# Research Highlights for awardees of
# GKII-Ashoka Breakthrough Grant

Ashoka University and Gupta-Klinsky India Institute at John Hopkins (GKII) launched the 1st round of Breakthrough Grant, a co-funded opportunity to support pilot projects led by faculty members from Ashoka University and Johns Hopkins University that involve interdisciplinary collaborations. The current research grant is focused on health data research.

## Leveraging Language Models and a Common Data Model to Unlock Real-World Evidence from Unstructured Electronic Health Record Data in India


*Gautam Ahuja*, BS[1,*], *Himani Balutia*, MPharm[2,*], *Siddhant Poudyal*, MS[2,*], Bableen Kaur, PhD[2], Ragul N, BS[1], Rudra Chinhara, BTech[7], Sivsanjai GA, BS[1], Hara Prasad Mishra, MBBS[4], Tamoghna Ghosh, MBBS[5], Sanjana Ahuja, BBA[2], Douluri Pushkala Devi, MSc[2], Shivangi Singh, MBBS[9], Upamanyu Das, MBBS[9], Tanisha Kohli, BS[8], Angad Singh, BS[1], Shrey Chhabra, BS[1], Aarush Kumbhakern, BS[1], Arundhati Mishra, MSc[2], Poorva Nandedkar, BDS[11], Claire Vania, MS[6], Suditi Arora, BDS[10], Ashiya Mahran, PGDHM[10], Divya Laroyia[10], Shyatto Raha[10], **Matthew Robinson**, MD[6], **Rintu Kutum**, PhD[1,2,3,11,#]

[1]Department of Computer Science, Ashoka University, Sonipat, Haryana, India
[2]Trivedi School of BioSciences, Ashoka University, Sonipat, Haryana, India
[3]Ashoka Mphasis Lab, Ashoka University, Sonipat, Haryana, India
[4]University College of Medical Sciences, Delhi, India
[5]All India Institute of Medical Sciences Delhi, Delhi, India
[6]Division of Infectious Diseases, Johns Hopkins School of Medicine, Baltimore, United States
[7]Department of Computer Science, Central University of Haryana, India
[8]Department of Biology, Ashoka University, Sonipat, Haryana, India
[9]Muzaffarnagar Medical College, Uttar Pradesh, India
[10]MyHealthcare Technologies Pvt. Ltd, Gurgaon, Haryana, India
[11]Lead contact; *Contributed equally;
[#]Correspondence: rintu.kutum@ashoka.edu.in; rintu.kutum@augmented-health-systems.org


### Objectives

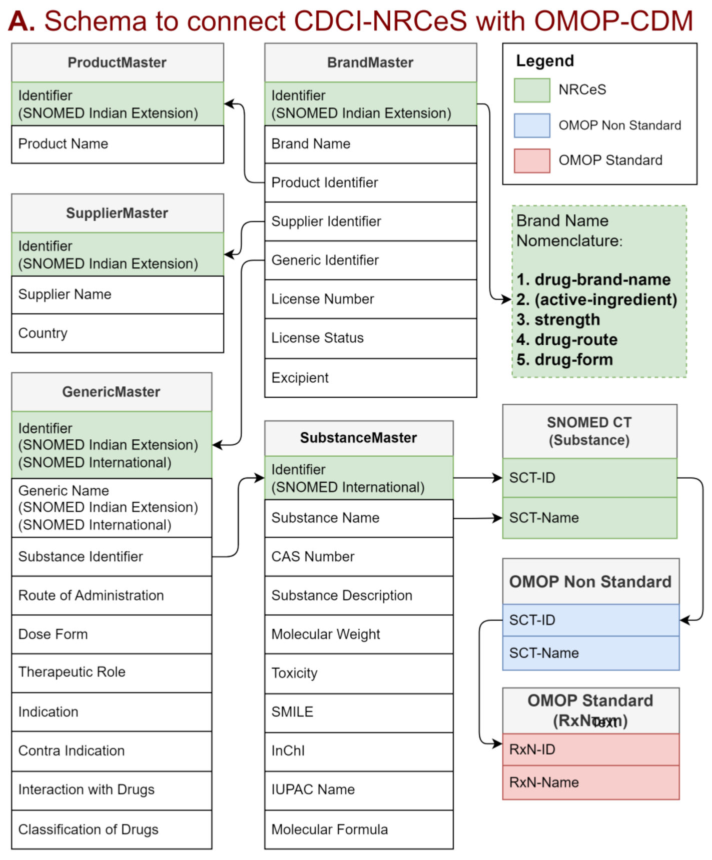
1. To develop natural language processing (NLP) algorithms and leverage open source large language models (osLLMs) tailored to the Indian context to extract and transform clinical features from unstructured clinical notes into the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM).
2. To determine the prevalence and choice of antibiotic prescription among outpatients with febrile respiratory illness in India and the US.
3. To assess the impact of antibiotic choice on treatment outcomes for inpatients with carbapenem-resistant organism pneumonia in India and the US.

### Methods

1. Bridging medication data of OMOP-CDM with the Common Drug Codes for India (CDCI), NRCeS with NLP and open source LLMs.
   - Design and development of schema to connect CDCI-NRCeS with OMOP-CDM.
   - Creation of high-quality annotation data (medication) to benchmark open-source LLMs with zero-shot and parameter-efficient fine-tuning.
   - Validation of the extracted medication-related information with rule-based clinical NLP.
2. Mapping medications related to the use of antibiotics with WHO AWaRe (Access, Watch, Reserve) classification.

### Results

**A.** Schema to connect CDCI-NRCeS with OMOP-CDM



ProductMaster
Identifier (SNOMED Indian Extension)
Product Name
SupplierMaster
Identifier (SNOMED Indian Extension)
Supplier Name
Country
GenericMaster
Identifier (SNOMED Indian Extension) (SNOMED International)
Generic Name (SNOMED Indian Extension) (SNOMED International)
Substance Identifier
Route of Administration
Dose Form
Therapeutic Role
Indication
Contra Indication
Interaction with Drugs
Classification of Drugs
BrandMaster
Identifier (SNOMED Indian Extension)
Brand Name
Product Identifier
Supplier Identifier
Generic Identifier
License Number
License Status
Excipient
SubstanceMaster
Identifier (SNOMED International)
Substance Name
CAS Number
Substance Description
Molecular Weight
Toxicity
SMILE
InChI
IUPAC Name
Molecular Formula
Legend
NRCeS
OMOP Non Standard
OMOP Standard
Brand Name Nomenclature:
1. drug-brand-name
2. (active-ingredient)
3. strength
4. drug-route
5. drug-form
SNOMED CT (Substance)
SCT-ID
SCT-Name
OMOP Non Standard
SCT-ID
SCT-Name
OMOP Standard (RxNorm)
RxN-ID
RxN-Name


**B.** Creation of high quality annotation data (medication)

**(i)** Prescription notes

Drez gargle 5 ml in 10ml water thrice a day for 3 days Tablet Mondeslor after dinner for 5 days Tablet Calpol 650 mg 3 to 4 times a day for 3 days Tablet Meftal 250 mg SOS for high grade fever or severe body pain

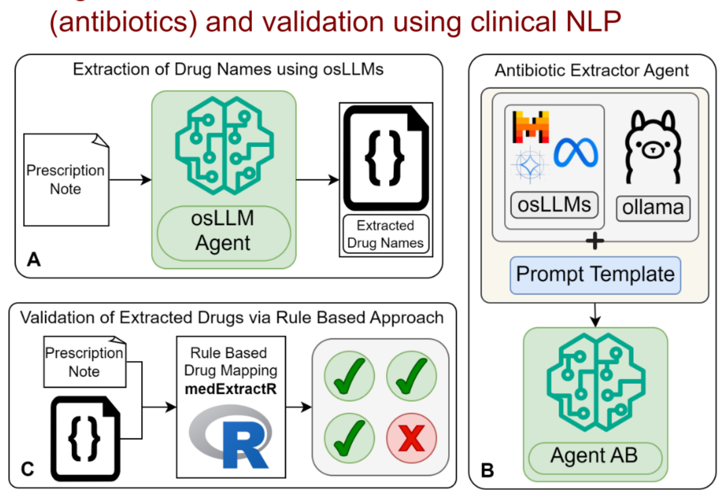
**(ii)** Medication annotation JSON format


{'drug_no_0': {'drug_name': 'Drez gargle', 'dose': '5 ml', 'dose_amount': '-', 'dose_change': '-', 'frequency': '-', 'intake_time': '-', 'route': '-', 'duration': '3 days'}, 'drug_no_1': {'drug_name': 'Tablet Mondeslor', 'dose': '-', 'dose_amount': '-', 'dose_change': '-', 'frequency': '-', 'intake_time': 'after dinner', 'route': '-', 'duration': '5 days'}, 'drug_no_2': {'drug_name': 'Tablet Calpol', 'dose': '650 mg', 'dose_amount': '-', 'dose_change': '-', 'frequency': '-', 'intake_time': '-', 'route': '-', 'duration': '3 days'}, 'drug_no_3': {'drug_name': 'Tablet Meftal', 'dose': '250 mg', 'dose_amount': '-', 'dose_change': '-', 'frequency': '-', 'intake_time': '-', 'route': '-', 'duration': '-'}}


**(iii)** Benchmark of osLLMs for medication extraction

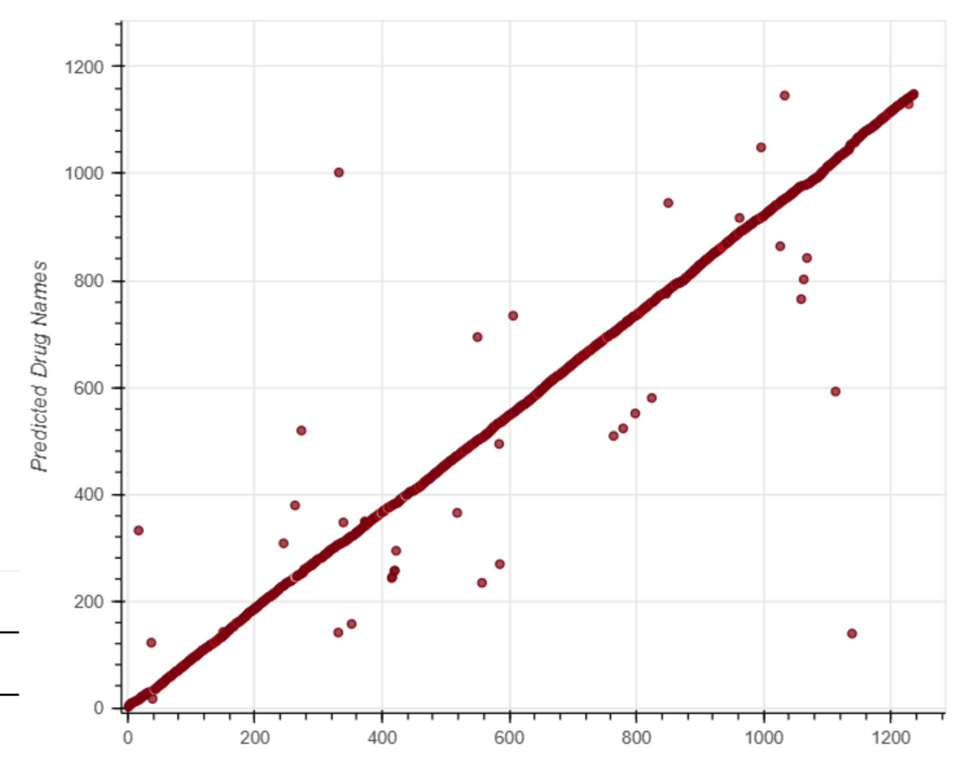
| Sl No. | osLLM | Precision | Recall | F1 | Hallucinated | Error rate |
|---|---|---|---|---|---|---|
| 1 | Gemma2-9b | 0.91 | 0.9 | 0.9 | 0.21 | 0.09 |
| 2 | Llama2-7b | **0.77** | **0.77** | **0.77** | 0.32 | **0.22** |
| 3 | Medalpaca-7b | 0.8 | 0.79 | 0.79 | **0.71** | 0.2 |
| 4 | Meditron-7b | 0.93 | 0.93 | 0.93 | 0.21 | 0.06 |
| 5 | Llama-3-8b | 0.93 | 0.91 | 0.92 | 0.18 | 0.08 |
| 6 | Llama-3.1-8b | 0.95 | 0.94 | 0.94 | **0.12** | 0.05 |
| 7 | Mistral-v0.1-7b | 0.93 | 0.92 | 0.92 | 0.26 | 0.07 |
| 8 | **Mistral-v0.3-7b** | **0.95** | **0.94** | **0.94** | 0.15 | **0.05** |
| 9 | Qwen2-7b | 0.95 | 0.93 | 0.93 | 0.18 | 0.06 |

**C.** Agent based workflow to extract medication (antibiotics) and validation using clinical NLP



Extraction of Drug Names using osLLMs
Prescription Note
osLLM Agent
Extracted Drug Names
A
Antibiotic Extractor Agent
osLLMs
ollama
Prompt Template
Agent AB
B
Validation of Extracted Drugs via Rule Based Approach
Prescription Note
Rule Based Drug Mapping medExtractR
C


**D.** Visual representation of the performance of agent based medication extraction with osLLM



Predicted Drug Names
True Drug Names

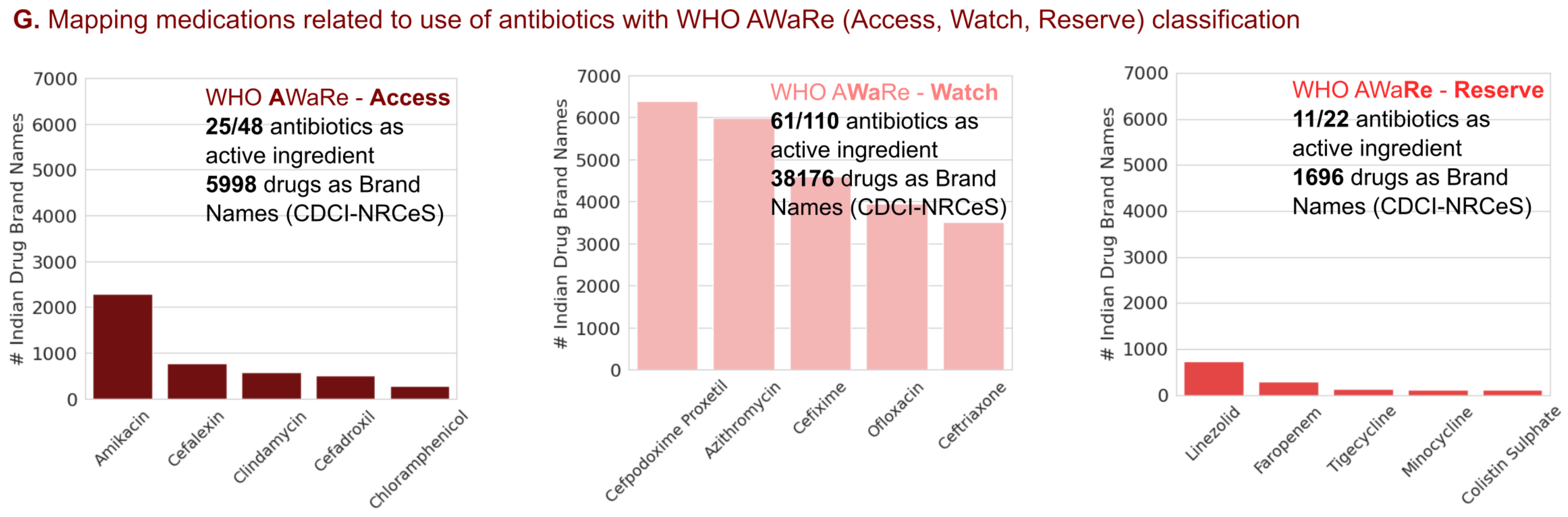

**E.** Confusion Matrix of the best model

| | Predicted Drug | Predicted Non-Drug |
|---|---|---|
| **Drug** | 1790 | 88 |
| **Non-Drug** | 166 | 0 |

**F.** Model Performance

| | Precision | Recall | F1 |
|---|---|---|---|
| **Drug vs Non-Drug** | 0.95 | 0.91 | 0.93 |

**G.** Mapping medications related to use of antibiotics with WHO AWaRe (Access, Watch, Reserve) classification



WHO AWaRe - Access
25/48 antibiotics as active ingredient
5998 drugs as Brand Names (CDCI-NRCeS)
WHO AWaRe - Watch
61/110 antibiotics as active ingredient
38176 drugs as Brand Names (CDCI-NRCeS)
WHO AWaRe - Reserve
11/22 antibiotics as active ingredient
1696 drugs as Brand Names (CDCI-NRCeS)
# Indian Drug Brand Names
Amikacin
Cefalexin
Clindamycin
Cefadroxil
Chloramphenicol
Cefpodoxime Proxetil
Azithromycin
Cefixime
Ofloxacin
Ceftriaxone
Linezolid
Faropenem
Tigecycline
Minocycline
Colistin Sulphate


### Future Direction

Mapping medications related to antibiotics with symptoms and disease conditions to determine the prevalence and use of antibiotics.


### Acknowledgments

RK would like to thank the GKII-Ashoka Breakthrough grant and Koita Centre for Digital Health-Ashoka (KCDH-A) for the funding. Also, RK acknowledges the compute infrastructure support from the Mphasis F1 Foundation.


## Machine Learning Methods for Enhanced Forecasting of Antiretroviral Therapy Demand in India


Bhavesh Neekhra[1], Swapnanil Mukherjee[1], Kshitij Kapoor[1], Debayan Gupta[1], Manish Bamrotiya[2], Sunil S. Solomon[2], Steven J. Clipman[2]

[1]Ashoka University, Sonipat, Haryana, India
[2]Johns Hopkins University School of Medicine, Baltimore, MD, USA


### Problem Statement

The primary objective is to develop and implement a machine learning-based model to enhance ART demand prediction in India. The specific aims are:

- **Aim 1:** Develop a machine learning model to predict the demand for each ART drug in India for the coming year.
- **Aim 2:** Validate the machine learning predictive model using historical data and compare its performance with current prediction methods.
- **Aim 3:** Develop a user-friendly interface for healthcare administrators to use the predictive model.

### Methods and Exploratory Data Analysis

- Exploratory data analysis (EDA) to understand the data patterns, identify missing values, outliers, relationship between features etc. to select the most suitable model.
- Experimented with various models including Support Vector Machine (SVM), Long Short Term Memory (LSTM), Simple Moving Average (SMA), AutoRegressive Integrated Moving Average (ARIMA), and Timesfm.

We also performed model validation by comparing our model's prediction with actual demand values for a period of 15 months.

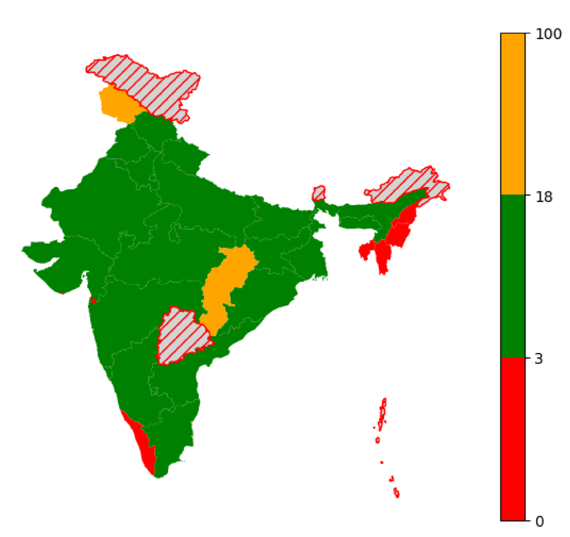


Figure 1. ABC/3TC Pediatric Stock, January 2017

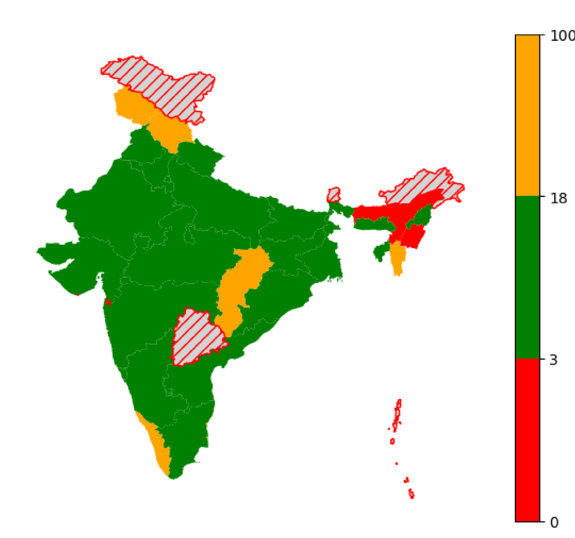


Figure 2. ABC/3TC Pediatric Stock, February 2017

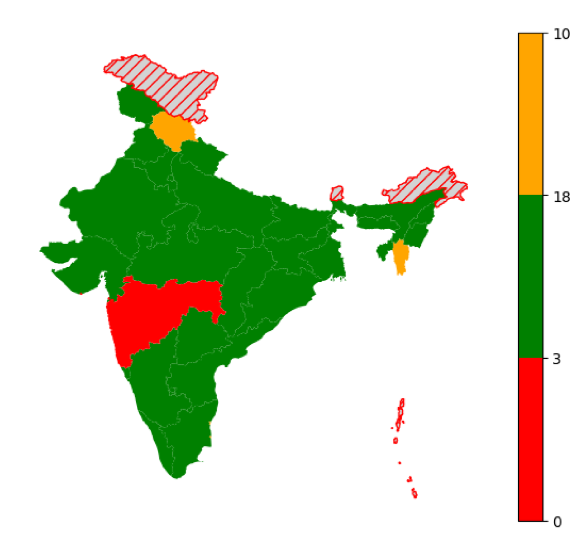
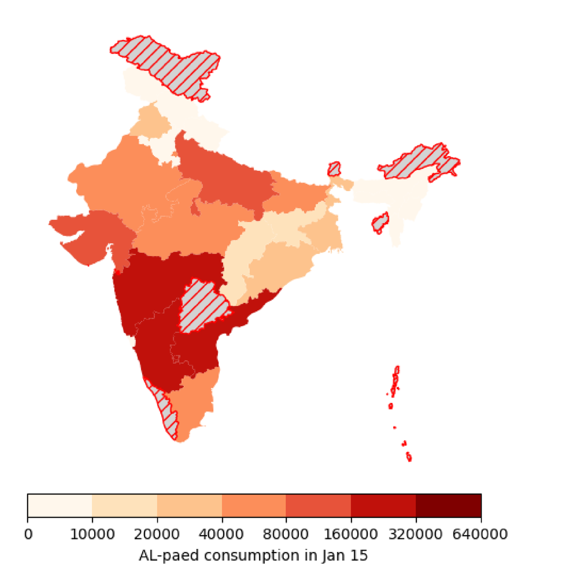
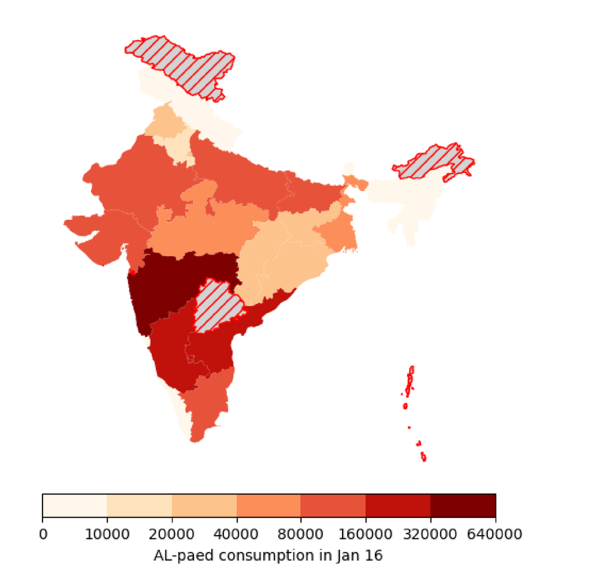


Figure 3. ABC/3TC Pediatric Stock, March 2017



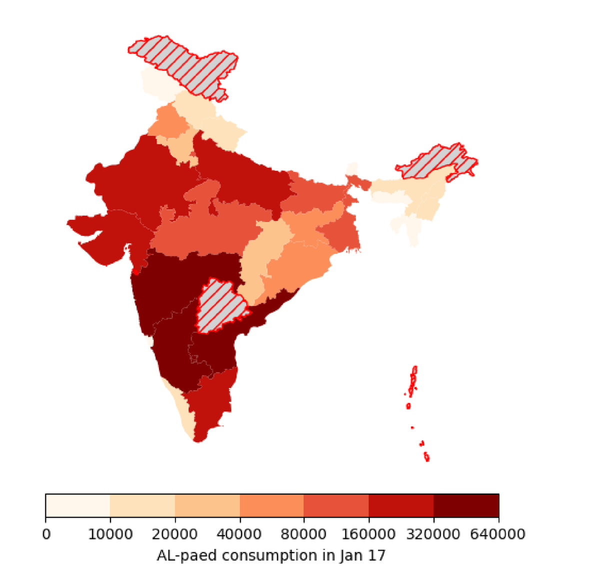
AL-paed consumption in Jan 15


Figure 4. ABC/3TC Pediatric Consumption, January 2015



AL-paed consumption in Jan 16

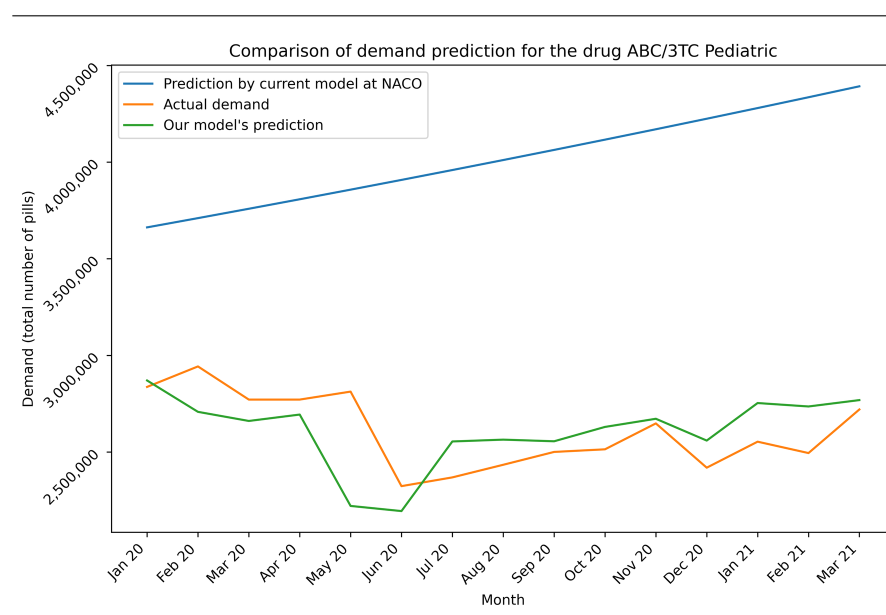

Figure 5. ABC/3TC Pediatric Consumption, January 2016



AL-paed consumption in Jan 17

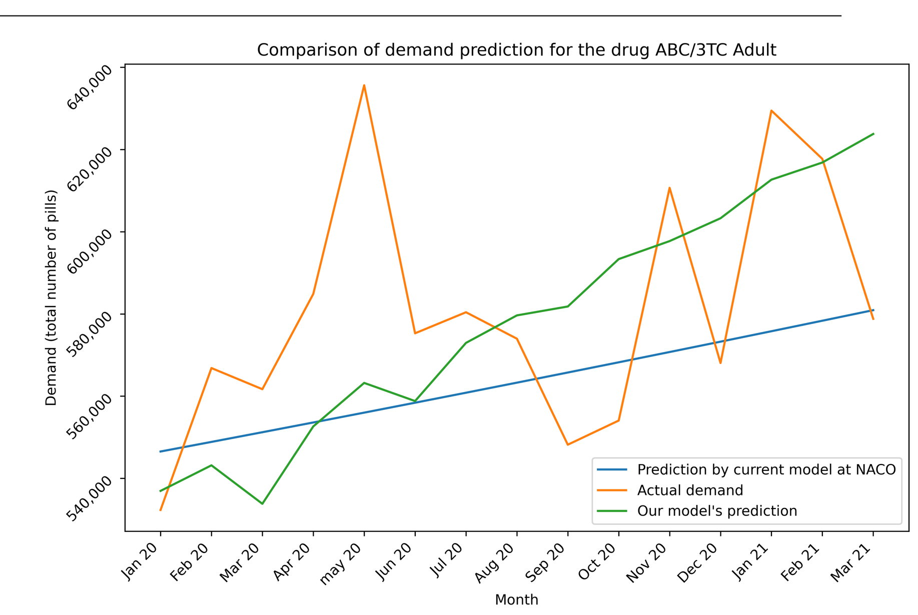

Figure 6. ABC/3TC Pediatric Consumption, January 2017

### Initial Results



Comparison of demand prediction for the drug ABC/3TC Pediatric
Prediction by current model at NACO
Actual demand
Our model's prediction
Demand (total number of pills)
Month


Figure 7. For ABC/3TC Pediatric: Comparison of actual demand, our model's prediction and current model at NACO



Comparison of demand prediction for the drug ABC/3TC Adult
Prediction by current model at NACO
Actual demand
Our model's prediction
Demand (total number of pills)
Month


Figure 8. For ABC/3TC Adult: Comparison of actual demand, our model's prediction and current model at NACO

| Drug Name | Actual | Our Model | NACO | MAPE (Actual, NACO) | MAPE (Actual, Our model) | % Change (Actual, NACO) | % Change (Actual, Our model) |
|---|---|---|---|---|---|---|---|
| ABC/3TC Adult | 8717970 | 8670610 | 8451718 | **4.15%** | 4.28% | 3.05% | **0.54%** |
| ABC/3TC Pediatric | 39109899 | 39140357 | 60255456 | 55.12% | **5.90%** | -54.07% | **-0.08%** |

Table 1. Model prediction error comparison: MAPE is calculated for 15 months period starting from January 2020 till March 2021

### Conclusion and Future Directions

Our results for ABC/3TC (both adult and pediatric) show better predictions than the current model at NACO. Specifically, for ABC/3TC pediatric, our predictions are within 0.08% while the current model over predicts by 54.07%. Our initial results suggest that to model drug demand for different drugs, we can not use the same model for all the drugs. We plan to extend our work in following ways:

- Model development and validation for all drugs.
- User interface and design and implementation, actively incorporating feedback from potential users.
- Applying our learning methodologies to enhance HIV care in India for targeted interventions including improving patient retention in care, personalized follow-up schedules and adherence support.


### Acknowledgments / Partnerships

We thank the National AIDS Control Organization (NACO) for providing us access to the necessary data and to the Johns Hopkins Gupta-Kilinsky India Institute for providing the funding to make this project possible. We also appreciate the efforts of Jaee and Saumya, who helped us clean and preprocess the data.



Poster Session