ADVANCED DATA ANALYTICS IN RARE DISEASE CLINICAL TRIALS USING R AND PYTHON (WITH AI/ML)

Biography:

Anbu Damodaran, with 25 years of experience in biometrics and clinical data operations, is a seasoned drug development professional. His expertise spans the entire drug development process, from first-in-human studies to post-market commitments, including regulatory submissions and compliance with ICH, GCP, CFR, and FDA regulations. He is currently the Associate Director of Statistical Programming at Alexion, AstraZeneca Rare Disease.

AI/ML-DRIVEN DATA ANALYTICS IN RARE DISEASE CLINICAL TRIALS USING R AND PYTHON

ABSTRACT: Rare disease clinical trials face formidable challenges, including limited patient populations, data scarcity, regulatory hurdles, and difficulty in defining meaningful endpoints. These “catch-22” scenarios often slow progress and hinder drug development. This presentation explores how artificial intelligence (AI) and machine learning (ML), implemented using open-source tools in R and Python, can help overcome these barriers across the clinical trial lifecycle. We highlight practical applications, theoretical foundations, and future directions, with a focus on open-source clinical NLP and cancer AI models that enhance data utility and trial efficiency.

DISCLAIMER

This presentation is intended solely for informational purposes and should not be interpreted as professional advice or guidance. The views and opinions expressed are those of the presenter and do not necessarily represent the views or policies of AstraZeneca.

AI/ML-DRIVEN DATA ANALYTICS IN RARE DISEASE CLINICAL TRIALS USING R AND PYTHON

Agenda:

- Rare Disease Pain Points & Catch-22 scenarios
- AI/ML Solutions
  - R/Python implementation
  - Theory, Praxis and Future
- AI/ML in Rare Disease Trials: Not a Panacea
- Open Source NLP models
  - Introduction
  - Rare disease applications
- Open Source Cancer AI models
  - Introduction
  - Rare cancer applications
- Use Cases of AI in Clinical/Statistical Programming
- Marketed Clinical Statistical Programming Solutions with AI APIs
- Conclusion/Questions

PAIN POINTS & CATCH-22 SCENARIOS IN RARE DISEASE CLINICAL TRIALS

Operational

- ❑ Patient Recruitment & Retention
- ❑ Regulatory & Compliance
- ❑ Site Management & Operations
- ❑ Funding & Budgeting
- ❑ Stakeholder Engagement

Analytical

- ❑ Study Design & Protocol Development
- ❑ Data Collection & Management
- ❑ Statistical & Analytical Challenges
- ❑ Technology & Innovation
- ❑ Real-World Data (RWD) & Evidence Generation

AI/ML SOLUTION FRAMEWORK

- Challenge: Define the rare disease trial barrier.
- AI/ML Approach: Describe the proposed solution.
- Tools: Key R/Python libraries.
- Theory: Core algorithms and assumptions.
- Praxis: Real-world adoption and limitations.
- Future: Experimental ideas and integration potential.

SMALL SAMPLE SIZE VS. STATISTICAL POWER

Challenge:

- Few eligible patients make statistical significance difficult to achieve.

AI/ML Solutions:

- Patient finder algorithms (EHR mining)
- External control arms with ML

Open-Source Tools:

- pandas, scikit-learn (Python)
- MatchIt (R), causalml, statsmodels

Theory, Praxis, and Future

- AI/ML methods effectively boost statistical power and group matching in rare disease trials.
- Digital twin modeling to simulate control arms or augment trial populations
- Growing use in select trials, with cautious regulatory and industry uptake.
- Emerging approaches include deep learning, multi-modal data, and federated learning.

ELIGIBILITY CRITERIA VS. FEASIBILITY

Challenge:

- Strict criteria exclude patients; looser criteria threaten rigor.

AI/ML Solutions:

- Eligibility optimization via ML
- Adaptive inclusion modeling

Open-Source Tools:

- PyCaret (Python), mlr (R)
- SimPy, custom simulation scripts

Theory, Praxis, and Future

- ML can optimize eligibility criteria, balancing scientific rigor with patient inclusion in trials.
- Adoption is limited but growing, mainly in pilot studies and feasibility analyses.
- Future advances could include adaptive, simulation-driven eligibility and real-time updates during recruitment.

CLINICAL ENDPOINTS VS. DISEASE UNDERSTANDING

Challenge:

- Scarcity of established endpoints for rare diseases.

AI/ML Solutions:

- Biomarker and digital endpoint discovery
- Digital phenotyping via wearables

Open-Source Tools:

- scikit-learn, keras, PyTorch
- tsfresh, pyts, lifelines (time series/survival)

Theory, Praxis, and Future

- AI/ML enables discovery of novel endpoints where few exist, addressing key gaps in rare disease trials.
- Adoption is early-stage; limited examples use wearables and digital phenotyping.
- Digital twin simulations for endpoint validation and personalized disease modeling
- Explore multi-modal data, federated learning, and adaptive digital endpoints.

NATURAL HISTORY STUDIES VS. URGENCY FOR TREATMENT

Challenge:

- Delays in natural history studies slow access to therapy.

AI/ML Solutions:

- Accelerated natural history modeling from RWD
- AI prediction of treatment effects
- Digital twin models to simulate untreated disease progression.

Open-Source Tools:

- lifelines, scikit-learn, Shiny (R dashboarding)

Theory, Praxis, and Future

- AI/ML can robustly accelerate natural history modeling and predict treatment effects by leveraging real-world data.
- Adoption is limited; a few studies use RWD and ML models, but regulatory and data quality concerns persist.
- Potential includes real-time learning dashboards, synthetic control arms, and AI-driven individualized outcome predictions.

STANDARD OF CARE VS. INNOVATION

Challenge:

- No established SOC blocks innovation and trial design.

AI/ML Solutions:

- Best-practice mining from global evidence
- AI clinical decision support

Open-Source Tools:

- networkx, spaCy, textacy

Theory, Praxis, and Future

- AI/ML can synthesize global evidence and optimize clinical decisions where no SOC exists.
- Adoption is early; some text mining of clinical literature, but limited integration into trial design.
- Digital twin-based scenario testing for protocol innovation.
- Explore dynamic best-practice recommendations and AI-driven protocol design adapting to emerging data.

AI/ML IN RARE DISEASE TRIALS: NOT A PANACEA

AI/ML can help address some rare disease research challenges (e.g., patient identification, disease modeling, data automation).

In some cases, AI/ML can transform "impossible" bottlenecks into more manageable, though still challenging, problems.

Core limitations—small datasets, clinical variability, and regulatory/ethical complexity—often remain.

AI/ML may mitigate certain barriers, but cannot fully overcome the toughest obstacles in rare disease trials.

OPEN SOURCE CLINICAL NLP MODELS

NLP Model

- A machine learning model trained to understand human language.
- In the clinical context, it can extract information from:
  - Electronic health records (EHRs)
  - Clinical notes and Discharge summaries
  - Radiology and Pathology reports

Clinical Focus

- These models are trained on **medical terminology**, **clinical abbreviations**, and **healthcare-specific language**.
- Tasks they perform include:
  - Named Entity Recognition (e.g., identifying diseases, medications, procedures)
  - Relation extraction (e.g., linking a medication to a diagnosis)
  - De-identification (removing personal information)
  - Text classification (e.g., categorizing patient notes)

OPEN SOURCE NLP TOOLS: RARE DISEASE APPLICATIONS (1/2)

Tool	Pain Point Addressed	How It Helps
LLM-Anonymizer	Privacy & compliance (HIPAA)	De-identifies patient data using local LLMs for secure sharing.
medSpaCy	Lack of domain-specific NLP	Customizable NLP for clinical notes (e.g., section detection, entity linking).
Apache cTAKES	Extracting structured data from clinical text	Identifies diseases, symptoms, meds, and procedures from narratives.
scispaCy	Need for lightweight biomedical NLP	Fast, spaCy-compatible models for biomedical text in resource-limited settings.
Stanza (Bio)	Multilingual & syntactic analysis	Biomedical models for parsing, NER, and cross-lingual documentation.

OPEN SOURCE NLP TOOLS: RARE DISEASE APPLICATIONS (2/2)

Tool	Pain Point Addressed	How It Helps
Transformer-DeID	High-accuracy de-identification	Deep learning-based PHI (Protected Health Information) redaction for safe data use.
Philter	Real-time de-identification	Filters PHI in live/streaming text for EHR or trial platforms.
BioSyn	Terminology inconsistency	Links synonyms to UMLS for better normalization and interoperability.
BioClinicalBERT	Understanding clinical language	Pretrained on clinical notes for cohort selection, AE detection, etc.
Ascle (MedGen)	Small datasets in rare diseases	Generates synthetic medical text for model training and augmentation.

OPEN CANCER AI MODELS

An **open source cancer AI model** is an artificial intelligence model that is:

Trained on cancer-related data (e.g., pathology reports, imaging, genomics, clinical notes),

Designed to assist in cancer detection, diagnosis, prognosis, or treatment planning, and

Typically developed using a robust stack of **open source programming languages, libraries, and tools.**

Released under an open source license, meaning it can be freely used, modified, and shared.

OPEN CANCER AI MODELS – IMAGING & PATHOLOGY

Radiology & Imaging Models

- **DeepLesion (NIH)**: Lesion tracking and progression monitoring.
- **MONAI**: Standardized preprocessing pipelines across institutions.
- **RadImageNet**: Transfer learning for rare tumor imaging.

Pathology & Digital Pathology Models

- **CLAM**: Weakly supervised learning for whole slide images (WSIs).
- **PathML**: Custom pipelines for rare histologies.

Applications

- Data harmonization, biomarker discovery, and rare cancer diagnosis.
- Example: Federated learning using MONAI and RadImageNet for rare sarcoma detection.

OPEN CANCER AI MODELS – GENOMICS & CLINICAL NLP

Genomics & Single-Cell Analysis Models:

- scVI (Single-cell Variational Inference)
- scANVI (semi-supervised extension of scVI)

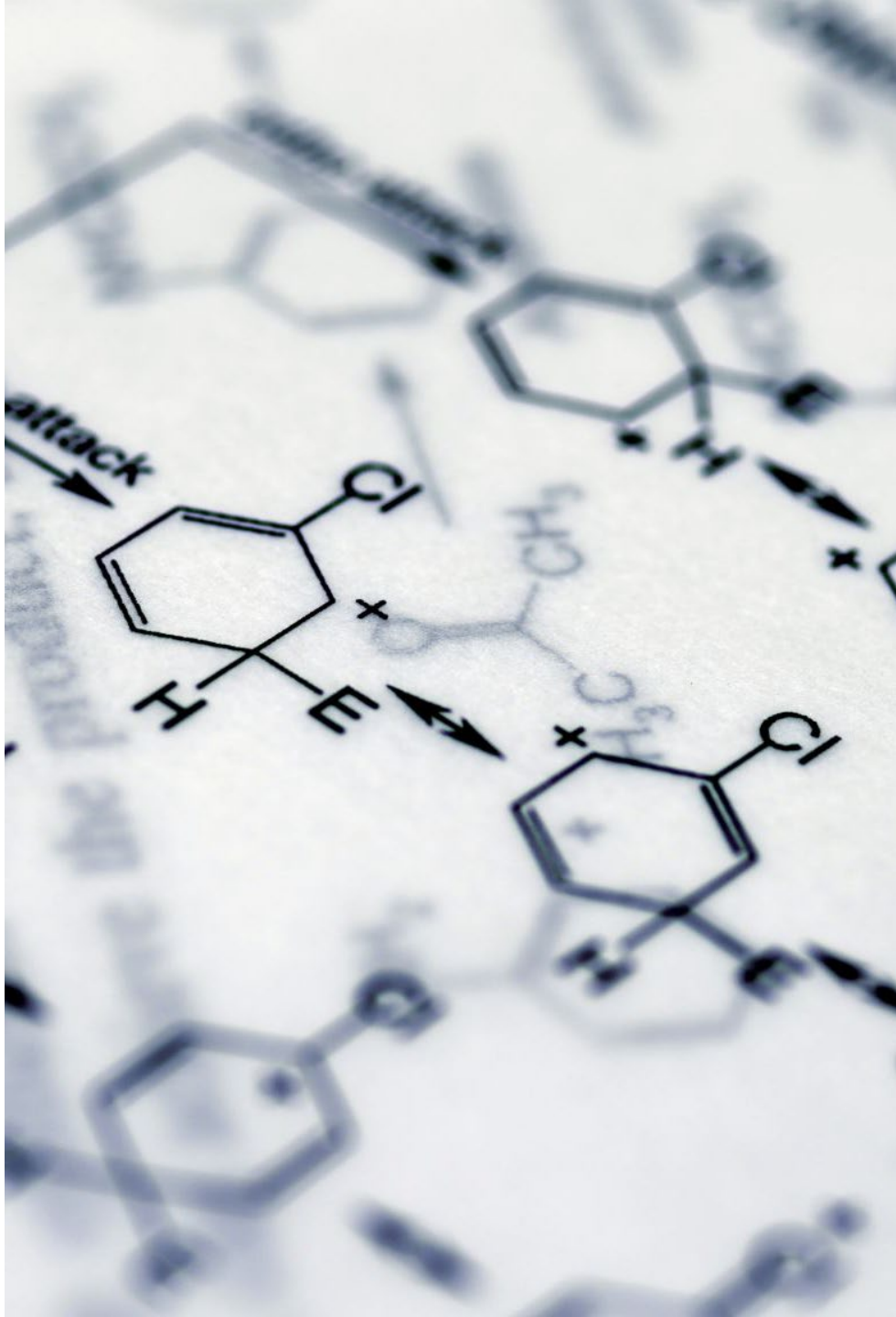
Clinical NLP Models

- **ClinicalBERT**, **BioBERT**, **GatorTron**: Extract phenotypes, mine literature, and stratify patients.

Applications

- Identify immune evasion signatures.
- Automate cohort selection and adverse event detection from clinical notes.

Example: scVI retrained on rare glioma scRNA-seq data reveals resistance pathways



OPEN CANCER AI MODELS – FRAMEWORKS & MULTIMODAL MODELS

Multimodal & Vision-Language Models

BioGPT: Biomedical text generation and understanding.

- **LLaVA-Med**: Combines imaging and text for robust predictions.

Frameworks & Toolkits

- **MONAI** and **PathML**: Modular, reproducible AI pipelines for oncology.

Applications

- Sparse data integration, semantic search, and collaborative model development.
- Enable cross-institutional sharing and reproducibility in rare cancer research.

OPEN CANCER AI MODELS

Discussion

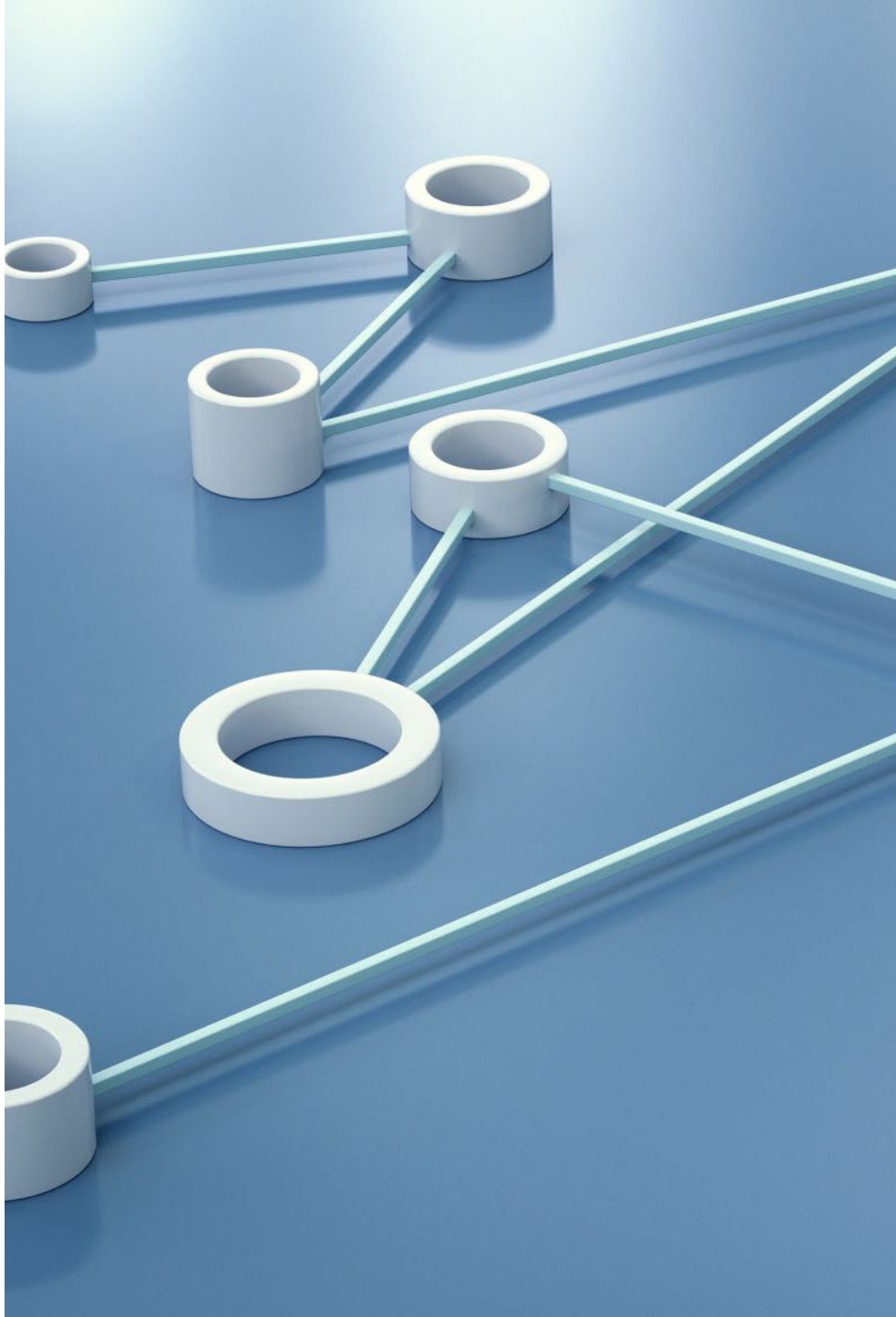
Open-source AI models offer a scalable and collaborative approach to addressing the challenges of rare oncology. Their modularity and adaptability make them ideal for low-data environments. However, clinical validation remains essential, and overfitting risks persist due to small sample sizes.

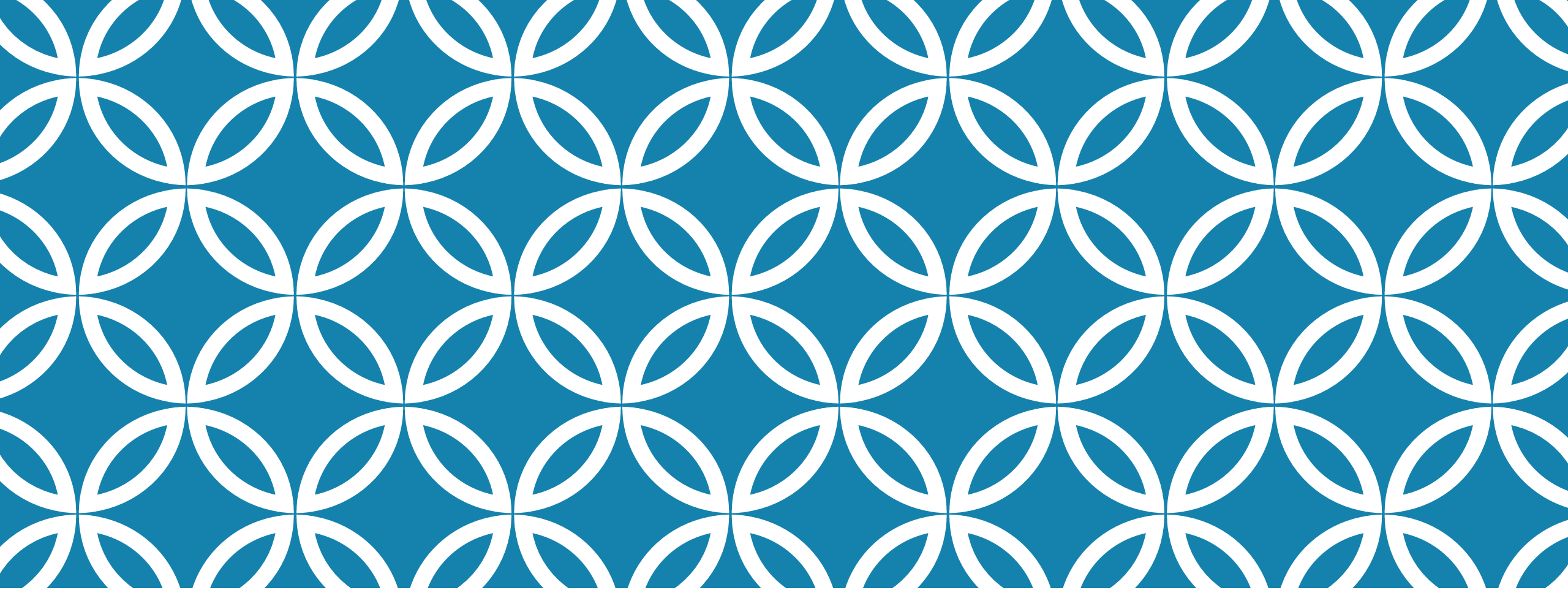
Limitations

- Many models require fine-tuning and validation in rare cancer contexts.
- Some referenced models lack public availability or documentation.
- Data heterogeneity in rare cancers may limit generalizability.

Future Directions

- Release de-identified rare cancer datasets for benchmarking (subject to data privacy or regulatory considerations).
- Expand multimodal learning approaches.
- Involve patient advocacy groups in model development and evaluation.





OPERATIONALIZING AI: USE CASES AND MARKET SOLUTIONS IN CLINICAL AND STATISTICAL PROGRAMMING

AI USE CASES IN CLINICAL STATISTICAL PROGRAMMING

Use Case and Company Example	Basic Functionality
Automated Data Cleaning & Validation – Novartis	AI flags data anomalies for review.
SDTM & ADaM Automation – IQVIA	AI maps raw data to standard formats.
Automated Statistical Analysis – Roche, SAS	AI generates and reviews outputs, reducing errors.
NLP for Clinical Narratives – AstraZeneca	NLP extracts insights and automates reporting.
Real-Time Data & Adaptive Trials – Pfizer	AI monitors trial data, enabling real-time adjustments.
Code Review – Janssen (J&J)	AI assists code quality and consistency

MARKETED AI SOLUTIONS FOR CLINICAL STATISTICAL PROGRAMMING

Solution/Platform	Provider/ Company	Main Use Case/ Functionality
IQVIA Smart Data Standardization & Mapping	IQVIA	Automated SDTM/ADaM mapping, data cleaning, and transformation using AI/ML models.
SAS Intelligent Decisioning	SAS Institute	ML-driven automation for programming, validation, and generation of TFLs.
Medidata AcORN	Medidata	Automated generation of TFLs using NLP and AI from clinical datasets.
Gen AI for Clinical Docs	Oracle Health Sciences	NLP-based drafting of narratives, medical summaries, and automated authoring from structured data.
Smart Clinical Data Review	Saama Technologies	Automated identification of data anomalies and QC checks using AI-driven analytics.

MARKETED AI SOLUTIONS FOR CLINICAL STATISTICAL PROGRAMMING

Solution/Platform	Provider/ Company	Main Use Case/ Functionality
DataRobot MLOps for Pharma	DataRobot	AI/ML workflow automation and monitoring for statistical and operational analytics models.
GxP-ready Clinical NLP API	Linguamatics (IQVIA)	NLP-based information extraction, AE signal detection, and medical curation from clinical text.
Metrics Workbench	Cytel	Automated TFL generation, statistical report templates, and QC/QM analytics enabled by AI components.
TriNetX Advanced Analytics	TriNetX	AI/ML-based cohort analytics, real-world data transformation, and automated summaries
Veeva Vault CDMS AI	Veeva Systems	AI-driven data discrepancy detection and medical review automation

CONCLUSION

AI, ML, and open-source tools can revolutionize rare disease trials by solving data and operational issues.

Advanced analytics enable smarter study design and patient selection, even with limited data.

Open-source NLP and AI models accelerate insights from complex data.

Digital twins improve outcome simulation and trial design, reducing control group size.

Adoption requires strong validation, regulatory support, and attention to privacy.

Progress relies on community-driven development and patient involvement for ethical deployment.

QUESTIONS?

Name: Anbu Damodaran

Affiliation: Alexion, AstraZeneca Rare Disease

E-mail: anbu.damodaran@alexion.com

REFERENCES

U.S. Food and Drug Administration. (2023, December). *Rare diseases: Considerations for the development of drugs and biological products*. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/rare-diseases-considerations-development-drugs-and-biological-products

U.S. Food and Drug Administration. (2023). *FDA requests input on rare disease clinical trial networks*. https://www.fda.gov/ news-events/fda-brief/fda-brief-fda-requests-input-rare-disease-clinical-trial-networks

Sun, X. (K.), & Wang, R. (2019). Machine learning approaches to identify rare diseases (Paper ST-325). In *Proceedings of PharmaSUG 2019*. Ultragenyx Pharmaceutical Inc.

Sun, X. (K.), & Wang, R. (2019). Bayesian methods for treatment design in rare diseases (Paper PO-324). In *Proceedings of PharmaSUG 2019*. Ultragenyx Pharmaceutical Inc.

Roman-Naranjo P, Parra-Perez AM, Lopez-Escamez JA. A systematic review on machine learning approaches in the diagnosis and prognosis of rare genetic diseases. J Biomed Inform. 2023 Jul;143:104429. doi: 10.1016/j.jbi.2023.104429. Epub 2023 Jun 22. PMID: 37352901.

Rare diseases: Still on the fringes of universal health coverage in Europe. (2023, December 11). *Lancet Regional Health – Europe, 37*, 100783.

RARE-X. (2022, June). *The power of being counted* [Report]. https://rare-x.org/case-studies/the-power-of-being-counted/

Hageman, I.C., van Rooij, I.A., de Blaauw, I. et al. A systematic overview of rare disease patient registries: challenges in design, quality management, and maintenance. Orphanet Journal of Rare Diseases 18, 106 (2023).

Shen X, Li X. Deep-learning methods for unveiling large-scale single-cell transcriptomes. Cancer Biol Med. 2023 Dec;20(12):972-980. doi: 10.20892/j.issn.2095-3941.2023.0436

Tan RSYC, Lin Q, Low GH, Lin R, Goh TC, Chang CCE, Lee FF, Chan WY, Tan WC, Tey HJ, Leong FL, Tan HQ, Nei WL, Chay WY, Tai DWM, Lai GGY, Cheng LT, Wong FY, Chua MCH, Chua MLK, Tan DSW, Thng CH, Tan IBH, Ng HT. Inferring cancer disease response from radiology reports using large language models with data augmentation and prompting. J Am Med Inform Assoc. 2023 Sep 25;30(10):1657-1664. doi: 10.1093/jamia/ocad133.

Mayo Clinic. (n.d.). RadOnc-Foundation-Models [GitHub repository]. GitHub. https://github.com/Mayo-Clinic-RadOnc-Foundation-Models

REFERENCES (CONT'D)

Wiest IC, Leßmann ME, Wolf F, et al. Deidentifying medical documents with local, privacy-preserving large language models: The LLM-Anonymizer. NEJM AI. 2025;2(4). doi:10.1056/AIdbp2400537

Eyre H, Chapman AB, Peterson KS, Shi J, Alba PR, Jones MM, Box TL, DuVall SL, Patterson OV. Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python. AMIA Annu Symp Proc. 2022 Feb 21;2021:438-447.

Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010 Sep-Oct;17(5):507-13. doi: 10.1136/jamia.2009.001560. PMID: 20819853; PMCID: PMC2995668.

Neumann, M., King, D., Beltagy, I., & Ammar, W. (2019). ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task* (pp. 319–327). https://doi.org/10.18653/v1/W19-5034

Zhang, Y., Zhang, Y., Qi, P., Manning, C. D., & Langlotz, C. P. (2021). Biomedical and clinical English model packages for the Stanza Python NLP library. *Journal of the American Medical Informatics Association, 28*(9), 1892–1899. https://doi.org/10.1093/jamia/ocab090

Dernoncourt, F., Lee, J. Y., & Szolovits, P. (2017). De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association, 24*(3), 596–606. https://doi.org/10.1093/jamia/ocw156

Yang R, Zeng Q, You K, Qiao Y, Huang L, Hsieh CC, Rosand B, Goldwasser J, Dave A, Keenan T, Ke Y, Hong C, Liu N, Chew E, Radev D, Lu Z, Xu H, Chen Q, Li I. Ascle-A Python Natural Language Processing Toolkit for Medical Text Generation: Development and Evaluation Study. J Med Internet Res. 2024 Oct 3;26:e60601. doi: 10.2196/60601. PMID: 39361955

Das A, Tariq A, Batalini F, Dhara B, Banerjee I. Exposing Vulnerabilities in Clinical LLMs Through Data Poisoning Attacks: Case Study in Breast Cancer. AMIA Annu Symp Proc. 2025 May 22;2024:339-348. PMID: 40417494; PMCID: PMC12099371.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. LLaVA-med: training a large language-and-vision assistant for biomedicine in one day. In Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23). Curran Associates Inc., Red Hook, NY, USA, Article 1240, 28541–28564.

Schuler A, Walsh D, Hall D, Walsh J, Fisher C; Critical Path for Alzheimer's Disease; Alzheimer's Disease Neuroimaging Initiative; Alzheimer's Disease Cooperative Study. Increasing the efficiency of randomized trial estimates via linear adjustment for a prognostic score. Int J Biostat. 2021 Dec 22;18(2):329-356.

APPENDIX – SAMPLE WORKFLOW

TITLE: Ensemble Modeling and Targeted Inference Pipeline for Outcome Analysis in Rare Diseases

Note: The author gratefully acknowledges Sarbesh Pandeya and Haodong Li, both affiliated with AstraZeneca Rare Disease, for the underlying concept and selection of R packages utilized in the sample code. This sample workflow is intended for informational purposes only.

```
# Load required libraries for modeling, preprocessing, and targeted learning

library(dplyr)        # Data manipulation

library(caret)        # Data partitioning and machine learning utilities

library(sl3)          # Super Learner ensemble modeling

library(xgboost)      # Extreme gradient boosting algorithm

library(earth)        # Multivariate Adaptive Regression Splines (MARS)

library(glmnet)       # Regularized regression (Lasso and Elastic-net)

library(sandwich)     # Robust variance/covariance estimators

library(tmle)         # Targeted Maximum Likelihood Estimation (TMLE)
```

APPENDIX

```
# Step 1: Data Processing
data_clean <- raw_data %>%
  filter(!is.na(outcome)) %>%                  # Remove rows with missing outcome
  mutate(across(where(is.character), as.factor))# Convert any character columns to factors
# Step 2: Data Splitting
set.seed(123)                                  # Ensure reproducibility of the split
train_index <- createDataPartition(
  data_clean$outcome, p = 0.8, list = FALSE     # Stratified split: 80% train, 20% test
)
train_data <- data_clean[train_index, ]         # Training set
test_data <- data_clean[-train_index, ]         # Test set
# Step 3: Super Learner Model Training
task <- make_sl3_Task(
  data = train_data,
  covariates = c("age", "sex", "biomarker1"),   # Model predictors
  outcome = "outcome"                           # Outcome variable
)
```

APPENDIX

```
# Define stack of base learners: XGBoost, MARS, and Elastic-net
learners <- Lrnr_stack$new(learners = list(
  Lrnr_xgboost$new(),
  Lrnr_earth$new(),
  Lrnr_glmnet$new()
))
# Build Super Learner ensemble using stacked learners
sl <- Lrnr_sl$new(learners = learners)
sl_fit <- sl$train(task)                    # Train Super Learner on the task


# Step 4: ANCOVA Modeling
PS = sl_fit$predict(TESTDATA)   # predict the prognostic score for the test data, using model trained from external/training data
x <- model.matrix(~ age + sex + biomarker1 + PS, data = train_data) # Model matrix for covariates
y <- train_data$outcome                                  # Outcome vector
ancova_model <- cv.glmnet(x, y, alpha = 0.5)             # Elastic-net regression (ANCOVA-like), alpha=0.5 blends Lasso and Ridge
```

APPENDIX

```
# Step 5: Robust Variance Estimation

robust_se <- sqrt(

  diag(vcovHC(ancova_model, type = "HC1"))      # Computes robust standard errors using heteroskedasticity-consistent estimator

)

# Step 6: Targeted Learning Estimation

tmle_result <- tmle(

  Y = train_data$outcome,                    # Outcome variable

  A = train_data$treatment,                  # Treatment variable

  W = train_data[, c("age", "sex", "biomarker1")], # Covariate matrix

  family = "gaussian"                        # Specify outcome as continuous

)

summary(tmle_result)                         # Summarize TMLE estimates
```