Can Simulated Data Unlock the Potential of Real World Evidence in Oncology?

Adam Reich
Adam.Reich@iqvia.com

Principal
Real World Insights

IQVIA

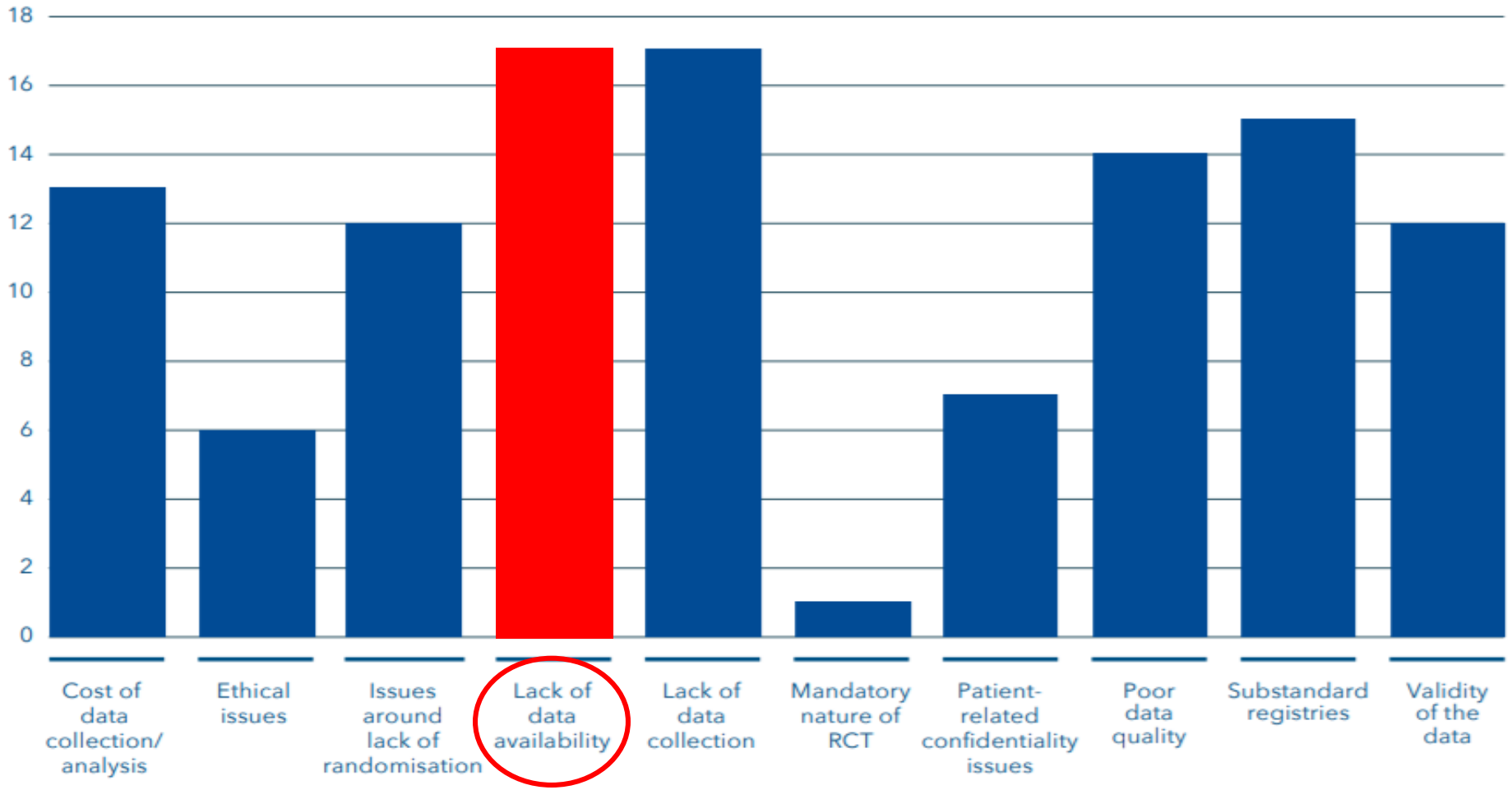
PhUSE is a global community of professionals, passionate about the advancement of clinical information.

Providing the industry with the premier platform for creating and sharing ideas, tools and standards around data and statistical & reporting technologies, PhUSEi s a non-profit membership society run by volunteers.

Today's event will focus on the handling and analysis of real-world data and the challenges surrounding this, from large datasets to dirty data, validation and standardisation.

What is the Largest Barrier to Expanded Use of Real World Evidence?

According to Key Stakeholders across Europe, the largest barrier to expanded use of Real World Evidence is the lack of data availability



In order to understand the use of RWE in Europe, the London School of Economics surveyed 260 stakeholders, including academics, healthcare professionals, patient representatives and HTA bodies, from 17 European countries. The above graph reflects the results of this survey.

RWE Challenges and Limitations

- **Data Source Identification**: For those of us working in real world evidence or seeking to utilise real world data, identifying the highest quality and most appropriate sources of data is a regular challenge
- **Data Access**: Even when those sources exist, finding ways to facilitate access and leverage the information can be an even more arduous task
- **Patient Privacy**: Due to rigorous information governance rules in place to maintain patient confidentiality, the challenge – especially in the UK – has always been to give researchers the detail they need in a way that protects the identity of patients

"It is essential that the use of RWE should not compromise patients' confidentiality. When using RWE, the rights of the individual must be balanced against the overall benefit"

Public Health England's Cancer Analysis System (CAS)

A valuable source of robust cancer data collected in England

The National Cancer Registration and Analysis Service (NCRAS) within Public Health England (PHE) collects data **on all patients diagnosed with cancer in England** and has permission without patient consent under Section 251 of NHS Act 2006. The data is used to support public health, improve healthcare outcomes and aid research.

The data collected feeds into the **Cancer Outcomes and Services Dataset (COSD)** and is then **linked with other datasets** to create the **Cancer Analysis Service (CAS)**

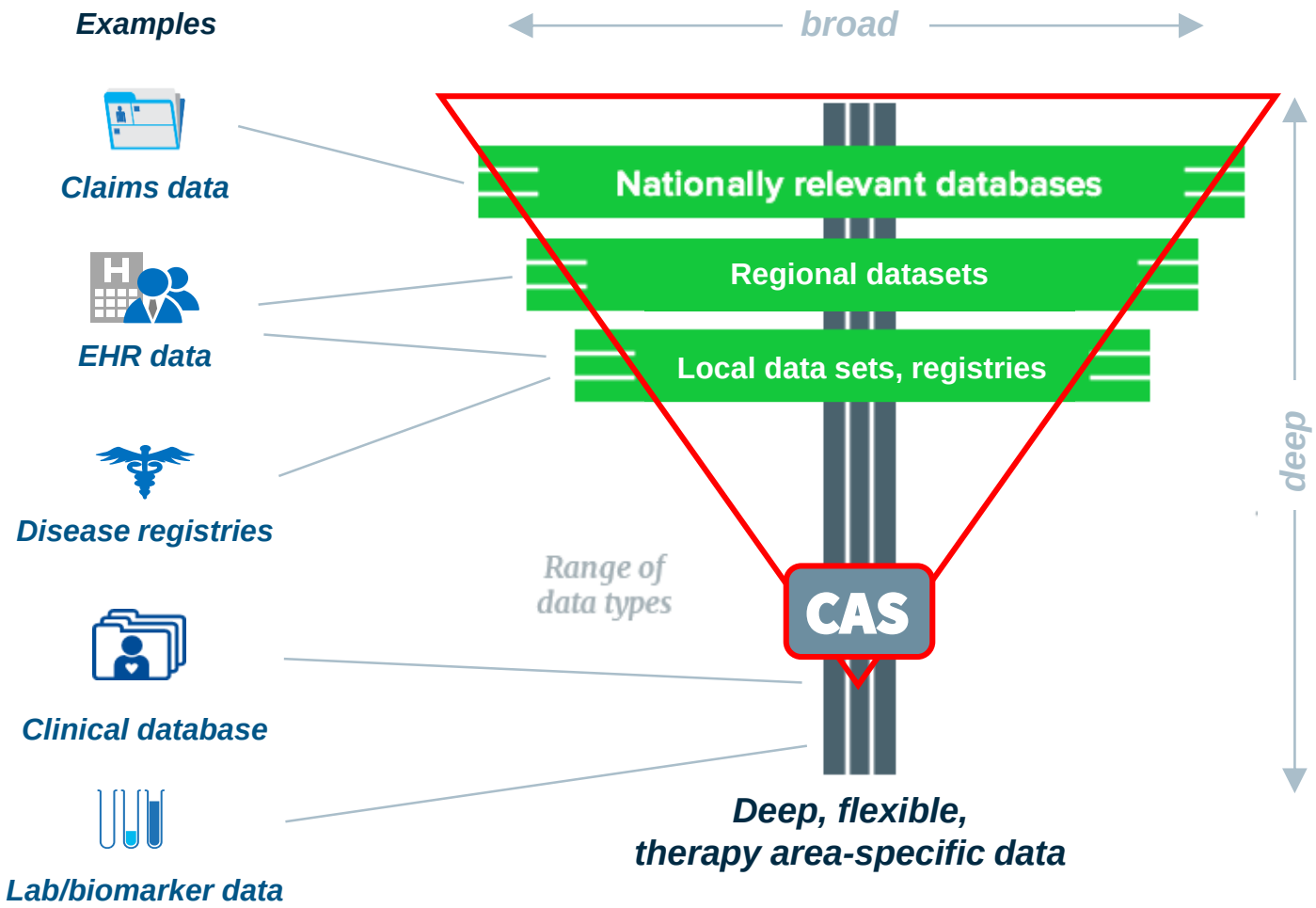
Of these linked datasets, the two most significant are:

- The **Systemic Anti-Cancer Therapy (SACT)** dataset, which tracks all systemic treatments (chemotherapy, biologics, immunotherapy, hormones, etc.) prescribed by NHS providers
- The **Hospital Episode Statistics (HES**) dataset, which contains details of all inpatient admissions and outpatient appointments at NHS hospitals

CAS is one of the most comprehensive **population-level cancer** databases in the world, containing robust information about **Patient and Tumour Characteristics**, **Treatments**, **Mortality** and **Hospitalisations** for all cancer patients in England

Finding data sources with both breadth and depth is challenging

Comparing breadth vs. depth in real world data sources



RWD sources spanning broad, near population level cohorts with limited clinical information about each patient

Clinically rich, deep data for a distinct sub-population of patients

A Comprehensive Source of Cancer Data at Our Fingertips…

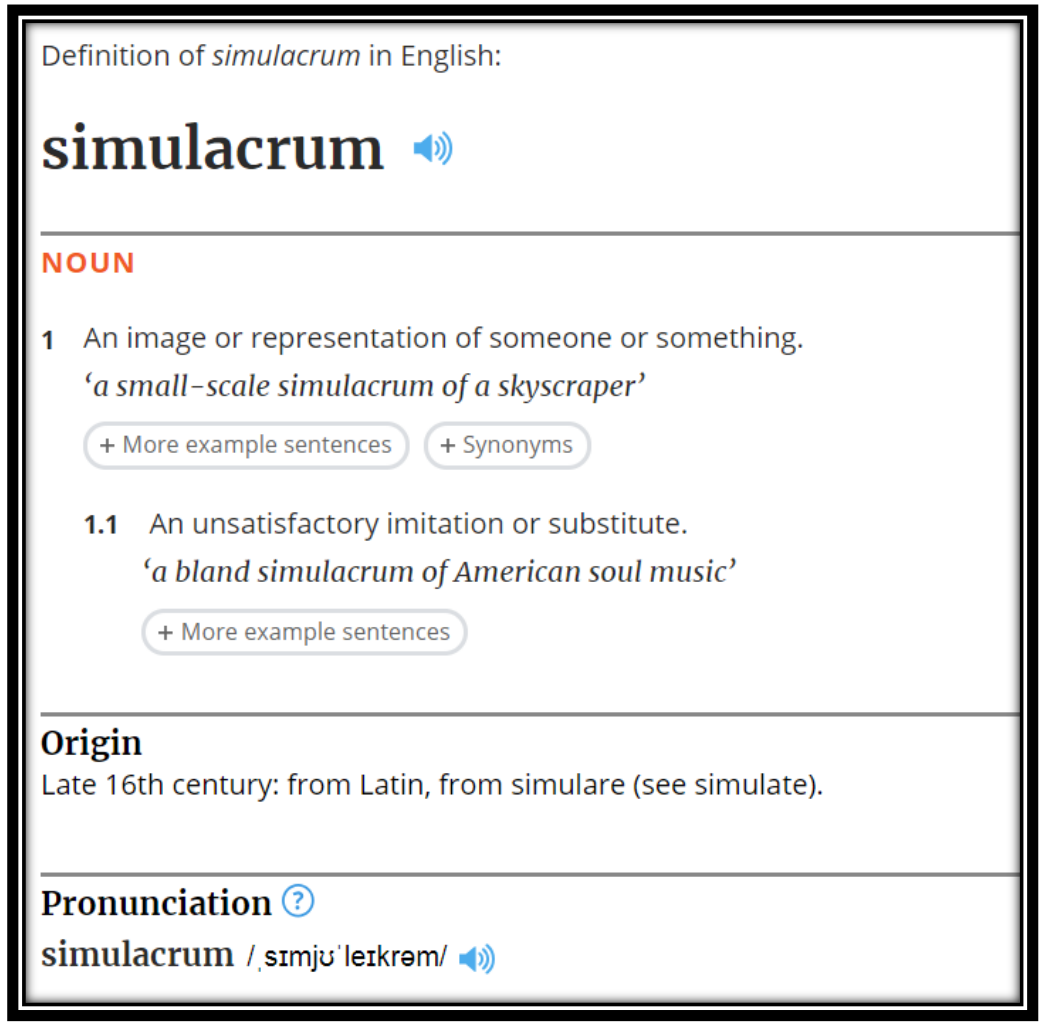
- CAS is an extremely important and valuable data source that can be used to better understand patient care, improve outcomes, and support drug development and research
- However, CAS is **sensitive and confidential**, so PHE has a responsibility to protect patient privacy and safeguard the use of the data
- For this reason, access has been heavily restricted and the data vastly underutilised

What is the Benefit of Having a Robust, World-Leading Cancer Registry if No One Can Use the Data?

That's Why the *Simulacrum* was Developed!

- The main purpose of the **Simulacrum** is to facilitate research based on the data held in the CAS whilst protecting patient confidentiality. By using synthetic data in place of the real data, researchers can work with the data without using identifiable information about any individual.

What is the Simulacrum? How Does it Facilitate Research?



Definition of *simulacrum* in English:

simulacrum

NOUN

1 An image or representation of someone or something.
'a small-scale simulacrum of a skyscraper'

+ More example sentences + Synonyms

1.1 An unsatisfactory imitation or substitute.
'a bland simulacrum of American soul music'

+ More example sentences

Origin
Late 16th century: from Latin, from simulare (see simulate).

Pronunciation
simulacrum /ˌsɪmjʊˈleɪkrəm/

What is the Simulacrum?

The Simulacrum is a synthetic oncology dataset structured to model the properties of the Cancer Analysis System (CAS) in England, but which contains no actual patient data. It was developed for the purpose of better understanding the structure and complexities of the CAS database without the potential risk of patient identification.

Available to the general public as of November 2018

"The data in the Simulacrum is entirely artificial. It does not contain data about real patients, so users can never identify a real person. It allows anyone who wants to use record-level cancer data to do so with no danger of breaching patient confidentiality. We have kept the data model – the shape of the data – the same as in CAS so that it can be used to write and test queries that would run on the real data."

- Dr. Jem Rashbass, Director for National Disease Registration at PHE

Simulacrum (n.d.) In *Oxford English Dictionary*. Retrieved from https://en.oxforddictionaries.com/definition/simulacrum

The Complex Nature of Cancer Data

Cancer data can be extremely complex and often times requires considerable assessment and exploration before being fully understood.

Some of the biggest challenges in working with cancer data are understanding:

- Strengths/Limitations
- Completeness
- Data Infrastructure
- Availability of Variables
- Clinical Definitions
- Analytic Complexities

Patient Groups Support Responsible Data Sharing

Collaboration with these organisations is critical to promoting the benefits of sharing and using data

use MY data

Is a movement for patients, relatives and carers

Harnessing the patient voice to build confidence in the use of patient data for research and analysis

Our vision

Our vision is of every patient willingly giving their data to help others, knowing that effective safeguards to maintain the confidentiality and anonymity of their data are applied consistently, transparently and rigorously.

"What's important is that we continue to develop research and continue to develop treatments for patients and the only way to do that is through research and through gaining access to appropriate data that is actually going to change policies, and change operational systems and change treatment for patients."

- Quotation from Use My Data Member

What Questions Do Patient Groups such as use MY data Have about Synthetic Data?

- Will this make any difference to researchers getting access to the real data?
- If this is so good, why hasn't it been done before? Or has it?
- Who "owns" the synthetic data and what usage controls are in place?
- Where is value created? And how is it measured?
- How close is the synthetic data to being "real"? Will the same disclosure controls (suppression rules) apply?
- What patient data was used to create this work [the Simulacrum]?

Questions above submitted by Chris Carrigan, Expert Data Advisor and Advocate, Use My Data. Presented during Pharmaphorum Webinar titled Can Simulated Datasets Unlock the Potential of Patient Data? Retrieved from https://pharmaphorum.com/debates-insight-patients/can-simulated-datasets-unlock-the-potential-of-patient-data/

How Was the Simulacrum Generated?

The Simulacrum was generated synthetically by mathematical and computational analysis of tens of thousand of anonymous extracts from the actual CAS data.

The Simulacrum was generated in several steps:

1. All data on cancer patients was compiled and anonymised by PHE before it was made available.
2. This data was divided into groups known as data cubes. They are completely anonymous and have been safely released to the public by PHE.
3. Using an algorithm, strong relationships between patient characteristics were identified. These relationships were identified so they could be replicated in the Simulacrum to make the data as realistic as possible.
4. Patient records which showed these strong relationships were gathered into groups of no fewer than 50. This was done to improve accuracy and add extra protections of privacy.
5. The synthetic data was then created by using these data cubes. For each synthetic patient record, a value was randomly chosen from the relevant data cube to represent a particular characteristic. This process was repeated for each new characteristic simulated. This random sampling further protects patient privacy because the procedure cannot be reversed to recreate real patient information.

Description of Simulacrum development comes from Health Data Insight CIC.
Retrieved from https://simulacrum.healthdatainsight.org.uk/developing-the-simulacrum/

An Environment Where Speed and Accuracy are Both Critical

Working with cancer data is challenging in many respects. And with some of the best sources (including CAS) reporting data with a lag of up to 12 months, the amount of time required to conduct an RWE study is even more critical.

The timing and efficiency of an RWE study is dependent on a number of factors, but is often influenced by the following:

- Data Access Process
- Study Complexity
- Study Iterations
- Scope Changes
- Quality Control
- Medical/Clinical Review

RWE Studies in CAS Require Applications and Approvals for Data Release which Prolong Timelines and Delay Research Results

Study Concept/Protocol Approved

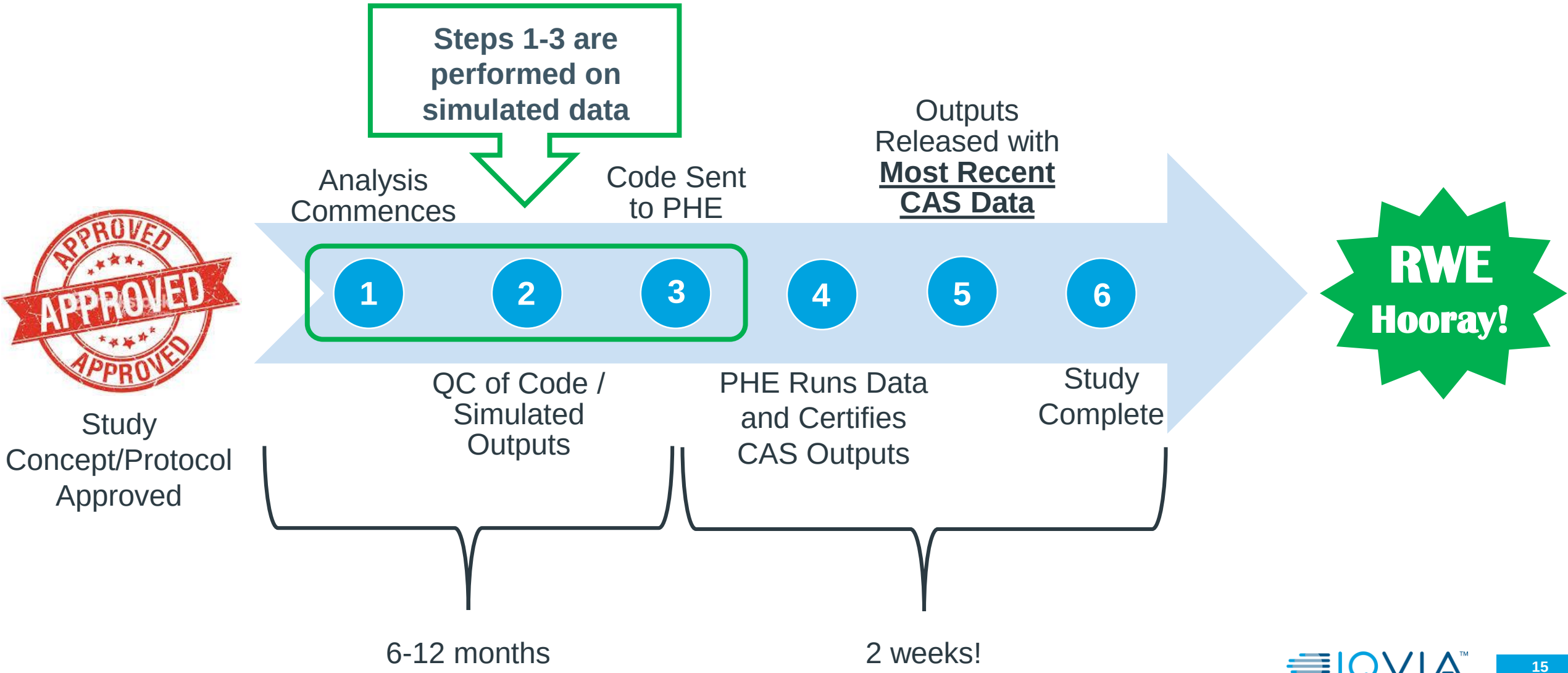
1. Application for Data
2. Application is Accepted
3. Data extract is released
4. Analysis Commences
5. QC of Outputs
6. Study Complete

RWE Hooray!

Up to 12 months

6-12 months

Simulated Data Changes the Framework of CAS Studies to Generate RWE Quicker and More Efficiently



Harnessing Simulacrum to Glean Real World Insights from CAS

Strengths of CAS Data

- Robustness and Completeness: Covers nearly all patients diagnosed with cancer in England
- Clinical Detail: Primary diagnosis, morphology, histology, staging, grade, performance status
- Comprehensive coverage of systemic anti-cancer therapy (chemotherapy, biologics, targeted agents, etc.)
- Mandatory death flag and death date for survival studies
- Linkage to Hospital Episode Statistics for coverage of hospitalisations, surgeries and adverse events

Benefits of Leveraging Simulacrum

- Conduct feasibility work and test coding without compromising patient confidentiality
- Develop study designs that are clinically relevant and statistically precise by refining methodology
- Encourages the use of anonymous, aggregated data
- Increases efficiencies and reduces the time needed to perform analyses
- To overcome the historically inaccessible nature of National Cancer Registry data in England

Unlocking the Potential of CAS Data to Support Drug Development and Health Outcomes Research

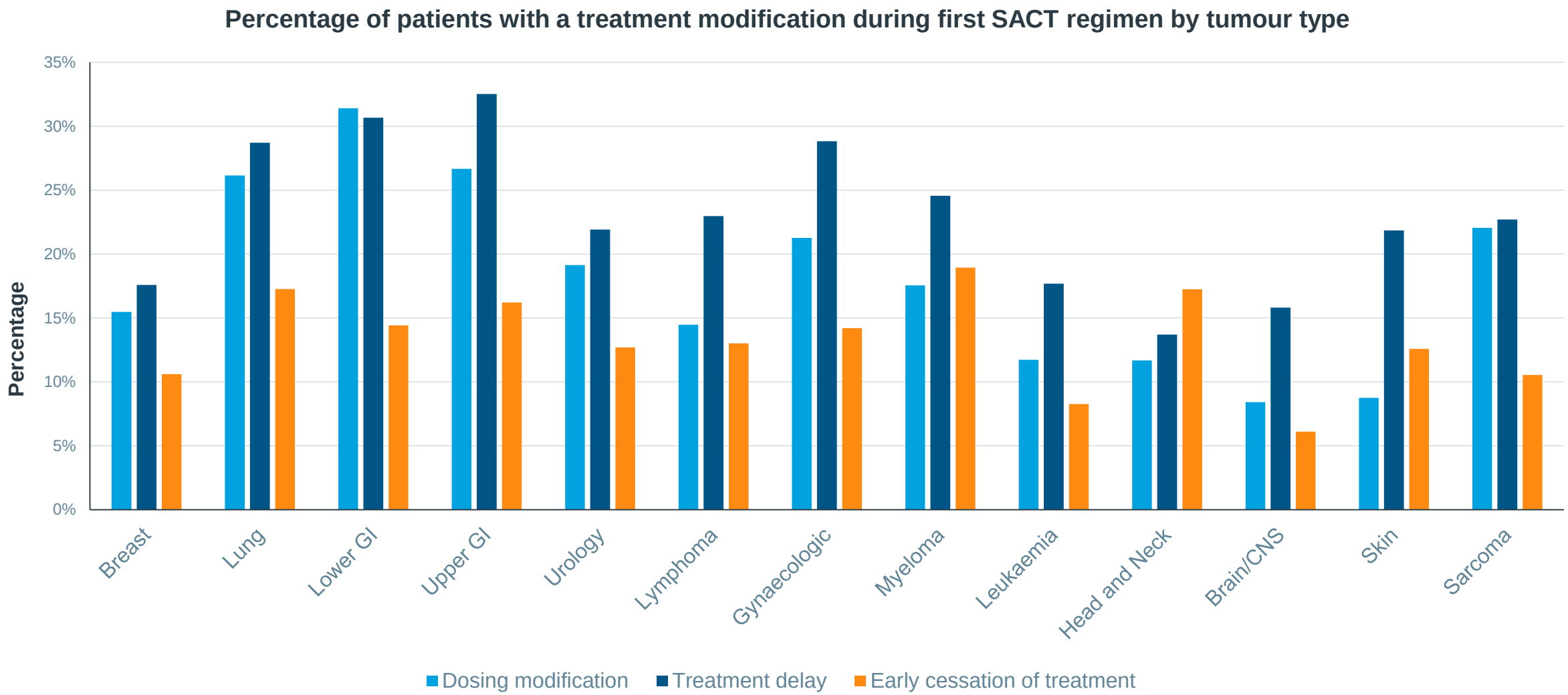
1. **Evaluate the Market:** How many people suffer from the condition? What drugs are currently used in the treatment of the condition and to what extent are clinical guidelines being followed?
2. **Identifying an Unmet Need:** To optimize inclusion and exclusion criteria for clinical trials, to help speed recruitment of patients and ensure relevant results; and to identify patients that are not benefiting from currently available medicines
3. **Understanding the Patient / Treatment Pathway:** To determine more effective patient pathways for specific populations; how might current pathways be adjusted following the launch of a new product
4. **Assessing Value / Patient Outcomes:** Helping to determine the value of treatment, which may reflect a variety of outcomes, including compliance, overall survival and patient response to treatment
5. **Protecting the Safety of Patients:** Using the data to augment clinical trial data in determining a medicine's safety profile including adverse events
6. **Expanding the Indications for Approved Therapies:** To provide persuasive evidence for expanding the approved uses of a drug to new types of patients and new diseases with limited options

Types of RWE Studies in CAS that Have Already Been Completed Through the Use of Simulated Data

Providing real-word oncology insights on patient pathways, treatment use, and clinical outcomes

- **Patient Profile by Tumor/Product**: Demographic (age, gender, ethnicity, ECOG) and tumor (TNM, met sites, morphology, grade, histology, staging) details for patients with a particular tumor type who are using a drug of interest
- **Time to Treatment Analysis**: Number of days from initial primary cancer diagnosis to initial treatment and/or treatment with drug(s) of interest for a cohort of patients meeting a basic set of inclusion criteria
- **Adoption of Medicines**: Over time, how quickly are physicians prescribing/using newly approved treatments? Are these new medicines being used among certain subpopulations or after a particular therapy fails?
- **Overall Survival**: Time from initial primary cancer diagnosis or specific SACT treatment to date of death for various patient cohorts (comparing cohorts with different profiles or different types treatments)
- **Drug Dosing and Duration of Therapy**: Following a patient from initial drug treatment through regimen completion and/or progression. Key metrics would include average dosing or dosing ranges, number of treatment cycles, and duration of therapy (regardless of line).
- **Line of Therapy Distribution**: Of patients using drug(s) of interest, in what line of therapy are they being treated? This requires building line of therapy where we must apply a pre-determined methodology/algorithm.
- **Treatment Pathway**: Following a patient from initial cancer diagnosis (or treatment) for a set amount of time. Key metrics would include number of treatment cycles, duration of therapy by line, drug switches, surgeries, etc.

Example: Understanding Treatment Modification Among Patients

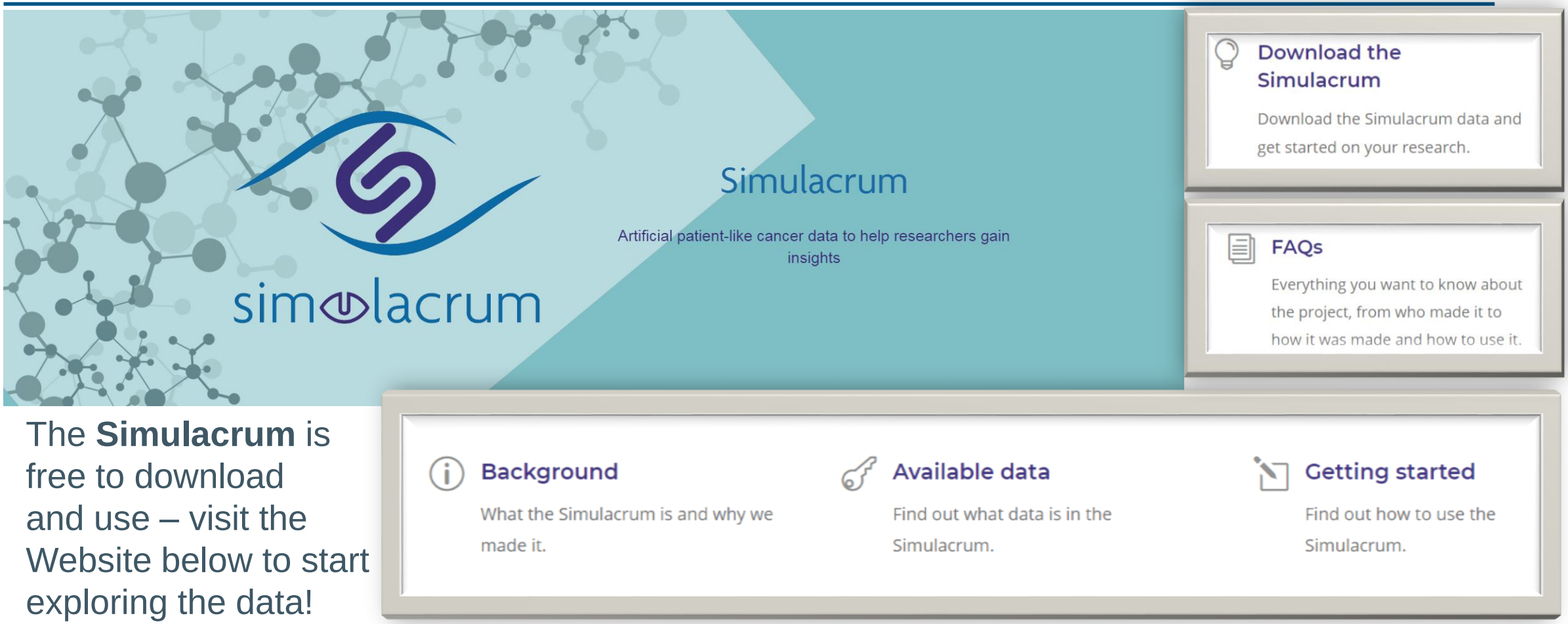


Experiences from Working in the Simulacrum

What are some of the key takeaways from using the Simulacrum to generate Real World Evidence?

- Working in the simulated data has provided a better understanding of the CAS data structure and completeness of important data fields
- We can test whether specific outputs from CAS may potentially be disclosive and adjust accordingly
- Properly linking information across data tables took time to sort out
- We've been able to establish standardised approaches to answering specific types of questions
- The Simulacrum, as expected, has limitations…but they can be overcome
- Studies run via the Simulacrum have brought attention to areas where data collection needs to be improved or amended

Want to Know More About the Simulacrum?



The **Simulacrum** is free to download and use – visit the Website below to start exploring the data!

https://simulacrum.healthdatainsight.org.uk/