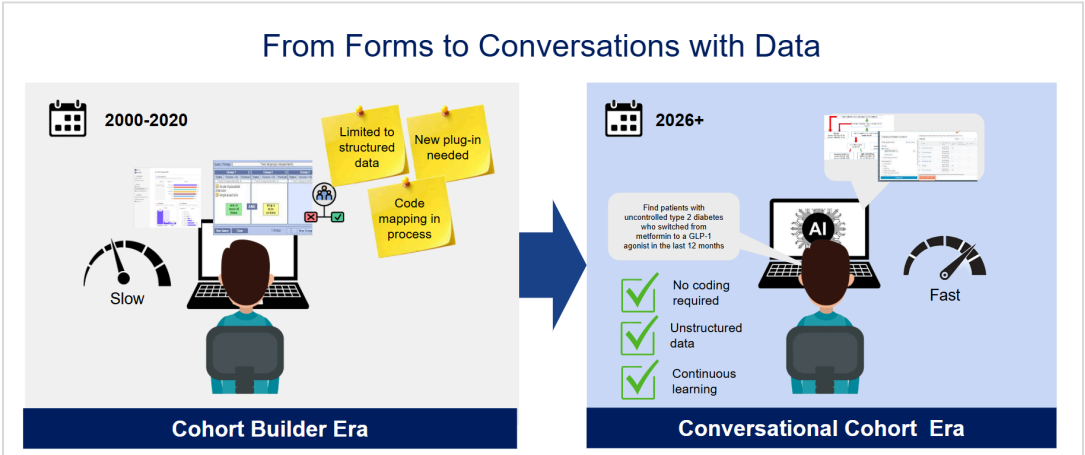
graticule

Overcoming Governance Challenges in Agentic AI Cohort Discovery

For decades, real-world data (RWD) researchers have relied on structured cohort discovery platforms—tools like i2b2, OMOP Atlas, commercial systems, python templates, or custom-built interfaces. I have worked with great teams extending these open source tools or building new ones like the work we did for tranSMART that extended i2b2 and the RWE tools we built at ConvergeHEALTH that extended the OMOP data model while I was at Deloitte. Other companies have done great work as well including TriNetX, Deep6, SignetAccel, Explorys, Truven/Marketscan, Amalga, transMed, ConvergeHealth Miner, and Epic's Slicer Dicer. Most real-world data aggregator companies like Holmusk and Loopback Analytics have also put some form of self service cohort builder on top of their data as well.

As an admission I vowed a few years ago to quit creating new cohort builders as if I was an addict. "Hi - I'm Dan. I design cohort builders and I have a problem."

These systems were a major achievement in the 2000s and have continuously improved and adapted as more data has become available. They allowed us to define populations of interest through inclusion and exclusion criteria, carefully mapped to ontologies and standardized terminologies. They gave us rigor, repeatability, and structure. But they've also been slow, rigid, and have had limited access and often could only get to a point where the analysis was stuck on a hard question out of scope of the tool. Also many had adoption challenges where white glove services were needed to allow intended users to make use of the tools because it required too much special knowledge to enter queries or interpret the results properly. Only those who were expert in programming extensions to the query tools or custom development capabilities could provide new views on the data such as plug-ins, new code branches, or custom views.



The shift from structured query tools to agentic AI systems has the potential to transform how we generate insights from real-world data

Now, something new is emerging. Across pharma companies, academic research groups, consulting firms, and data aggregators, I am seeing multiple early pilots and production implementations of **natural**

language interfaces for real-world data. It falls into this category broadly speaking of the buzzword-worthy Agentic AI applications. Instead of relying on analysts to translate research questions into SQL, i2b2, or OMOP concept set queries, researchers are starting to do something remarkable: simply talk conversationally to agents that query the data and get custom views iteratively generated.

In my role as CTO of Graticule I have seen demonstrations of the first generation of these tools that will start to be unleashed shortly in trade shows, marketing materials, AMIA presentations, VC pitches, consulting vision introductions and podcasts.

This is a prediction - *We will see an explosion of agentic cohort explorers in 2026 as development teams learn how to build and secure generative AI to support the cohort selection and qualification workloads.*

At Graticule we will be involved in some of them in different ways and yes this is a partial relapse for me into cohort building tools. But Generative AI is as big of a paradigm shift as the original bomb of public internet and the world wide web. This sudden movement toward agentic approaches to query patient data reminds me of when we first began to use cloud computing.

Another prediction and warning to innovators: *Agentic cohort exploration is going to offer incredible benefits but will be slowed significantly in early phases because of the unchartered risks and uncertainty for proceeding.*

My prior company, Recombinant Data, had been working with AWS on many early patient data and translational research projects. When we were acquired by a larger consulting firm the larger firm worked hard to quickly restrict the ‘cloud computing use’ out of concern for the many risks of cloud. These included data leaving the internal physical client infrastructure, potential IP ownership claims by the cloud vendor, and the unknown risks for security of computers with highly sensitive data connected to an open internet infrastructure. All of those were and are valid concerns. Things quickly risked grinding to a halt.

So with restricted access to the cloud and clients who wanted the benefits of compute and storage as a service I started putting the cloud computing bills for clients on my corporate credit card. This work around went on for a year or so until the risk and compliance department and the business all aligned on going forward in a big way with cloud. It was probably not the best solution but it worked.

What I’m trying to say with this example is that, like the shift to cloud, the vision for genAI and patient data becoming realized is inexorable but the path there is likely to be blocked for longer than anyone expects and primarily with governance issues. This will make building trust the key to the speed of adoption and realization of value of agentic cohort building and analysis systems. But more about the vision first.

A Vision of What’s Possible

Imagine asking a question like:

> “Find patients with uncontrolled type 2 diabetes who switched from metformin to a GLP-1 agonist in the last 12 months, and compare their hospitalization rates within 6 months.”

That’s the kind of question that happens often in an RWE team before even doing detailed feasibility for a project. And now imagine instead of days of back-and-forth with analysts, the system responds within

seconds with a count and summary of the attrition table for the population with a breakdown for each of the primary criteria.

Behind the scenes, the AI leverages embedded medical knowledge, concept sets, and published evidence. It maps “uncontrolled diabetes” into codes and lab thresholds, interprets “GLP-1 agonist” into a rapidly evolving drug list of incretins, and constructs the right SQL queries against the underlying data models for the source data. It can even pull in context from the literature to explain the assumptions it made for what to include in the concept sets.

Then after you receive the cohort count you then ask refining questions about those patients such as - “show me a summary the outcomes for those patients that are relevant to diabetes and obesity” or “summarize drug switches among those patients and potential adverse events from notes or other changes in their medical records that were likely to lead to the drug switches”. You could ask about attrition among the variables relevant to the analyses constructed thus far. You can keep iterating and the tools will rapidly query data and return contextually relevant answers that you can refine with plain english requests. Each response leads to the ability for more intelligent next questions based on what you see and what you understand as potential gaps. And this is before we even consider if we gave the model an abstract question like - “What is an efficient study design based on data in this data set for a trial to test the novel incretin we have in phase 1 with properties from the file I just uploaded?”

This isn’t science fiction anymore—it’s happening in early prototypes and pilots. We generally all know it’s possible because we have played with the publicly available natural language tools like Microsoft Co-Pilot, Claude, and Chat-GPT. They do amazing things for general use all the time now. They have become mainstream and featured in South Park episodes as a character! But now these tools are being tuned to support our investigation of complex medical data that is both sensitive and requires a lot of specialized knowledge to query it successfully.

Breaking Free from the Old Rigid Structures

Legacy cohort builders were powerful but narrow. They worked well for structured data and predefined logic, but they came with real frustrations that slowed research and limited discovery:

- **Rigid workflows**: Queries had to be built in fixed, stepwise sequences.
- **Schema dependency**: Data had to be painstakingly mapped into common ontologies like OMOP, PCORNet, i2b2, or proprietary structures before it was even queryable. Entire categories of data were often unavailable until mapping was complete.
    - For example, i2b2 projects often struggled with laboratory data: historical results were not coded in LOINC, but the query ontology required LOINC, which meant researchers couldn’t use those data until LOINC mapping caught up to the backlog of uncoded concepts.
    - In Europe, many RWD sources face similar challenges today with medications. Poorly coded drug tables make it nearly impossible to answer basic questions such as *“How many patients with condition X have ever taken treatment Y?”*

- **Unstructured data blind spots**: Free-text notes, imaging reports, and pathology narratives remained largely inaccessible because NLP was also rigid and a separate process from cohort development. It operated in awkward ‘plug in‘ approaches that didn’t scale or intersect well with the structured data.

- **Manual optimization**: Analysts had to tune and restructure queries to get them to run efficiently

Form Based vs AI-Driven Cohort Builders

Dimension	Form Based Cohort Builder	Conversational AI Cohort Builder
Workflow	Rigid, stepwise query building	Flexible, dynamic query design
Data Model	Dependent on schema mapping	Reduced dependency; can query raw data directly
Concept Model	Requires mapping to concepts	Limited reliance on manual mapping
Data Type	Mostly structured data; unstructured data ignored	Structured + unstructured (notes, imaging, pathology)
Iteration	Slow; requires rework for each change	Real-time iteration & immediate feedback
Optimization	Manual query tuning required	The system adapts based on performance
Engagement Model	User relies on informatics analyst to build cohorts	User works directly with the tool, replacing the analyst step
Knowledge	Limited to predefined criteria; researchers must know what to ask in advance	Surfaces new concepts, relationships, and hypotheses using embedded medical knowledge

Natural Language AI Driven Cohort Builders: Flexibility, Agility, Discovery

AI-driven cohort builders flip this model. They can:

- **Design queries on the fly**, adapting structure dynamically instead of forcing users into predefined boxes.

- **Optimize automatically**, rewriting or tuning queries to improve performance without user intervention.

- **Interrogate unstructured data**, including free-text clinical notes, in ways that structured tools could never achieve and can integrate the unstructured outputs with structured data in the output tables.

- **Reduce dependency on common data models**, translating conceptual questions to fit heterogeneous databases. Queries can run across multiple raw data sources without months of mapping work and aligning the data to model the query tool was built against.

- **Enable agility through real-time iteration**. Researchers can refine queries through multiple cycles, getting immediate feedback and clarifying definitions as they learn from the data. Issues

can be identified and resolved rapidly

- **Discover relevant concepts automatically**. AI knows which treatments are relevant for a disease, what outcomes to track, and can propose new avenues for exploration—turning cohort building from a search exercise into a discovery exercise.

This combination of flexibility, agility, and discovery shifts cohort building from a “fill out the form” exercise into a **conversational, hypothesis-generating partnership** with the data.

From Querying to Discovery

Traditional cohort builders forced researchers to know what they were looking for in advance. AI changes that:

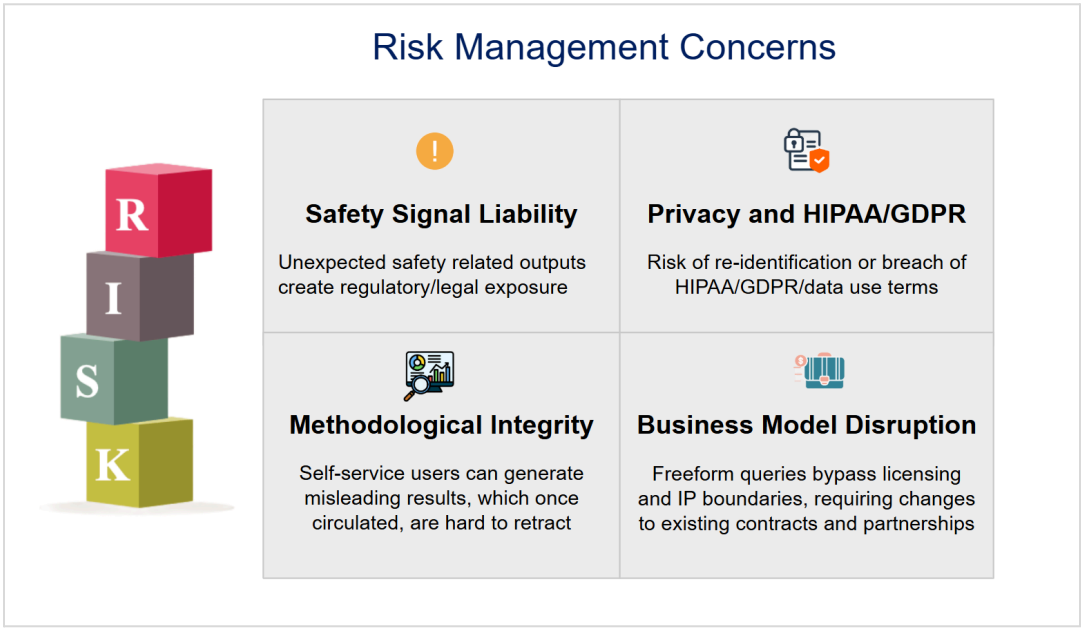
- **Iterative refinement**: Users can experiment and learn in real time, with the system suggesting improvements or alternative approaches.
- **Dynamic optimization**: The system adapts queries for performance and compatibility across different datasets.
- **Conceptual linking**: Embedded medical knowledge allows AI to connect drug classes, disease definitions, and outcomes automatically.
- **Uncovering the unknown**: By exploring known treatments and outcomes, the AI surfaces relationships and insights researchers might not have anticipated.

In short, Generative AI allows teams to **discover patterns, relationships, and insights** that were previously hidden behind rigid query structures by opening the aperture on how the questions can be asked. But here is the issue.

But… what happens when we first say - “You can ask the query tool for anything!”

But With Great Power Comes Real Risk

The flexibility of natural language interfaces is compelling, but risky. Legal teams often assume generative AI creates IP infringement risks with third-party data vendors and these misconceptions can stall deployment until clarified. Here are some simple example objections that quickly come to mind for what a risk management group will flag halting live use of the new AI cohort builder approach:



1. Safety Signal Liability
A casual question like *"Does Drug X increase mortality?"* could return preliminary safety signals. Within pharma, that triggers strict reporting obligations, compliance reviews, and potential legal exposure.

2. Privacy and HIPAA/GDPR
Natural language queries could inadvertently isolate small groups—or even individuals. Queries like *"Outcomes for Hispanic women in Houston under 25 with cystic fibrosis"* might cross re-identification boundaries. While source data may be deidentified, the use of data may be outside of what regulations allow for the data being used from a data use agreement established to manage ethics and research constraints.

3. Methodological Integrity
Without proper guardrails, self-service users could generate misleading results, skip confounder adjustments, or misinterpret outcomes. People may publish or share insights with colleagues or customers that really are hallucinations or that have methods that have been proven biased or statistically insignificant. Once circulated, flawed insights are hard to retract and can cause harm or wasted time refuting.

4. Business Model Disruption
Freeform querying can undermine licensing agreements, enabling access beyond what was intended. This creates risk for traditional data monetization, existing partnerships, ownership structures of assets and Intellectual Property.

Governance: Policies, Training, and Publication

Responsible deployment requires **more than technology**. It requires establishing Governance and controls around the governance both through education and technical controls.

- **User Training**: Educate users on appropriate query types, boundaries of their role, and approved research protocols.

- **Policies and Restrictions**: Define clear rules for how AI-generated insights can be used, published, or communicated.

- **Publication Processes**: If an insight may have external value, it should be escalated to an evidence generation team to design a validated study independent of the AI tool.

- **Role-Based Access**: Tier users according to expertise, ensuring exploratory users cannot perform analyses beyond their scope.

Combining technology, policy, and training builds a foundation for safe, responsible AI-driven cohort discovery. Given the advanced nature of the tools as well as the environment for building them it is likely that engineering these controls into the system will become a large component of the AI engineering problem. Basically the next generation of systems will rely on agents or AI monitoring prompts to stop or redirect the generation of disallowed or risky operations. These systems will prevent or suppress surfacing outputs that appear to be 'illegal' from a policy perspective.

Business Model Innovation

AI-driven tools don't just change workflows they create **strategic opportunities**. The advent of reinventing access to the data through a conversational interface changes the nature of the business being offered. For licensors or groups that offer source data as health systems seeking to support collaborations these new tools will likely mean figuring out how to use them to break the monolithic licensing models such as 'license all the data' and enable creative frameworks closer to answering questions or tiering access to specific enriched information carved out by the tool. Here are some innovations we will likely see :

- **Subscriptions and Tiered Access**: Data providers can offer AI tools as a service, with tiered capabilities. Higher tiers could allow subcohort creation or enriched cohorts for licensed use.

- **Enriched Cohorts as Products**: Generate and license cohorts dynamically, monetizing insights rather than raw data alone.

- **Clinical Trial Feasibility**: Health systems could use the AI to rapidly assess patient populations, supporting faster, smarter site selection.

- **Disintermediation & Federated Models**: Health systems may reduce reliance on third-party aggregators, or new federated AI aggregators could emerge to query distributed datasets without centralizing them.

- **Metered Abstraction Plugins:** Well trained models for specific subject areas will be used to process data. Such as an oncology or cardiology LLM. These will be trained on specialized data and then made available with metering for use in terms of how much data is processed through them or be used in place of a human normally doing abstraction.

graticule

In essence, AI-driven cohort discovery reshapes **who captures value** and **how**.

Engage Early, Not Late With Compliance

This is an old lesson and one to write 1000 times on the chalk board with regards to new innovations with the use of patient data. Organizations must involve **privacy, compliance, and business teams early**. Too often, technical teams build tools only to find deployment blocked by governance hurdles and then need to re-engineer. This will be even more important and complex with something as invisible to the compliance teams as a magical box that listens to inputs from a user and provides answers against whatever data is available to it. Some early steps to consider:

- Take the time to educate compliance and business teams on what AI can now do. Give them demos and generate slides that clearly communicate how they work at a conceptual level. Communicate how they can be secured or controlled to meet the known compliance requirements

- Work closely with compliance and legal teams through early education and engagement- address IP infringement misconceptions, map governance and business requirements alongside features. Including the compliance and risk teams early and often will save time in avoiding building components that will be wasteful because they will never see the light of day.

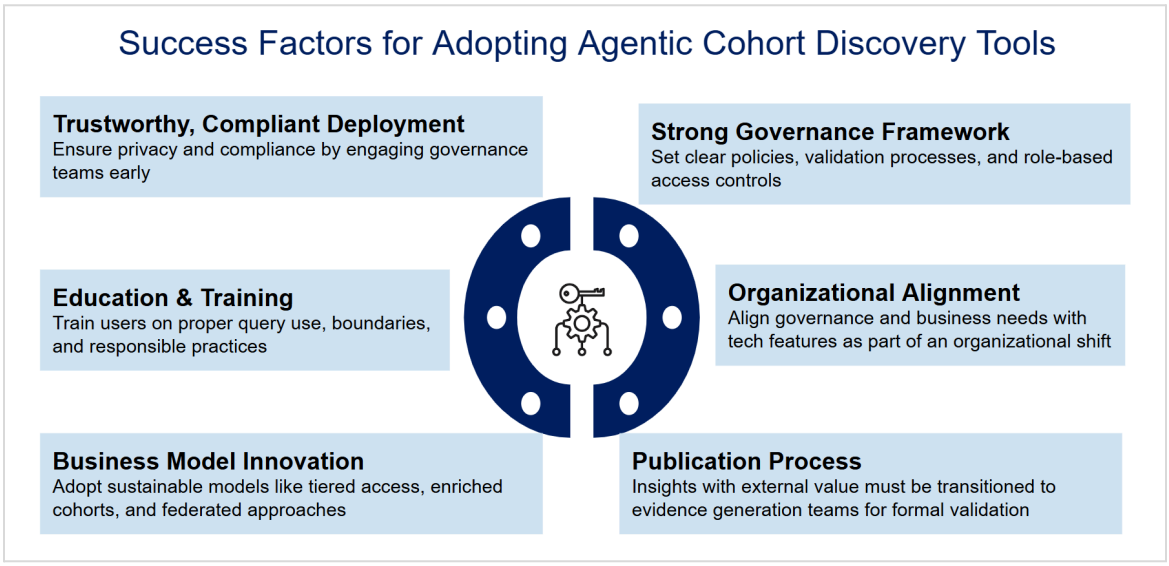
- Show the contrast between rigid structured queries and AI-optimized discovery to accelerate alignment.

- Establish regular check-ins including demos and access to use the tools during development to provide ample opportunity to receive feedback and redirection as the capabilities.

- Author sufficient documentation to describe the approach, risks and controls. This may include creating a protocol and IRB document that covers the typical components that describe a study. While the technology may not require an IRB approval to operate from an organizational or compliance perspective the structure of ethics reviews are effective at forcing a clear design for addressing risks.

- Don’t think that technology can solve for everything. It may take time to fully become aware of either the risks and compliance needs or have the resources to overcome them. Create or use existing SOPs and training to handle the fact that people are the ones interacting with these powerful tools. They will need to be given appropriate skills and approaches to use them in an appropriate manner to avoid generating risks.

Transitioning from prototyping to deployment is not just a technical challenge—it’s an organizational shift. Be prepared and make compliance and risk one of the first partners you build in the journey.



The Decisive Moment

In short, natural language/generative IA/agentic (insert your buzzword here) cohort discovery could be the most transformative leap in real-world evidence since the first standardized data models and query tools. Success won't come from technical breakthroughs alone—it will come from:

- **Trustworthy, compliant deployment**
- **Well-trained, well-governed users**
- **Business model innovation that captures and distributes value responsibly**

The first organizations to balance **agility, discovery, governance, and monetization** will set the standard for the next decade of real-world evidence.

The real question isn't just: *"Can we talk directly to the data?"*

The question is: *"Can we trust what comes back—and build a sustainable, compliant, and innovative ecosystem around it?"*

*Curious to see agentic AI in action? Tune in to our **latest podcast episode** featuring a live demo of Deloitte's Gen AI solution, RWE Agent, and explore what's possible with agentic AI in real-world evidence generation.*

Interested in learning more?
Reach out to us at info@graticule.life