Clinical trial risk tool

Fast Data Science

Clinical Trial Risk Tool

We have developed a tool called the Clinical Trial Risk Tool, which predicts the risk of a trial failing to deliver uninformative results. You can drag and drop the PDF of a protocol and the tool calculates the risk level (low, medium, or high) using natural language processing.
The initial version of the tool is available for free online at https://app.clinicaltrialrisk.org/

The tool is being improved to extract more features from the protocol text, and to predict trial cost in dollars as well as risk.

We are expanding from HIV and TB, which were the first areas we focused in, to other areas such as vaccine trials, Covid, and oncology trials. We are interested in talking to users to identify key areas of interest and we can add features on request.

The development was funded by the Gates Foundation.

Target users are trial sponsors, CROs, and investigators.

Fast Data Science

Natural language processing consulting in London, UK in Microsoft Partner Network.

Fast Data Science offer a range of consulting and implementation solutions in Data Science, Machine Learning, including Deep Learning, and Artificial Intelligence. If you would like to start we can assist at various points in your project. If you would like assistance or advice in one of these areas please get in touch with us.

Read more

Fast Data Science Story

In 2008, **Thomas Wood** finished his Master degree in Computer Speech, Text and Internet Technology at Cambridge University. He then worked for a series of startups and large multinationals in areas of natural language processing and machine learning. He noticed that certain industries such as healthcare, pharmaceuticals, and insurance, were sitting on a goldmine of unstructured text data, but company-internal initiatives to use this data often failed.

In 2018, **Thomas founded Fast Data Science,** aiming to help companies and organisations use their unstructured data. Clients included **the Gates Foundation, Tesco, White Ribbon Alliance, and Ulster University.**

Since then, the company has grown to a team of four, with regular partners. We have clients across multiple countries and industries. **We specialise in natural language processing (NLP), healthcare and pharma.**

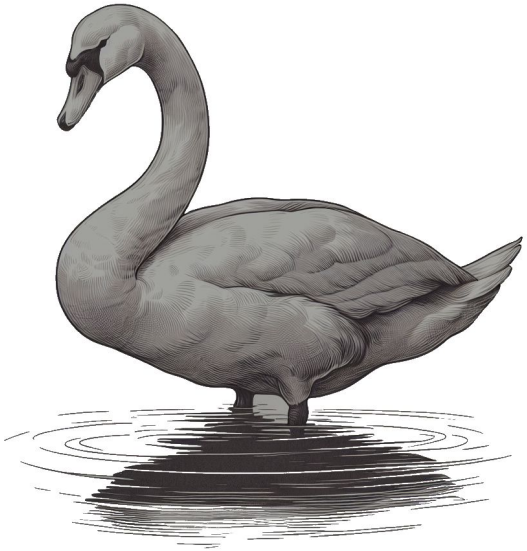
Important questions that AI can help with in planning clinical trials

Running a clinical trial

- can we predict likely cost?
- can we predict likely duration of enrollment?
- can we predict likelihood of trial failure?

A huge input on business processes is predicting an unlikely but undesirable event: the “grey swan” of a trial ending uninformatively.

An informative trial is a trial which delivers answers to research questions and helps to advance medical science. Individuals participating in clinical trials expect that their efforts will help to bring about these advances, but sometimes poor trial design results in preventable uninformativeness.



What happens with a study that ends uninformatively?

It never finishes, often because insufficient participants were recruited, or

It is never published, because it ended underpowered, or

It is never published, due to poor design or an inadequate analysis plan, or

It is published but focuses on a question other than the original research question, or

It is published only after many years' delay, or

It is published promptly and stakeholders must accept criticism for wasted money and resources.

How to determine the risk of trial failure uninformatively?

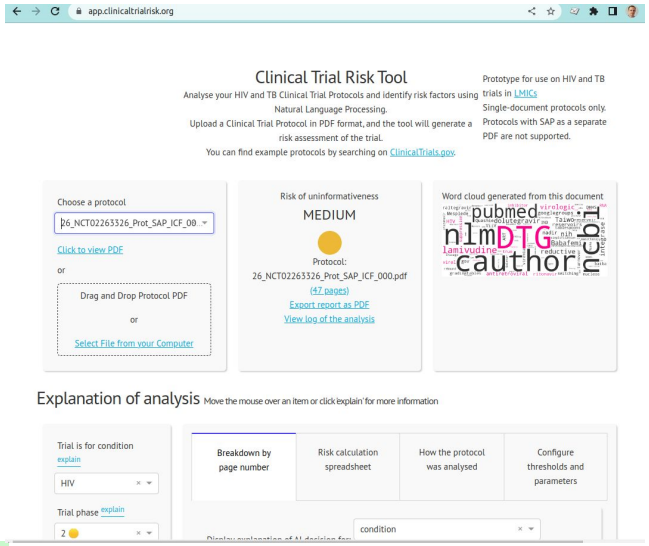
Table 1. Results of a qualitative survey of feature importance for determining risk.

Weighting informativeness features	Mean score
Has an Statistical Analysis Plan	100%
Effect estimate not disclosed or unreliable	84%
tertile_of_sample_size by domain by phase	75%
Tertile of number of sites by domain by phase	72%
Composite product of tertile of Primary Duration times tertile of Sample Size	72%
tertile of number of (co-)primary endpoints by domain by phase	72%
Number of endpoints	66%
Multiple countries (Y/N)	56%

We have used a combination of quantitative and qualitative methods to develop a risk score of the trial.

See Wood, McNair, *Clinical Trial Risk Tool: software application using natural language processing to identify the risk of trial uninformativeness* (2023)

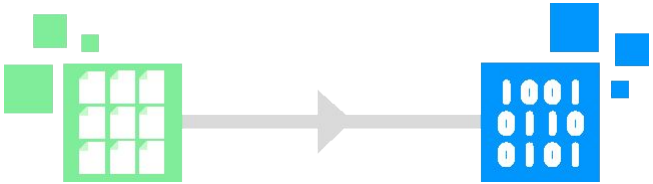
Clinical trial risk tool



For the Bill and Melinda Gates Foundation, we developed and trained a deep learning tool using natural language processing (NLP) to predict the risk of running a clinical trial.

The tool is online at https://clinicaltrialrisk.org/ and published in Gates Open Research https://gatesopenresearch.org/articles/7-56/v1

User uploads a clinical trial protocol in PDF format, and the tool identifies high/medium/low risk of the trial failing (ending uninformatively).



Before running a clinical trial, the investigator writes the trial protocol, often 200+ pages in PDF format.

Fast Data Science developed an ML model which extracts important data from the protocol: type of treatment, pathology, number of subjects, etc.

app.clinicaltrialrisk.org

Clinical Trial Risk Tool

Analyse your HIV and TB Clinical Trial Protocols and identify risk factors using Natural Language Processing.

Upload a Clinical Trial Protocol in PDF format, and the tool will generate a risk assessment of the trial.

You can find example protocols by searching on ClinicalTrials.gov.

Prototype for use on HIV and TB trials in LMICs

Single-document protocols only.

Protocols with SAP as a separate PDF are not supported.

Choose a protocol

26_NCT02263326_Prot_SAP_ICF_00...

Click to view PDF

or

Drag and Drop Protocol PDF

or

Select File from your Computer

Risk of uninformativeness

MEDIUM

Protocol:

26_NCT02263326_Prot_SAP_ICF_000.pdf

(47 pages)

Export report as PDF

View log of the analysis

Word cloud generated from this document

Explanation of analysis

Move the mouse over an item or click 'explain' for more information

Trial is for condition

explain

HIV

Breakdown by page number	Risk calculation spreadsheet	How the protocol was analysed	Configure thresholds and parameters

Clinical trial risk tool

https://app.clinicaltrialrisk.org

Trial is for condition explain

HIV

Trial phase explain

2

Has the Statistical Analysis Plan been completed? explain

Yes

Has the Effect Estimate been disclosed? explain

Yes

Number of subjects ⚠ Low confidence! explain

90

Possible sample sizes found: 90, 45, 41

Sample size tertile: 0 (small trial) set values of tertiles

Number of arms explain

2

Countries of investigation explain

United States

Tool extracts key features from trial text and puts them into risk model

Risk of uninformativeness

MEDIUM

Protocol: 26_NCT02263326_Prot_SAP_ICF_000.pdf

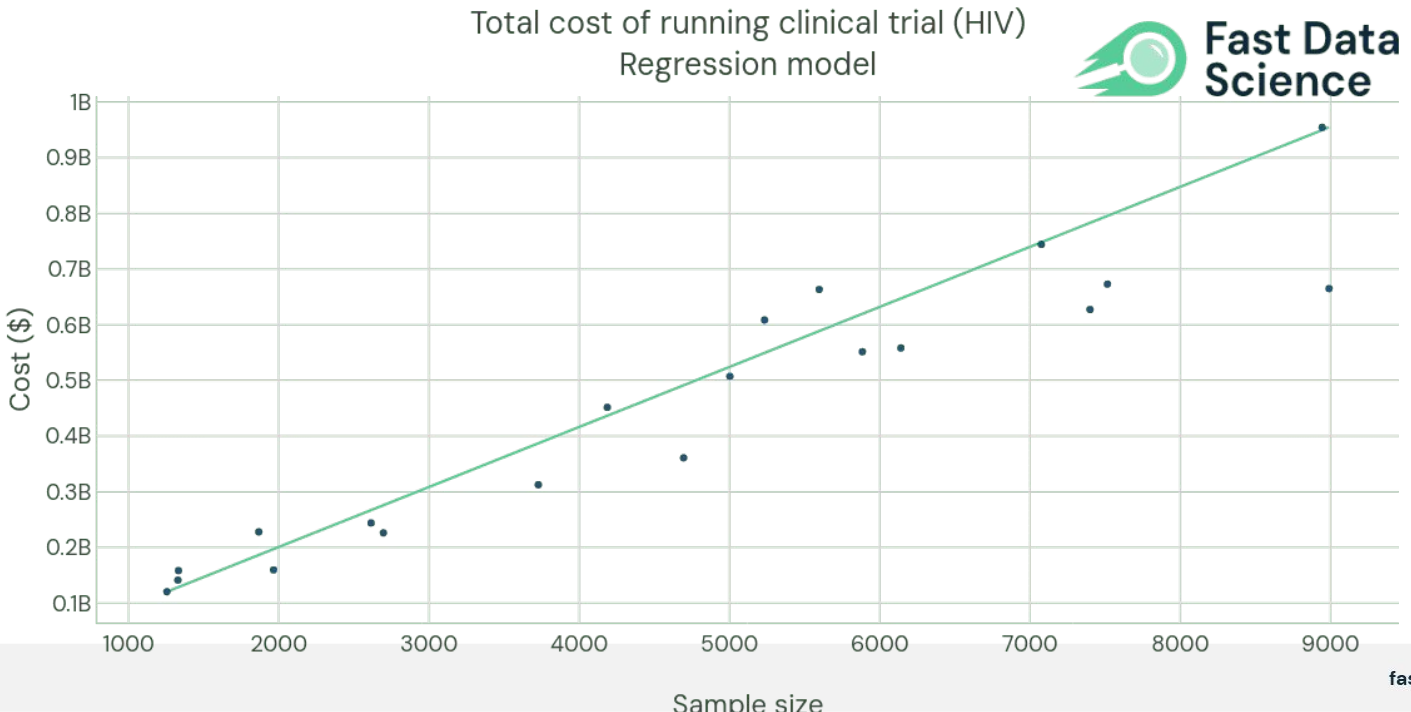
(47 pages)

Export report as PDF

View log of the analysis

Predicting cost in $ of running trial

We are working on cost modelling to predict the dollar value of a trial based on the raw protocol. This is due for completion in Q3 of 2024.



Predicting costs, risk, or duration

There is open data we can use to build cost models

	A	B	C	D	E	F	G	H	I	J	K	
1	indication (longer)	Cancer	Genetic disorder	Infectious	Technology	CT.gov URL	Tech II	Enrollment	Trial Phase	Total Cost	Per Patient Cost ($PP)	Source for this Data
2	Advanced Myeloid Malignan	Cancer			biologic drug	-		30	Phase 1	$328,000.00	$10,933	https://docs.google.com/
3	Blood Cancer	Cancer			biologic drug	https://clinicaltrials.gov/ct2/show/NCT03483324		9	Phase 1	$5,000,000.00	$555,556	https://docs.google.com/
4	Blood Cancer	Cancer			biologic drug	https://clinicaltrials.gov/ct2/show/NCT03925935		24	Phase 1	$6,192,579.00	$258,024	https://docs.google.com/
5	Severe Combined Immunodeficiency, X-linked (X-SCID)				biologic drug	https://clinicaltrials.gov/ct2/show/NCT02963064		90	Phase 1/2	$19,068,382.00	$211,871	https://docs.google.com/
6	B cell cancers, Leukemia	Cancer			biologic drug	https://clinicaltrials.gov/ct2/show/NCT03088878		156	Phase 1/2	$18,292,674.00	$117,261	https://docs.google.com/
7	Blood Cancer	Cancer			biologic drug	https://clinicaltrials.gov/ct2/show/NCT02222688		26	Phase 1	$4,179,598.00	$160,754	https://docs.google.com/
8	Colon Cancer	Cancer			biologic drug	https://clinicaltrials.gov/ct2/show/NCT02953782		112	Phase 1/2	$10,234,048.00	$91,375	https://docs.google.com/
9	Leukemia, Acute Myeloid (AML)	Cancer			biologic drug	https://clinicaltrials.gov/ct2/show/NCT03248479		96	Phase 1	$5,000,000.00	$52,083	https://docs.google.com/
10	Blood Cancer, Solid Tumors	Cancer			biologic drug	https://clinicaltrials.gov/ct2/show/NCT02216409		88	Phase 1	$6,505,568.00	$73,927	https://docs.google.com/
11	Breast Cancer	Cancer			biologic drug	https://clinicaltrials.gov/ct2/show/record/NCT0078		720	Phase 3	-	$104,186.49	https://drive.google.com
12	Stage IV Melanoma	Cancer			cell therapy	https://ClinicalTrials.gov/show/NCT00438984		11	Phase 1	$936,164.00	$85,106	https://docs.google.com/
13	Stage IV Breast Cancer	Cancer			cell therapy	https://clinicaltrials.gov/ct2/show/NCT00791037		23	Phase 1/2	$2,236,359.00	$97,233	https://docs.google.com/
14	Malignant Glioma				cell therapy	https://clinicaltrials.gov/ct2/show/NCT00612001		8	Phase 1	$1,275,311.00	$159,414	https://docs.google.com/
15	Non-Small Cell Lung Cancer	Cancer			cell therapy	https://clinicaltrials.gov/ct2/show/NCT00850785		6	Phase 1	$653,850.00	$108,975	https://docs.google.com/
16	Stage IV Melanoma	cancer			cell therapy	https://clinicaltrials.gov/ct2/show/NCT01189383		20	Phase 1/2	$1,410,939	$70,547	https://docs.google.com/
17	Amyotrophic Lateral Sclerosis				cell therapy	https://clinicaltrials.gov/ct2/show/NCT03280	stem cell	261	Phase 3	$15,912,390.00	$60,967	https://docs.google.com/
18	Brain Cancer	Cancer			cell therapy	https://clinicaltrials.gov/ct2/show/NCT02546102		414	Phase 3	$5,391,016.00	$13,022	https://docs.google.com/
19	Spinal Cord Injury				cell therapy	https://clinicaltrials.gov/ct2/show/NCT02302	stem cell	25	Phase 1/2	$14,323,318.00	$572,933	https://docs.google.com/
20	Leukemia, Acute Myeloid (AML)	Cancer			cell therapy	https://clinicaltrials.gov/ct2/show/NCT03301	stem cell	146	Phase 2	$4,310,000.00	$29,521	https://docs.google.com/
21	Melanoma	Cancer			cell therapy	https://clinicaltrials.gov/ct2/show/NCT01875653		4	Phase 3	$3,000,000.00	$750,000	https://docs.google.com/
22	Type 1 diabetes				cell therapy	https://clinicaltrials.gov/ct2/show/NCT02691	CAR T	113	Phase 2	$8,568,363.00	$75,826	https://docs.google.com/
23	Heart disease associated with Duchenne muscular dystrophy				cell therapy	https://clinicaltrials.gov/ct2/show/NCT02485938		25	Phase 2	$3,376,259.00	$135,050	https://docs.google.com/
24	Pulmonary Hypertension				cell therapy	https://clinicaltrials.gov/ct2/show/NCT03145	stem cell	26	Phase 1/2	$7,354,772.00	$282,876	https://docs.google.com/
25	Blood Cancer, Bone Marrow Tra	Cancer			cell therapy	https://clinicaltrials.gov/ct2/show/NCT03475	virus-specific T	60	Phase 1/2	$4,825,587.00	$80,426	https://docs.google.com/
26	Brain Cancer	Cancer			cell therapy	https://clinicaltrials.gov/ct2/show/NCT02208	CAR T	92	Phase 1	$12,753,854.00	$138,629	https://docs.google.com/
27	Brain Cancer, Breast Cancer	Cancer			cell therapy	https://clinicaltrials.gov/ct2/show/NCT03696	CAR T	39	Phase 1	$9,015,149.00	$231,158	https://docs.google.com/
28	Sickle Cell Disease				cell therapy	https://clinicaltrials.gov/ct2/show/NCT03249	stem cell	6	Phase 1	$5,742,180.00	$957,030	https://docs.google.com/
29	Kidney Failure				cell therapy	https://clinicaltrials.gov/ct2/show/NCT03363945		75	Phase 3	$11,217,155.00	$149,562	https://docs.google.com/
30	Multiple Myeloma	Cancer			cell therapy	https://clinicaltrials.gov/ct2/show/NCT03288	CAR T	180	Phase 1	$19,813,407.00	$110,074	https://docs.google.com/
31	Beta Thalassemia				cell therapy	https://clinicaltrials.gov/ct2/show/NCT03432364		6	Phase 1/2	$8,000,000.00	$1,333,333	https://docs.google.com/
32	B cell cancers, Leukemia	Cancer			cell therapy	https://clinicaltrials.gov/ct2/show/NCT03233	CAR T	57	Phase 1	$11,034,982.00	$193,596	https://docs.google.com/
33	Retinitis Pigmentosa				cell therapy	https://clinicaltrials.gov/ct2/show/NCT02320812		28	Phase 1/2	$17,144,825.00	$612,315	https://docs.google.com/
34	Lung Cancer	Cancer			cell therapy	https://clinicaltrials.gov/ct2/show/NCT03546361		36	Phase 1	$11,815,315.00	$328,203	https://docs.google.com/
35	Melanoma, Skin cancer	Cancer			cell therapy	https://clinicaltrials.gov/ct2/show/NCT03240	CAR T	12	Phase 1	$14,144,221.00	$1,178,685	https://docs.google.com/
36	Sarcoma	Cancer			cell therapy	https://clinicaltrials.gov/ct2/show/NCT03240	CAR T	12	Phase 1	$4,693,839.00	$391,153	https://docs.google.com/
37	Sickle Cell Disease				cell therapy	https://clinicaltrials.gov/ct2/show/NCT02247	stem cell	6	Phase 1	$13,145,465.00	$2,190,911	https://docs.google.com/

Trial cost modelling from protocol text

We are developing simple and understandable models which calculate cost per subject based on PDF of protocol or other study document in unstructured format… based on real past trials

Total cost of trial is $2100851.00 and model used for weights was model_0_weights

Settings and weights

You can adjust the weights. The top few rows tell you when a set of weights is applicable.

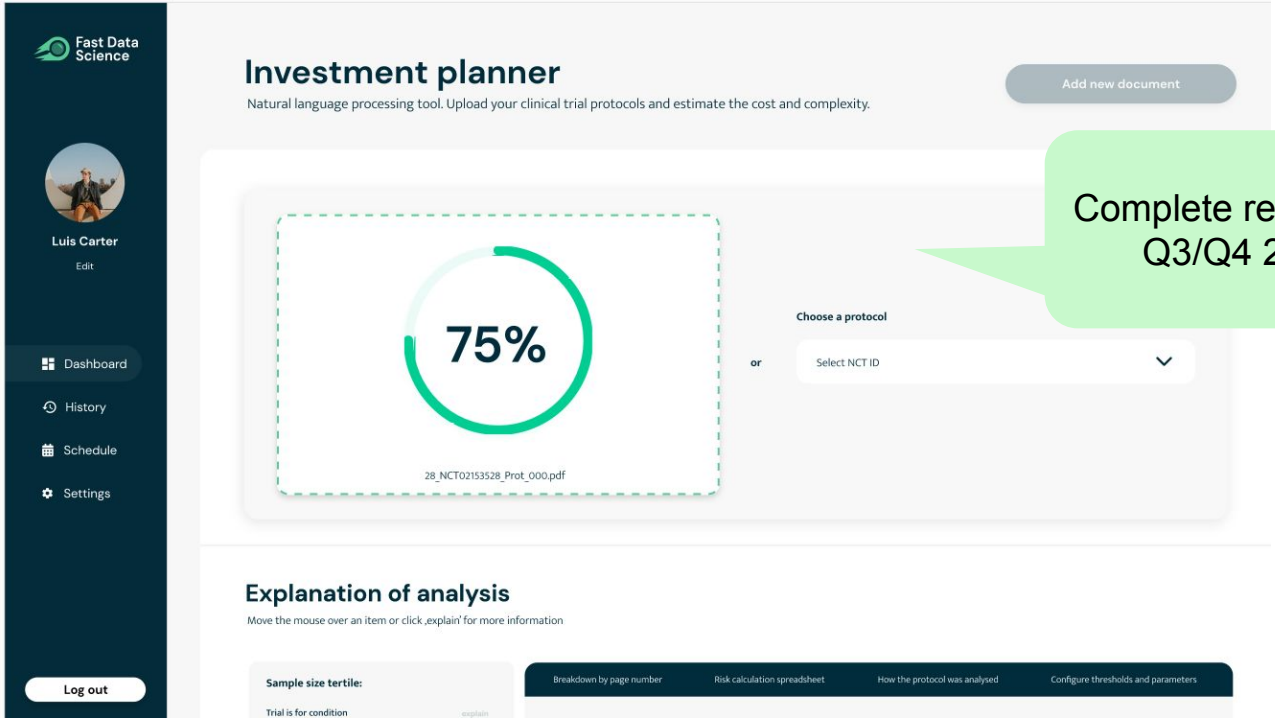
EXPORT

	feature	description		2_weights
×	valid_for_condition	what conditions do these weights apply to?		
×	valid_for_vaccine	what vaccine values do these weights apply to?		
×	valid_for_phase	what phase do these weights apply to?	early_phase_1,1	
×	valid_for_intervention_type	what intervention type do these weights apply to?	behavic	
×	constant	Constant		455572276
×	document_type_protocol	Document Type is Protocol		0
×	document_type_sap	Document Type is SAP		0

```
21096.903997661182
+ -2 * Total Enrollment
+ 86,134 * HIV
+ 42,955 * TB
+ 14,217 * Malaria
+ 30,075 * HAT
+ 5,628 * COVID-19
+ 64,232 * Phase 1
+ 5,431 * Phase 2
```

Fast Scie

New user interface in Q3/Q4 2024



Complete redesign in Q3/Q4 2024.

Data Science Services

AI for healthcare

We have undertaken large projects in healthcare for clients such as the NHS. We have developed clinical named entity recognition models and predictive models for healthcare workforce management.

AI in pharmaceuticals

The pharmaceutical industry is moving towards widespread adoption of AI. We have worked on projects to extract data from pharma KOL insights, NLP models to extract data from clinical trials.

Data strategy consulting

Fast Data Science can assist with your entire data strategy, from opportunity identification through to stakeholder workshops, opportunity prioritisation, and infrastructure planning.

Machine learning consulting

A valuation of your machine learning processes and strategy and recommendations to build durable and maintainable machine learning systems, avoiding vendor lock in

Public sector procurement for AI projects

If you are responding to an RFI or project out for tender, or a funding grant for a research project, we would be glad to help draft your application and be listed as the technical partner in your project. We have applied for and been awarded projects with universities and public bodies such as the NHS, Office of Rail and Road, and Tarion (Canadian housing regulator). We have an ongoing partnership with Ulster University and University College London.

Data analytics consulting

Optimise your business intelligence processes, leverage existing data, and identify opportunities to extract value from your data.

AI due diligence

If you are considering investing in or acquiring a company in the AI space, we can perform a due diligence exercise. The director, Thomas Wood, has the CUBS (Cardiff University/Bond Solon) certificate for expert witness work in England and Wales (civil cases).

Deploying machine learning models

Machine learning model deployment is an often overlooked aspect of data science. Often, the majority of effort required in a project is invested in deployment. We have experience with hosting, all three major cloud providers (AWS, Google, Azure), continuous integration and deployment (CI/CD) tools such as Github Actions, and infrastructure management such as Terraform.

Training and upskilling analytics teams in data science

We can run customised workshops and produced video tutorials to boost an organisation's data science capabilities.

Clients and past projects of Fast Data Science

National Health Service

The publicly funded healthcare system of the UK and the country's largest employer with nearly 2 million on the payroll.

White Ribbon Alliance

Washington, DC based childbirth charity.

Dashboard

Boehringer Ingelheim

One of the major European pharmaceutical companies, and manufacturer of well known drugs for respiratory diseases, oncology and diabetes among others.

Tesco plc

The most well known supermarket chain in the UK, and a multinational retailer with presence in several countries.

CV-Library

Currently the UK's third largest job board, founded in 2000.

Wellcome Data

Wellcome is a global charitable foundation founded in 1936. Through their work, they support science in solving urgent health problems facing everyone.

Information Commissioner's Office

ICO is an executive non-departmental public body, sponsored by the Department for Science, Innovation and Technology.

Ulster University

A university with a national and international reputation for excellence, innovation and regional engagement.

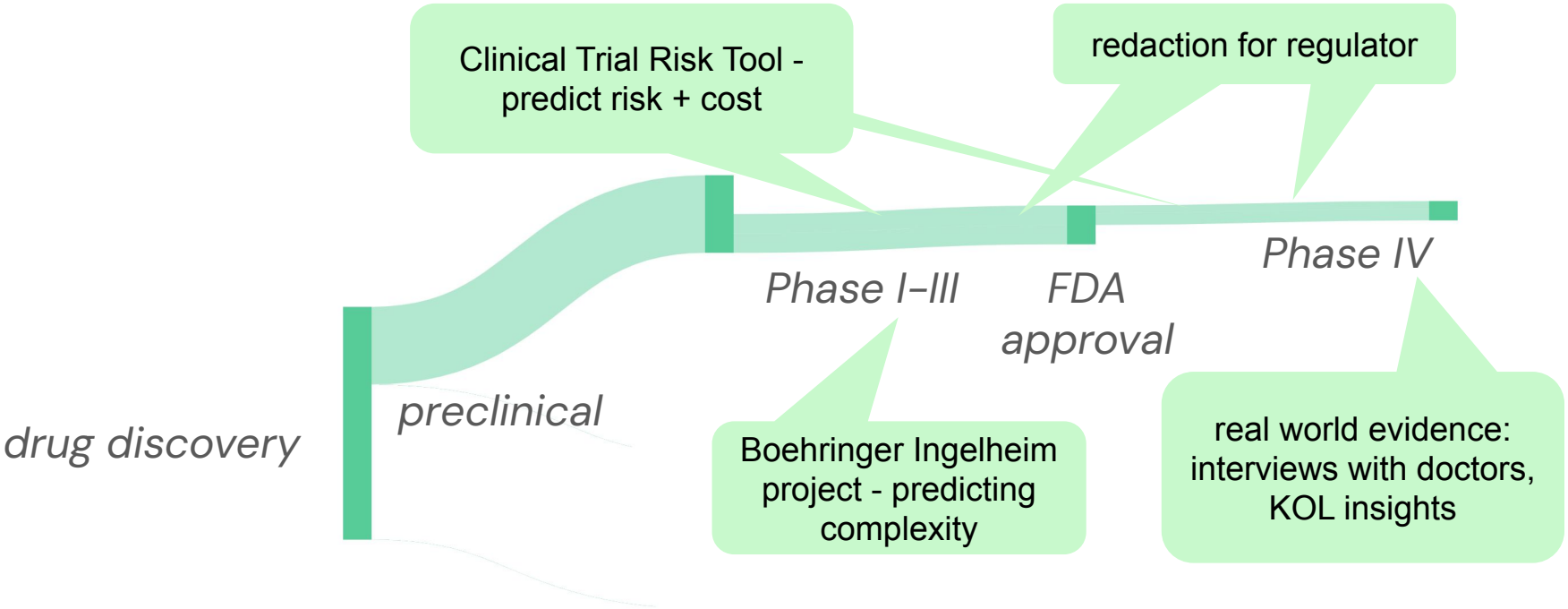
cbtclinics

CBT Clinics is a UK-based company offering mental healthcare practitioners.

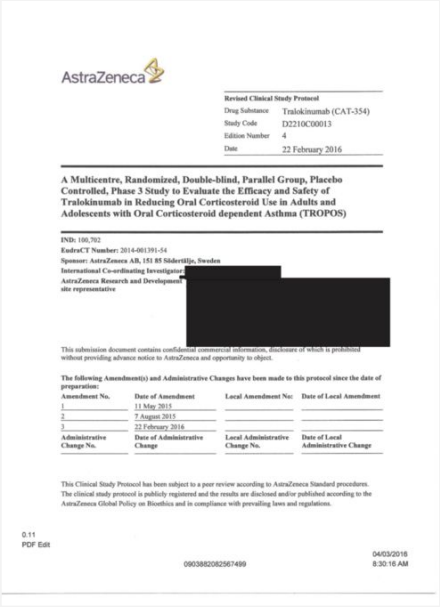
Ordnance Survey

Ordnance Survey (OS) is the national mapping agency for Great Britain.

Where our past projects fit in drug development pipeline



Boehringer Ingelheim – complexity modelling



AstraZeneca

Revised Clinical Study Protocol

Drug Substance	Tralokinumab (CAT-354)
Study Code	D2210C00013
Edition Number	4
Date	22 February 2016

A Multicentre, Randomized, Double-blind, Parallel Group, Placebo Controlled, Phase 3 Study to Evaluate the Efficacy and Safety of Tralokinumab in Reducing Oral Corticosteroid Use in Adults and Adolescents with Oral Corticosteroid dependent Asthma (TROPOS)

IND: 100,702
EudraCT Number: 2014-001391-54
Sponsor: AstraZeneca AB, 151 85 Södertälje, Sweden
International Co-ordinating Investigator: [redacted]
AstraZeneca Research and Development site representative [redacted]

This submission document contains confidential commercial information, disclosure of which is prohibited without providing advance notice to AstraZeneca and opportunity to object.

The following Amendment(s) and Administrative Changes have been made to this protocol since the date of preparation:

Amendment No.	Date of Amendment	Local Amendment No:	Date of Local Amendment
1	11 May 2015		
2	7 August 2015		
3	22 February 2016		

Administrative Change No.	Date of Administrative Change	Local Administrative Change No.	Date of Local Administrative Change

This Clinical Study Protocol has been subject to a peer review according to AstraZeneca Standard procedures. The clinical study protocol is publicly registered and the results are disclosed and/or published according to the AstraZeneca Global Policy on Bioethics and in compliance with prevailing laws and regulations.

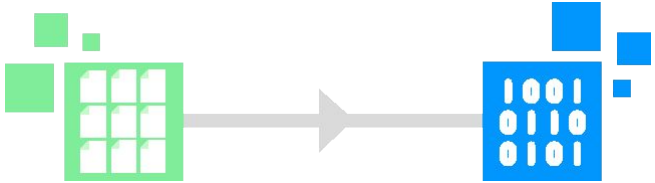
0.11
PDF Edit

090388208256749 9

04/03/2016
8:30:16 AM

Boehringer Ingelheim

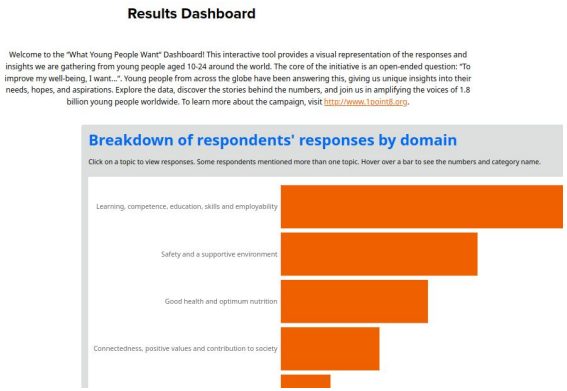
For the German pharma company Boehringer Ingelheim, we developed and trained a deep learning tool using natural language processing (NLP) to predict more than 50 output variables from a clinical trial protocol. This allows pharma companies and regulators to analyse and quantify large numbers of clinical trial protocols, allowing more accurate cost estimation.



User drags and drops protocol PDF

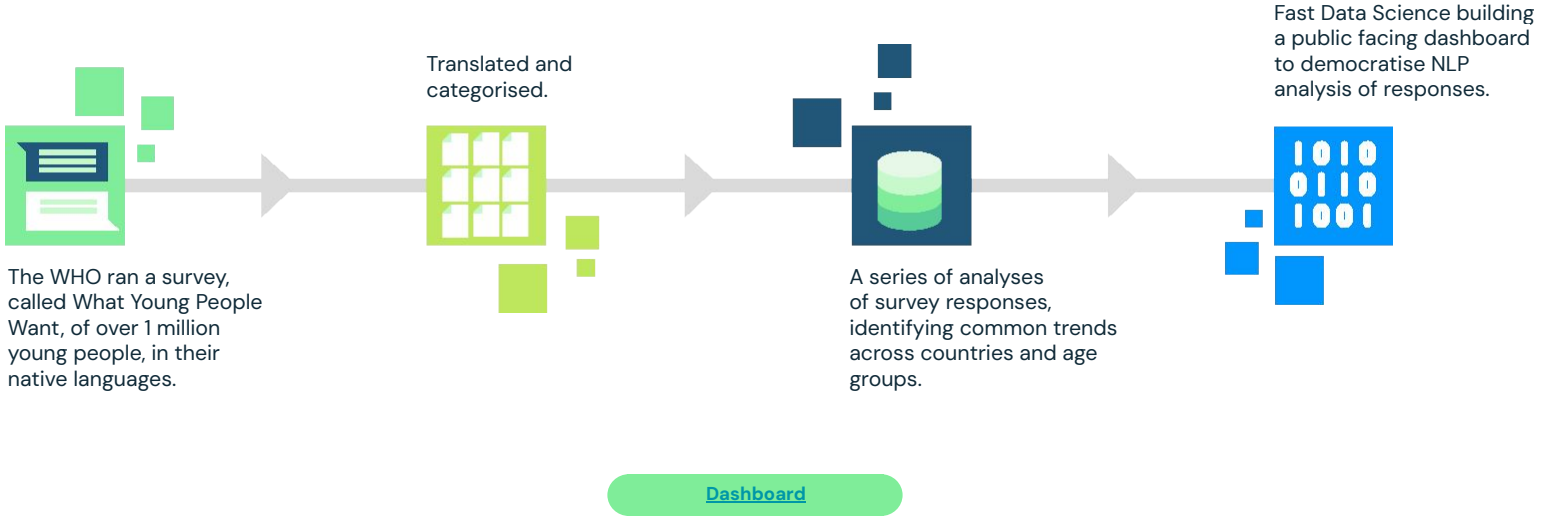
Tool identifies features such as sample size, number of cycles

Site complexity score: 1, 2 or 3
Subject complexity score: 1, 2 or 3
Study complexity score: 1, 2 or 3



World Health Organization

The WHO hired Fast Data Science to analyse multilingual text responses to their campaign to reach 1.8 billion young people around the world
https://wypw.1point8b.org/



Confidential client: Regulatory pharma

For one company in the regulatory space in the pharmaceutical industry, we are developing rule based and machine learning models to redact and categorise sensitive information in clinical trial narrative reports for the regulator

- adverse events
- concomitant medications
- medical history
- etc

Mr. Smith is a 62-year-old male with a history of Stage III pancreatic cancer who is participating in a stage 3 clinical trial for a novel treatment regimen. He presented to the emergency department with complaints of [redacted]. Mr. Smith reported that he [redacted].
On examination, Mr. Smith was found to have [redacted] revealed a [redacted]
Given Mr. Smith's history of pancreatic cancer and ongoing participation in a clinical trial, caution was taken in managing his pain and determining the most appropriate treatment plan for [redacted] Consultation with his oncologist was sought to ensure that any interventions would not interfere with his ongoing cancer treatment or compromise his overall health.

Confidential client: KOL insights

For another client we have developed NLP tools to process transcripts of interviews and text data gathered at Phase IV, identifying drugs mentioned and relationships between them e.g. suspected interaction, connection to adverse events, reluctance to prescribe, etc.

Drug recognition library free online:
https://fastdatascience.com/drug-named-entity-recognition-python-library
(interaction models are proprietary)

In my opinion, based on my experience and understanding of these medications, I believe that using Tenelomab and Acometimab together may result in decreased effectiveness of Tenelomab. Acometimab is a medication commonly used to treat conditions such as rheumatoid arthritis and cytokine release syndrome, by targeting specific pathways in the immune system. On the other hand, Acometimab is a calcium channel blocker often used to treat high blood pressure and chest pain.

When these two medications are used together, there is a potential for drug interactions that could impact the effectiveness of Tenelomab. Amlodipine may interfere with the

interaction

Coming in 2024

- Increase to 30 features identified in text, including schedule of events, regimen, chemotherapy cycles, and more
- Ability to model cost and for user to customise their in-house cost model
- Improved user interface with login and support for multiple documents
- We are exploring other use cases of the technology e.g. Wellcome Trust is interested in predicting duration of enrollment period - this is of particular interest to CROs also.


س
知