



Data-driven clinical trial feasibility and study design

How a new approach to data management can
boost efficiency and lead to better
decision-making in drug development



ONTOFORCE

☰ Summary

Drug development is highly competitive and clinical studies are a crucial and expensive step in the high-risk, high-reward route to ground-breaking treatments. Assessing a clinical study's feasibility, designing a study or deciding to move forward with a study requires processing a huge amount of information and taking multiple dependencies into account.

Finding precise, comprehensive and relevant data is a challenging job for study designers, researchers and data scientists. While there has never been more information available, **never has such a large fraction of that data been so intangible**. Bits and pieces of relevant data are scattered across registries, companies, departments, databases and many other siloes. The overall stream of information isn't just too vast to plow through manually, it's also fragmented beyond comprehension and exists in too many hard-to-digest formats.



“On average, research and data scientists spend 20% of their time looking up the right information.”

Unsurprisingly, research and data scientists spend an estimated one day each week on average just finding, filtering through and collecting the right information – before they can start their actual work. Logically, data issues have a direct impact on the efficiency of a knowledge worker, and chances are that it also directly impacts overall study timelines. Spending too much time searching for and gathering information impedes decision making, and in a highly competitive environment, **this can jeopardize the project**. After all, patients are kept waiting for a cure.

All of the above is not a new problem. For the last few decades, we have been dealing with an apparent paradox. The constant increase in availability of data due to technological advancements goes hand in hand with the burden of being unable to easily use that data for a multitude of purposes. It's striking that although most of the technological challenges can be addressed nowadays, a rollout is often very difficult in practice.



How data should be treated in an ideal world

In an ideal world, data should be generated and treated with not just the primary endgoal in mind, **but as an asset that needs proper attention**.

In this paper, we'll go deeper into the current state of the art of the management of clinical data and metadata that resides inside or outside an organization. Based on practical use cases, we explain the immediate and longer-term impact of treating data as a valuable asset to achieve concrete business goals.

Clinical study data and where to find it

As described above, clinical studies are an essential part of the drug development process. A clinical study is the moment when a company needs to put the public spotlight on the result of years of developmental work performed behind the scenes. It is only when the early research and preclinical work shows promising results that a company will consider moving into the clinical research phase.

Typically, pharmaceutical clinical studies are divided into four phases, three of which must be successfully validated by regulatory bodies before a drug is approved. In Europe, the green light is given by the European Medicines Agency (EMA), while the US Food & Drug Administration (FDA) grants approvals in the United States. Other countries have their own authorization bodies.

Regulators publish information about previously or currently approved clinical studies. The information is available on dedicated websites, which give users the capability of searching and retrieving subsets of interest.

Because the clinical research path is littered with obstacles and only a fraction of the new drugs entering phase 1 will be approved, competition is fierce and many companies are reluctant to share much information about a study. This **conflicts with the need to stay informed** of the status of other similar studies. This clear tension is managed by regulators, which have their own guidelines about what kind of study information must be publicly available.

Major clinical trial repositories such as clinicaltrials.gov in the US, the EU Clinical Trials Register, the UMIN Clinical Trials Registry (UMIN-CTR) in Japan and the Chinese Clinical Trial Registry (ChiCTR), gather data and metadata related to clinical studies. In addition, the WHO aggregates a subset of the data from these registries in the International Clinical Trials Registry Platform (ICTRP).

Clinical study registry data integration

Concretely, if one would like to have an overview of current and past clinical studies for a specific condition, time period or a combination of other search criteria, **all of these registries must be consulted** to get a comprehensive overview. Bringing these databases together is a logical step toward improving search activities. Unfortunately, the data in these databases can't be just copied and pasted at once.

Studies are often published in more than one registry. Logically, if studies are conducted in multiple countries or continents, they need approval from different regulators. In addition, there isn't a standardized method of cross-referencing clinical study registries (see Fig. 1). It requires a special extraction of other identifiers to be able to map these cases.

EU Clinical Trials Register

EudraCT Number	2015-002060-17
Sponsor's Protocol Code Number:	2015/077/HP
National Competent Authority:	France-ANSM
Clinical Trial Type:	EEA CTA
Trial Status:	Completed
Date on which this record was first entered in the EudraCT database:	5/08/2015

Clinicaltrials.gov

ClinicalTrials.gov Identifier:	NCT02584439 <i>History of Change</i>
Other Study ID Numbers:	2015/077/HP 2015-002060-17 (EudraCT Number)
First Posted:	October 22, 2015 <i>Key record dates</i>
Last Update Posted:	December 7, 2016
Last Verified:	December 9, 2016

Figure 1: Example of a duplicate registration of a clinical study.

The above figure shows a duplicate registration of a clinical study; initially registered in the EU Clinical Trials Register with ID 2015-002060-17 (left)¹ and subsequently in clinicaltrials.gov with ID NCT02584439 (right)².



A next step in the data merging and harmonization process is the mapping and consolidation of properties. A simple yet significant example is the way study phases are defined and annotated in different repositories.

Looking again at the EU Clinical Trials Register and clinicaltrials.gov, the former uses Roman numerals (I, II, III) and the latter uses Arabic numerals (1, 2, 3) (see Fig 2).

Phase

The stage of a clinical trial studying a drug or biological product, based on definitions developed by the [U.S. Food and Drug Administration \(FDA\)](http://www.fda.gov). The phase is based on the study's objective, the number of participants, and other characteristics. There are five phases: [Early Phase 1 \(formerly listed as Phase 0\)](#), [Phase 1](#), [Phase 2](#), [Phase 3](#), and [Phase 4](#). [Not Applicable](#) is used to describe trials without FDA-defined phases, including trials of devices or behavioral interventions.

Phase I

Phase I is the first stage of the clinical development of a medicinal product. It includes the first administration of that product in human subjects. Phase I Clinical Trials include a small number of subjects (up to 30) and frequently involve healthy volunteers, but may also involve patients. They generally have no therapeutic intent and are initial studies of the safety and tolerability of an IMP. They also include pharmacokinetic and sometimes pharmacodynamic studies.

Phase II

Phase II Clinical Trials are conducted in patients with the intended target disease. They are conducted to investigate the safety and efficacy of an IMP, and to determine the doses to be used in the larger Phase III trials. They usually involve 100-300 people in total.

Phase III

Phase III Clinical Trials are the large Clinical Trials (several hundred to several thousand subjects) used to determine the safety and efficacy of the IMP. They are usually multicentre and very often multi-country Clinical Trials. It is usually on completion of the Phase III Clinical Trials that a Marketing Authorisation application is made.

Phase IV

These are studies carried out after the Marketing Authorisation has been granted, and are carried out within the terms of the marketing authorisation (therapeutic use and dose) and using the authorised product Post marketing studies to delineate additional information including the medicine's risks, benefits, and optimal use. These studies are designed to monitor effectiveness of the approved intervention in the general population and to collect information about any adverse effects associated with widespread use.

Figure 2: Excerpts from the glossaries of clinicaltrials.gov (top)³ and EU Clinical Trial Register (bottom)⁴ defining the different study phases

The next step in processing the data and improving its usability is the challenge of correctly identifying concepts such as diseases or conditions, therapeutic interventions (drugs, devices, etc.) and organizations (sponsors, collaborators, study sites) or people (principle investigators, etc.) involved. **Semantic matching is a major step** towards improving the interoperability of the data and a critical step in making the data ready for new ways of searching. For example, searching by sponsor reveals all studies sponsored by a specific company or research institute; searching by a condition term that is part of a disease classification system allows the user to exploit the parent and child relationships between disease terms.

Integrating clinical study registry data by taking these steps into account is part of the process of making data more FAIR (findable, accessible, interoperable and reusable)⁵. We strongly believe that **translating the FAIR principles into practice** is an essential step towards a more efficient and elegant usage of clinical study data. Making data meaningful by mapping concepts, standardization and improving FAIRness impacts the user experience directly.

No pain, no gain?

Bringing data to the next level of interoperability and unlocking more of its potential comes with a cost. A quick and dirty patch-up job won't cut it. But how big should the investment be?

It is key to strike a balance and avoid diving into the rabbit hole of solving every data problem. Smart data science is about optimizing data handling and management without omitting domain expertise. The goal is to solve a concrete problem while always keeping the FAIR data principles in mind. Similar to the DRY (don't repeat yourself) paradigm in software engineering, making data more FAIR means avoiding messes and losing future opportunities. At ONTOFORCE, we were already applying a lot of the FAIR principles avant-la-lettre by respecting twenty-year-old semantic web principles and the linked data philosophy and embedding these into an industry-strength software platform.

This platform helps you get a grip on the data chaos by focusing on automation, scalability and extendability. Today, with semantic search, smart filter and compare capabilities, and intelligent visualization, we offer **the technology required to extract maximum value from data** – and do it fast.

From data to insights

The downstream benefits of this are significant on many levels. When the relevant information becomes easy to find, filter and compare, researchers and data scientists can spend more time developing high-quality, data-driven insights that benefit the drug development process, the companies they work for – and ultimately, patients.

Study designers can optimize their trials for maximum recruitment and retention rates, avoid pitfalls, and drastically increase the chances of success. Moreover, this approach also yields tangible results in the fast-developing field of precision medicine, where the need to identify specific patient populations during trial is even more critical.



Making data manageable: an introduction to search, filter and compare

Researchers derive value from data aggregation and analysis: it allows them to discover new insights and opportunities, and to make well-founded recommendations to study designers, chief medical officers, therapeutic business owners, etc. To do this, however, the data must first be made manageable.

Bridging the gaps: unifying data across platform

Ask any life sciences professional, and they will agree that the diversity and complexity of classification methods for human conditions and diseases is unparalleled. To make matters even more complicated, these classifications coexist and are widely accepted by various stakeholders in life sciences, from lab scientists and clinicians to pharmaceutical and biotech firms, regulatory bodies, governments, etc. With each classification comes a different identifier, even if they are highly similar or have exactly the same meaning.

Obviously, this makes it extremely difficult to search, compare and analyze similar data where different diseases classifications are applied. As a result, the information can't be merged automatically on the level of diseases.

- Clinicaltrials.gov, the central clinical trial repository of the United States, uses the Medical Subject Headings (MeSH) to encode the condition under study.
- The EU Clinical Trials Register uses the Medical Dictionary for Regulatory Activities (MedDRA).

How can we overcome this?

To enable the interoperability, re-use, and merging of data, the DISCOVER knowledge platform brings classifications like MeSH, MedDRA, CT, and UMLS

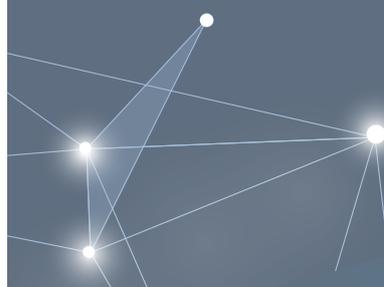
together. This means users can search and analyze data in both registries at the same time, through one intuitive interface, and by using the system they are most familiar with.

DETERMINING RELEVANT CLINICAL DATA SETS

Having access to the right information is a prerequisite for effective clinical trial design, as we have seen. But researchers often lack the time to process all the data – resulting in missed opportunities. Data modeling and new technologies like semantic search allow clinical researchers and data scientists to aggregate the data bits and pieces they need, and bundle them for future processing and analysis.

Integrating real-world data

One key element in this is the possibility to re-use data that was originally generated for another primary purpose. In some cases, the combined data from previous studies can lead to new hypothesis generation or insights, without requiring the set-up of a new clinical trial. Other examples include the integration of real-world data to make cohort selections, or the use of population data to design studies that avoid certain adverse effects or that target the precisely correct patient population subset.



The basics: search, filter, and compare

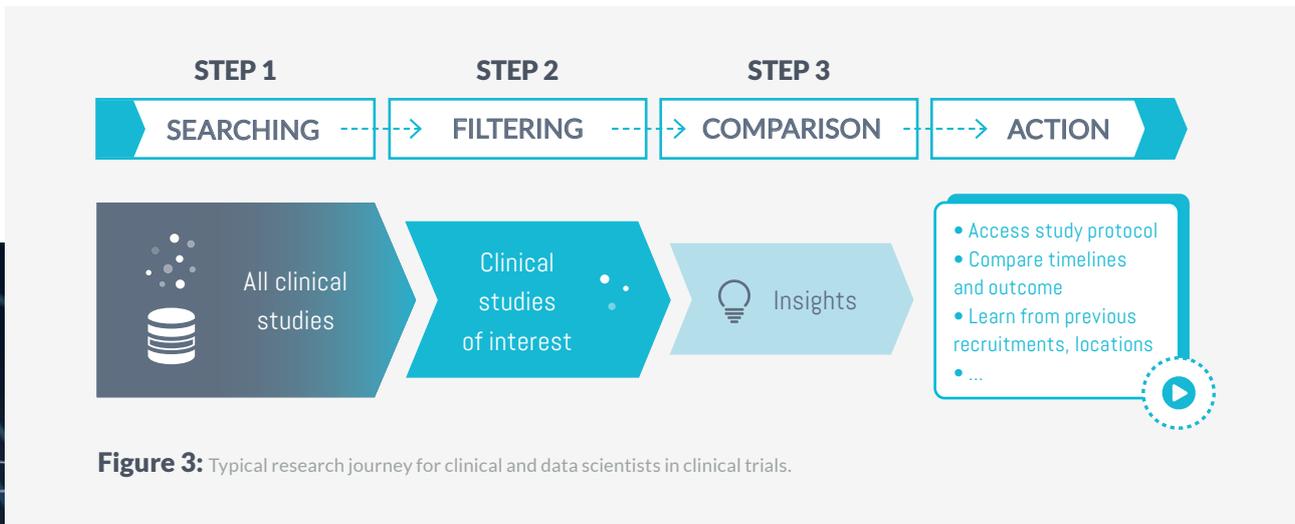


Figure 3: Typical research journey for clinical and data scientists in clinical trials.



STEP 1 SEARCHING FOR A SPECIFIC SUBSET OF CLINICAL STUDIES

In many cases, researchers looking for new trial opportunities, new endpoints, competitive insights, etc. will begin by delving into a specific subset of clinical studies. This query can be based on various search and filter criteria, such as:

- condition or therapeutic area
- type of study
- sponsor
- trial phase
- development pipeline
- target



STEP 2
VISUALIZE AND COMPARE RELEVANT DATA POINTS

Ideally, step 1 results in a manageable shortlist of relevant studies and reports. Subsequently, the researcher will want to select the relevant data points in each and visualize the results in a list or table for easy comparison. These visualizations can incorporate numerous points of view: disease centric, company centric, treatment centric, etc.

To make this possible, data from multiple public, internal and third-party sources needs to be integrated, standardized and merged together.

In addition, important filter criteria need to be mapped to concepts and further enriched with context.



STEP 3
REPORT USABLE FINDINGS

Finally, the researcher will want to be able to save relevant clinical studies and/or data point collections and export them in an accessible format that can be shared with colleagues down the value chain.

A SIMPLE SOLUTION FOR A WIDE RANGE OF COMPLEX PROBLEMS

The three-step process above sounds as simple as can be, but is nearly impossible without powerful tools for data standardization and integration, smart filtering, and semantic search capabilities.

In the next few chapters, we'll explore more in-depth use cases and examples of how smart data and knowledge management can help researchers get to reliable insights faster, and optimize the critical path of clinical trial design.



Application 1: How to quickly find new endpoints

According to a study by the MIT Sloan School of Management⁶, only 14% of all drugs undergoing clinical trials eventually get FDA approval. The reasons for failed trials are myriad, but more often than not, they have to do with poor planning or misunderstandings of key biological and/or drug development principles, leading to inadequate study design, inappropriate efficacy markers, etc.

Minimizing the chances of failure

To avoid failure, one of the most important things to do is ensure that your study design is optimized from the start. Apart from implementing an efficient design that fits your purpose and is compliant with regulations, this also includes:

- making sure the study population is appropriately sized;
- applying the right inclusion and exclusion criteria;
- selecting the right endpoints.

Selecting the right endpoints

Let's zoom in on the latter, as endpoints play a key role in setting up an efficient clinical trial. The best endpoints are, of course, those that are compliant with regulations and that present an unbiased readout of clinical benefits. Sometimes, however, it pays to find alternative endpoints to get to a speedy approval as well.



USE CASE:

Discovering new endpoints in an early-stage cancer study

To shorten the time to readout in an early-stage cancer study, an oncology clinical research team was looking for new endpoints. In this case, traditional endpoints like disease-free survival (DFS), progression-free survival (PFS) and event-free survival (EFS) would not be appropriate.

As a starting point, researchers began by looking for example oncology studies where circulating tumor DNA (ctDNA) levels and minimal residual disease (MRD) were primary or secondary endpoints. This results in a longlist of potential studies of interest. These are then compared in detail to make a selection.

A touch of DISCOVER

Using DISCOVER, search terms with known synonyms, like 'ctDNA' or 'circulating tumor DNA', can be easily expanded. This ensures that every relevant study is displayed. In the results list, the search terms are highlighted, and a full record of the study can be accessed with one click. The list of relevant studies can then be easily exported in the appropriate file format.

LEARN MORE



[Clinical trials opportunity use case page](#)

Application 2: How to improve recruitment rates – with data

In addition to study design and selecting the right endpoints, the success of every clinical trial hinges on its ability to enroll a statistically significant set of patients according to a highly specific set of criteria. Studies show that 85% of clinical trials fail to retain enough patients, while 11% of study sites fail to enroll even a single patient. Unsurprisingly, patient recruitment (and retention) is widely recognized as one of the largest bottlenecks in the clinical trial process – and one of the costliest.

Collecting historical data

The key question here is: can our study recruit the required number of patients within the assigned time? To reach an answer, researchers can start by **collecting historical data**: how many patients have been recruited in similar studies in the past? How many dropped out? More specific data points include:

- recruitment start date;
- duration of enrollment;
- enrollment per country, site or arm;
- predicted vs. real recruitment rate.



Inclusion and exclusion criteria

Even more interesting is a search across databases for information on **inclusion and exclusion criteria** in similar studies. Data-driven study design practices that look at patient availability relative to the inclusion/exclusion criteria indicate which variables may cause the greatest challenge to recruitment. Finding a set of criteria that perfectly balances the need for unbiased outcomes and ensures proper recruitment rates is key here. However, aggregating this type of unstructured data is very challenging, and can be done with a powerful knowledge platform, such as DISCOVER.

Recruitment performance as criterium for finding partner sites

In addition, the efficiency of subject recruitment is an important criterium of partner site selection, as well as an indication of the potential test population. Larger institutions with good reputations, for example, are more likely to enroll subjects in a study faster.

Application 3: Finding the right clinical study sites

Selecting the right clinical sites that fit your specific study needs is another important factor for clinical trial success. Research shows that selecting underperforming sites can cost pharmaceutical companies millions per year. However, the diversity and inconsistency of information available makes choosing high performance sites difficult, leading to significant study delays.

Researchers and study designers have a lot to gain from efficient site identification, including:

- lower risk of choosing low-performing sites;
- avoiding collaborating with debarred clinical investigators;
- faster study start-up and site cycle times;
- overall increased efficiency.

For optimum site identification, researchers need to have access to multiple data sources and tools that allow them to easily search, filter, and compare features. Additionally, intuitive visualization is indispensable in making well-founded decisions quickly. Map visualizations can also help in selecting well-dispersed sites for optimum access to the right patient population.

Identifying key opinion leaders

In the previous chapter, we discussed the importance of recruitment rates in site selection. Another important criterium is the presence of key opinion leaders (KOLs): renowned experts in therapeutic areas. Finding these and mapping them onto sites of potential interest can help speed up the clinical trial and improve its probability of success.

USE CASE:

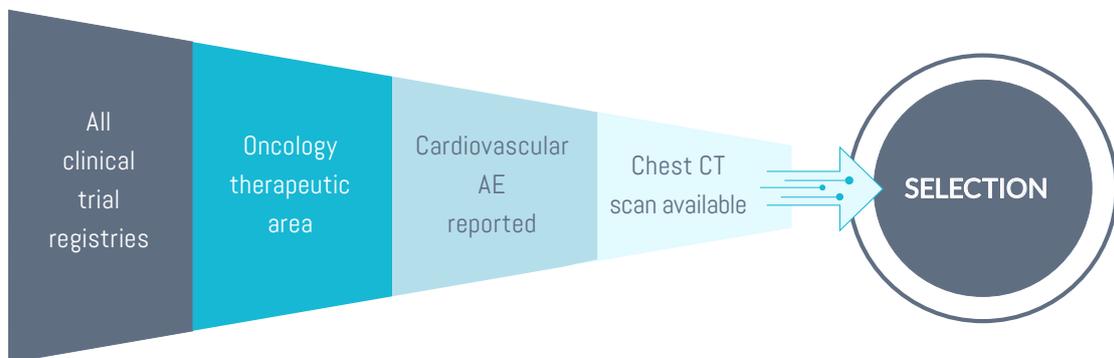
Finding predictive factors for interstitial lung disease

When setting up a clinical trial, it can be interesting to know whether certain patients are prone to develop an adverse event not uncommon in trials for a specific disease and/or treatment . To do this, however, it's first necessary to acquire a better understanding of the predictive factors.

In this example, using internal data, our researcher could start a search with filtering for:

- all oncology studies where Intertitial Lung Disease (ILD) is observed as adverse event;
- next, filter for all studies have imaging data, more specifically CT scans of the chest region.

THE SEARCH JOURNEY WOULD LOOK LIKE THIS⁷:

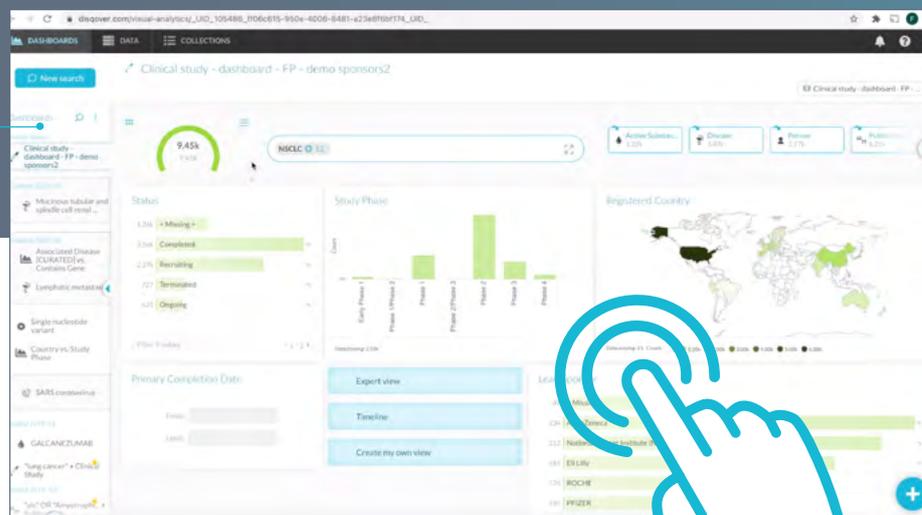




A touch of DISCOVER

Starting from all clinical studies, the researcher filters them down to a manageable subset that offers the data needed for further analysis. DISCOVER enables them to create customized filter templates with set parameters to significantly speed up this process – in this example the first steps are: selecting all studies that are linking to the Oncology Therapeutic Area followed by selecting the concept 'Interstitial Lung Disease' (ILD) in the filter widget 'adverse event'. The final selection criteria are 'CT' and 'chest' as available 'imaging type' and 'body part'. Within the intuitive interface, users can easily select the studies that are relevant for further analysis and export the required data in the appropriate format.

LEARN MORE



[Clinical trial value chain page](#)

Application 4: Unlocking key competitive insights

The go-to-market for new treatments is often a neck-and-neck affair. With so much time and resources invested in it, losing the race can have a major impact on the organization as a whole. A comprehensive, competitive assessment at the country, therapeutic, drug class, development pipeline and even individual site levels is an absolute must.

Assessing the competitive field

When looking at clinical trials from a competitive perspective, there's a lot of information that could be of interest. For example:

- Data on the timing of planned and ongoing trials vying for the same patients;
- Information on pending drug approvals and access to marketed products;
- Information about the mode of action of a treatment;
- Demographic data that could reveal the prevalence of certain conditions in a region;
- Epidemiological data offering a historical snapshot of disease incidence and prevalence within a patient population;
- Insights on the standard of care for a condition by country.



These insights can help decision makers shape their strategies and enable clinical and data researchers to answer pressing questions, like:

- What's the market potential for the treatment under review?
- Which adverse events are reported for a specific combination of clinical studies?
- Based on competitors' progress, should we halt further investment in clinical trials for this treatment, or press on?

Some of this data is available via clinical trial registries such as clinicaltrials.gov, or can be found in internal, commercial and subscription databases. Tools like DISCOVER bring these different data sets together and make the information interoperable, easy to search through, and ready for insightful comparison.

Adverse events investigation from a competitive perspective

Apart from enabling study design optimization and recruitment performance, adverse events investigation also offers value for competitive mapping. Studies that report the high occurrence of adverse events, for example, are an indication that the treatment isn't safe enough to proceed to the next phase. In addition, insights into adverse effects encountered by competitors could be important predictors for success of your own treatment.

Unlocking these insights allow researches to

- search and filter for clinical studies and check the reported adverse events, or;
- search for adverse events and follow the link to clinical trials where these adverse events are reported.



Conclusion: how data-driven insights move clinical trials forward

Without access to powerful knowledge platforms to gain data-driven insights, researchers and study designers rely on subjective experiences, repetition of previous, unproven trial design strategies and guesswork. As a result, enrollment in clinical trials suffers, and resources are allocated suboptimally.

Without access to semantic search and data interoperability, many of the use cases mentioned in this white paper would simply be infeasible.



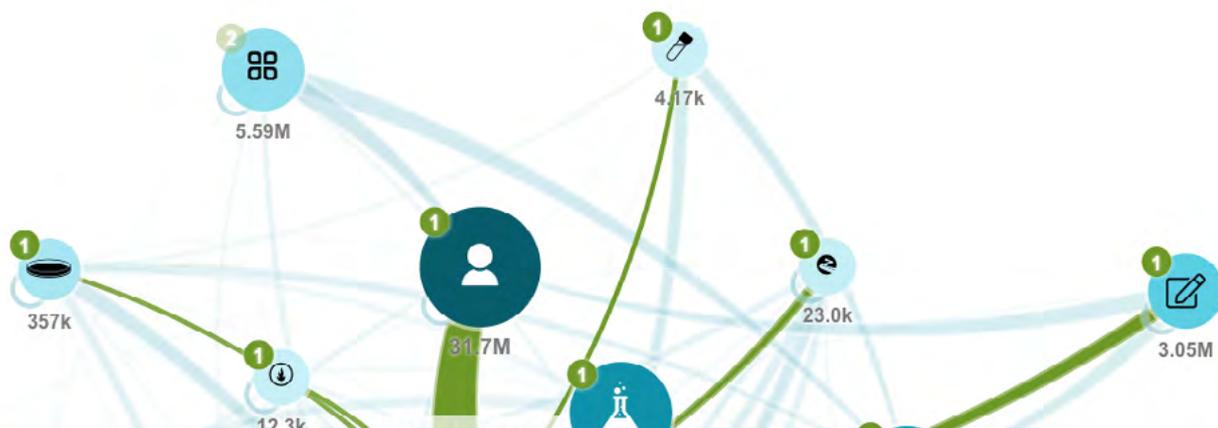
“We didn’t have the resources to comb through huge amounts of data, and it was too dispersed and too hard to find.”



“Sometimes we could do it, but it was just too much time and resource intensive. As a result, we didn’t do it as often as we should have. We based our decisions on outdated data, missed key opportunities, and were blind to new developments.”

Assessing a trial’s feasibility and setting up a trial for success don’t have to be time-consuming exercises. With the right data and intelligent tools, they can be completed in a matter of days, giving you a roadmap to follow and enabling you to anticipate exactly what will be required and how long it is all likely going to take.

A data-driven feasibility assessment can help ensure that your clinical research plan is designed to enroll the right patients, rely on the right investigators, and takes place in the right locations for success. With a tool like DISCOVER, you can act on key moments in clinical trials, increase the probability of success, and shorten the time to readout.



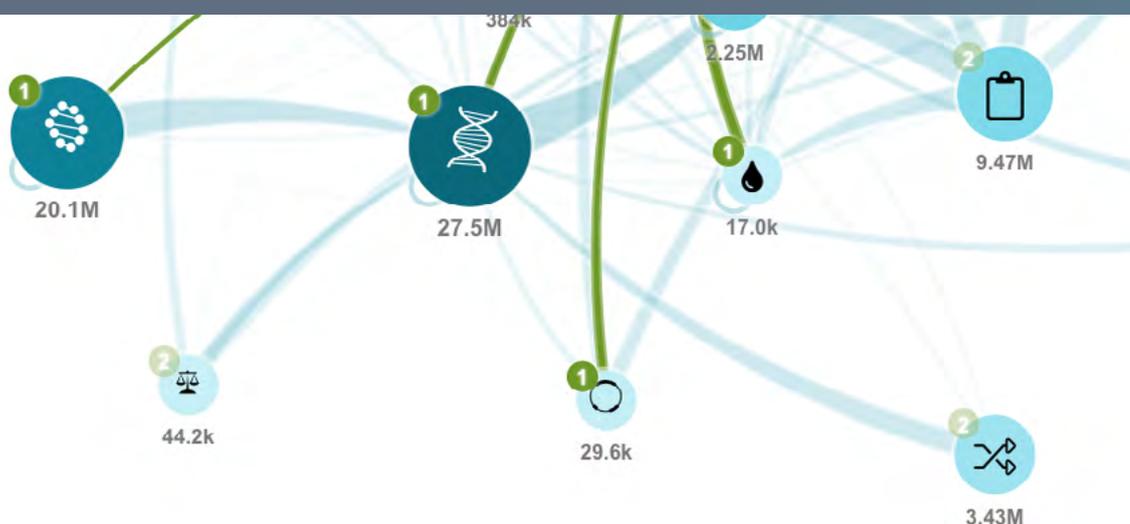
Enter the world of DISCOVER

DISCOVER, ONTOFORCE'S linked data platform, gathers, transforms and orchestrates information from internal, third-party and public sources, unlocking self-service knowledge discovery. DISCOVER delivers actionable insights by fostering data literacy and enhancing search accuracy, making the platform an integral part of your enterprise ecosystem.

Massive volumes of information spread across multiple sources. New analytics tools popping up every day. The daunting tasks of data harmonization and integration. From drug discovery and clinical research to literature analysis and chemistry, every aspect of life sciences is fraught with data-related challenges.

Powered by semantic search and an intuitive user interface, DISCOVER enables you to explore and connect data from disparate sources to uncover new insights in a matter of minutes. This empowers you to:

- master the flood of information in healthcare and life sciences
- speed up the research and go-to-market of new treatments and products
- homogenize, link and reveal hidden correlations.



Sources used

Example of a duplicate registration of a clinical study:

1 <https://www.clinicaltrialsregister.eu/ctr-search/search?query=2015-002060-17>

2 <https://clinicaltrials.gov/ct2/show/NCT02584439>

Example of phases' description:

3 <https://clinicaltrials.gov/ct2/about-studies/glossary>

4 <https://eudract.ema.europa.eu/help/Content/Glossary.htm>

5 Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>

6 Wong CH, Siah KW, Lo AW. Estimation of clinical trial success rates and related parameters. *Biostatistics*. 2019 Apr 1;20(2):273-286. doi: 10.1093/biostatistics/kxx069. Erratum in: *Biostatistics*. 2019 Apr 1;20(2):366. PMID: 29394327; PMCID: PMC6409418.

7 Raschi E, Fusaroli M, Ardizzoni A, Poluzzi E, De Ponti F. Cyclin-dependent kinase 4/6 inhibitors and interstitial lung disease in the FDA adverse event reporting system: a pharmacovigilance assessment. *Breast Cancer Res Treat*. 2020 Nov 5:1–9. doi: 10.1007/s10549-020-06001-w. Epub ahead of print. PMID: 33150548; PMCID: PMC7641870.

About ONTOFORCE

ONTOFORCE TRANSFORMS DATA INTO KNOWLEDGE

For more than a decade, ONTOFORCE has addressed the problem that many Life Science companies struggle with: bringing together structured and unstructured data to create new insights. These insights lead to accelerated drug discovery, more in-depth insights into real-world evidence, optimized clinical trial research and faster go-to-market.

Do you wish to operate analytically, exploratively, or collaboratively? DISCOVER, the knowledge platform of ONTOFORCE, provides these insights quickly, clearly and efficiently. Combine internal data or commercial data with the public data sources of DISCOVER, and you take the lead.

We already work for customers such as AstraZeneca, UCB and BMS and numerous other life science colleagues. Thanks to the intense collaboration with renowned research institutes such as IMEC, VIB, UGent, KULeuven and Stanford University and international research and industrial consortia such as ELIXIR, FAIRplus and Pistoia Alliance, you have the guarantee of engaging with a global player that has made translating data into insights its primary objective.

-  ontoforce.com
-  twitter.com/ONTOFORCE
-  linkedin.com/company/ontoforce

MAIN OFFICE

Technologiepark 122
AA Tower, 3rd Floor
9052 Ghent, Belgium

-  ontoforce.com
-  +32 9 292 80 37
-  info@ontoforce.com