Renal) can be matched with any of the general surgical terms (surgery, resection, removed, cryoablation, etc.) to indicate an abdominal surgery. Thus for a single criterion out of a single trial specification, the initial effort for vocabulary collection is undoubtedly the biggest step that any commercial solution must deal with efficiently.

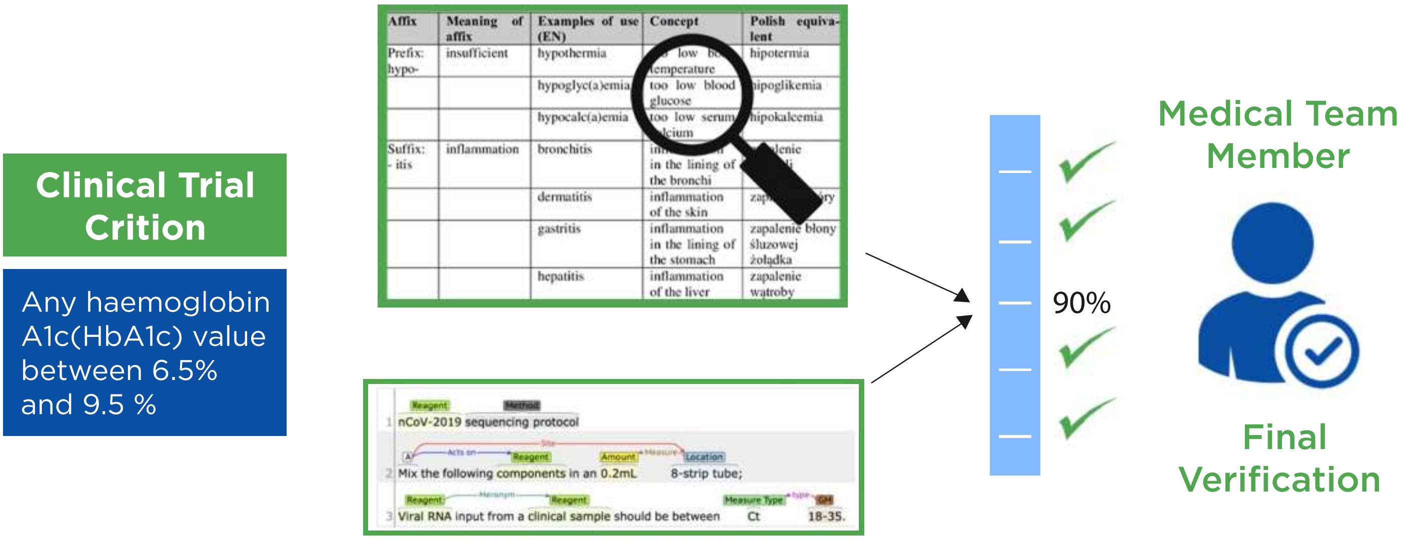
In our latest product offering Bayezene for efficient modeling of clinical trials, this is exactly the emphasis we placed in our design as we set out to create a first of its kind solution.

VOCABULARY EXTENSION IS THE KEY

To this end, we opted for an exploration technique using an iterative, semi-supervised method, with a man in the middle for approval / rejection. We fully embrace the state-of -art pre-trained neural word embedding trained on Biomedical text for this process. Our algorithm takes advantage of term similarity and word context in our iterations, to discover more and more suitable terms relevant to the original intended list of terms. Given a sizable corpus of clinical notes, it is an exercise that could be completed in hours of a non-medically trained developer / admin person's time, as long as the staff member is apt in Google Health Knowledge Graphs (GHKG) and Universal Medical Language System (UMLS) searches. The process continues until no more new terms can be found out of the (unannotated) training corpus.

The medically trained team member is only consulted after the list is deemed complete for a final validation. This is a very important aspect of our solution which overcomes a huge efficiency gap in reported approaches.

VOCABULARY EXPLORER



Any haemoglobin A1c(HbA1c) value between 6.5% and 9.5 %

Affix	Meaning of affix	Examples of use (EN)	Concept	Polish equiva-lent
Prefix: hypo-	insufficient	hypothermia	[illegible] low [illegible] temperature	hipotermia
		hypoglyc(a)emia	too low blood glucose	hipoglikemia
		hypocalc(a)emia	too low serum calcium	hipokalcemia
Suffix: - itis	inflammation	bronchitis	in[illegible] in the lining of the bronchi	[illegible]lenie [illegible]li
		dermatitis	inflammation of the skin	zap[illegible]ry
		gastritis	inflammation in the lining of the stomach	zapalenie blony śluzowej żołądka
		hepatitis	inflammation of the liver	zapalenie watroby

nCoV-2019 sequencing protocol

Mix the following components in an 0.2mL 8-strip tube;

Viral RNA input from a clinical sample should be between Ct 18-35.

Both our vocabulary explorer and annotation free Relation Extraction modules are designed for high efficiency, leading to rapid turnaround even for brand new clinical trial criteria.

SOCIAL DETERMINANTS

More and more, patient attributes in the area of social determinants are considered during qualification. A patient with severe mobility disadvantages is much more likely to miss the trial study sessions and negatively impact the data collection.

However, conditions such as a patient must speak English, or live with a caretaker, would not benefit from searching a medical specialty database such as UMLS. Such patient attributes will especially be reliant on the exploration process described above with a carefully selected training corpus which is more focused on lifestyle elements. This consideration highlights the importance of NLP tool innovation over simple lookups of the various medical term dictionaries.

TAXONOMY

In certain cases, even the most complete collection of synonyms is insufficient to capture all occurrences of a particular condition. Take for example, the condition ketoacidosis which may be expressed as metabolic acidosis, the parent concept of the former. This would be a completely normal way for a physician to express an evaluation but would baffle the best efforts in NLP auto concept extraction. In such a case, a survey of the parent concept is needed for completeness. While much of this is captured in the very mature UMLS database, our solution must demonstrate great dexterity in leveraging these advanced features.

Overtime, our system is continuously enriched with a collection of knowledge graphs (KG) to encapsulate the ever-expanding taxonomy we discover along the way. However, our approach does not at all count on having a mature KG before tackling real world data. Our key strength is the simplicity and the rapidity in which our solution can process new clinical trial specifications using exploration. Nevertheless, this ever-growing taxonomy will continue to accelerate the processing of future trial criteria and clinical notes corpus.

RELATION EXTRACTION WITHOUT MANUAL LABELING

While the vocabulary extension step is likely most of the battle in any given trial study, there are occasions when a criterion spells out a non-obvious constraint between two entities.

Think of specific relations that may exist between a drug and a disease:

- Drug cured disease
- Drug treat disease
- Drug caused adverse effect on disease
- Drug unrelated to disease

Or the many ways an implant can be related to the patient's jaw:

- The implant is placed into the jaw bone
- The implant is blocking the jaw bone
- The implant is pushing against the jaw bone

Again, smart RE models are plentifully available in the literature. The real bottleneck still lies in the exorbitant effort in text annotation. As in the case of concept extraction, this is inarguably the real showstopper in any trial recruitment, considering the cost involved for just a single trial design.

To deal with this efficiency gap, we must come up with a pipeline which eliminates the manual annotation step, particularly one that employs medically trained personnel.

AN ULTRA EFFICIENT RE PIPELINE

By tapping into the state-of-the-art research in RE and semi-supervised learning, we have adopted a combination of best of breed solutions, which completely do away with line by line annotation. The overall technique incorporates advanced concepts such as soft rules (rules that mostly work and can lead to better conclusion when used in consensus), dark knowledge (uses statistical information discarded by conventional classifiers), and fine tuning of biomedical word representations.

This improvement completely changes the math of RE production workflow efficiency. A neural NLP model typically requires a few 100,000 trained samples. Manual annotation involves identifying individual sentences, out of a possibly astronomical corpus, that meet a specific semantic condition. Unlike synonym searches, no particular keywords that can effectively filter out relation mismatches. Labeling relations also requires more contemplation per case than synonyms. Furthermore, whereas a new soft rule can be immediately tested against the entire corpus, any given existing relation label does not help seeking out more samples. Most relations can be quite adequately captured with just a handful of such rules.

NUMERICAL EFFICIENCY GAIN

Collectively, we leverage NLP methods that give the best accuracy to date, as compared to previous academic publications. We achieved the utmost processing efficiency both in vocabulary extension and novel relation extraction, with the emphasis on turn around speed. For instance, we found that an experienced physician took 20 minutes of time reading through each clinical note document (around 2500 words) for a single trial criterion. The N2C2 example has 13 trial criteria, which is a typical minimum length in the real world. Assuming a search pool of 10,000 documents in pursuit of 500 cohorts, we are looking at 20 x 13 x 10000 = 2,600,000 minutes or 43,333 hours, or 1,805 days.

And this is assuming an effective filter has been applied to achieve a 5% hit rate. A lower hit rate would necessitate an even larger corpus. This is to be compared to our vocabulary exploration method which experiments have shown to be accomplished within one to two days of effort for each new criterion involving a brand new medical or social concept.

BIGRIO COHORT SELECTION SOLUTION

BigRio is a specialty AI consultancy with particular emphasis on pharmaceutical and healthcare industries. We have long maintained a state-of-the-art expertise in neural NLP technology. Our proposed solution is always developed based on up-to-date peer reviewed research reports, cutting edge methods and easily deployable frameworks developed through rigorous refinement by our seasoned data scientists.

We also drastically differ from the pack in the unique way we formulate our 'beat the world' solution. Instead of emphasizing the tuning of classifiers, we tackle the biggest obstacle in NLP automated cohort selection, which is the annotation step, proposing the completion of new studies in days, not months or years required using the traditional process.

Our method starts with intuitively correct linguistic rules which can be quickly applied to a large corpus and feedback the positive candidate percent. Each time a new pattern is discovered, a new rule is generated to capture an unlimited number of additional matching sentences. The rest of the optimization process is all computerized in our machine learning pipeline. Our rigorously tested framework and method brings three orders of magnitude acceleration into clinical trial design efforts.

Our approach is the only one, which is considered generalizable and completely eliminates the manual annotation step for both building the medical and social concept dictionary and resolving new and novel relations in clinical notes text. For more information write to us at info@bigr.io