PMSA

PHARMACEUTICAL MANAGEMENT SCIENCE ASSOCIATION

Use of AI/ML and generative AI for prognosis in Myelodysplastic Syndrome using de-identified Market Clarity Database

Presenter: Vikash Verma

Disclosure

- The author is an employee of Optum.
- The content is solely the responsibility of the author and does not necessarily represent the official views of Optum or any other parties.
- The author doesn’t have any financial and personal relationships with other people or organizations that could inappropriately influence (bias) the work discussed in this presentation.
- No additional compensation was provided for the current work.

Presentation Flow

1	**Introduction – about the speaker**
2	**Disease background and challenge**
3	**Proposed solution**
4	**Conclusions**
5	**Appendix**

Presentation Flow

1	**Introduction – about the speaker**
2	Disease background and challenge
3	Proposed solution
4	Conclusions
5	Appendix

Introduction – about the speaker

Vikash Verma, Director
Optum Lifesciences

Vikash has 15 years of experience in forecasting, HEOR, managed market analytics, portfolio/data strategy, sales force excellence, and KOL management. He has worked in the US and Europe with leading pharmaceutical companies for new product launches, building digital strategies for patient engagement, and building next-generation analytical platforms for forecasting and managed market analytics.

Currently, he leads the Optum Life Sciences India team, comprising consultants, doctors, data scientists, and data engineers to support projects on healthcare commercial effectiveness and real-world insights solutions. He has presented/authored over 80 posters/whitepapers in international journals/conferences.

Before joining Optum, Vikash worked with companies like TCS, GSK Knowledge Centre and Pharmarc (part of IQVIA now). He received his BS degree in pharmaceutical science from Manipal College of Pharmaceutical Sciences, an MBA (marketing) from Manipal Institute of Management, an Executive program in business analytics from IIM Calcutta and the CXO Programme - in Driving Growth from XLRI.

Team – includes a mix of doctors, business consultants, and data scientists

Dr. Shailaja Daral, MBBS, MD, MBA,
Senior Consultant, Optum

Abhimanyu Roy, BS (pharma), MBA,
Senior *Consultant, Optum*

Shashi Bhushan Khan, MS (Statistics),
Senior Data Scientist, Optum

Riddhi Markan, MS (Economics),
Data Scientist, Optum

Sudhanshu Chawla, B. Tech,
Data Scientist, Optum

Abhishek Gaur, BS (pharma), MBA,
Senior Consultant, Optum

Dr. Abhinav Nayyar, MBBS, MBA,
Consultant, Optum

Dr. Shikha Anand, MBBS, DNB,
Consultant, Optum

Kulbhushan Tanwar, B. Tech,
Data Scientist, Optum

Presentation Flow

1	Introduction – about the speaker
2	**Disease background and challenge**
3	Proposed solution
4	Conclusions
5	Appendix

Myelodysplastic syndrome (MDS) is a type of blood cancer that affects the bone marrow's ability to produce mature blood cells

- Bone marrow produces blood stem cells which eventually develop into mature blood cells over time.
- These blood stem cells can either become:
  - Lymphoid stem cells – which mature into white blood cells
  - Myeloid stem cells – which mature into one of three types of blood cells(red blood cells, platelets, or granulocytes)
- In myelodysplastic syndrome (MDS), the blood stem cells fail to develop into mature red blood cells, white blood cells, or platelets.
- These immature blood cells do not function effectively. A decrease in healthy blood cells can lead to the development of anemia, infections, or easy bleeding.

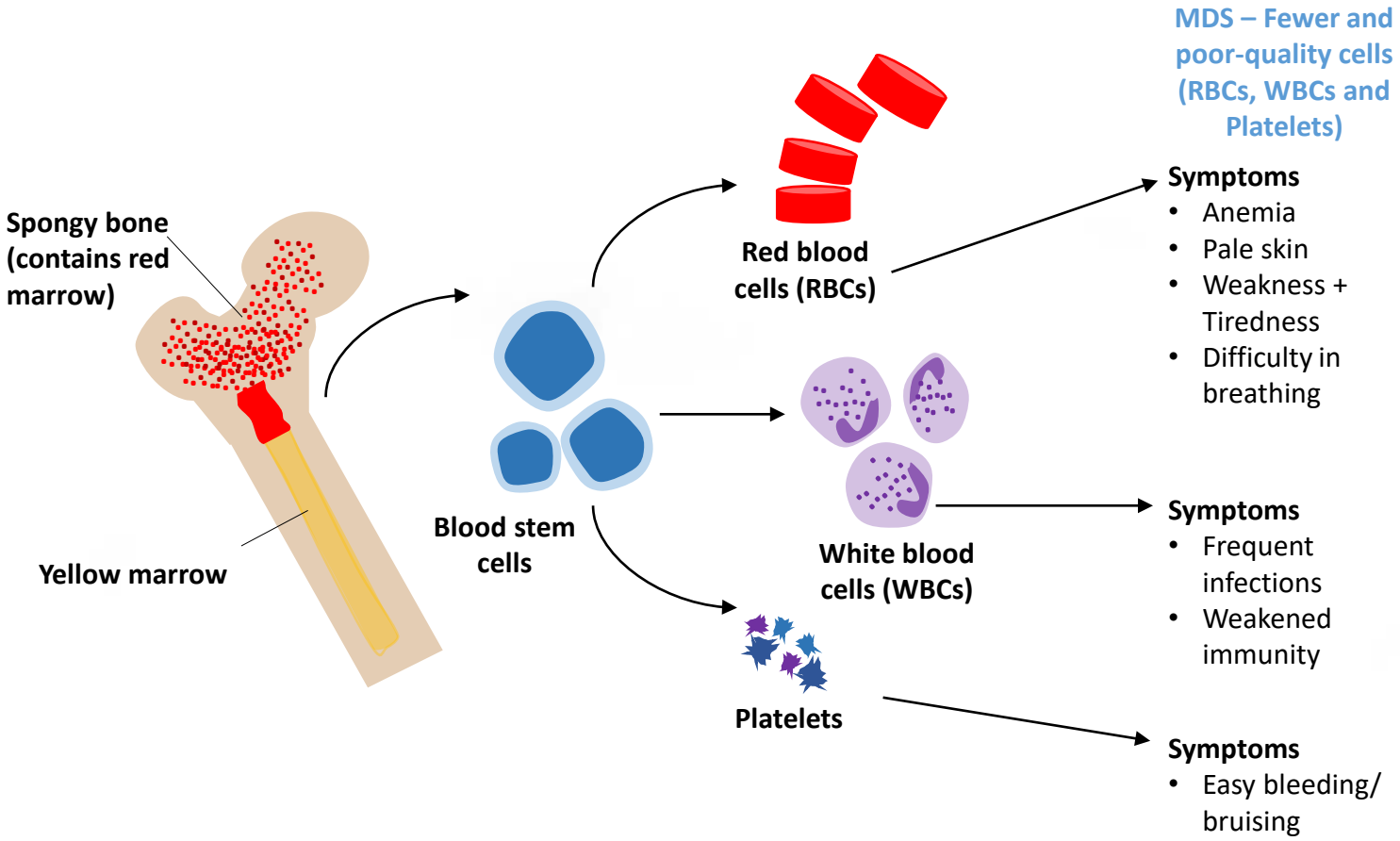


Fig 1: Bone Anatomy and Different Blood Cells
Source: Centre for Clinical Haematology

IPSS - is a commonly used tool to predict the long-term outcome of patients with MDS

- MDS staging is performed differently than the staging for any other type of cancer. Traditional staging systems are based on the size and extent of a solid tumor; MDS develop in the bone marrow and do not produce solid tumors that can be assessed in conventional ways.
- Physicians developed the International Prognostic Scoring System (IPSS) to stage myelodysplastic syndromes. IPSS includes the percentage of blasts in the bone marrow, karyotype, and the number of cell lineages with cytopenia.

Variable[1]	Score				
	0	0.5	1.0	1.5	2.0
Bone marrow blasts (percent)	<5	5 to 10	-	11 to 20	21 to 30
Karyotype*	Good	Intermediate	Poor	-	-
Cytopenia**	0/1	2/3	-	-	-

Risk group[1]	IPSS score	Median Survival (years) without therapy
Low	0	5.7
Intermediate - 1	0.5 to 1.0	3.5
Intermediate – 2	1.5 to 2.0	1.2
High	2.5 to 3.5	0.4

*Karyotype – Chromosomal abnormality:
Good: Normal;-Y; deletion of long arm of chromosome 5q and 20q
Poor: Complex (23 abnormalities); monosomy 7
Intermediate: All others chromosome abnormality
**Cytopenia -
Red blood cells : Hemoglobin <10 g/dL
(100 g/L)
White blood cells: Absolute neutrophil count <1800/microL
Platelets: Platelet count <100,000/microL

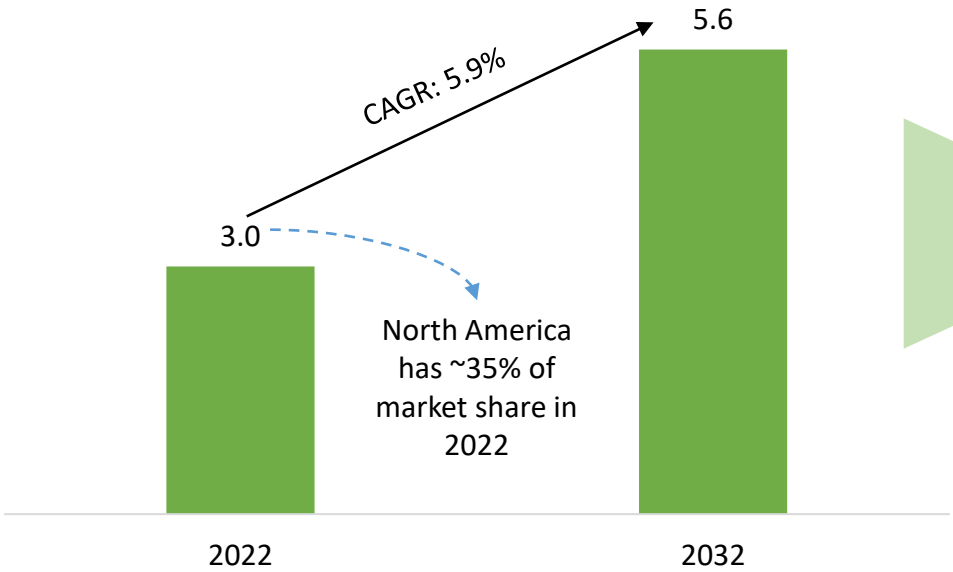
Source: 1. MDS Risk Assessment – IPSS – The Hematologist

MDS is a rare disease and there is a need for more effective treatment options for MDS

- **Incidence (US): 20,000 new cases** are reported every year[1]

- **Prevalence (US)** - estimated to be between **60,000 – 170,000 patients**[1]

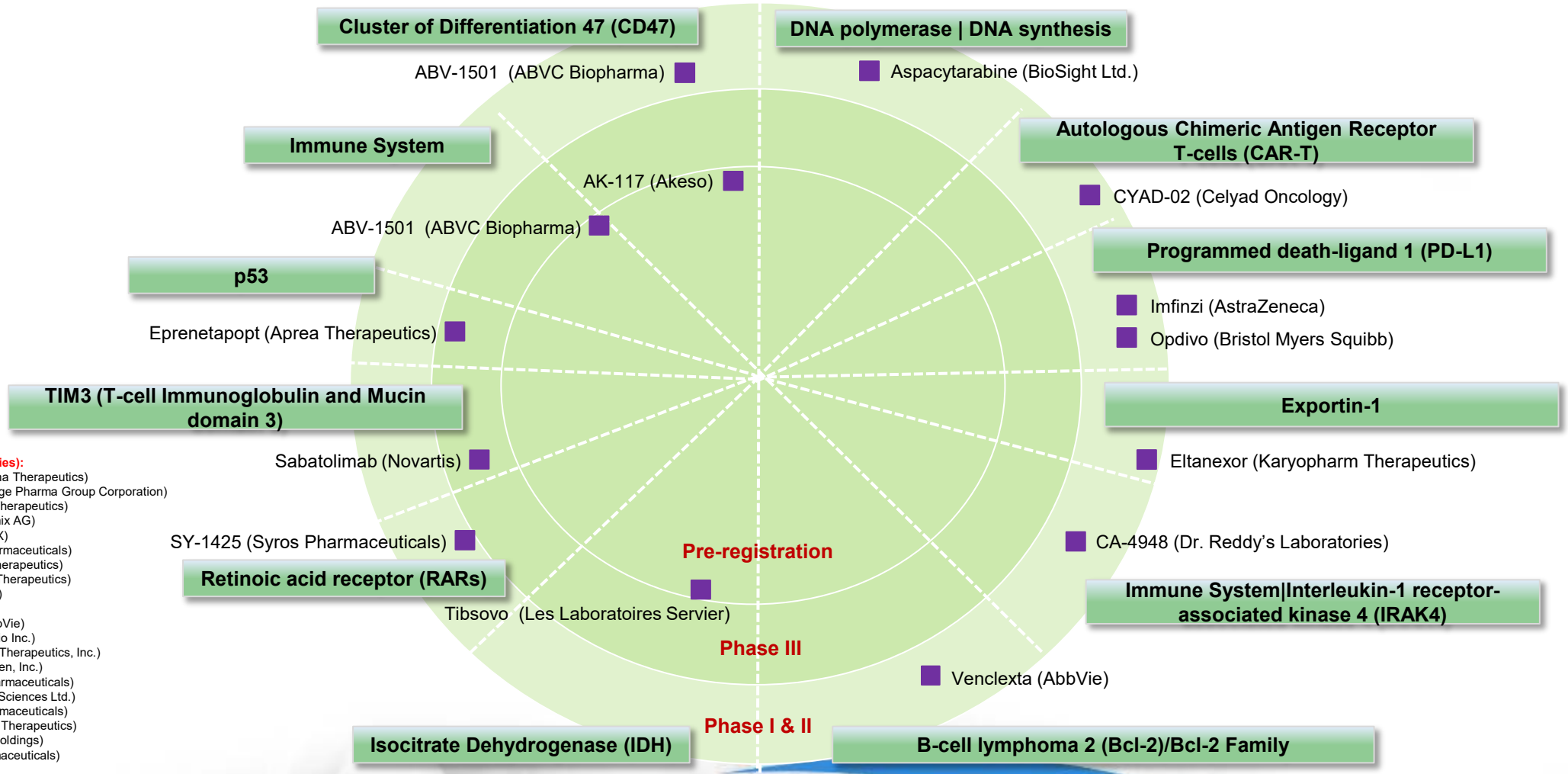
Global MDS Treatment Market (US$ B)[2]



- Hypomethylating agents (azacitidine and decitabine) are the most effective medications for treating MDS that are currently used
- Drivers for MDS market:
  - Growing number of patient pool of age 45 and above, and rising government support for cancer treatment
  - Launch of combination therapies for higher response rate. Eg: Azacitidine in combined with lenalidomide or vorinostat (a deacetylase inhibitor currently FDA approved for treating cutaneous T-cell lymphoma)

Sources: 1. Myelodysplastic Syndrome (MDS) Research - Leukemia & Lymphoma Society (LLS); 2. Myelodysplastic Syndrome Treatment Market Forecasts (2022– 2032) – Future Market Insights

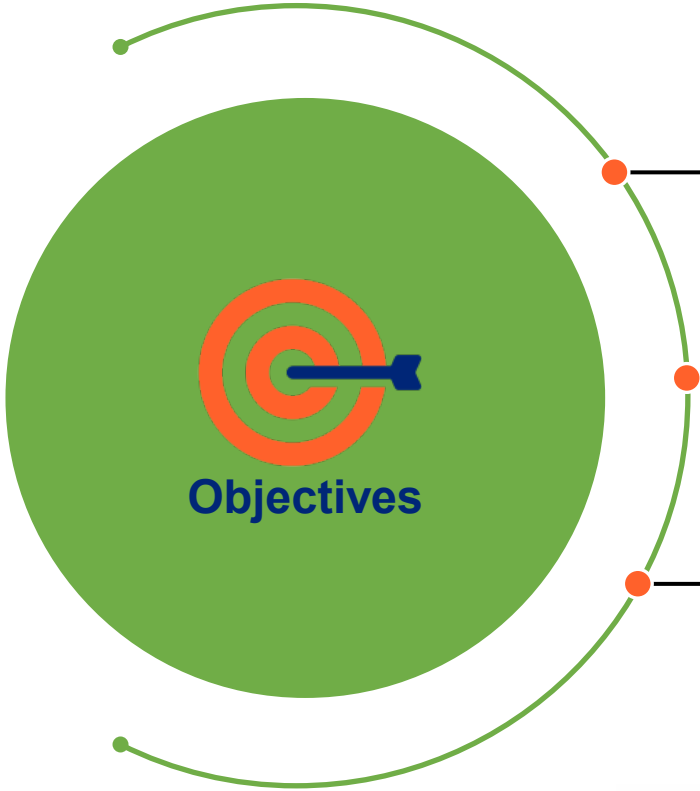
Upcoming product pipeline is strong



Phase I candidates (companies):

1. AMV-564 (Amphivena Therapeutics)
2. APG-2575 (Ascentage Pharma Group Corporation)
3. APVO436 (Aptevo Therapeutics)
4. Asunercept (Apogenix AG)
5. BTX-A51 (BioTheryX)
6. CB-5339 (CASI Pharmaceuticals)
7. DVX201 (Coeptis Therapeutics)
8. FHD-286 (Foghorn Therapeutics)
9. JNJ-67571244 (JNJ)
10. K-NK003 (Sanofi)
11. Lemzoparlimab (AbbVie)
12. MB-102 (Mustang Bio Inc.)
13. OXi4503 (Oncotelic Therapeutics, Inc.)
14. PRGN-3006 (Precigen, Inc.)
15. Rezlidhia (Rigel Pharmaceuticals)
16. RVT-2001 (Roivant Sciences Ltd.)
17. AG-946 (Agios Pharmaceuticals)
18. Briquilimab (Jasper Therapeutics)
19. ASTX030 (Otsuka Holdings)
20. Vyxeos (Jazz Pharmaceuticals)

Prognosis in Myelodysplastic Syndrome patients by using of AI/ML and generative AI technique

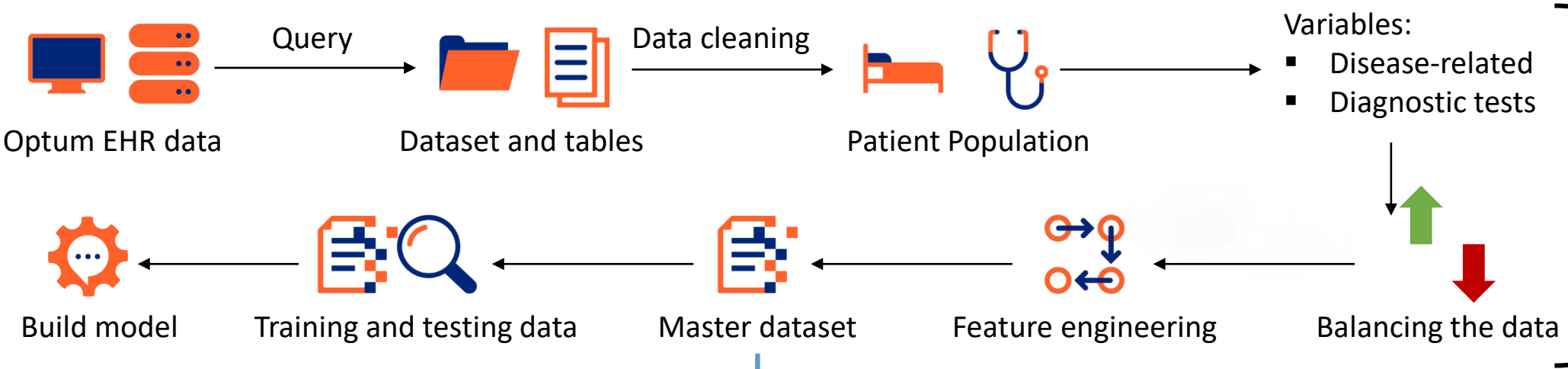


- Evaluation of the AI/ML technique to predict MDS prognosis
- Use of clinical notes to enhance the results
- Use of generative AI while analyzing clinical notes

Presentation Flow

1	Introduction – about the speaker
2	Disease background and challenge
3	**Proposed solution**
4	Conclusions
5	Appendix

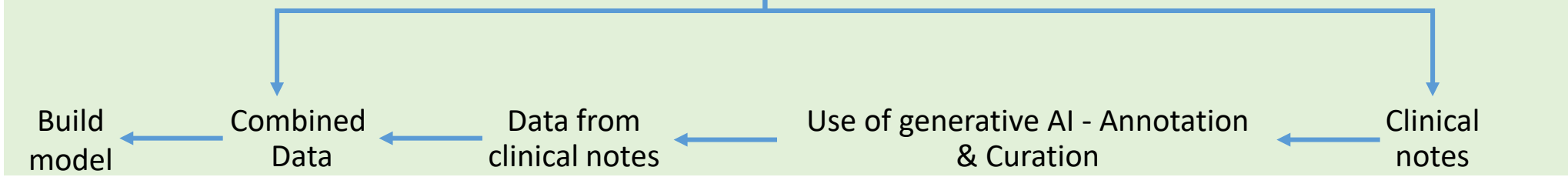
AI/ML technique applied to predict MDS prognosis; results were enhanced by using clinical notes and generative AI technique



Two models –

1. **Based on EHR structured data* only**
2. **Based on EHR structured + unstructured data using standard NLP** from clinical notes**

Enhancements



Based on EHR structured + unstructured data using standard NLP + customized NLP* & GI from clinical notes**

*Based on the ICD-10, CPT, and NDC codes
**It includes data for signs and symptoms from clinical notes – standard practice at Optum
***Customized NLP on clinical notes was performed for exposure to radiations and chemical

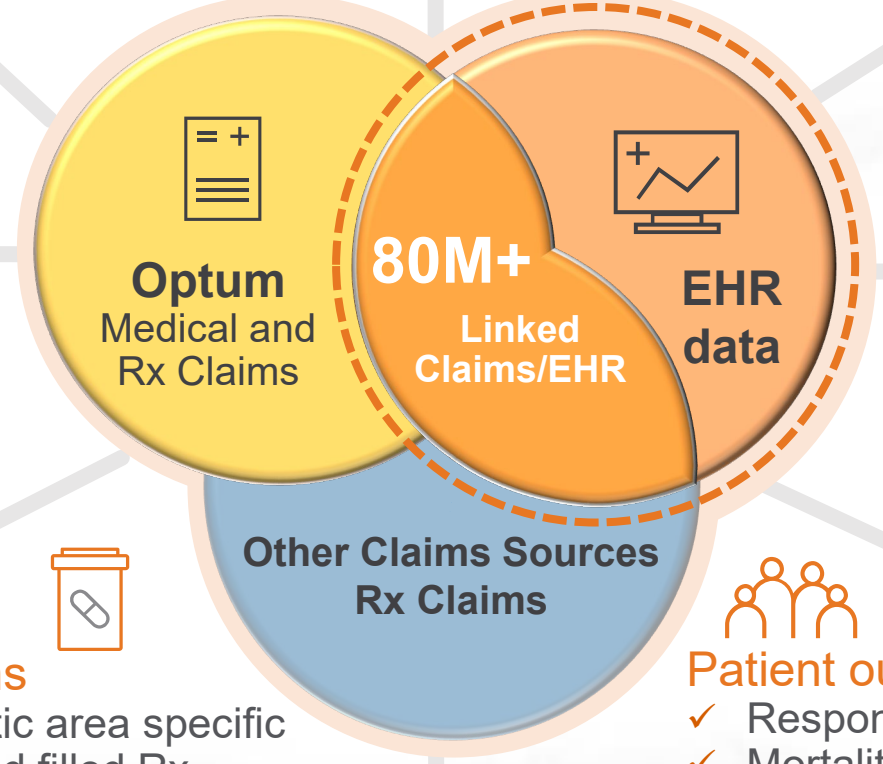
Market Clarity provides a variety of insights into all healthcare encounters, physician notes and associated costs

Cost
Disease burden for pharmacy, provider, and facility spending

400+ Lab values
A1C, eGFR, FEV etc.

Vital signs
BMI, blood pressure, pain scores etc.

Provider notes
- ✓ Treatment rationale
- ✓ Lifestyle observations
- ✓ Signs and symptoms
- ✓ Biomarkers



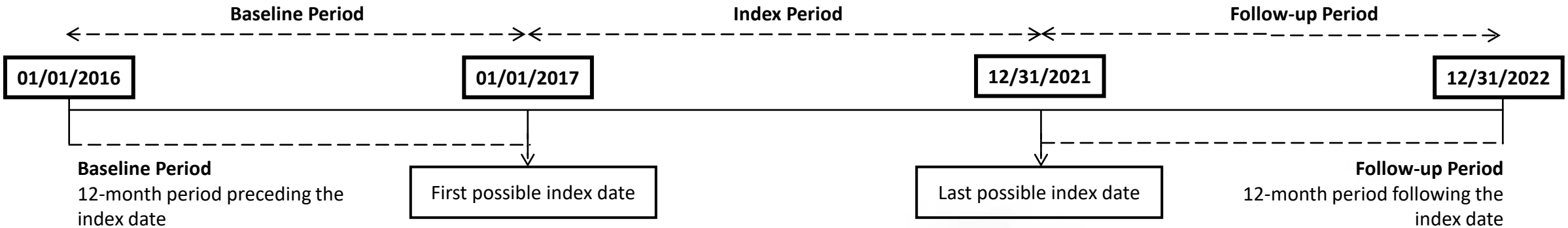
Patient and provider details
- ✓ Physician speciality
- ✓ Age, gender, race
  Smoking status, race
- ✓ Geography

Medications
- ✓ Therapeutic area specific
- ✓ Written and filled Rx

Patient outcomes
- ✓ Response to treatment
- ✓ Mortality outcomes'

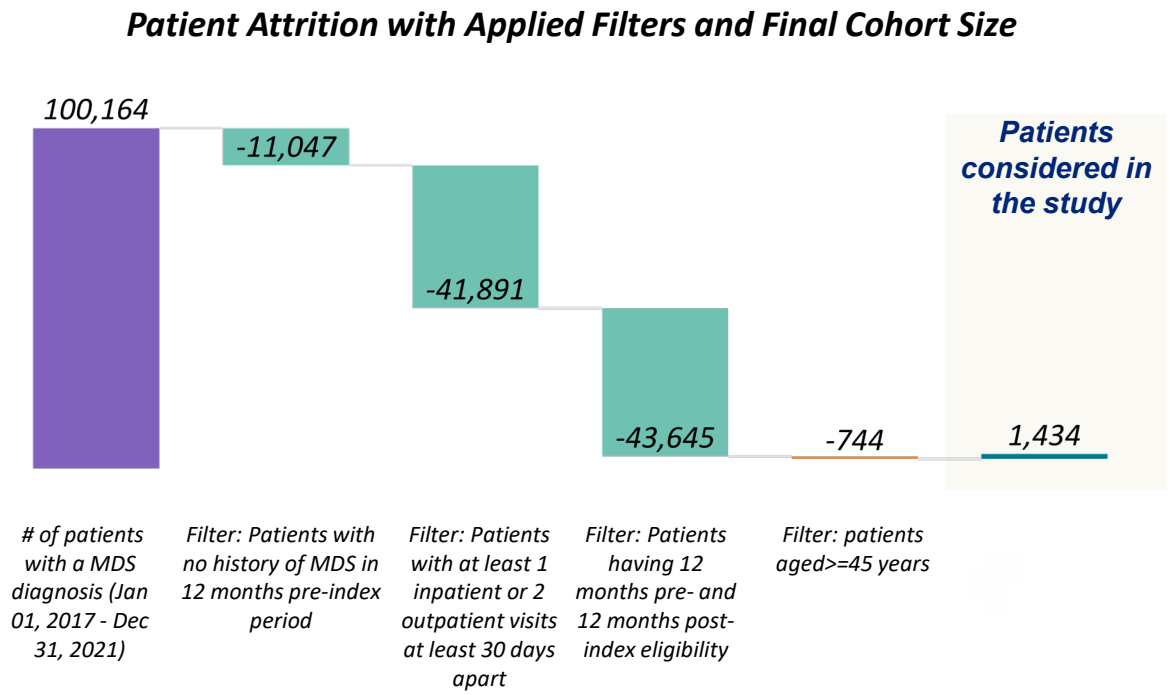
Medical procedures and conditions
- ✓ Indications
- ✓ Treatments
- ✓ Comorbidities and complications

Study design for the ML model



Data Sources	Optum EHR data (de-identified and fully compliant with the Health Insurance Portability and Accountability Act)
Index Period	January 1, 2017 – December 31, 2021
Look Forward (Follow-up) Period	12 months post-index date
Look Back (Baseline) Period	12 months prior to the index date
Inclusion Criteria	Patients aged ≥45 years who have continuous clinical activity in pre- and post-index along with the following event in the index period ▪ ≥2 outpatient diagnoses on two different dates at least 30 days apart or ▪ ≥1 inpatient diagnosis or ▪ ≥2 clinical notes with positive mention of MDS
Exclusion Criteria	Patients with the above criteria in the Look Back (Baseline) Period
Index Event	First occurrence of a diagnosis of interest
ICD-10 codes	D46/C94.6 for MDS identification

Patient population for the ML model

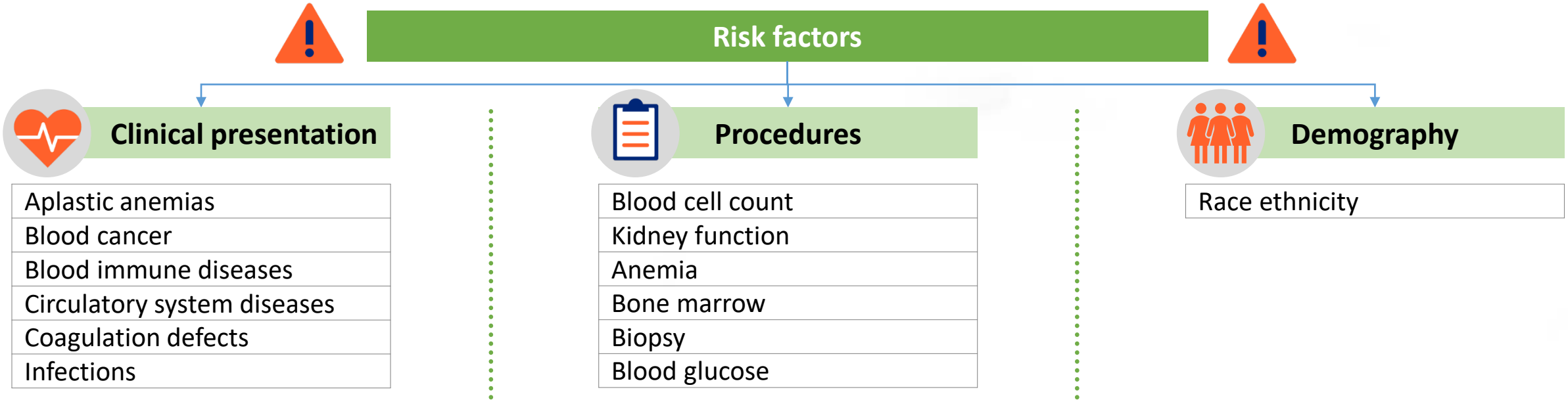


Step	Waterfall for Cohort Identification	Patient Count
S0	Number of patients with a MDS diagnosis from Jan 01, 2017 - Dec 31, 2021	100,164
S1	Number of patients with no history of MDS in 12 months pre-index period	89,117
S2	Number of patients with at least 1 inpatient or 2 outpatient visits at least 30 days apart	47,226
S3	Number of patients having 12 months pre- and 12 months post-index eligibility	3,581
S4	Number of patients aged>=45 years	2,837
S5	Number of patients considered after data preparation	1,434
Cohort	**S5**	**1,434**

Patients who had MDS diagnosis were considered as case group and control was created by random sampling. To reduce the confounding effects of age, gender, and region, the study matched cases to controls using propensity score matching (PSM) method

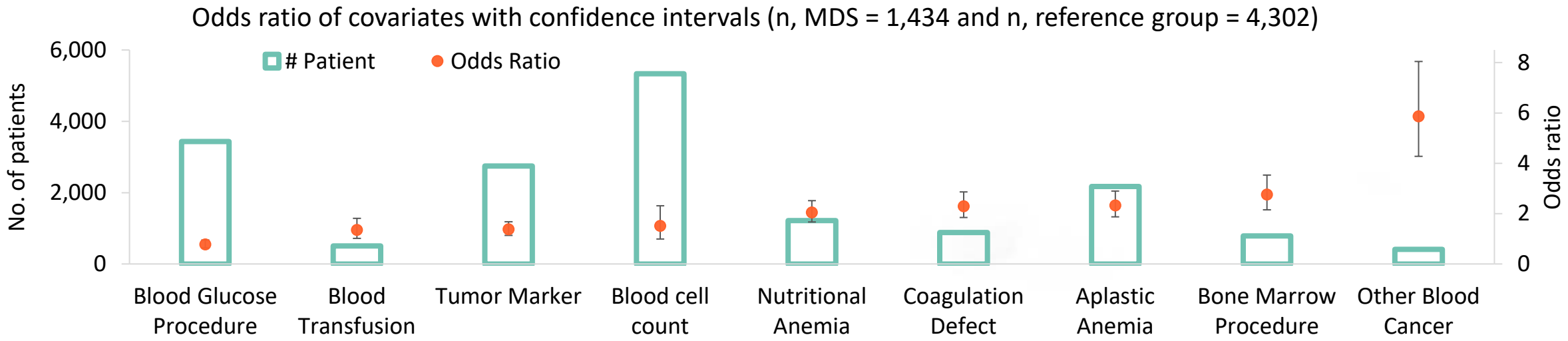
Risk factors that were used as predictor variable

Brainstormed, performed primary research with the clinical SMEs and literature review to determine the important predictor variables for each stage of MDS



Complete list of risk factors is given in the appendix section

Highest odds ratio is seen in 'Other Blood Cancer', indicating it has the strongest association compared to the reference group



Aplastic anemia

Blood cancer

Coagulation defects

Blood cell count

Blood immune diseases

Neoplasm

Anemia

Radiation exposure

Chemotherapy

Bone marrow

Nutritional anemia

Blood transfusion

Process of predicting MDS prognosis: Model training and model selection and Choose cutoff

Divided datasets into train-dataset and test-dataset on a ratio of 70:30

- Random sampling of 70% for patients into a training dataset
- Allocated the remaining 30% into testing data
- Used balancing technique to balance test data

Built models on the training dataset

- Logistic Regression (LR, recursive feature elimination)
- Random Forest (RF)
- XGBoostClassifier (XGB)

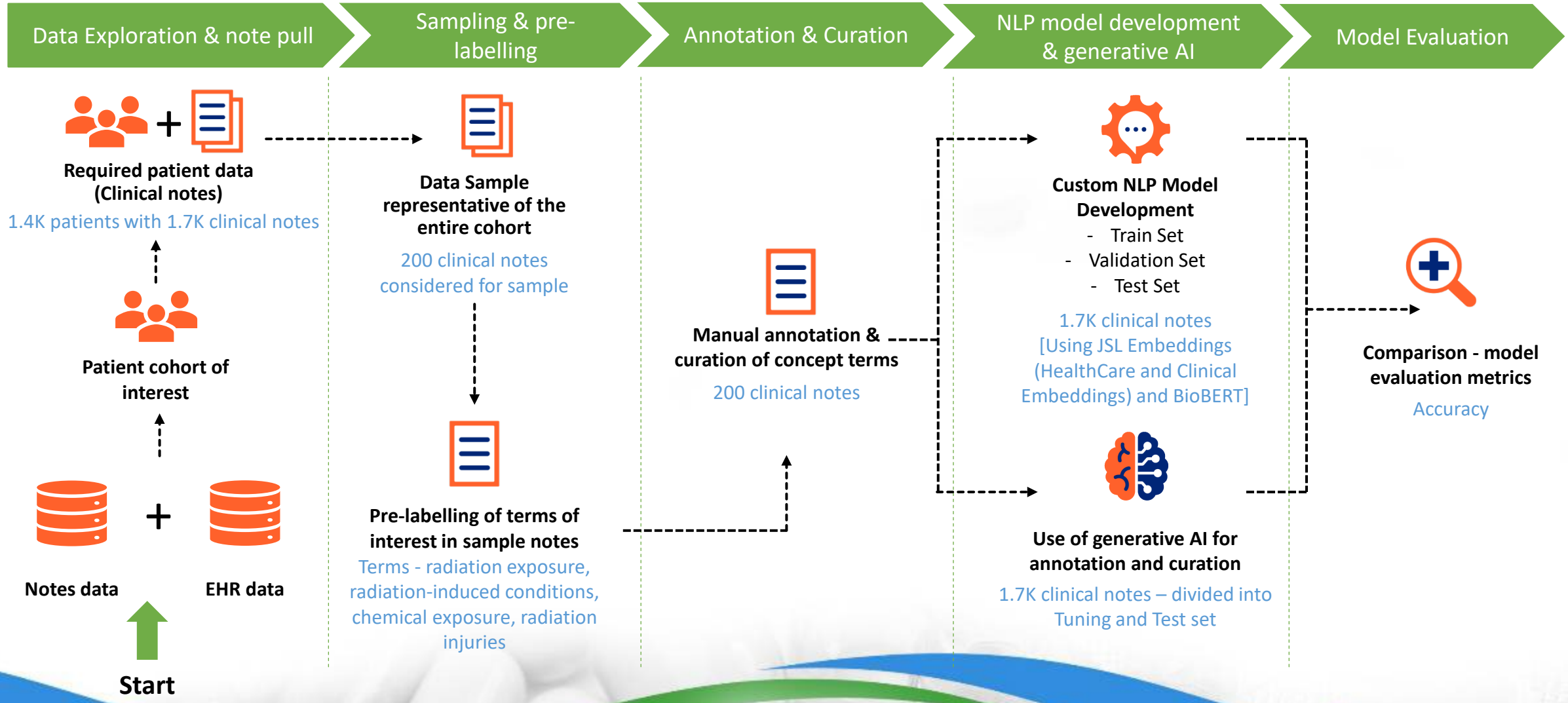
Selected the model with high sensitivity and AUC

- Applied each model to the testing data and pick the one with the best sensitivity and AUC
- **Logistic Regression** is selected for early prediction
- **K-means clustering** with MCA for segmentation of MDS patients. This was based on significant variables from LR model

Enhancements:

1. Customized data from clinical notes (Annotation & Curation)
2. Generative AI to increase the accuracy and speed to customize data from clinical notes

Methodology – Customization of data from clinical notes and use of generative AI



Pre-labelling terms of annotation of the clinical notes

Detailed review of the clinical notes for the MDS patients reviled that exposure to radiation and certain chemicals can increase the risk of developing MDS. Primary research with the clinical SMEs and literature review was conducted to determine the key terms that should be searched in the clinical notes.

1. Radiation Exposure
- Initial Encounter with Radiation
- Subsequent Encounter with Radiation
- Sequela from Radiation Encounter

2. Chemical Exposure
- Exposure to Benzene
- Exposure to Pesticides
- Exposure to Other Industrial Chemicals

3. Radiation-Induced Conditions
- Acute Pulmonary Manifestations due to Radiation
- Chronic Pulmonary Manifestations due to Radiation
- Radiation-Induced Skin Changes (acute, chronic, unspecified)

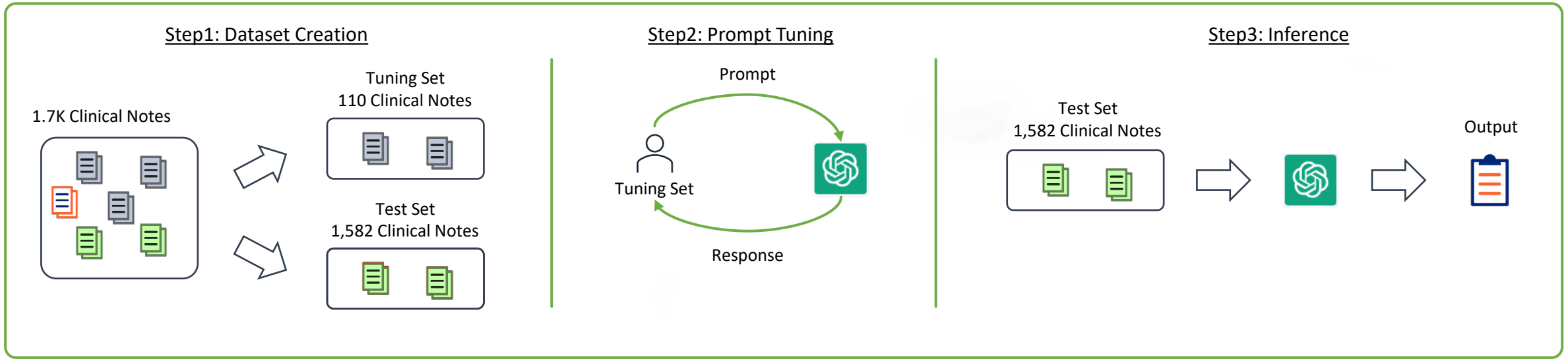
4. Radiation Injuries
- Radiation Injury (unspecified)
- Radiation Cystitis
- Radiation Proctitis

Complete list of key terms is given in the appendix section

Use of generative AI to increase the accuracy and speed to customize data from clinical notes

Objectives of using generative AI technique:

1. **Can we do it faster?**
2. **Can we get more accuracy?**



Used generative AI for named entity recognition (NER) extraction of terms such as radiation exposure, radiation-induced conditions, chemical exposure, and radiation injuries for creating structured features which could be fed into the predictive model solutions for the Model finetuning

1. **Solution developed using generative AI is ~10X faster as compared to manual effort**
2. **Generative AI provided accurate output as compared to manual effort**

Entity Type	Generative AI Accuracy	Manual effort Accuracy
Exposure to radiation	94.8%	90.5%
Benzene	96.6%	92.3%
Radioactive isotopes	98.2%	98.1%
Chemical exposure	98.2%	98.0%
Pesticide	94.2%	90.7%

NLP Modeling: NER Classification Embeddings

Embeddings refer to a technique used to represent words or phrases as vectors in a high-dimensional space which are created from large amounts of text data and learn to map words or phrases to vectors based on their context and meaning

Outperforming

HealthCare Embeddings

This model contains a pre-trained weights of a language representation model for Healthcare domain which are trained PubMed + ICD10 + UMLS + MIMIC III corpora

Clinical Embeddings

This model contains a pre-trained weights of a language representation model for Clinical domain which are trained on PubMed corpora

BioBERT

This model contains a pre-trained weights of BioBERT, a language representation model for biomedical domain, especially designed for biomedical text mining tasks such as biomedical named entity recognition, relation extraction, question answering, etc..

Label	Precision	Recall	F-1 Score
Radiation Exposure	0.83	0.85	0.84
Chemical Exposure	0.84	0.82	0.82
Radiation-Induced Conditions	0.85	0.86	0.85
Radiation Injuries	0.86	0.85	0.85
Overall	**0.85**	**0.84**	**0.84**

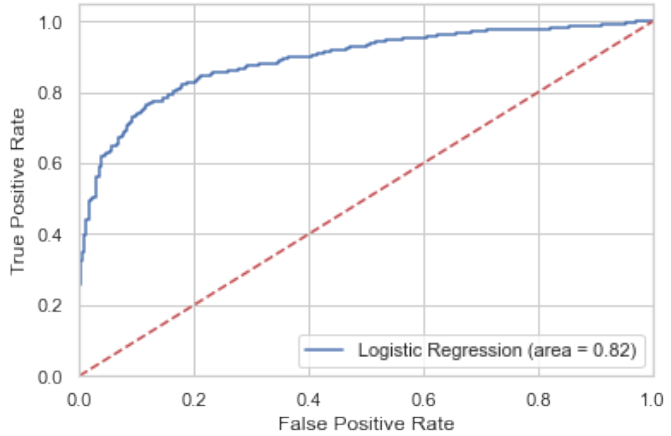
Comparison of models based on different data sets

Logistic regression (LR), random forest (RF), XGBoost (XG) models were used to evaluate the prediction of MDS prognosis. LR was the best performing model

Model based on EHR data* only

Model	Precision	Recall	f1-score
LR	84%	79%	82%

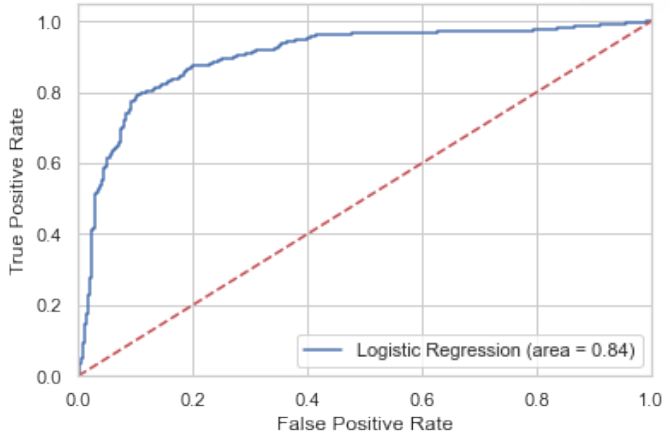
ROC curve for model based on EHR data only



Model based on EHR structured + unstructured data using standard NLP from clinical notes**

Model	Precision	Recall	f1-score
LR	86%	81%	83%

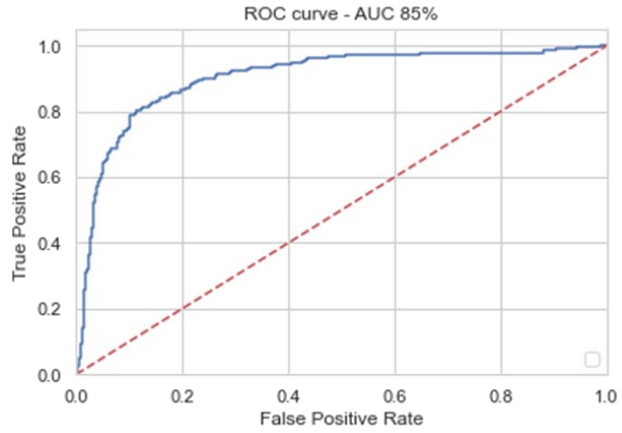
ROC curve for model based on EHR and standard data from clinical notes



Model based on EHR structured + unstructured data using standard NLP + customized NLP* from clinical using generative AI notes**

Model	Precision	Recall	f1-score
LR	87%	81%	84%

ROC curve for model based on EHR, standard and customized data from clinical notes



Model based on EHR, standard and customized data from clinical notes are a better indication of predicting MDS prognosis

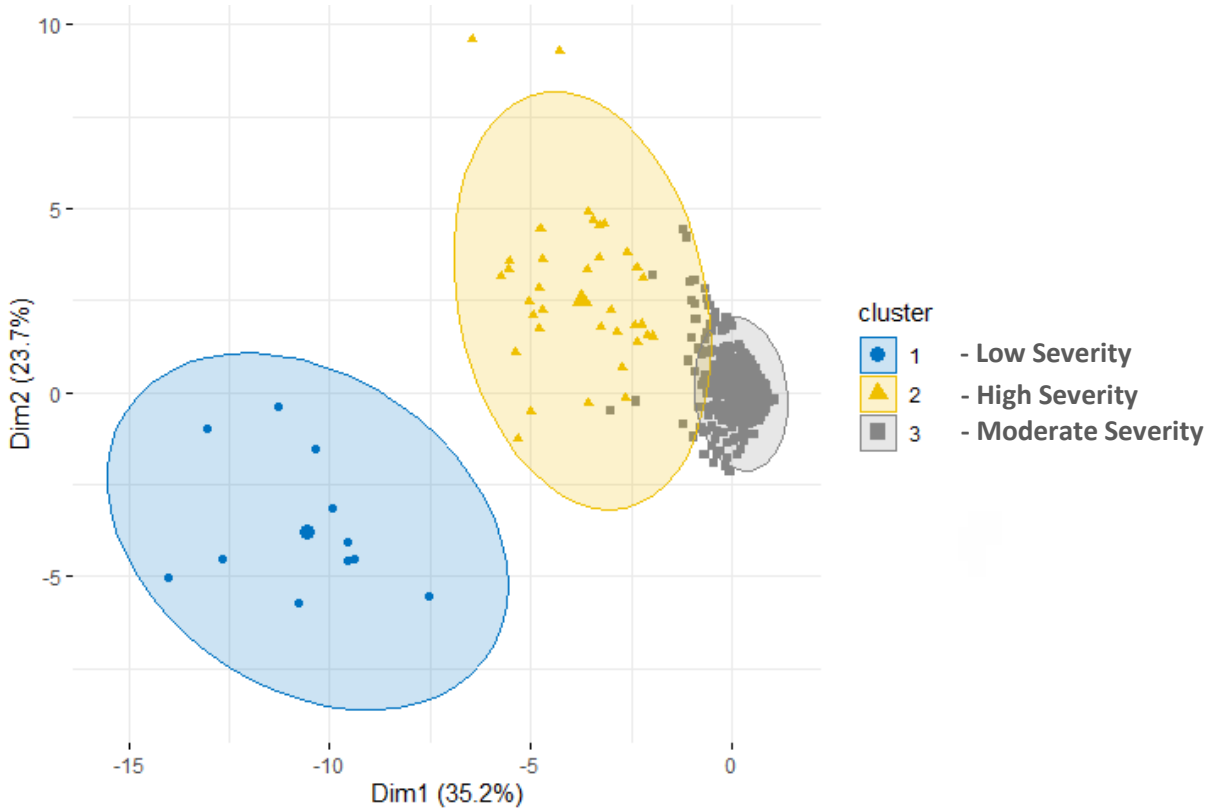
*Based on the ICD-10, CPT, and NDC codes

**It includes data for signs and symptoms from clinical notes – standard practice at Optum

***Customized NLP on clinical notes was performed for exposure to radiations and chemical

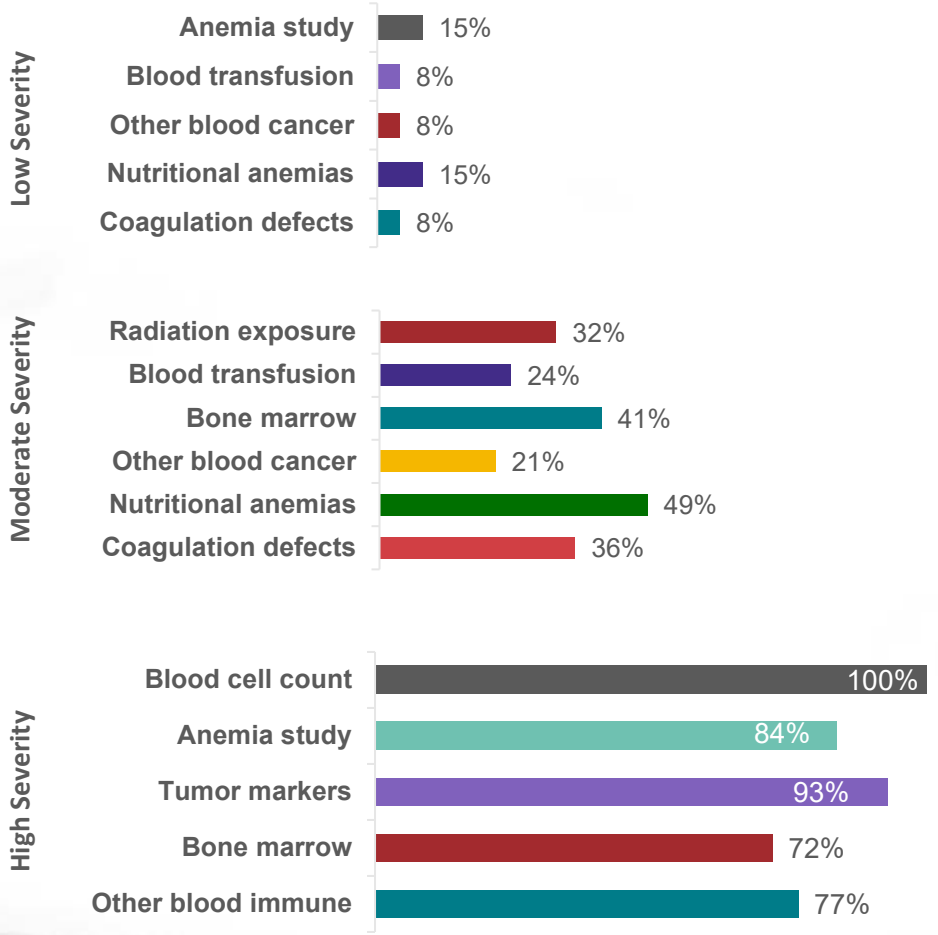
Segmentation of MDS patients based on pre-index risk factors

MDS patient classification



K-means clustering with MCA for segmentation of MDS patients

Percentage of patients with abnormal lab values



Presentation Flow

1	Introduction – about the speaker
2	Disease background and challenge
3	Proposed solution
4	**Conclusions**
5	Appendix

Conclusions

The research findings indicate that the terminology used in clinical notes presents a more precise indicator of prognostic symptoms than the structured data within Electronic Health Records (EHR). This suggests that physicians tend to document symptoms related to Myelodysplastic Syndromes (MDS) and other risk factors in a patient's medical records prior to the disease's onset.

When these predictive features are incorporated into the models, there is an enhancement in the models' accuracy. Our research findings corroborate this assertion.

The utilization of generative AI in the development of solutions has proven to be approximately ten times more efficient than manual efforts in the annotation of clinical notes.

Early prediction and segmentation of MDS patients offer significant benefits to payers, providers, and pharma

Payers

Allocate resources effectively, manage costs, and assess risk accurately by identifying high-risk patients and tailoring coverage

Providers

Develop personalized treatment plans, proactively manage patients at higher risk, and optimize resource utilization based on segmentation

Pharma

Target drug development to specific patient subgroups, design more effective clinical trials, and demonstrate value to support market access and reimbursement discussions

Thank You

VIKASH VERMA
Director – Data Science| Optum RX & LS l United Health Group
Vikash.Verma@Optum.com

pmsa
PHARMACEUTICAL MANAGEMENT SCIENCE ASSOCIATION

Presentation Flow

1	Introduction – about the speaker
2	Disease background and challenge
3	Proposed solution
4	Conclusions
5	**Appendix**

Risk factors that were used as predictor variable

Risk factors

Clinical presentation

Aplastic anemias
Blood cancer
Blood immune diseases
Circulatory system diseases
Coagulation defects
Digestive system diseases
Endocrine metabolic diseases
Genitourinary system diseases
Hemolytic anemias
Infections
Leukemia
Multiple myeloma
Musculoskeletal disorders
Neoplasm
Nervous system diseases
Neuro developmental disorders
Nutritional anemias
Radiation exposure
Respiratory system diseases
Secondary neoplasm

Procedures

Blood cell count
Kidney function
Anemia
Bone marrow
Biopsy
Blood glucose
Blood transfusion
Chemotherapy
Coagulation profile
Infection screening
Inflammatory markers
Liver function test
Lipid profile
Metabolic panel
Physical rehabilitation
Protein studies
Tumor markers
Urinalysis
Vitamin B
Vitamin D

Demography

Race ethnicity

Terms searched in the clinical notes

Exposure to different chemicals, pesticides, radioactive materials, and radiations.

1. Radiation Exposure
- Initial Encounter with Radiation
- Subsequent Encounter with Radiation
- Sequela from Radiation Encounter
- Exposure to Radioactive Isotopes
- Retained Radioactive Fragments

2. Chemical Exposure
- Exposure to Benzene
- Exposure to Pesticides
- Exposure to Other Industrial Chemicals

3. Radiation-Induced Conditions
- Acute Pulmonary Manifestations due to Radiation
- Chronic Pulmonary Manifestations due to Radiation
- Radiation-Induced Gastrointestinal Disorders (including Gastroenteritis, Colitis, Proctitis, Enteritis)
- Radiation-Induced Skin Changes (acute, chronic, unspecified)
- Radiation-Induced Disorders of the Skin and Subcutaneous Tissue
- Irradiation Cystitis
- Erectile Dysfunction Post-Radiation Therapy
- Radiation Sickness (unspecified, subsequent encounter, sequela)

4. Radiation Injuries
- Radiation Injury (unspecified)
- Radiation Cystitis
- Radiation Proctitis
- Radiation Dermatitis
- Radiation Fibrosis
- Radiation Necrosis
- Radiation Burn
- Radiation-Induced Organ Disorders (including Colitis, Esophagitis, Gastroenteritis, Pneumonitis)