GPT-Powered Data Insights: Advancing Clinical Data Analytics with Large Language Models (LLMs)

PHUSE US Connect 2025 - ML25

Mel Hullings, Director, Data Management & Analytics

March 19, 2025

Formation Bio

Agenda

- Intro to AI in Analytics
- GPT-Powered Workflows
    1. Data Cleaning & Transformation
    2. Data Modeling & Analysis
    3. QC & Validation of AI-Generated Outputs
- Key Takeaways & Future Directions

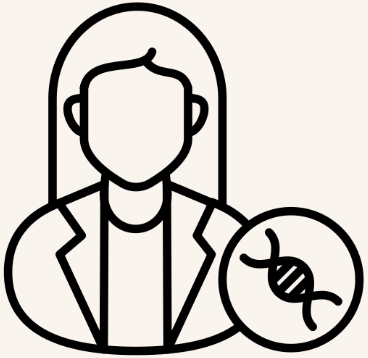
AI-Powered Clinical Data Geneticists

Why LLMs in Clinical Data Analytics?

Much like geneticists decode DNA to understand biological functions, LLMs can decode clinical data to extract meaningful insights while overcoming challenges such as:

- **Data Growth & Complexity:** exponential growth of data (structured & unstructured)
- **Data Quality:** cleaning and standardizing data for analysis is slow and prone to error
- **Regulatory Requirements:** FDA & EMA require structured, traceable, and validated data reporting

💡 *Manual processes are no longer scalable.*



FDA Good Machine Learning Practices for Medical Device Development. U.S. Food and Drug Administration (FDA). October 2021.

Transforming Clinical Data Analytics with LLMs

✅ **Natural Language Understanding (NLU):** AI can put data into context to interpret study protocols, medical narratives, and investigator comments.

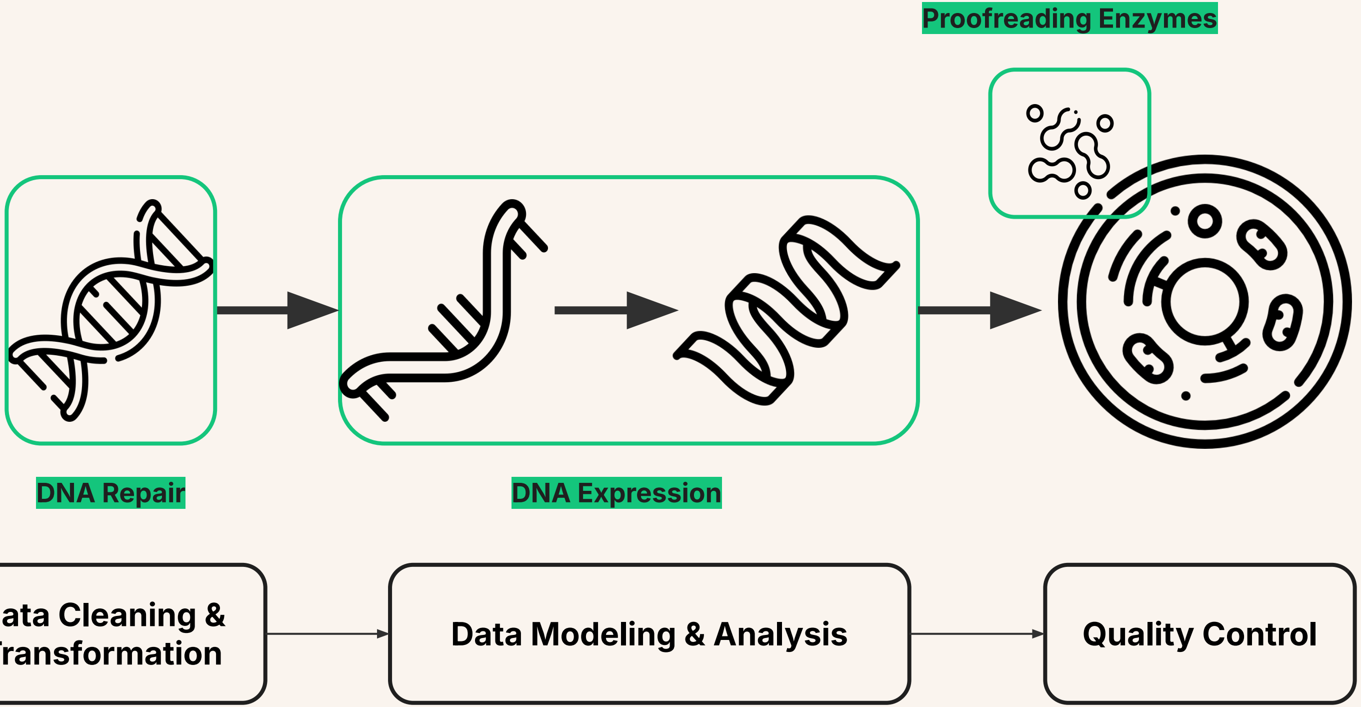
✅ **Pattern Recognition in Structured Data:** GPT models can flag inconsistencies in case report forms (CRFs) and clinical datasets.

✅ **Automated Report Generation:** AI can summarize safety trends, efficacy endpoints, and statistical analysis results.

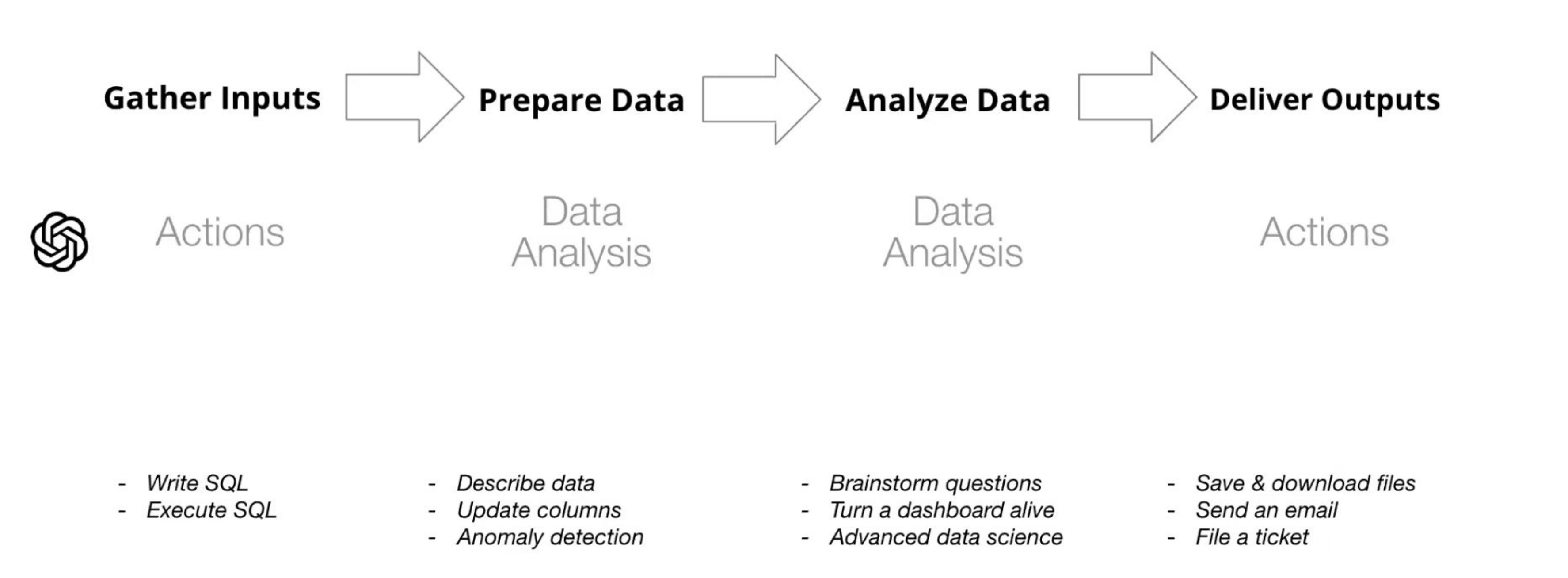
📌 *Example: A recent study demonstrated that GPT-4 outperformed traditional rule-based NLP methods in summarizing unstructured EHR data and extracting clinical phenotypes*

Bhattarai et al. (2024). LLMs for Predictive Analytics & Pharmacovigilance in Clinical Trials.

GPT-Powered Workflows



Data Analysis Process



Gather Inputs → Prepare Data → Analyze Data → Deliver Outputs

Gather Inputs	Prepare Data	Analyze Data	Deliver Outputs
Actions	Data Analysis	Data Analysis	Actions
- *Write SQL* - *Execute SQL*	- *Describe data* - *Update columns* - *Anomaly detection*	- *Brainstorm questions* - *Turn a dashboard alive* - *Advanced data science*	- *Save & download files* - *Send an email* - *File a ticket*

Data analysis is an iterative process where you continually explore questions, data, and results. There are opportunities for AI to support you throughout this process!

OpenAI (2024). OpenAI Presents AI-Empowered Data Analysis.

ChatGPT's "Data Analyst" Custom GPT

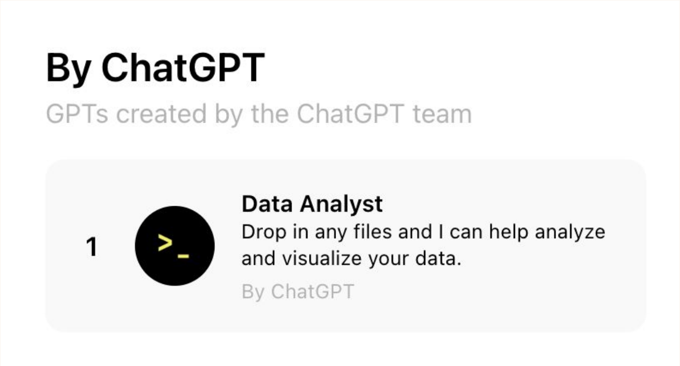
Collaborate with ChatGPT's AI Data Analyst Custom GPT to get from data to insights faster.

Capabilities:

- Chat with your data
- Get proactive insights and suggestions
- Create customizable, presentation-ready charts

Why ChatGPT is recommended for data analysis:

- Automates repetitive analyses
- Advanced data exploration and insights
- Enterprise-grade data privacy and security
- No training the LLM based on your data, ever



By ChatGPT

GPTs created by the ChatGPT team

1 Data Analyst

Drop in any files and I can help analyze and visualize your data.

By ChatGPT

***Note:** Reference your company's data governance guidelines before uploading data.*

Case Study: Using AI to Analyze the Osteoarthritis Initiative (OAI) Public Database

Seminars in Arthritis and Rheumatism 58 (2023) 152148

Contents lists available at ScienceDirect

Seminars in Arthritis and Rheumatism

journal homepage: www.elsevier.com/locate/semarthrit

ELSEVIER
Check for updates

The natural history of end-stage knee osteoarthritis: Data from the osteoarthritis initiative☆

Jeffrey B. Driban[a,*,1], Lori Lyn Price[b,c], Bing Lu[d], Klaus Flechsenhar[e], Grace H. Lo[f,g], Timothy E. McAlindon[a]

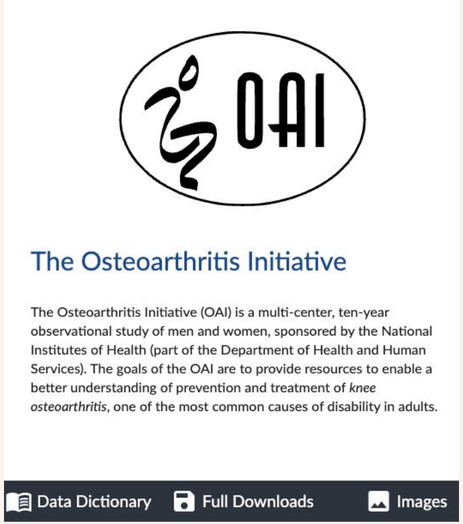
[a] Division of Rheumatology, Allergy, and Immunology, Tufts Medical Center, 800 Washington Street, Box 406, Boston, MA 02111, United States
[b] The Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, MA, United States
[c] Tufts Clinical and Translational Science Institute, Tufts University, Boston, MA, United States
[d] Department of Public Health Sciences, University of Connecticut Health Center, Farmington, CT, United States
[e] Immunology and Inflammation Therapeutic Area, Type 1/17 Immunology Cluster, Sanofi, Frankfurt am Main, Germany
[f] Medical Care Line and Research Care Line, Houston VA HSR&D Center for Innovations in Quality, Effectiveness and Safety, Michael E. DeBakey VAMC, Houston, TX, United States
[g] Section of Immunology, Allergy, and Rheumatology, Baylor College of Medicine, Houston, TX, United States

ARTICLE INFO

Keywords:
Pain
Osteoarthritis
Knee
Patient-reported outcome measures

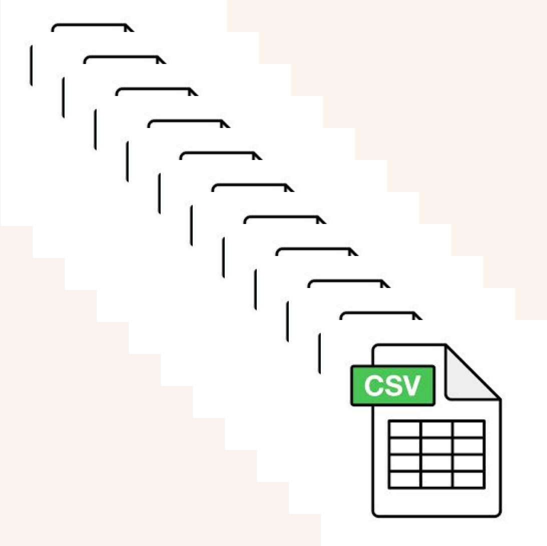
ABSTRACT

Objective: We aimed to describe the natural history leading to end-stage knee osteoarthritis (esKOA), focusing on knee symptoms, radiographic severity, and the presence of limited mobility or instability.
Methods: We performed knee-based analyses of Osteoarthritis Initiative data from 7691 knees (4165 participants). We used a validated definition of esKOA that relied on meeting one of two criteria: (1) severe radiographic knee osteoarthritis (Kellgren-Lawrence [KL] grade=4) with moderate-to-intense pain (Likert WOMAC pain+function>11/88) or (2) KL grade<4 with intense or severe pain (WOMAC pain+function>22) and limited mobility (flexion contracture≥5°) or instability (based on a varus and valgus stress test). We also introduced an alternate definition of esKOA that relied on meeting one of two criteria that omitted physical exam findings:(1)



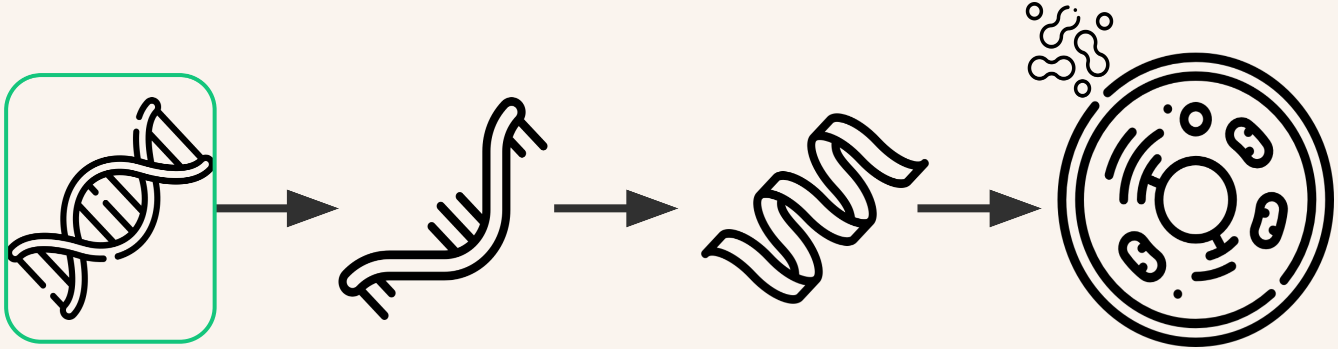
The Osteoarthritis Initiative

The Osteoarthritis Initiative (OAI) is a multi-center, ten-year observational study of men and women, sponsored by the National Institutes of Health (part of the Department of Health and Human Services). The goals of the OAI are to provide resources to enable a better understanding of prevention and treatment of *knee osteoarthritis*, one of the most common causes of disability in adults.



Osteoarthritis Initiative. (2025). Osteoarthritis Initiative (OAI) Database. National Institute of Mental Health Data Archive.

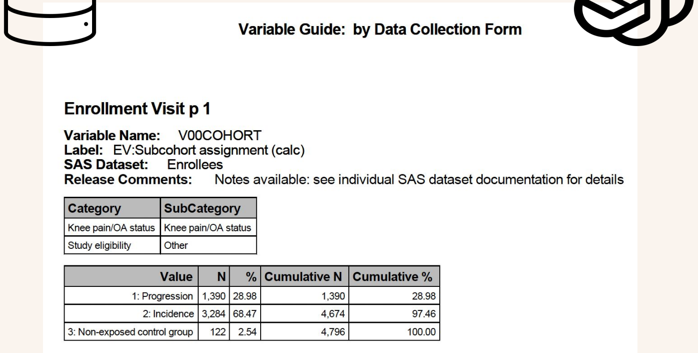
1. Data Cleaning & Transformation



DNA Repair

Data Cleaning & Transformation → Data Modeling & Analysis → Quality Control

Prompts to assess data quality and add columns



Variable Guide: by Data Collection Form

Enrollment Visit p 1

Variable Name: V00COHORT
Label: EV:Subcohort assignment (calc)
SAS Dataset: Enrollees
Release Comments: Notes available: see individual SAS dataset documentation for details

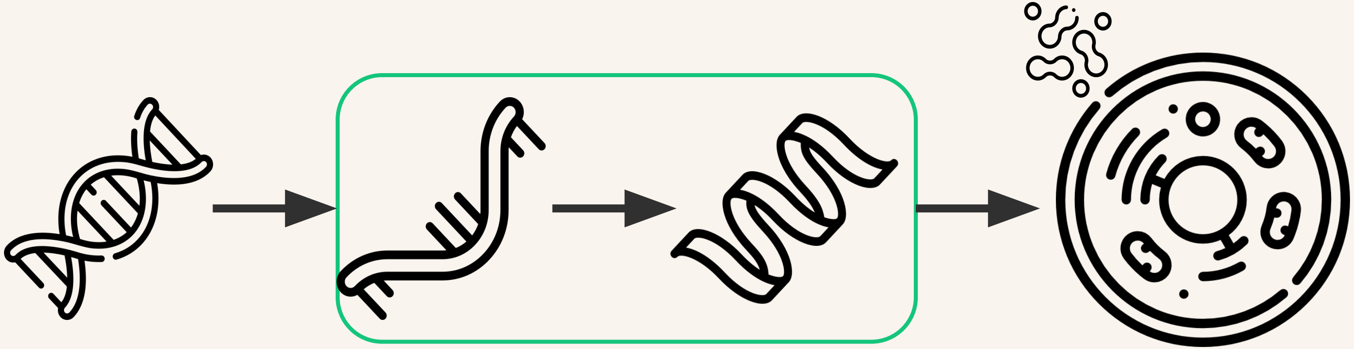
Category	SubCategory
Knee pain/OA status	Knee pain/OA status
Study eligibility	Other

Value	N	%	Cumulative N	Cumulative %
1: Progression	1,390	28.98	1,390	28.98
2: Incidence	3,284	68.47	4,674	97.46
3: Non-exposed control group	122	2.54	4,796	100.00

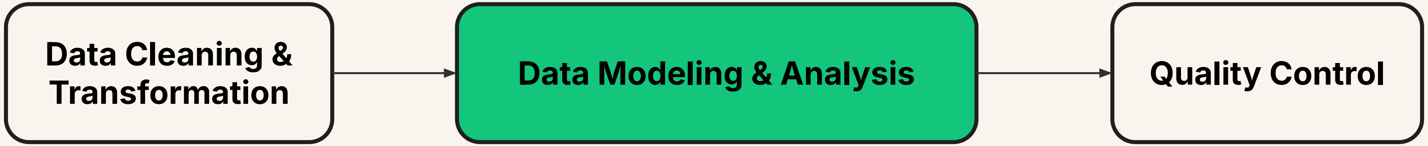
➔ **Data cleaning:** What are some anomaly checks to run in this dataset?

➔ **Data transformation:** Add a column after “V00COHORT” to the enrollees dataset labeled V00COHORT_Verbatim” and in that column for each subject’s ID, enter the text “Progression” if “V00COHORT” is a 1, “Incidence” if “V00COHORT” is a 2, and “Control” if “V00COHORT” is a 3.

2. Data Modeling & Analysis



DNA Expression



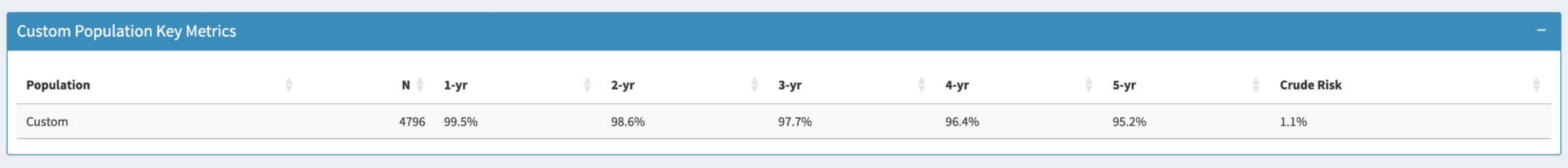
Data Cleaning & Transformation

Data Modeling & Analysis

Quality Control

Prompts to analyze and interact with data

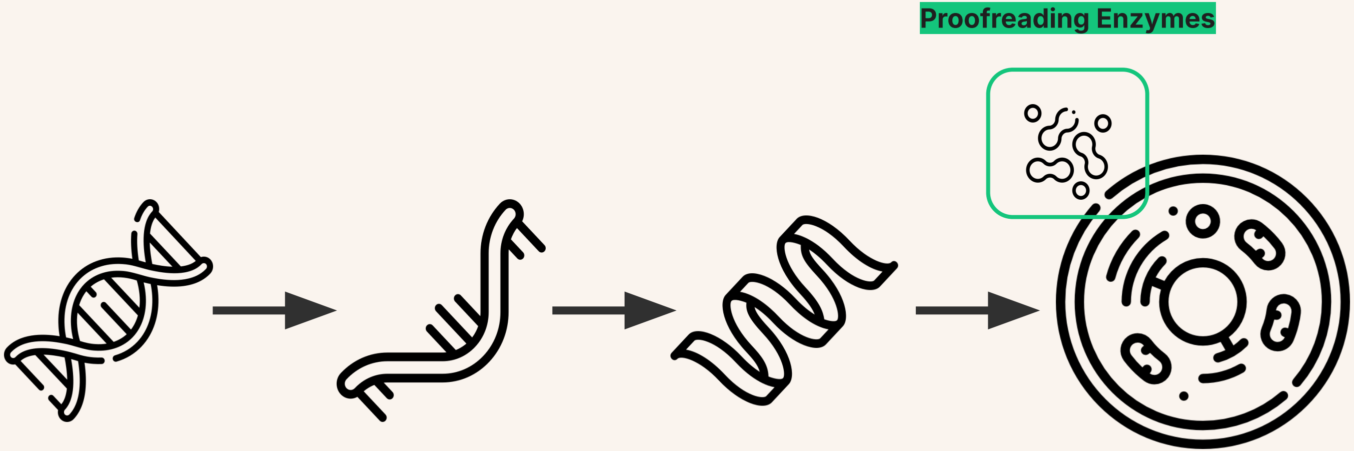
- ➔ **Data summarization:** Will perform upon upload and then specify further, such as asking how many subjects in the “outcomes 99” table had a knee replacement
- ➔ **Data viz:** Create a bar chart to show the breakdown of race and ethnicity for each cohort.
- ➔ **Data interactivity:** Generate R code for a Shiny app that allows you to explore data.



Custom Population Key Metrics

Population	N	1-yr	2-yr	3-yr	4-yr	5-yr	Crude Risk
Custom	4796	99.5%	98.6%	97.7%	96.4%	95.2%	1.1%

3. QC & Validation of AI-Generated Outputs



Data Cleaning & Transformation → Data Modeling & Analysis → Quality Control

Quality Control

Regardless of the classification of data, critical thinking is required to evaluate risk based on the data, information, and decisions being made for each use case, including the ability to quality control (QC) outputs because AI can be inconsistent, factually incorrect, and hallucinate.

Tips to check quality of AI analyses:

- Logical checks
- Ask for sources
- Output verification (i.e. human-in-the-loop)
- Awareness of bias in the data

Key Takeaways

1️⃣**LLMs enhance data processing** by automating cleaning, transformation, and modeling, improving efficiency of data insights.

2️⃣**AI-powered analytics facilitate data quality and exploration**, making complex clinical trial data more interpretable and actionable.

3️⃣**Shiny apps with LLM integration** allow for real-time interaction with clinical datasets.

4️⃣**Validation & QC remain critical**—AI-generated insights must be rigorously verified to ensure accuracy and compliance.

Future Directions

🚀 Advancements in Multi-Modal AI

- Next-gen models (e.g., GPT-5, Med-PaLM 2) will integrate **text, images, and structured data**, enabling deeper insights from multi-source clinical datasets.

📜 Regulatory Landscape Evolution

- As AI adoption grows, **regulatory bodies (FDA, EMA, MHRA) will refine guidance** on AI-driven analytics, emphasizing transparency and validation.

🧠 Ethical & Interpretability Challenges

- While AI augments clinical decision-making, **explainability remains a challenge**—future AI models must improve interpretability and trustworthiness.

Questions?