Correlation Between Breastfeeding Practices & Gluten Sensitivity Risk

Samira Kazimi
DSCI510

PROJECT INTRODUCTION

Research Question: Are early breastfeeding duration and exclusivity correlated with celiac disease risk later in life?

Motivation: Understanding breastfeeding's protective effects on gluten sensitivity can inform evidence-based public health policy

Approach: Multi-dataset integration with correlation analysis and statistical visualization

Impact: Contribute to infant nutrition and health outcomes research

DATA SOURCES & APPROACH

Dataset 1 – Celiac Disease Lab Data 2206 patients, 15 variables (Age, IgA, IgG, IgM, Disease Type)
Dataset 2 – Infant Feeding & Growth 720 infant records, 5 variables (Weight, Height, Month, Feeding Method)
Dataset 3 – Monkeypox Research 71 research projects, 10 variables (Topic, Status, Country, Completion)

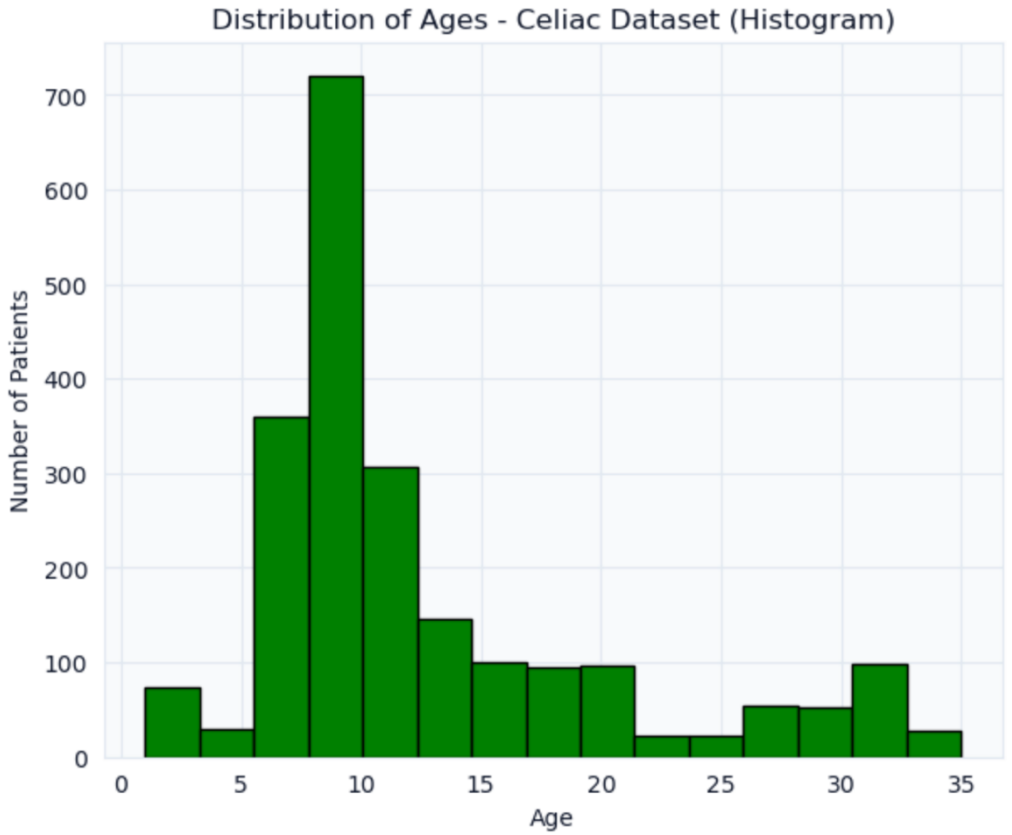
DATA PROCESSING & METHODS

1. **Data Integration:** Merged three datasets from API and Kaggle sources using pandas
2. **Cleaning:** Handled missing values, normalized demographic variables
3. **Analysis Techniques:** Correlation coefficients, heatmaps, and demographic breakdowns
4. **Visualization:** Seaborn for statistical plots and data relationships

CELIAC DISEASE: AGE DISTRIBUTION

Celiac disease diagnosis most common in elementary school age children around 9–11 years old. This means doctors should pay special attention to this age group.
The immune blood tests (IgA, IgG, IgM) do not change much with age, so high levels can signal celiac disease in both younger and older kids.

Peak Age Range: 9-11 years (~700 patients)
Secondary Peak: 6-8 years (~350 patients)
Median Age: ~10 years

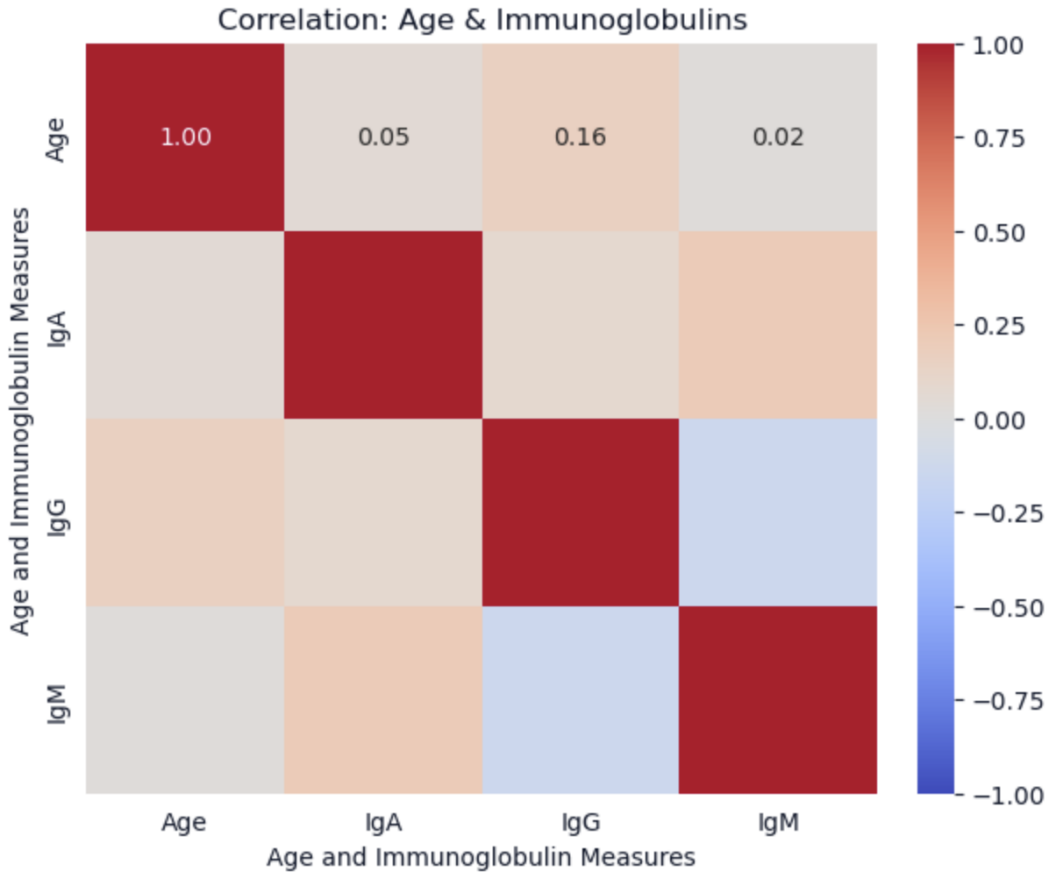


AGE & IMMUNOGLOBULINS CORRELATION

Heatmap shows weak to moderate correlations between age and immune markers. Immunoglobulin patterns are largely independent of age.

Age-IgA: No linear correlation ($r \approx 0.05$)
Age-IgG: Weak positive correlation ($r \approx 0.16$)
Age-IgM: Near-zero correlation ($r \approx 0.02$)

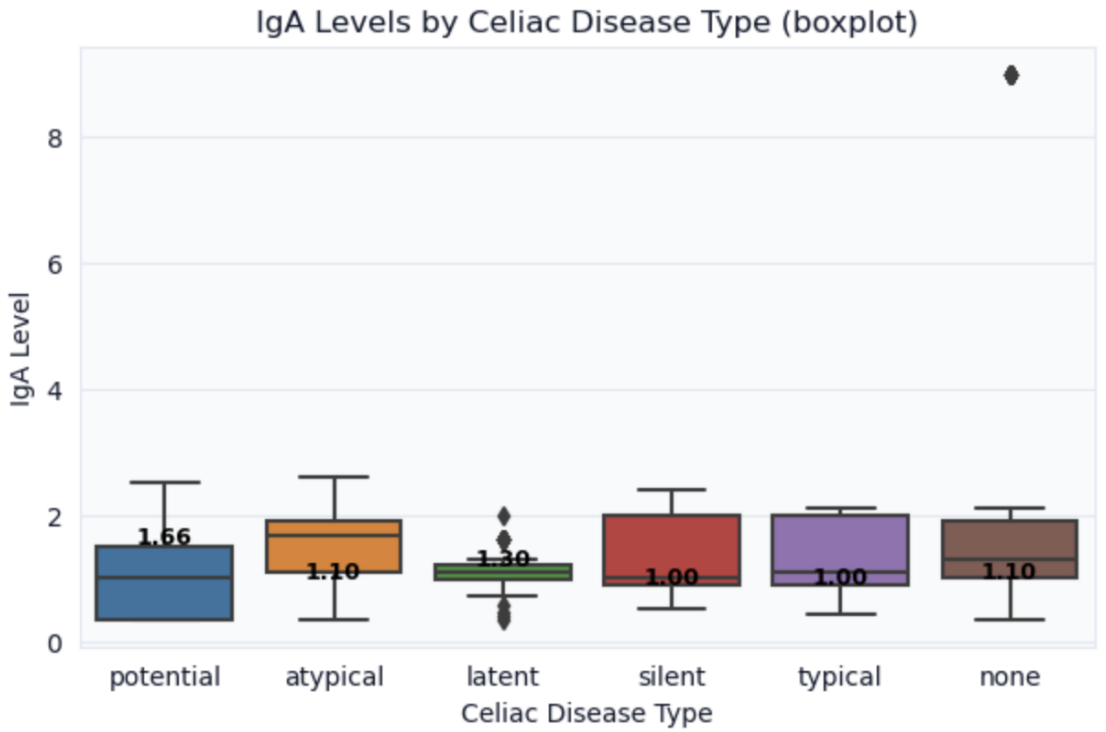
Description:
IgA: Antibody that protects body surfaces like the gut, lungs, saliva, and tears.
IgG: Main antibody in the blood that gives long-term protection against infections.
IgM: First antibody your body makes when a new infection starts.



IGA LEVELS BY CELIAC TYPE

IgA levels vary significantly by celiac disease classification. Because each celiac type shows a different IgA pattern, doctors can use the IgA value as a clue to which kind of celiac disease the patient probably has.

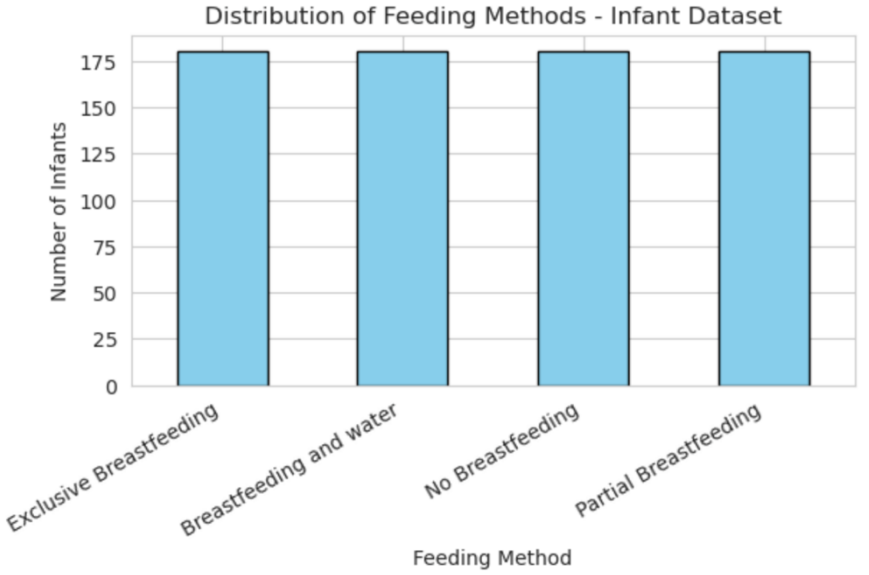
Potential: Median IgA ≈ 1.66 U/mL (highest)
Atypical: Median IgA ≈ 1.10 U/mL (intermediate)
Typical: Median IgA ≈ 1.00 U/mL (lowest)



INFANT FEEDING METHOD DISTRIBUTION

The study has the same number of babies in each feeding group, so it is fair to compare growth between the different feeding methods. All babies gain weight over 6 months, but babies with extra feeding (like breastfeeding plus water or formula) usually gain weight a little faster.
Equal distribution across four distinct feeding methods (n=180 per group).

Exclusive Breastfeeding: ~180 infants
Breastfeeding + Water: ~180 infants
No Breastfeeding (Formula): ~180 infants
Partial Breastfeeding: ~180 infants



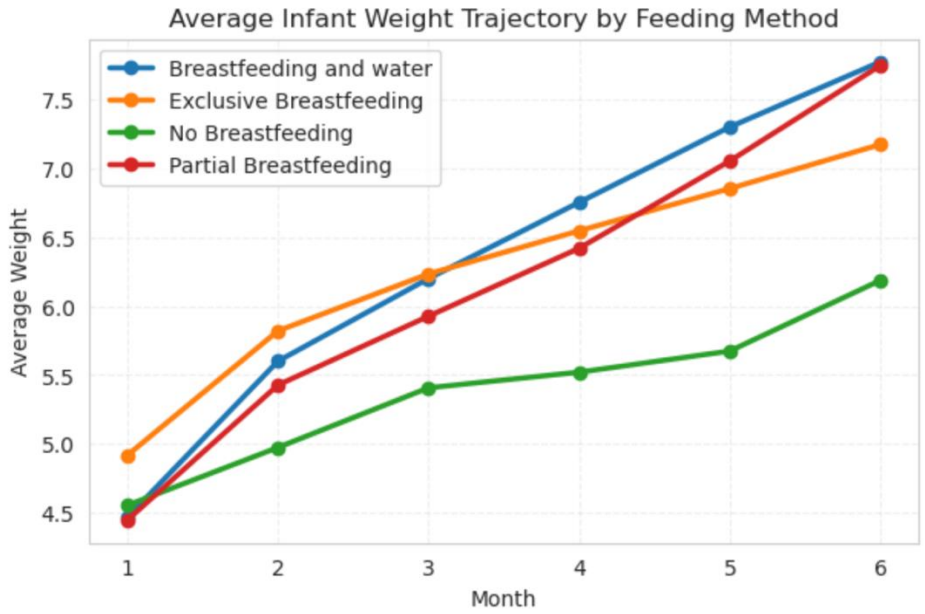
WEIGHT TRAJECTORY BY FEEDING METHOD

Babies who get breast milk plus extra drinks (like water or formula) usually gain weight the fastest. Giving extra milk or water on top of breastfeeding is linked to higher weight gain.

Month 1 (~4.5 kg, all groups aligned)
All babies start at almost the same average weight, about 4.5 kilograms. There is no real difference between feeding groups yet.

Month 3 (5.8–6.2 kg, divergence begins)
By the third month, babies in different feeding groups start to separate. Some groups are a little heavier than others, around 5.8–6.2 kilograms.

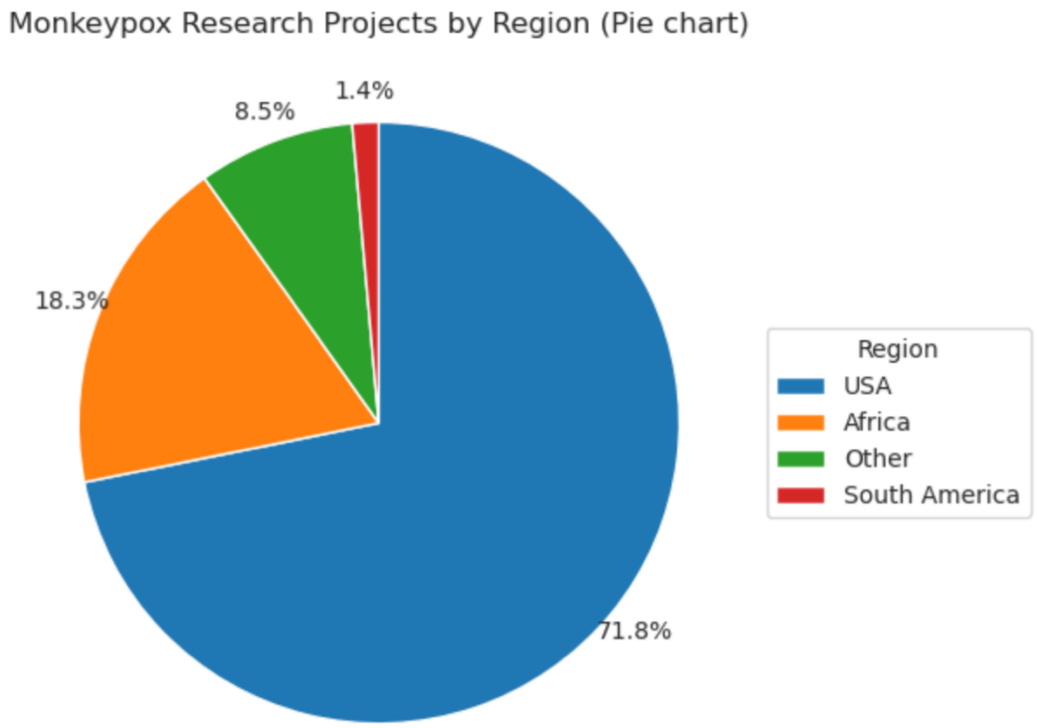
Month 6 (6.1–7.6 kg, feeding-dependent peak)
At six months, the differences are clear. Average weight now depends on how the babies are fed, from about 6.1 to 7.6 kilograms, with some feeding methods leading to more weight gain than others.



MONKEYPOX PROJECTS: GEOGRAPHIC DISTRIBUTION

Most monkeypox research projects are in the USA, with fewer projects in other countries. This means research is mainly done in places that already have strong health systems. This graph shows that most work happens in rich, developed countries that have good laboratories and can produce vaccines, while poorer countries have only a few projects. Many of these studies test how well vaccines protect people and how well different medicines work against monkeypox, so the main focus is vaccine strength and drug effectiveness.

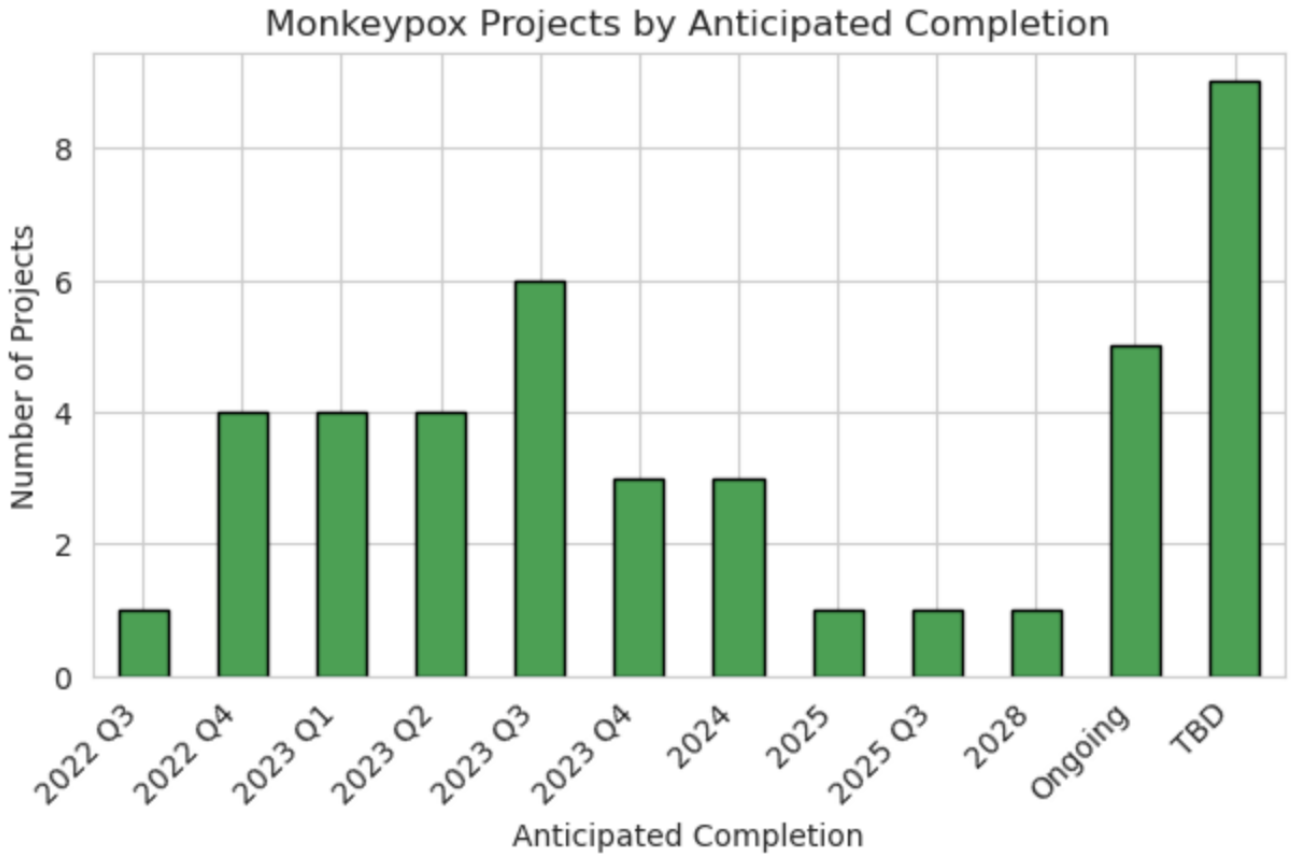
USA (Domestic): ~40 projects (71%)
International: ~6-8 projects per region



MONKEYPOX: COMPLETION TIMELINE

Most projects are planned to finish in 2024 and 2025, so many results and new vaccines are expected in those years. After 2025, some projects will continue, but the largest wave of findings comes in that 2024–2025 period. Many of these studies focus on developing and testing vaccines, so vaccine projects are the main reason the research is moving quickly in the short term.

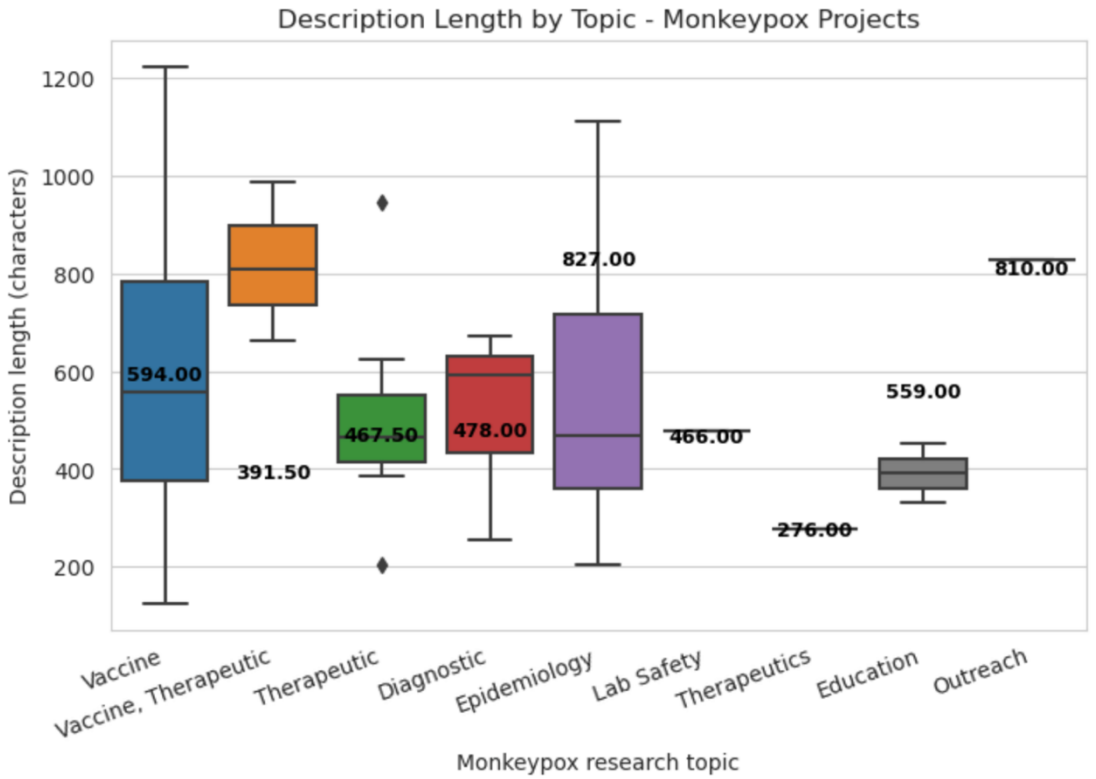
2022–2023: ~15 projects initiated
2024–2025: ~30 projects (peak activity)
2026+: ~15 projects (long-term studies)



RESEARCH SCOPE: DESCRIPTION LENGTH BY TOPIC

Projects about vaccines usually have the longest written descriptions. This suggests these studies are more complex and have many details to explain. Longer descriptions therefore signal that the project is more complicated than others.

Vaccine/Therapeutic: 400-1200 characters (most detailed)
Epidemiology: 300-700 characters (moderate detail)
Outreach/Lab Safety: 250-400 characters (concise)



CONCLUSION

The celiac disease data show that many diagnoses happen between ages 10 and 14, and blood tests that measure immune markers (IgA, IgG, IgM) are very helpful for finding the disease.

The infant feeding data show that all four feeding methods let babies grow, but by month 3 the weight lines start to separate, so early feeding choices already change growth patterns.

The monkeypox data show how research and vaccines are planned and funded over time, with most projects led by the USA and many finishing around 2024–2025.

Together, these datasets suggest that early life feeding like breastfeeding patterns in the first 6 months, may shape immune development and gut health, which later connects to diseases such as celiac that appear in childhood and adolescence.

This project uses real infant growth data plus celiac lab data to support the idea that healthy, well-planned early feeding could help protect long-term digestive and immune health.

CHALLENGES

- Integrating three heterogeneous health datasets (infant growth, celiac labs, and monkeypox research summaries) required extensive cleaning, standardization, and careful joining before any reliable analysis was possible.
- Extracting clear clinical insights was challenging because key variables (feeding patterns, immunoglobulin levels, research regions and timelines) were encoded inconsistently, forcing custom mappings, derived features, and multiple validation steps.

THANK YOU