



LET THE USE CASE DECIDE:

The Zelta way on AI and supervised machine learning



Walker Bradham
Senior Director, Product Management & Design, Zelta by Merative

DJ McCloskey
Sr Principal Developer, Software Development Engineering



Table of Contents

03	Introduction
04	Generative AI in clinical trials
05	The pros and cons of LLMs
06	Generative AI hallucinations
07	Why supervised machine learning for Zelta
08	The benefits of supervised machine learning
09	The Zelta track record on creating value with AI
10	What's next for Zelta automation and ML
11	Letting use cases decide the way on AI



Introduction

For decades, when people thought of artificial intelligence (AI) what usually came to mind first were science-fiction movies. But today, the pop culture touchstone for AI isn't Skynet, it's ChatGPT. You'd be hard pressed to find someone who hasn't heard of the large language model (LLM) chatbot generating essays, screenplays, and entire conversations around any given prompt; passing 100 million users in 5 days (TikTok took 9 months), as of this writing, it has been tried out by over 180 million users.

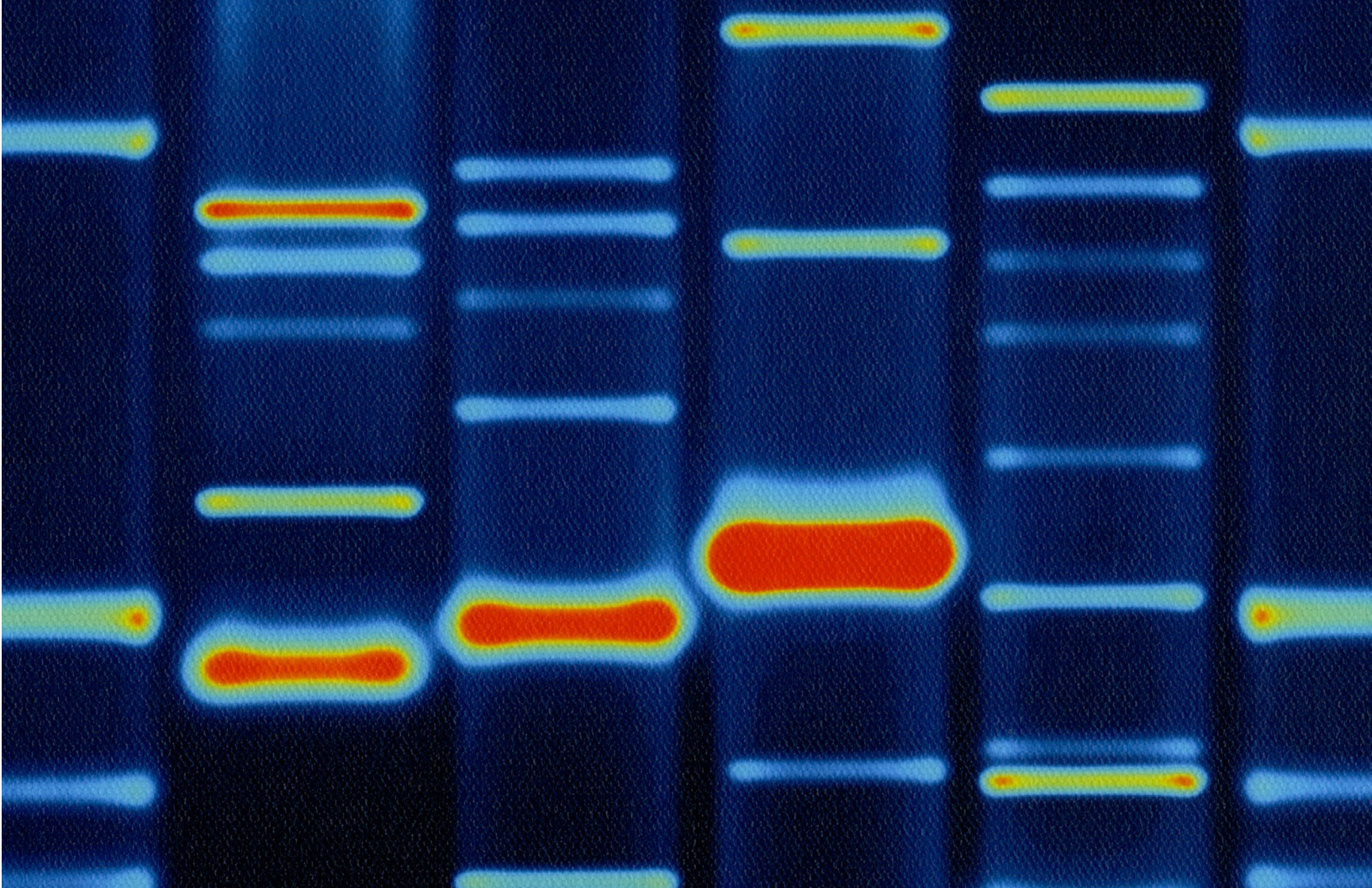
You would also be hard pressed to attend any conference about clinical data management, clinical trials, or clinical development right now without hearing data managers, medical coders, IT administrators, or clinical trial researchers talk about the potential applications of generative AI in their work – and for good reason. We've been exploring the viability of AI in healthcare and medical research, and its ability to augment human efforts in diagnosing diseases, running clinical trial studies, and accelerating creation of treatments and vaccines, since the 1970s. But is generative AI really the key to unlocking that potential?

At Zelta, we're very intrigued about the possibilities for generative AI in clinical trial studies. As an eClinical platform provider that has been working for over a decade to help trial teams build, execute, and accelerate clinical trials with greater control and confidence, AI and automation are central to our strategy today and our vision for the future. But we believe that realizing the benefits of generative AI, and implementing the given application in a dependable, trustworthy manner, requires careful consideration and validation. That's why our strategy remains **value** driven and not technology driven. Through careful selection of the technology based on the objective, we prioritize performance, reliability, privacy, security, and serviceability. In other words, we let the use case decide.

That's why our strategy is focused on more targeted, value-adding deployments – namely, through scalable supervised machine learning (ML) purpose built directly within our eClinical platform, made available to all Zelta users. Our prediction models are built on a proven, simpler and earlier neural network design than the transformer neural network approach that underpins today's LLMs.

In this paper, we'll delve into some of the basic compare-and-contrast features of generative AI vs. purpose-built supervised ML models, how we evaluate the spectrum of these options to solve user challenges, and the history of how Zelta has leveraged automation and supervised machine learning to help our clients achieve measurable ROI in cycle-time reductions – with a preview of where we'll be pushing the envelope next.





Generative AI in clinical trials

What is generative AI, really?

Generative AI is a subset of AI which distills patterns from training data and then, based on what it has learned, can be prompted to generate new data appropriate to the prompt context. This can mean generating text, images, audio, or even video.

One of the most famous examples of generative AI today is OpenAI's ChatGPT. ChatGPT is based on a Large Language Model (LLM), trained on trillions of tokens of human language text taken from the web and other sources. It can be used to generate anything from (mostly) reliable answers to popular questions, translations, summaries, even resignation letters in the style of Homer Simpson pretending to be Shakespeare – or, more relevantly, a medical diagnosis based on a History of Present Illness section from a patient record.

It is important to understand how it works. At a very high level, the training of an LLM compresses the language drawn from its training data such that it can optimize its ability to predict the next word in a sequence of input words. This has surprised everyone in the AI industry with its power to cause the training to internalize abstractions and generate results that can seem miraculous at times. These models are extremely creative. But, coupled with their immense scale and the opaqueness of their abstractions, this creativity is both their greatest strength and its most troubling weakness.

Such is the success of ChatGPT and other models like Gemini (Google), Claude (Anthropic), and Llama (Facebook), that there is now a very real problem with expectations and understanding the limitations of these systems.

The pros and cons of LLMs

Benefits

Large language models can understand and generate human language text that is able to remain coherent over long passages, without losing the context being drawn from the input data. To name a few examples: translation, open domain question answering, content creation, storytelling, and conversational agents like automated customer support bots. Traditional and supervised machine learning simply cannot compete with LLMs in these areas, particularly when it comes to coherent text generation and semantic interpretation of that text.

Traditional natural language processing (NLP) is another benefit. LLMs are able to perform many standard NLP tasks like classification, entity extraction, or relationship extraction at state-of-the-art levels. While traditional ML methods can compete nearly at this same level, and might even be preferred in some use cases, that approach still entails more narrow domains with appropriate training data sets.

Current concerns and challenges

Hallucination

LLMs can return unexpected or incorrect results in a very convincing manner. This undermines their reliability, especially in applications requiring high accuracy and factual correctness. See next section for more detail and examples.

Inconsistency and unreliability

The same prompt can generate different responses on subsequent attempts. This inconsistency can be problematic in applications that require reproducible and stable outputs. It also means outputs may not be reliable, given the risk that responses might ignore important context, leading to it give an incorrect answer.

Regulatory and compliance

Some types of LLMs present challenges caused by the lack of interpretability, which in turn are exacerbated by all of the above considerations. Many of the [emerging AI governance standards](#) explicitly classify LLMs as a class of model that requires the highest levels of scrutiny. Regulation is something that any LLM model provider, or user, will have to give careful consideration to when planning their AI use cases.

Time consuming

Depending on the task, response time expectations must be adjusted significantly as LLMs do not always respond as most other APIs that developers might be used to.

Cost

LLMs are very costly per usage, relative to most other APIs used by developers. Users are often charged based on the number of words in their input (prompt) and the number of words in the response. Additionally, LLMs require very specialized hardware to train and run models at scale. While those costs will come down over time as the models become more efficient, we are still a long ways off from that equipment being equivalent in cost to traditional ML technology.

Although all these aspects are drawbacks, this does not mean LLM-based solutions cannot be made viable. But the costs and risks associated with LLMs must be taken into account as part of any detailed clinical trial solution design and implementation.

You
Can you provide MedDRA code for heart attack?

ChatGPT
The Medical Dictionary for Regulatory Activities (MedDRA) is a standard terminology used for the classification of adverse event information from the use of medical products. The specific term "heart attack" is not represented in MedDRA. Instead, it is categorized under broader cardiovascular events.

In MedDRA, the preferred term for a heart attack is "myocardial infarction." The corresponding MedDRA code for myocardial infarction is 10027409. This code is used to classify adverse events and medical conditions related to myocardial infarction in clinical and pharmacovigilance databases.

LLT Code	10027409
LLT Name	Mesothelioma malignant localised
PT Code	10027407
PT Term	Mesothelioma malignant
HLT Code	10025681
HLT Term	Malignant musculoskeletal and connective tissue neoplasms
HLGT Code	10028389
HLGT Term	Musculoskeletal and connective tissue neoplasms
SOC Code	10028395
SOC Term	Musculoskeletal and connective tissue disorders
Primary	N

Generative AI hallucinations

Let's look at a real-life example of what we mean by "hallucination" in the context of LLMs.

We asked ChatGPT to provide the MedDRA code for "heart attack." The response is impressively fluent, clear, and readable. It's very authoritative and believable. And in that response we do get a real MedDRA code returned. The only problem is that it's the wrong code!

Worse still, the full response gives a real sense that ChatGPT both understood the question and has deep knowledge of MedDRA and the medical domain, when it actually has not. It seamlessly uses the correct language from MedDRA ("preferred term" is a central concept in MedDRA), which communicates authority to those of us familiar with it. It conveys its apparently deep knowledge of MedDRA's medical conditions terminology, correctly explaining that "heart attack" is in fact "myocardial infarction." So, we are naturally inclined to believe the answer given. But despite a response confidently written to convince us that its answer is correct, the code given is actually the MedDRA code for mesothelioma, not myocardial infarction.

This is an example of generative AI hallucinating: returning a response that looks right but is just factually not correct. Was this some intentional sleight of hand? No, the LLM is simply choosing a highly probable next word at each position in the text as it generates the response. This probabilistic aspect means that anything that looks like a valid MedDRA code can be an output; it does not have to actually be correct.

Let's take a step back here and assess. The LLM did a fine job of identifying that "heart attack" is synonymous with "myocardial infarction" in the given context. In fact, here we start to see a core strength of LLMs: they really do internalize an excellent representation of how concepts expressed in language relate to one another. Call it an "equivalence of terms." This gives a hint as to how we can use LLMs effectively today. Instead, we could leverage this strength of the LLM to map the input natural language term to its most likely synonyms given the context (e.g., mapping an ambiguous term like "heart attack" to its more formal term "myocardial infarction") and then, as a separate step in the solution implementation, perform a specific verbatim lookup in a dictionary to get a guaranteed correct code for the mapped terms (in this case "myocardial infarction").

There is definite value to LLMs, but clearly careful consideration is needed in how we build our solutions around generative AI. If used naively, the authoritative tone of an LLM-generated response can be much riskier than it first appears. Even the common mitigation for any implementation of machine learning in the clinical application space, of having a human in the loop to validate the output of a model, can be biased by this confident sounding LLM-generated language.

Strict validation protocols need to be put in place for LLM use cases where the cost of error is high. Due to this issue, presenting answers without an authoritative narrative is preferable. Better still, it should draw a user's attention to the fact that the answer may be incorrect. This is a case where supervised machine learning models have an inherent advantage: they are assumed to be unreliable and have no ability to "convince" the user via a persuasive narrative. Rather, we gain trust in the model through repeated real world experiences.

Why Zelta still uses supervised machine learning for many AI use cases

For most use cases, supervised ML models more than suffice, often at a tiny fraction of the cost of LLMs – both to run and to implement. When they meet the requirements, and sufficient real world training data is available, supervised ML implementations are typically more effective, performant, scalable, maintainable, and explainable than LLMs.

Generative AI has a very interesting future ahead of it, and its potential applications for clinical trials are still largely untapped. For those organizations with large amounts of historical trial data and the resources to build their own LLMs, or invest in fine-tuning models that are commercially available, generative AI may bear fruit sooner than later. However, given the concerns discussed, the hype around these generally available generative AI models shouldn't overshadow the fact that there often is a more effective and lower risk way for achieving AI innovation in clinical trial study execution. Favor simpler alternatives to LLMs in any use cases where their unique strengths are not absolutely required, and which do not have a robust validation phase or where tolerance to error is higher – for example, chatbot scenarios where users are clearly warned about the veracity of the responses.

That's why at Zelta we have extended our core philosophy of using the right tool for the right job to include LLMs and generative AI more broadly – in other words, letting the use case decide. However, the tool is always secondary to the job, where focus on customer value drives everything including selection of the right technology and tools. Why use a costly, unreliable, opaque, and compute-hungry LLM when you can achieve the same results with an interpretable, tiny, highly efficient supervised ML model?

It sounds obvious but determining the right thing to do (“the right job”) is the most important starting point. From there, we very carefully select the best technology to implement the solution. It should be no surprise that different ML technologies have different strengths, and so a methodology for selecting AI and ML approaches is central to how we implement value for Zelta users. No buzzwords and hype; just clear, measurable value statements backed by well thought-out pragmatic solutions.

That's why, even at this incredibly exciting time for generative AI, our approach leans toward supervised machine learning. LLMs are a form of broad AI. Foundational and non-task specific, they are adaptable to different tasks but currently still are very much a work in progress when it comes to being optimal for real-world applications. Supervised ML, on the other hand, spans a long and proven history from simple regression models to deep neural networks, and the smaller language model predecessors to today's large language models.

Today, supervised machine learning is the cornerstone of Zelta's AI strategy, with generative methods and modern LLMs providing ever increasing value in our applications.

The screenshot shows a web interface for a clinical trial dictionary. At the top, it displays metadata: Section: Revision Name: Version 1.0, Study Event Group Name: AddtlForms, Study Event Name: AE, Form Name: AE, Section Name: AEEVNT. Below this is a search bar with the text 'Meningitis due to Escherichium' and a 'Search' button. To the right of the search bar is a toggle for 'Search with Artificial Intelligence'. Below the search bar, it shows '3 suggestions' and '0 other results'. A table below lists the search results with columns for '1', '2', and '3'. Each result row includes an 'Apply Term' button and various codes and terms.

	1	2	3
LLT Code	10027199	10003458	10004049
LLT Name	Meningitis	Aseptic meningitis	Bacterial meningitis
PT Code	10027199	10027201	10027202
PT Term	Meningitis	Meningitis aseptic	Meningitis bacterial
HLT Code	10007935	10007935	10004047
HLT Term	Central nervous system and spinal infections	Central nervous system and spinal infections	Bacterial infections NEC
HLGT Code	10021879	10021879	10004018
HLGT Term	Infections - pathogen unspecified	Infections - pathogen unspecified	Bacterial infectious disor
SOC Code	10021881	10021881	10021881
SOC Term	Infections and infestations	Infections and infestations	Infections and infestation

The benefits of supervised machine learning

Control over the training data: With large language models like ChatGPT, you don't know what data the AI has been trained on, or what limits may be placed on it. Supervised ML approaches take only the training data that is supplied by the user, targeted at the objective at hand, and vetted for correctness, compliance, and usage rights. That said, one caveat of this approach is that curating this data can be very expensive. It is worth noting, though, that LLMs can help in augmenting sparse data sets when done carefully with human quality control.

Reliability: There's no risk of random unrelated MedDRA codes being presented back to you in an authoritative-sounding response. A supervised machine learning model allows you to control the input data and optimize results for your specific use case requirements. And although all machine learning is prone to errors and can never be guaranteed 100% correct, errors with supervised ML models can be readily understood and addressed. Your training data determines the results of the model at runtime, and in this case, you know exactly what data was used for training. No third party will have included training data that exposes you to legal action or renders interpretation of your results unreliable.

Consistency: With supervised ML, results are reproducible and addressable (i.e., we have a much greater ability to fix issues with model predictions). This reliability and consistency is essential for clinical trials where consistency and traceability are paramount.

Excels at specific tasks: Where some users may see LLMs as one-size-fits-all AI solutions, often different types of models are more appropriate for accomplishing a specific task. Both the data and the model can be tailored and deployed for specific applications. This makes supervised ML ideal for meeting the needs of unique situations with specific requirements, like clinical trial studies.

Faster: Supervised ML models are very fast and can run on non-specialized hardware.

Cheaper: Due to their size and speed, supervised ML is orders of magnitude cheaper to use at runtime, and, due to their much simpler development trajectory, are cheaper to develop and maintain.

Built into the workflow: Smaller, more focused ML models can be integrated directly into the workflow at key decision points for users. This makes them better able to augment workflows and add measurable value. This also allows for a strong, perpetual improvement feedback loop, reinforced by human expert knowledge input through the user interaction data exhaust. Performance data from specific tasks being executed is fed back as further training data to better improve the models.

Bring your own tool (BYOT): In addition to offering in-app ML and automation meant to address common data management challenges, it is also important to offer dynamic integration options. For example, vendors in the advanced analytics and risk-based monitoring space offer exciting engagement-based AI capabilities that show great promise for assisting with operational and clinical decision making. Unlike traditional APIs that have strictly

defined parameters and predictable behavior, LLMs rely on human language text as input. This introduces a level of fuzziness and unpredictability, making integration cumbersome and non-deterministic. The more focused approach of supervised machine learning means that the API is more stable and deterministic, and reduces solution complexity and maintainability.

That's the Zelta approach to AI: innovating clinical data management through focused, efficient, and uniquely scalable supervised ML deployments that can transform long-standing data challenges into new strategic advantages.

The screenshot displays a user interface for editing a question in a clinical trial system. At the top, it shows '1 of 2 AE- Adverse Events | 2 References' with an information icon and an 'OID' label. The main area contains a text input field with the question 'Did the subject have any AEs?'. Below this is a rich text editor toolbar with options for bold, italic, underline, link, and text color, along with a 'Normal' style dropdown and mathematical symbols. A text input field below the toolbar is labeled 'Enter text for a new question'. Further down, there are sections for 'Instructions (optional)' with a text input field containing 'Please carefully review the EHR narratives for any adverse ev', and 'Question type' with a dropdown menu set to 'Horizontal radio list'. Below these are fields for 'Short name reference' (containing 'AEYN'), 'OID (unique ODM identifier)' (containing 'I.AEYN'), and 'Reporting column' (containing 'AEY|'). An 'AI Suggestions' dropdown is open below the reporting column, showing 'AEYN' and 'CDASH' as options.

The Zelta track record on creating value with AI and automation

Zelta uniquely scales supervised machine learning and automation across the platform to fundamentally change the way medical device, biotech, pharmaceutical companies, and CROs manage their clinical data. These include a point-and-click electronic data capture (EDC) designer, automatic design validation, study design library, system-generated test cases, predictive resupply, med coding with AI, and zero programming system integrations.

AI and advanced automation aren't new features for the Zelta platform – we've adopted and leveraged them to real success for our customers' clinical trial needs for well over a decade.

2012: Automation – system generated artifacts

- Zelta eClinical platform goes live with novel study design automation features out of the box.
- Applications: Point and click EDC designer, automatic design validation, study design library, system generated test cases.
- **Results:** Tens of thousands of study documents created; thousands of design components replicated from study to study; 400+ studies tested with system-generated test cases.

2015: Algorithm – predictive resupply

- Predictive resupply introduced into RTSM offering.
- **Results:** >60 studies, spanning 22 customers, 84% of which are in phase II and III; 13% Phase III of RTSM trials; 57% increase in Phase III go-live within 18 months of introduction.

2015: Automation – data migrator (ETL)

- Zero programming in-system extract, load transformation tool (ETL) for mapping and exchanging source and third-party data.
- **Results:** +400 studies, 100M imported data points, +20 trials rescued, +15 partner integrations.

2019-2021: Supervised machine learning – med coding with AI

- In-workflow AI assistant (for both MedDRA and WHODrug) for medical coder when coding verbatims are not captured by auto-coding rules, i.e. omissions.
- **Results:** 40,000 terms coded, 100 studies, 1,000 hours saved, 39% coded with zero searches. 51% decrease from manual coding to AI-assisted searches.

The screenshot shows a software interface for medical coding. At the top, it displays metadata: Revision Name: Version 1.0, Study Event Group Name: AddtlForms, Study Event Name: Concomitant Medications, Form Name: Concomitant Medications, and Section Name: MedNameEng. Below this is a search bar containing 'Amoxicillin-clavulanate' and a toggle for 'Search with Artificial Intelligence'. The results section shows '2 suggestions' and '0 other results'. A table displays two columns of results, each with an 'Apply Term' button. The first result has a code of 02043401313 and the drug name 'AMOXICILLIN CLAVULANIC ACID'. The second result has a code of 02043401442 and the drug name 'AMOXICILLIN AND CLAVULANATE POTASSIUM'. The table lists various codes and names including Preferred Code, Preferred Name, ATC Code 4, ATC 4, ATC Code 3, ATC 3, and ATC Code 2.

Code	1	2
Drug Name	AMOXICILLIN CLAVULANIC ACID	AMOXICILLIN AND CLAVULANATE POTASSIUM
Preferred Code	02043401001	02043401001
Preferred Name	SPEKTRAMOX	SPEKTRAMOX
ATC Code 4	J01CR	J01CR
ATC 4	COMBINATIONS OF PENICILLINS, INCL. BETA-LACTAMASE INHIBITORS	COMBINATIONS OF PENICILLINS, INCL. BETA-LACTAMASE INHIBITORS
ATC Code 3	J01C	J01C
ATC 3	BETA-LACTAM ANTIBACTERIALS, PENICILLINS	BETA-LACTAM ANTIBACTERIALS, PENICILLINS
ATC Code 2	J01	J01

“Using AI in our medical coding, we can get through the volume more quickly but also more intelligently.”

David Provenghi
 Director of Clinical Data Management
 ProSciento



What's next for Zelta automation and AI

With over a decade invested into AI and process automation, we're just beginning to scratch the surface of what these solutions can do to enhance clinical trial processes and provide greater confidence and control over the outcomes. AI isn't going anywhere, which is why Zelta is firmly committed to expanding its footprint into innovative new features in the near term, including:

- **Triggers for monitoring levels:** Built-in, risk-based controls for tracking and throttling monitoring activities up and down as needed.
- **Audit trail scans for risk and fraud:** In-platform study metadata and administrative data scans to detect and alert for potential data collection access and behavioral concerns.
- **In-workflow AI and supervised ML enhancements that reduce tedious manual work:** For example, ML-driven quality checks, ML-driven design assistants, and system-generated test cases – allowing study managers to spend more time on value-adding, quality-driving work that requires the human touch.
- **Digital protocol to EDC design:** Automated EDC generation based on digital protocols and biomedical concepts.

Conclusion: Letting use cases, not hype, define our way on AI

At the end of the day, AI, machine learning, and advanced automation promise to fundamentally change the way people manage clinical data. We believe that and it's why at Zelta we have been incorporating AI and machine learning into our platform for over a decade.

It's also important to be clear about what AI can and can't do, as well as what it actually means when we say "AI." Generative AI and LLMs hold a lot of promise for the future of clinical trial research. We are working hard to address their current shortcomings and ensure these technologies deliver pragmatic value and return on investment for all Zelta users. We are doing this in concert with our current supervised machine learning capabilities where the use cases require it. The real value in AI for clinical research today is in its application for defined, scalable use cases that significantly aid in productivity across everyday workflow tasks – enhancing daily work through deployments specifically tailored to your needs. You shouldn't have to reinvent your processes to take advantage of AI. With supervised machine learning, you don't need to.

That's why Zelta has taken the path of innovating clinical data management through our uniquely scalable machine learning approach, adding value to your current workflows in a way that finally generates a real return on investment. It's not a brand-new tool to master; it's not a major new disruption to how you've been doing things for years. It's purely a value add, and one that you have the power to turn on and off as you need it.

You don't have to take our word for it. See for yourself how Zelta has used AI to empower medical coders and help them process data more efficiently and accurately.

READ

[How Zelta and AI have made medical coding faster, more precise at ProSciento](#)

WATCH

[Make medical coding easier with Zelta](#)

About Zelta

Zelta by Merative is a clinical trials solution business that includes both a clinical data management and acquisition platform and consulting, enablement, and extension services. Zelta's unified cloud-hosted platform supports all phases and complexities of research, including more than 500 phase III trials.

Learn more at merative.com/clinical-development

About Merative

Merative is a data, analytics and technology partner for the health industry, including providers, health plans, employers, life sciences companies and governments. With trusted technology and human expertise, Merative works with clients to drive real progress. Merative helps clients orient information and insights around the people they serve to improve decision-making and performance. Merative, formerly IBM Watson Health, became a new standalone company as part of Francisco Partners in 2022.

Learn more at merative.com



© Merative US L.P. 2024. All Rights Reserved.

Produced in the United States of America
September 2024

Merative and the Merative logo are trademarks of Merative US L.P. Other product and service names might be trademarks of Merative or other companies.

The information contained in this publication is provided for informational purposes only. While efforts were made to verify the completeness and accuracy of the information contained in this publication, it is provided AS IS without warranty of any kind, express or implied. In addition, this information is based on Merative's product plans and strategy as of the date of this publication, which are subject to change by Merative without notice. Nothing contained in this publication is intended to, nor shall have the effect of, creating any warranties or representations from Merative, or stating or implying that any activities undertaken by you will result in any specific performance results. Merative products are warranted according to the terms and conditions of the agreements under which they are provided.

MCD-7369671779 Rev 1.0