Probabilistic Risk Prediction of Therapy-Related Myeloid Neoplasms_v0.6_draft

Leveraging Hematological Profiles for Early Detection in Oncology Survivors

MD Anderson - Leukemia Department

Zijun Wu

THE UNIVERSITY OF TEXAS
MD Anderson Cancer Center
Making Cancer History®

Agenda

Problem Statement

Data Profile

Data Processing

Data problems and difficulties & solutions

19 features introduction

General Tendency and Individual Images

Proposed Solution and Model Selection

Conclusion & Future Steps

Problem Statement

0.5%-1% cancer will become tMN patients, the longest latency could be 7 years.

Find out how to target these tMN patients before the diagnosis

Therapy related leukemia progresses rapidly, and patients may face a life-threatening situation within a month.

Devise a time-efficient prediction model capable of delivering predictive outcomes prior to diagnosis

tMN prediction is a race of time

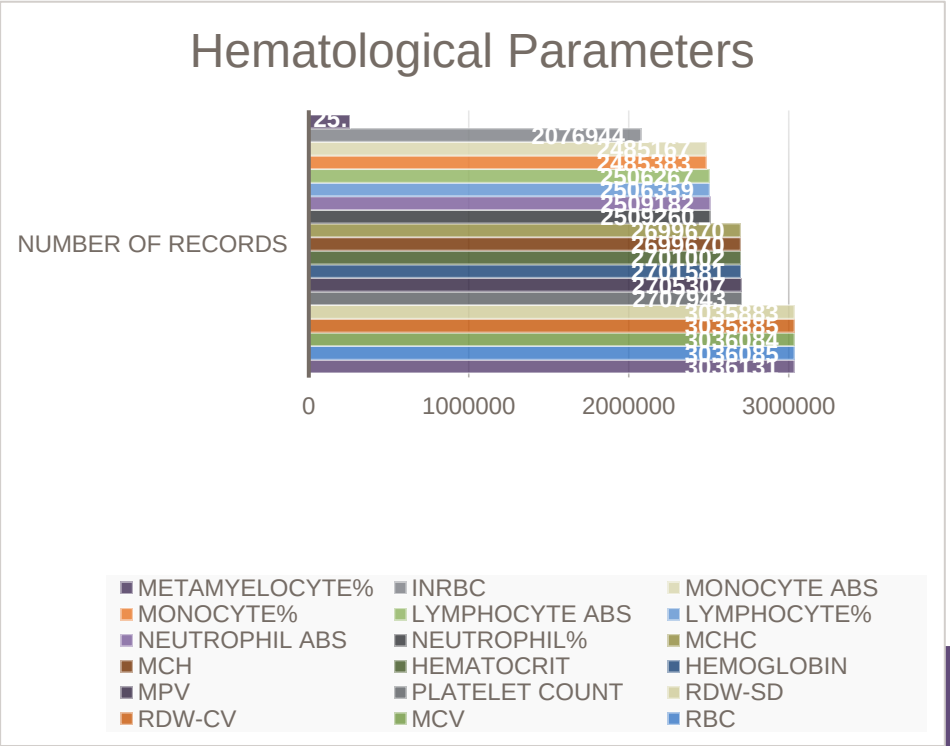
- tMN Is a race of time to predict the acute disease before
- As the potential patients approaching to the critical transition point, more signs are showing on their body, but they becoming more and more vulnable to the transition to tMNs.
- The primary objective of analyzing this dataset is to understand the hematological profiles of patients, particularly in relation to the diagnosis of tMN.
- This involves exploring trends, patterns, and potential indicators within the hematological parameters leading up to the diagnosis date.
- The lab data provides a high-level overview of the data, including patient information, hematological parameters, and temporal aspects, setting the stage for further exploration and analysis.

Data Profile

The dataset comprises 48,732,425 records capturing diverse hematological parameters across numerous patient encounters.

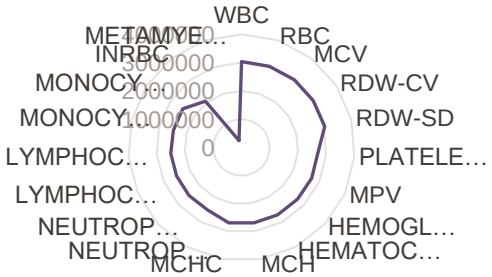
Each record provides insight into a patient's health status based on hematological measurements during a specific encounter.

Time-based information allows for the exploration of how hematological parameters relate to the diagnosis date and the time intervals leading up to the diagnosis.



Data Dimension	Descriptions
Data Variables	from 97,000 cancer patients, including 217 tMN patients
Data Size	5 million records after cleaning duplicates
Data Type	Time-based numerical data, categorical data
Target	19 categories hematological parameters

Number of Records



Pre-Data Processing

Data Problems:

1. No CHIP category, filtering out using indicators or other features

2. Don’t have much before and after chemotherapy indicator differences or comparison (lack of relevant key features)

3. Dates of each leukemia treatment

Multi-layer filtering to get tMN patients.

No classification of tMN patients. Filtering out based on conditional filtering on 514600 diagnosis records.

Filtering 217 out of 97,000 patients

Create relevant key features

Difficulties

Limited Temporal Data:

- Rapidly progressing diseases often have a short timeframe between diagnosis and critical events (e.g., death).
- This results in a limited amount of temporal data available for model training and challenges to capture the dynamics of disease progression adequately.

Imbalanced Data:

- The dataset may be imbalanced, with a small number of instances representing the critical events (e.g., death) compared to non-critical events.
- This imbalance can affect model performance, as the algorithm may be biased towards predicting the majority class.

Dynamic Nature of the Disease:

- Rapidly progressing diseases can exhibit dynamic and unpredictable patterns, making it difficult to model and predict the disease trajectory accurately.
- Changes in patient condition may occur rapidly, and capturing these nuances in the data is challenging.

Limited Feature Stability:

- Features that are relevant early in the disease progression might lose their significance as the disease advances.
- Identifying stable features that provide consistent predictive power throughout the progression becomes a challenge.

Raw Data

	Patient MRN	Encounter Epic CSN	Encounter Date	Component Common Name	Result Text Value	tMN_DiagnosisDate	YearsBeforeDiagnosis	SixMonthIntervalsBeforeDiagnosis	tMN_Patient	First Date of Chemo Administered
0	8668	1.026804e+09	2016-05-12 08:07:00	MCV	89.00	2023-09-12	-7	-14	0	NaN
1	8668	1.026804e+09	2016-05-12 08:07:00	RBC	4.12	2023-09-12	-7	-14	0	NaN
2	8668	1.026804e+09	2016-05-12 08:07:00	RDW-CV	13.30	2023-09-12	-7	-14	0	NaN
3	8668	1.026804e+09	2016-05-12 08:07:00	RDW-SD	43.10	2023-09-12	-7	-14	0	NaN
4	8668	1.026804e+09	2016-05-12 08:07:00	WBC	6.40	2023-09-12	-7	-14	0	NaN
...	...	...	...	...	...	...	...	...	...	...
48732421	2706623	1.110458e+09	2023-09-11 08:30:00	PLATELET COUNT	2.00	2023-07-18	0	0	1	2023/02/02 14:13:00
48732422	2706623	1.110672e+09	2023-09-15 06:45:00	HEMATOCRIT	27.10	2023-07-18	0	0	1	2023/02/02 14:13:00
48732423	2706623	1.110672e+09	2023-09-15 06:45:00	HEMOGLOBIN	9.00	2023-07-18	0	0	1	2023/02/02 14:13:00
48732424	2706623	1.110672e+09	2023-09-15 06:45:00	MPV	10.40	2023-07-18	0	0	1	2023/02/02 14:13:00
48732425	2706623	1.110672e+09	2023-09-15 06:45:00	PLATELET COUNT	24.00	2023-07-18	0	0	1	2023/02/02 14:13:00

48732426 rows × 10 columns

- Data quantity: 5 million records, cleaned and combined from 1 billion records
- Sample Quantity: 91,402 patients
- Time window: 2016 to 2023

Data Transform

Patient MRN	Encounter Date	tMN_Patient	HEMATOCRIT	HEMOGLOBIN	INRBC	LYMPHOCYTE %	LYMPHOCYTE ABS	MCH	MCHC	MCV	METAMYELOCYTE %	MONOCYTE %	MONOCYTE ABS	MPV	NEUTROPHIL %	NEUTROPHIL ABS	PLATELET COUNT	RBC	RDW-CV	RDW-SD	WBC
35938	2018-04-20 07:00:00	0	40.1	13.1	0.0	8.3	0.70	28.9	32.7	88.0	NaN	12.1	1.02	10.5	69.6	5.89	192.0	4.54	14.3	45.5	8.5
35938	2018-05-17 12:08:00	0	38.0	13.0	0.0	18.2	1.36	30.7	34.2	90.0	NaN	13.4	1.00	10.0	54.7	4.09	196.0	4.24	16.4	54.2	7.5
35938	2018-06-22 07:35:00	0	39.0	12.9	0.0	21.0	1.64	30.9	33.1	94.0	NaN	11.0	0.86	10.4	52.0	4.06	221.0	4.17	15.5	53.2	7.8
35938	2018-07-20 00:00:00	0	36.5	12.0	0.0	9.3	0.71	31.7	32.9	96.0	NaN	11.8	0.90	10.6	63.2	4.80	197.0	3.79	14.6	52.2	7.6
35938	2018-08-15 00:00:00	0	35.4	12.1	0.0	9.0	0.33	32.4	34.2	95.0	NaN	15.0	0.56	10.1	54.0	2.00	202.0	3.74	13.1	45.6	3.7
35938	2018-09-13 10:30:00	0	36.3	11.9	0.0	9.3	0.67	31.6	32.8	96.0	NaN	13.9	1.00	9.7	62.0	4.45	234.0	3.76	13.7	48.1	7.2
35938	2018-12-13 11:17:00	0	38.5	12.8	0.0	18.0	1.26	31.9	33.2	96.0	1.0	9.0	0.63	10.4	69.0	4.83	194.0	4.01	13.5	47.7	7.0
35938	2019-06-27 09:35:00	0	39.7	12.9	0.0	15.7	1.51	31.0	32.5	95.0	NaN	10.9	1.05	9.6	67.2	6.46	228.0	4.16	13.0	45.1	9.6
35938	2020-03-08 10:10:00	0	41.3	13.0	0.0	22.4	1.95	31.1	31.5	99.0	NaN	11.3	0.98	10.7	58.5	5.08	211.0	4.18	13.7	49.7	8.7

```
Number of null values in each column:
Patient MRN              0
Encounter Date           0
tMN_Patient              0
HEMATOCRIT          131265
HEMOGLOBIN          131036
INRBC               486455
LYMPHOCYTE %        160490
LYMPHOCYTE ABS      160626
MCH                 131482
MCHC                131482
MCV                  21540
METAMYELOCYTE %    1273941
MONOCYTE %          175332
MONOCYTE ABS        175571
MPV                 187679
NEUTROPHIL %        157898
NEUTROPHIL ABS      158022
PLATELET COUNT      130002
RBC                  21539
RDW-CV               26402
RDW-SD               26439
WBC                  21497
Birth Date              10
Sex                     10
Marital Status          61
```

- Transform into column values
- Add age and several indicators
- Null value cleaning

Data Cleaning

	Patient MRN	tMN_Patient	HEMATOCRIT	HEMOGLOBIN	INRBC	LYMPHOCYTE %	LYMPHOCYTE ABS	MCH	MCHC	MCV	...	NEUTROPHIL %	NEUTROPHIL ABS	PLATELET COUNT	RBC	RDW-CV	RDW-SD	WBC	Female	Married	Age_at_Encounter
0	35938	0	40.1	13.1	0.0	8.3	0.70	28.9	32.7	88.0	...	69.6	5.89	192.0	4.54	14.3	45.5	8.5	1	0	84.652407
1	35938	0	38.0	13.0	0.0	18.2	1.36	30.7	34.2	90.0	...	54.7	4.09	196.0	4.24	16.4	54.2	7.5	1	0	84.726915
2	35938	0	39.0	12.9	0.0	21.0	1.64	30.9	33.1	94.0	...	52.0	4.06	221.0	4.17	15.5	53.2	7.8	1	0	84.824958
3	35938	0	36.5	12.0	0.0	9.3	0.71	31.7	32.9	96.0	...	63.2	4.80	197.0	3.79	14.6	52.2	7.6	1	0	84.900753
4	35938	0	35.4	12.1	0.0	9.0	0.33	32.4	34.2	95.0	...	54.0	2.00	202.0	3.74	13.1	45.6	3.7	1	0	84.971937
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1427975	2773820	0	27.7	8.9	0.2	11.1	2.72	28.3	32.1	88.0	...	76.6	18.76	327.0	3.14	16.6	50.4	24.5	0	1	57.343243
1427979	2773829	0	29.9	10.0	0.0	29.0	1.19	29.4	33.4	88.0	...	54.0	2.21	358.0	3.40	16.7	47.4	4.1	0	1	63.371289
1427980	2773829	0	29.7	9.8	0.0	37.0	1.63	29.3	33.0	89.0	...	55.0	2.42	352.0	3.35	17.1	52.3	4.4	0	1	63.376572
1428009	2775104	0	31.4	10.2	0.0	6.0	0.05	28.2	32.5	87.0	...	92.0	0.74	36.0	3.62	14.5	44.0	0.8	0	1	59.705662
1428010	2775104	0	23.7	8.2	0.6	4.0	0.19	28.5	34.6	82.0	...	89.0	4.27	64.0	2.88	15.0	41.8	4.8	0	1	59.724760

1178603 rows × 23 columns

- Broken records
- Null values
- Falsely calculated results

Sample Patient Record - MRN xxxxx

	Patient MRN	Encounter Date	Component Common Name	Result Text Value	tMN_Patient
470	35938	2018-04-20 07:00:00	MCV	88.00	0
471	35938	2018-04-20 07:00:00	RBC	4.54	0
472	35938	2018-04-20 07:00:00	RDW-CV	14.30	0
473	35938	2018-04-20 07:00:00	RDW-SD	45.50	0
474	35938	2018-04-20 07:00:00	WBC	8.50	0
475	35938	2018-05-17 12:08:00	MCV	90.00	0
476	35938	2018-05-17 12:08:00	RBC	4.24	0
477	35938	2018-05-17 12:08:00	RDW-CV	16.40	0
478	35938	2018-05-17 12:08:00	RDW-SD	54.20	0
479	35938	2018-05-17 12:08:00	WBC	7.50	0
480	35938	2018-06-22 07:35:00	MCV	94.00	0
481	35938	2018-06-22 07:35:00	RBC	4.17	0
482	35938	2018-06-22 07:35:00	RDW-CV	15.50	0
483	35938	2018-06-22 07:35:00	RDW-SD	53.20	0
484	35938	2018-06-22 07:35:00	WBC	7.80	0
485	35938	2018-07-20 00:00:00	MCV	96.00	0
486	35938	2018-07-20 00:00:00	RBC	3.79	0
487	35938	2018-07-20 00:00:00	RDW-CV	14.60	0
488	35938	2018-07-20 00:00:00	RDW-SD	52.20	0
489	35938	2018-07-20 00:00:00	WBC	7.60	0
490	35938	2018-08-15 00:00:00	MCV	95.00	0
491	35938	2018-08-15 00:00:00	RBC	3.74	0
492	35938	2018-08-15 00:00:00	RDW-CV	13.10	0
493	35938	2018-08-15 00:00:00	RDW-SD	45.60	0
494	35938	2018-08-15 00:00:00	WBC	3.70	0
495	35938	2018-09-13 10:30:00	MCV	96.00	0
496	35938	2018-09-13 10:30:00	RBC	3.76	0
497	35938	2018-09-13 10:30:00	RDW-CV	13.70	0
498	35938	2018-09-13 10:30:00	RDW-SD	48.10	0
499	35938	2018-09-13 10:30:00	WBC	7.20	0
500	35938	2018-12-13 11:17:00	MCV	96.00	0
501	35938	2018-12-13 11:17:00	RBC	4.01	0
502	35938	2018-12-13 11:17:00	RDW-CV	13.50	0
503	35938	2018-12-13 11:17:00	RDW-SD	47.70	0
504	35938	2018-12-13 11:17:00	WBC	7.00	0
505	35938	2019-06-27 09:35:00	MCV	95.00	0
506	35938	2019-06-27 09:35:00	RBC	4.16	0
507	35938	2019-06-27 09:35:00	RDW-CV	13.00	0
508	35938	2019-06-27 09:35:00	RDW-SD	45.10	0
509	35938	2019-06-27 09:35:00	WBC	9.60	0
510	35938	2020-03-08 10:10:00	MCV	99.00	0
511	35938	2020-03-08 10:10:00	RBC	4.18	0
512	35938	2020-03-08 10:10:00	RDW-CV	13.70	0
513	35938	2020-03-08 10:10:00	RDW-SD	49.70	0
514	35938	2020-03-08 10:10:00	WBC	8.70	0
15180478	35938	2018-04-20 07:00:00	INRBC	0.00	0
15180479	35938	2018-04-20 07:00:00	MCH	28.90	0
15180480	35938	2018-04-20 07:00:00	MCHC	32.70	0
15180481	35938	2018-04-20 07:00:00	NEUTROPHIL %	69.60	0
15180482	35938	2018-04-20 07:00:00	NEUTROPHIL ABS	5.89	0
15180483	35938	2018-05-17 12:08:00	INRBC	0.00	0
15180484	35938	2018-05-17 12:08:00	MCH	30.70	0
15180485	35938	2018-05-17 12:08:00	MCHC	34.20	0
15180486	35938	2018-05-17 12:08:00	NEUTROPHIL %	54.70	0
15180487	35938	2018-05-17 12:08:00	NEUTROPHIL ABS	4.09	0
15180488	35938	2018-06-22 07:35:00	INRBC	0.00	0
15180489	35938	2018-06-22 07:35:00	MCH	30.90	0
15180490	35938	2018-06-22 07:35:00	MCHC	33.10	0
15180491	35938	2018-06-22 07:35:00	NEUTROPHIL %	52.00	0
15180492	35938	2018-06-22 07:35:00	NEUTROPHIL ABS	4.06	0
15180493	35938	2018-07-20 00:00:00	INRBC	0.00	0
15180494	35938	2018-07-20 00:00:00	MCH	31.70	0
15180495	35938	2018-07-20 00:00:00	MCHC	32.90	0
15180496	35938	2018-07-20 00:00:00	NEUTROPHIL %	63.20	0
15180497	35938	2018-07-20 00:00:00	NEUTROPHIL ABS	4.80	0
15180498	35938	2018-08-15 00:00:00	INRBC	0.00	0
15180499	35938	2018-08-15 00:00:00	MCH	32.40	0
15180500	35938	2018-08-15 00:00:00	MCHC	34.20	0
15180501	35938	2018-08-15 00:00:00	NEUTROPHIL %	54.00	0
15180502	35938	2018-08-15 00:00:00	NEUTROPHIL ABS	2.00	0
15180503	35938	2018-09-13 10:30:00	INRBC	0.00	0
15180504	35938	2018-09-13 10:30:00	MCH	31.60	0
15180505	35938	2018-09-13 10:30:00	MCHC	32.80	0
15180506	35938	2018-09-13 10:30:00	NEUTROPHIL %	62.00	0
15180507	35938	2018-09-13 10:30:00	NEUTROPHIL ABS	4.45	0
15180508	35938	2018-12-13 11:17:00	INRBC	0.00	0
15180509	35938	2018-12-13 11:17:00	MCH	31.90	0
15180510	35938	2018-12-13 11:17:00	MCHC	33.20	0
15180511	35938	2018-12-13 11:17:00	NEUTROPHIL %	69.00	0
15180512	35938	2018-12-13 11:17:00	NEUTROPHIL ABS	4.83	0
15180513	35938	2019-06-27 09:35:00	INRBC	0.00	0
15180514	35938	2019-06-27 09:35:00	MCH	31.00	0
15180515	35938	2019-06-27 09:35:00	MCHC	32.50	0
15180516	35938	2019-06-27 09:35:00	NEUTROPHIL %	67.20	0
15180517	35938	2019-06-27 09:35:00	NEUTROPHIL ABS	6.46	0
15180518	35938	2020-03-08 10:10:00	INRBC	0.00	0
15180519	35938	2020-03-08 10:10:00	MCH	31.10	0
15180520	35938	2020-03-08 10:10:00	MCHC	31.50	0
15180521	35938	2020-03-08 10:10:00	NEUTROPHIL %	58.50	0
15180522	35938	2020-03-08 10:10:00	NEUTROPHIL ABS	5.08	0
27675126	35938	2018-04-20 07:00:00	LYMPHOCYTE %	8.30	0
27675127	35938	2018-04-20 07:00:00	LYMPHOCYTE ABS	0.70	0
27675128	35938	2018-04-20 07:00:00	MONOCYTE %	12.10	0
27675129	35938	2018-04-20 07:00:00	MONOCYTE ABS	1.02	0
27675130	35938	2018-05-17 12:08:00	LYMPHOCYTE %	18.20	0
27675131	35938	2018-05-17 12:08:00	LYMPHOCYTE ABS	1.36	0
27675132	35938	2018-05-17 12:08:00	MONOCYTE %	13.40	0
27675133	35938	2018-05-17 12:08:00	MONOCYTE ABS	1.00	0

Sample Patient Record - MRN xxxxx

- Each date and time specific records
- Hematological test classification
- Test result values – Average repetition values

27675134	35938	2018-06-22 07:35:00	LYMPHOCYTE %	21.00	0
27675135	35938	2018-06-22 07:35:00	LYMPHOCYTE ABS	1.64	0
27675136	35938	2018-06-22 07:35:00	MONOCYTE %	11.00	0
27675137	35938	2018-06-22 07:35:00	MONOCYTE ABS	0.86	0
27675138	35938	2018-07-20 00:00:00	LYMPHOCYTE %	9.30	0
27675139	35938	2018-07-20 00:00:00	LYMPHOCYTE ABS	0.71	0
27675140	35938	2018-07-20 00:00:00	MONOCYTE %	11.80	0
27675141	35938	2018-07-20 00:00:00	MONOCYTE ABS	0.90	0
27675142	35938	2018-08-15 00:00:00	LYMPHOCYTE %	9.00	0
27675143	35938	2018-08-15 00:00:00	LYMPHOCYTE ABS	0.33	0
27675144	35938	2018-08-15 00:00:00	MONOCYTE %	15.00	0
27675145	35938	2018-08-15 00:00:00	MONOCYTE ABS	0.56	0
27675146	35938	2018-09-13 10:30:00	LYMPHOCYTE %	9.30	0
27675147	35938	2018-09-13 10:30:00	LYMPHOCYTE ABS	0.67	0
27675148	35938	2018-09-13 10:30:00	MONOCYTE %	13.90	0
27675149	35938	2018-09-13 10:30:00	MONOCYTE ABS	1.00	0
27675150	35938	2018-12-13 11:17:00	LYMPHOCYTE %	18.00	0
27675151	35938	2018-12-13 11:17:00	LYMPHOCYTE ABS	1.26	0
27675152	35938	2018-12-13 11:17:00	METAMYELOCYTE %	1.00	0
27675153	35938	2018-12-13 11:17:00	MONOCYTE %	9.00	0
27675154	35938	2018-12-13 11:17:00	MONOCYTE ABS	0.63	0
27675155	35938	2019-06-27 09:35:00	LYMPHOCYTE %	15.70	0
27675156	35938	2019-06-27 09:35:00	LYMPHOCYTE ABS	1.51	0
27675157	35938	2019-06-27 09:35:00	MONOCYTE %	10.90	0
27675158	35938	2019-06-27 09:35:00	MONOCYTE ABS	1.05	0
27675159	35938	2020-03-08 10:10:00	LYMPHOCYTE %	22.40	0
27675160	35938	2020-03-08 10:10:00	LYMPHOCYTE ABS	1.95	0
27675161	35938	2020-03-08 10:10:00	MONOCYTE %	11.30	0
27675162	35938	2020-03-08 10:10:00	MONOCYTE ABS	0.98	0
37916933	35938	2018-04-20 07:00:00	HEMATOCRIT	40.10	0
37916934	35938	2018-04-20 07:00:00	HEMOGLOBIN	13.10	0
37916935	35938	2018-04-20 07:00:00	MPV	10.50	0
37916936	35938	2018-04-20 07:00:00	PLATELET COUNT	192.00	0

37916937	35938	2018-05-17 12:08:00	HEMATOCRIT	38.00	0
37916938	35938	2018-05-17 12:08:00	HEMOGLOBIN	13.00	0
37916939	35938	2018-05-17 12:08:00	MPV	10.00	0
37916940	35938	2018-05-17 12:08:00	PLATELET COUNT	196.00	0
37916941	35938	2018-06-22 07:35:00	HEMATOCRIT	39.00	0
37916942	35938	2018-06-22 07:35:00	HEMOGLOBIN	12.90	0
37916943	35938	2018-06-22 07:35:00	MPV	10.40	0
37916944	35938	2018-06-22 07:35:00	PLATELET COUNT	221.00	0
37916945	35938	2018-07-20 00:00:00	HEMATOCRIT	36.50	0
37916946	35938	2018-07-20 00:00:00	HEMOGLOBIN	12.00	0
37916947	35938	2018-07-20 00:00:00	MPV	10.60	0
37916948	35938	2018-07-20 00:00:00	PLATELET COUNT	197.00	0
37916949	35938	2018-08-15 00:00:00	HEMATOCRIT	35.40	0
37916950	35938	2018-08-15 00:00:00	HEMOGLOBIN	12.10	0
37916951	35938	2018-08-15 00:00:00	MPV	10.10	0
37916952	35938	2018-08-15 00:00:00	PLATELET COUNT	202.00	0
37916953	35938	2018-09-13 10:30:00	HEMATOCRIT	36.30	0
37916954	35938	2018-09-13 10:30:00	HEMOGLOBIN	11.90	0
37916955	35938	2018-09-13 10:30:00	MPV	9.70	0
37916956	35938	2018-09-13 10:30:00	PLATELET COUNT	234.00	0
37916957	35938	2018-12-13 11:17:00	HEMATOCRIT	38.50	0
37916958	35938	2018-12-13 11:17:00	HEMOGLOBIN	12.80	0
37916959	35938	2018-12-13 11:17:00	MPV	10.40	0
37916960	35938	2018-12-13 11:17:00	PLATELET COUNT	194.00	0
37916961	35938	2019-06-27 09:35:00	HEMATOCRIT	39.70	0
37916962	35938	2019-06-27 09:35:00	HEMOGLOBIN	12.90	0
37916963	35938	2019-06-27 09:35:00	MPV	9.60	0
37916964	35938	2019-06-27 09:35:00	PLATELET COUNT	228.00	0
37916965	35938	2020-03-08 10:10:00	HEMATOCRIT	41.30	0
37916966	35938	2020-03-08 10:10:00	HEMOGLOBIN	13.00	0
37916967	35938	2020-03-08 10:10:00	MPV	10.70	0
37916968	35938	2020-03-08 10:10:00	PLATELET COUNT	211.00	0

Detail Data Profile

- Lab test categories are mainly numerical after transformations.
- Age is detailed with decimals to capture changes within one year, resulting in the highest number of unique values.
- HGB has highest SD
- Top frequency in lab test category item is INRBC for having 75,8234 values for having most 0 values. Second highest is MPV, we expect a lot of repetition values.

	Data Type	Nulls	Zeros	Min	Median	Max	Mean	Standard Deviation	Unique	Top Frequency
Patient MRN	int64	0	0	35938	2298881	2775104	2007516.31	696148.33	62571	359
tMN_Patient	int64	0	0	0	0	1	0.0068	0.082	2	1170607
HEMATOCRIT	float64	0	0	5.70	35.50	70.80	35.22	6.00	1265	7771
HEMOGLOBIN	float64	0	0	2.50	11.60	23.20	11.55	2.052	671	21489
INRBC	float64	336886	0						617	758234
LYMPHOCYTE %	float64	0	0	0	21.90	99.30	23.97	14.14	1783	10121
LYMPHOCYTE ABS	float64	0	0	0	1.080	429.66	1.32	2.83	3836	8041
MCH	float64	0	0	15.30	30.50	124.20	30.53	2.87	899	19633
MCHC	float64	0	0	22.80	32.80	59.80	32.79	1.32	604	41216
MCV	float64	0	0	9	93	137	93.073	7.47	359	77638
MONOCYTE %	float64	0	0	0	9.90	97	10.62	5.38	956	24562
MONOCYTE ABS	float64	0	0	0	0.52	134.46	0.59	0.50	1738	19862
MPV	float64	0	0	6.80	9.80	18.20	9.96	1.030	469	50358
NEUTROPHIL %	float64	0	0	0	62.50	100	60.84	15.71	1902	8198
NEUTROPHIL ABS	float64	0	0	0	3.16	229.44	3.90	3.93	6440	3102
PLATELET COUNT	float64	0	0	0	192	4477	199.16	106.11	1996	5514
RBC	float64	0	0	0.53	3.83	8.46	3.81	0.73	1976	6053
RDW-CV	float64	0	0	9.80	14.70	36.60	15.36	2.59	1149	35869
RDW-SD	float64	0	0	28.60	49.70	131.60	51.63	9.35	2227	10266
WBC	float64	0	0	0.30	5.20	615.70	6.11	5.65	2385	20243
Female	int64	0	0	0	0	1	0.50	0.50	2	593320
Married	int64	0	0	0	1	1	0.70	0.46	2	830776
Age_at_Encounter	float64	0	0	0.15	62.63	102.62	59.76	15.63	688197	54

Scaling

	Patient MRN	tMN_Patient	HEMATOCRIT	HEMOGLOBIN	INRBC	LYMPHOCYTE %	LYMPHOCYTE ABS	MCH	MCHC	MCV	...	NEUTROPHIL %	NEUTROPHIL ABS	PLATELET COUNT	RBC	RDW-CV	RDW-SD	WBC	Female	Married	Age_at_Encounter
0	35938	0	0.812973	0.754167	0.0	-1.108270	-0.217688	-0.568844	-0.065099	-0.679162	...	0.557265	0.504992	-0.067447	0.991351	-0.408402	-0.655118	0.423513	1	0	1.592354
1	35938	0	0.462865	0.705433	0.0	-0.408308	0.015196	0.059229	1.073510	-0.411428	...	-0.390876	0.047193	-0.029749	0.582710	0.401727	0.275269	0.246409	1	0	1.597121
2	35938	0	0.629583	0.656700	0.0	-0.210339	0.113996	0.129015	0.238530	0.124039	...	-0.562686	0.039563	0.205866	0.487360	0.054529	0.168328	0.299540	1	0	1.603394
3	35938	0	0.212788	0.218099	0.0	-1.037567	-0.214160	0.408158	0.086715	0.391773	...	0.150010	0.227769	-0.020324	-0.030252	-0.292669	0.061387	0.264119	1	0	1.608243
4	35938	0	0.029398	0.266833	0.0	-1.058778	-0.348245	0.652409	1.073510	0.257906	...	-0.435419	-0.484363	0.026799	-0.098359	-0.871333	-0.644424	-0.426586	1	0	1.612798
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1427975	2773820	0	-1.254331	-1.292636	0.2	-0.910301	0.495080	-0.778202	-0.520543	-0.679162	...	1.002700	3.778255	1.204875	-0.915641	0.478882	-0.131107	3.257177	0	1	-0.154943
1427979	2773829	0	-0.887551	-0.756569	0.0	0.355287	-0.044789	-0.394379	0.466252	-0.679162	...	-0.435419	-0.430953	1.497038	-0.561485	0.517460	-0.451930	-0.355745	0	1	0.230744
1427980	2773829	0	-0.920895	-0.854035	0.0	0.920912	0.110467	-0.429272	0.162623	-0.545295	...	-0.371785	-0.377543	1.440491	-0.629592	0.671770	0.072081	-0.302613	0	1	0.231082
1428009	2775104	0	-0.637474	-0.659102	0.0	-1.270888	-0.447044	-0.813095	-0.216914	-0.813029	...	1.982658	-0.804822	-1.537687	-0.261815	-0.331247	-0.815529	-0.940188	0	1	-0.003791
1428010	2775104	0	-1.921203	-1.633770	0.6	-1.412294	-0.397644	-0.708416	1.377140	-1.482363	...	1.791757	0.092973	-1.273797	-1.269797	-0.138359	-1.050799	-0.231772	0	1	-0.002569

tMN_Patient	1.000000
MCV -1	0.075232
MCH -1	0.067886
MPV -1	0.057469
Age_at_Encounter	0.042973
RDW-SD -1	0.042081
NEUTROPHIL % -1	0.014081
Married	0.013559
RDW-CV -1	0.010632
MCHC -1	0.005398
MONOCYTE % -1	0.003502
NEUTROPHIL ABS -1	-0.006886
LYMPHOCYTE ABS -1	-0.010224
MONOCYTE ABS -1	-0.011115
WBC -1	-0.011775
Female	-0.014026
LYMPHOCYTE % -1	-0.014747
HEMOGLOBIN -1	-0.020647
HEMATOCRIT -1	-0.022401
PLATELET COUNT -1	-0.041275
RBC -1	-0.047014

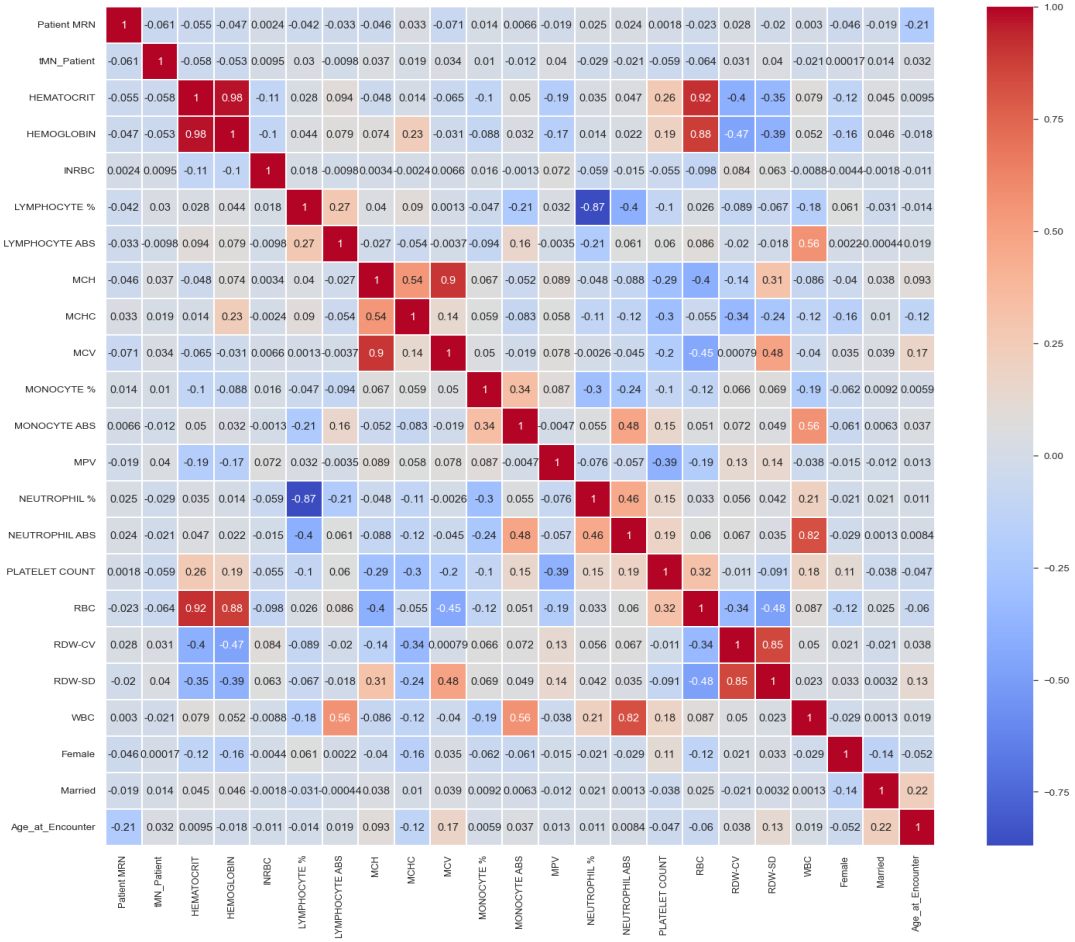
Correlation Matrix

Relationship Strength and Directions:

- ➢ HCT, HGB, PLT, RDW-CV

Multicollinearity:

- Positive correlation (closer to 1):
  - ➢ WBC & NEUT ABS
  - ➢ RDW-SD
  - ➢ MCV & MCH
  - ➢ HCT & RBC & HGB
- Negative correlation (closer to -1):
  - ➢ NEUT% & LYMPH%



12.18

	tMN_Patient	HEMATOCRIT -1	HEMOGLOBIN -1	LYMPHOCYTE % -1	LYMPHOCYTE ABS -1	MCH -1	MCHC -1	MCV -1	MONOCYTE % -1	MONOCYTE ABS -1	MPV -1	NEUTROPHIL % -1	NEUTROPHIL ABS -1	PLATELET COUNT -1	RBC -1	RDW-CV -1	RDW-SD -1	WBC -1	Married	Female	Age_at_Encounter
tMN_Patient	1.000000	-0.022401	-0.020647	-0.014747	-0.010224	0.067886	0.005398	0.075232	0.003502	-0.011115	0.057469	0.014081	-0.006886	-0.041275	-0.047014	0.010632	0.042081	-0.011775	0.013559	-0.014026	0.042973
HEMATOCRIT -1	-0.022401	1.000000	0.978345	0.015161	0.152595	-0.037471	-0.073034	-0.011512	-0.106707	0.078067	-0.241554	0.059564	0.101237	0.384050	0.918723	-0.331921	-0.246108	0.144175	0.009961	-0.068738	-0.001753
HEMOGLOBIN -1	-0.020647	0.978345	1.000000	0.027905	0.138916	0.063697	0.130307	0.007597	-0.095276	0.060965	-0.226003	0.043286	0.078094	0.322383	0.888242	-0.391151	-0.290946	0.118314	0.010131	-0.095801	-0.026479
LYMPHOCYTE % -1	-0.014747	0.015161	0.027905	1.000000	0.275564	0.010142	0.072534	-0.023260	-0.068093	-0.225525	0.009289	-0.853475	-0.384195	-0.084046	0.022904	-0.114407	-0.091742	-0.233895	-0.036292	0.080813	0.047056
LYMPHOCYTE ABS -1	-0.010224	0.152595	0.138916	0.275564	1.000000	-0.025281	-0.061169	-0.000587	-0.094940	0.159378	-0.028497	-0.202916	0.085878	0.113384	0.141688	-0.032195	-0.019513	0.479400	-0.016013	0.015175	0.022546
MCH -1	0.067886	-0.037471	0.063697	0.010142	-0.025281	1.000000	0.476501	0.908991	0.046368	-0.049489	0.099036	-0.002634	-0.070697	-0.235014	-0.389445	-0.075257	0.360378	-0.076534	0.057823	-0.018224	0.121925
MCHC -1	0.005398	-0.073034	0.130307	0.072534	-0.061169	0.476501	1.000000	0.070687	0.055532	-0.087171	0.077376	-0.087799	-0.118277	-0.301093	-0.110125	-0.309188	-0.244932	-0.129074	-0.002194	-0.126550	-0.122705
MCV -1	0.075232	-0.011512	0.007597	-0.023260	-0.000587	0.908991	0.070687	1.000000	0.028308	-0.015184	0.079011	0.037497	-0.025262	-0.130088	-0.394292	0.055684	0.521776	-0.026838	0.065302	0.038947	0.196669
MONOCYTE % -1	0.003502	-0.106707	-0.095276	-0.068093	-0.094940	0.046368	0.055532	0.028308	1.000000	0.365778	0.104116	-0.320127	-0.224805	-0.118083	-0.112142	0.048395	0.034173	-0.176513	0.014263	-0.051616	0.008058
MONOCYTE ABS -1	-0.011115	0.078067	0.060965	-0.225525	0.159378	-0.049489	-0.087171	-0.015184	0.365778	1.000000	-0.015058	0.043977	0.436720	0.149558	0.076228	0.059069	0.036435	0.556816	0.020682	-0.053155	0.007040
MPV -1	0.057469	-0.241554	-0.226003	0.009289	-0.028497	0.099036	0.077376	0.079011	0.104116	-0.015058	1.000000	-0.071655	-0.073815	-0.409369	-0.251100	0.131174	0.132554	-0.067728	-0.001837	-0.018314	0.004027
NEUTROPHIL % -1	0.014081	0.059564	0.043286	-0.853475	-0.202916	-0.002634	-0.087799	0.037497	-0.320127	0.043977	-0.071655	1.000000	0.450748	0.159835	0.043128	0.081392	0.081156	0.264516	0.012515	-0.053727	-0.056778
NEUTROPHIL ABS -1	-0.006886	0.101237	0.078094	-0.384195	0.085878	-0.070697	-0.118277	-0.025262	-0.224805	0.436720	-0.073815	0.450748	1.000000	0.211342	0.103173	0.064609	0.044333	0.895207	0.011238	-0.031274	-0.033764
PLATELET COUNT -1	-0.041275	0.384050	0.322383	-0.084046	0.113384	-0.235014	-0.301093	-0.130088	-0.118083	0.149558	-0.409369	0.159835	0.211342	1.000000	0.409835	-0.021512	-0.061628	0.223954	-0.036519	0.114063	-0.065810
RBC -1	-0.047014	0.918723	0.888242	0.022904	0.141688	-0.389445	-0.110125	-0.394292	-0.112142	0.076228	-0.251100	0.043128	0.103173	0.409835	1.000000	-0.304717	-0.408004	0.143263	-0.015212	-0.077959	-0.079376
RDW-CV -1	0.010632	-0.331921	-0.391151	-0.114407	-0.032195	-0.075257	-0.309188	0.055684	0.048395	0.059069	0.131174	0.081392	0.064609	-0.021512	-0.304717	1.000000	0.850903	0.048986	0.003296	-0.006288	0.011191
RDW-SD -1	0.042081	-0.246108	-0.290946	-0.091742	-0.019513	0.360378	-0.244932	0.521776	0.034173	0.036435	0.132554	0.081156	0.044333	-0.061628	-0.408004	0.850903	1.000000	0.031566	0.034877	0.013036	0.116127
WBC -1	-0.011775	0.144175	0.118314	-0.233895	0.479400	-0.076534	-0.129074	-0.026838	-0.176513	0.556816	-0.067728	0.264516	0.895207	0.223954	0.143263	0.048986	0.031566	1.000000	0.007972	-0.026583	-0.017254
Married	0.013559	0.009961	0.010131	-0.036292	-0.016013	0.057823	-0.002194	0.065302	0.014263	0.020682	-0.001837	0.012515	0.011238	-0.036519	-0.015212	0.003296	0.034877	0.007972	1.000000	-0.134824	0.236318
Female	-0.014026	-0.068738	-0.095801	0.080813	0.015175	-0.018224	-0.126550	0.038947	-0.051616	-0.053155	-0.018314	-0.053727	-0.031274	0.114063	-0.077959	-0.006288	0.013036	-0.026583	-0.134824	1.000000	-0.028861
Age_at_Encounter	0.042973	-0.001753	-0.026479	0.047056	0.022546	0.121925	-0.122705	0.196669	0.008058	0.007040	0.004027	-0.056778	-0.033764	-0.065810	-0.079376	0.011191	0.116127	-0.017254	0.236318	-0.028861	1.000000

12.18

	precision	recall	f1-score	support
0	1.00	0.76	0.86	4708734
1	0.02	0.65	0.05	43259
accuracy			0.76	4751993
macro avg	0.51	0.70	0.45	4751993
weighted avg	0.99	0.76	0.85	4751993

	precision	recall	f1-score	support
0	0.68	0.75	0.72	43439
1	0.72	0.65	0.68	43385
accuracy			0.70	86824
macro avg	0.70	0.70	0.70	86824
weighted avg	0.70	0.70	0.70	86824

12.18

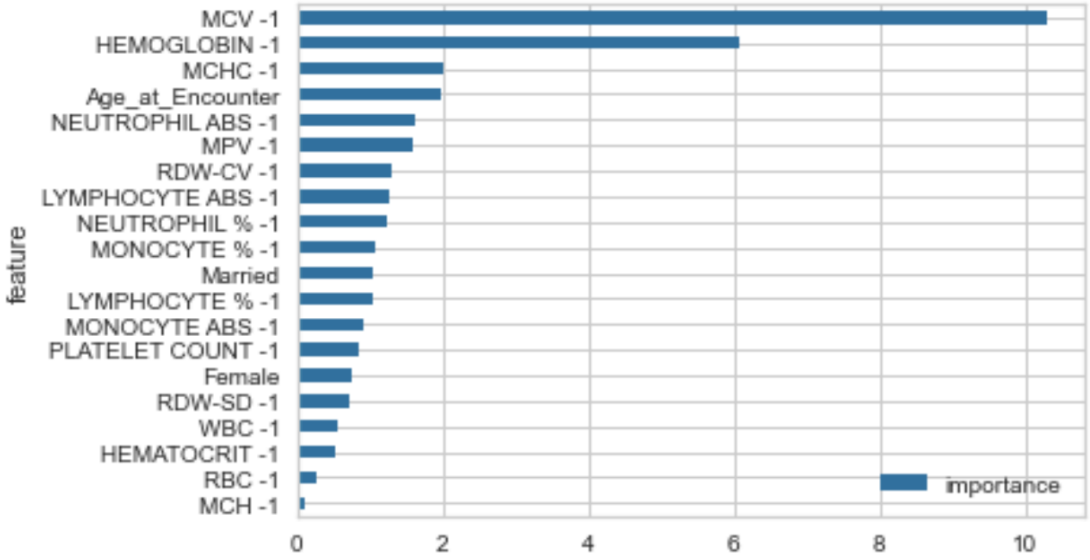
GLM MODEL

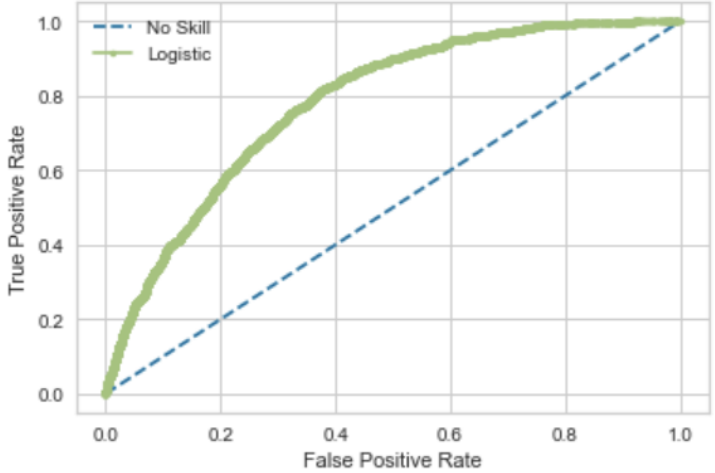
```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:             tMNPatient   No. Observations:               434116
Model:                            GLM   Df Residuals:                   434095
Model Family:                Binomial   Df Model:                           20
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:            -2.4025e+05
Date:                Tue, 12 Dec 2023   Deviance:                   4.8049e+05
Time:                        12:55:01   Pearson chi2:                 4.31e+05
No. Iterations:                     5   Pseudo R-squ. (CS):             0.2438
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
Intercept            -0.6022      0.009    -69.155      0.000      -0.619      -0.585
HEMATOCRIT1          -1.2158      0.119    -10.224      0.000      -1.449      -0.983
HEMOGLOBIN1           2.3557      0.112     21.025      0.000       2.136       2.575
LYMPHOCYTEPCT1        0.0457      0.015      2.952      0.003       0.015       0.076
LYMPHOCYTEABS1        0.2163      0.028      7.784      0.000       0.162       0.271
MCH1                 -2.2633      0.075    -30.274      0.000      -2.410      -2.117
MCHC1                 0.6039      0.038     16.031      0.000       0.530       0.678
MCV1                  2.3362      0.068     34.289      0.000       2.203       2.470
MONOCYTEPCT1          0.0796      0.009      9.341      0.000       0.063       0.096
MONOCYTEABS1         -0.1063      0.013     -8.317      0.000      -0.131      -0.081
MPV1                  0.4637      0.004    113.775      0.000       0.456       0.472
NEUTROPHILPCT1        0.2174      0.017     12.761      0.000       0.184       0.251
NEUTROPHILABS1        0.4609      0.056      8.158      0.000       0.350       0.572
PLATELETCOUNT1       -0.1591      0.005    -29.207      0.000      -0.170      -0.148
RBC1                 -1.3680      0.054    -25.311      0.000      -1.474      -1.262
RDWCV1                0.2414      0.016     15.569      0.000       0.211       0.272
RDWSD1               -0.3271      0.017    -18.693      0.000      -0.361      -0.293
WBC1                 -0.5287      0.070     -7.509      0.000      -0.667      -0.391
Married               0.0491      0.008      5.829      0.000       0.033       0.066
Female               -0.2587      0.008    -34.315      0.000      -0.273      -0.244
AgeatEncounter        0.6762      0.005    131.054      0.000       0.666       0.686
=====================================================================================

Model AIC: 480536.6900461577
```

12.18 Logist

	precision	recall	f1-score	support
0	0.68	0.75	0.72	43439
1	0.72	0.65	0.68	43385
accuracy			0.70	86824
macro avg	0.70	0.70	0.70	86824
weighted avg	0.70	0.70	0.70	86824

	precision	recall	f1-score	support
0	0.72	0.69	0.70	43439
1	0.70	0.73	0.72	43385
accuracy			0.71	86824
macro avg	0.71	0.71	0.71	86824
weighted avg	0.71	0.71	0.71	86824





12.18 Comparison

All - Correlation Matrix

Multicollinearity:

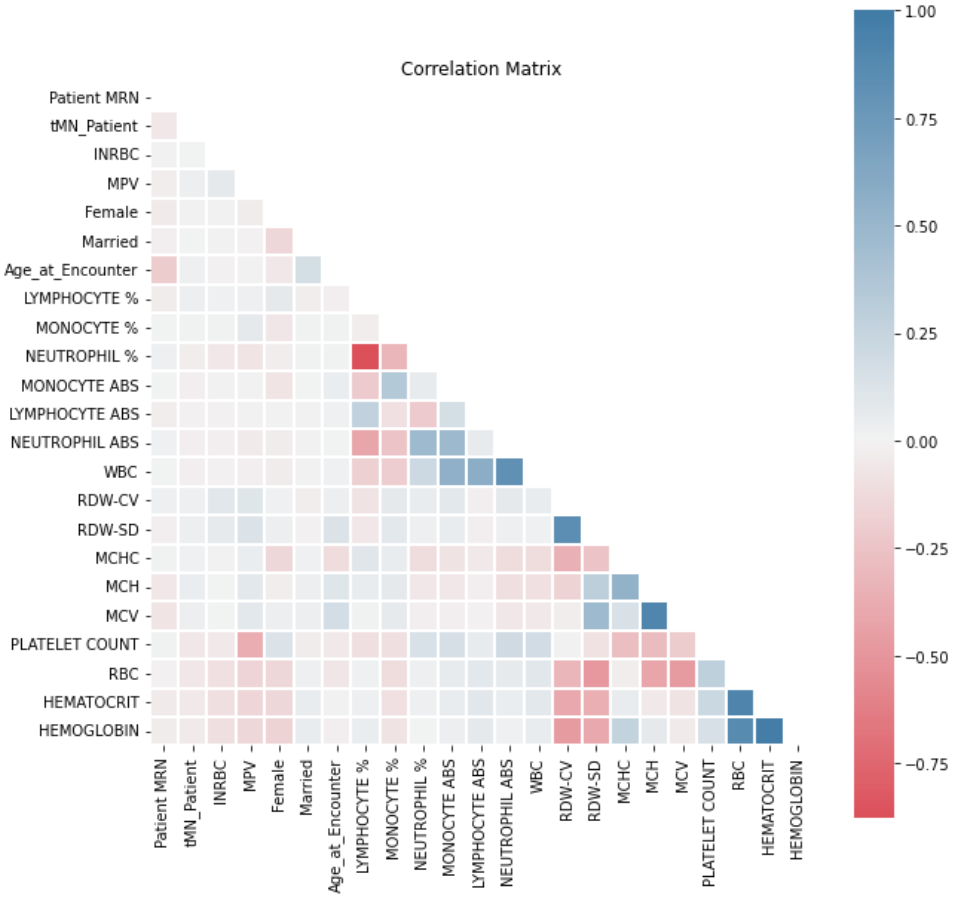
- Positive and Negative correlation pair exist

Identification of Patterns:

- Clusters or patterns of high correlation among groups of features may indicate underlying structures in the data.

Potential Redundancy:

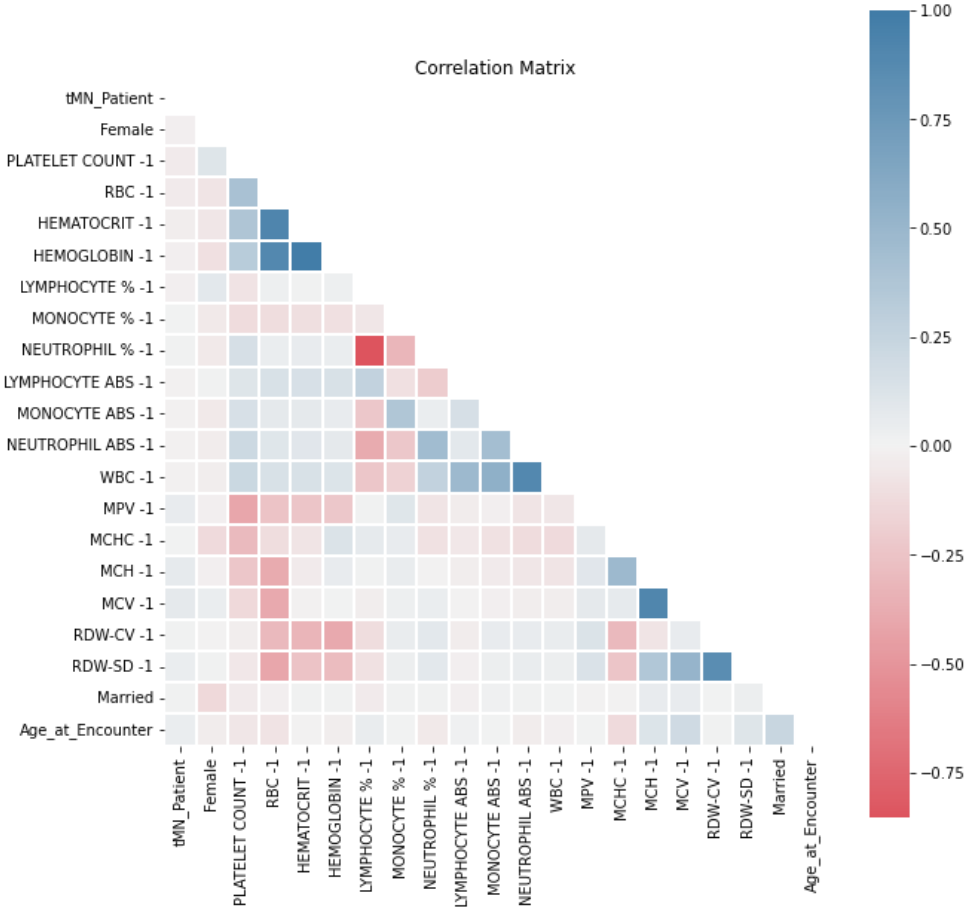
- High correlation between two features may indicate redundancy, and one of them could potentially be excluded to simplify the model.



N1 - Correlation Matrix

Stronger Multicollinearity:

- -1 period element correlation has more positive and negative correlation pairs exist
- The latest 6-month data shows comparatively stronger correlations between feature pairs compared to the feature pairs in the whole 7-year dataset
- Indicates a change or a shift in the underlying patterns or relationships between those features



N1 - Correlation Matrix

Potential Reasons

Intervention or Treatment Effects:

- Significant events or interventions within the last 6 months, such as chemotherapy treatments. Positive correlations may arise as a result of the therapeutic interventions.

Pre-Diagnostic Changes:

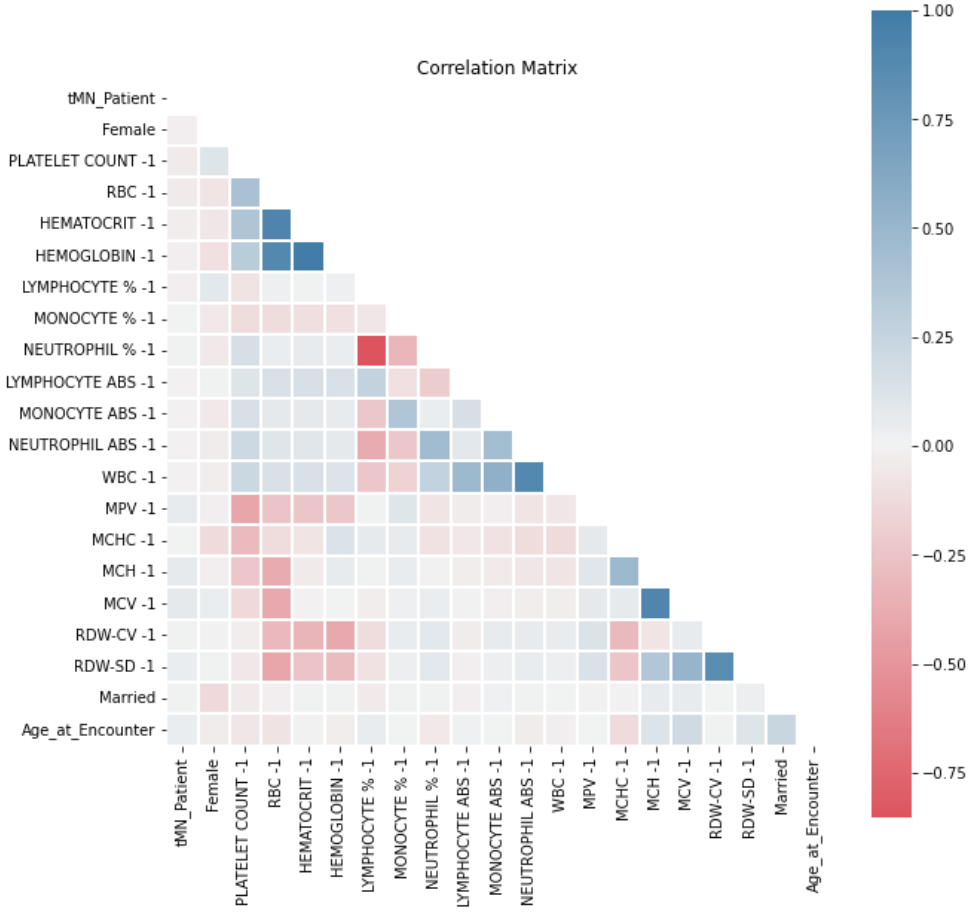
- Changes in blood indicators occurring in the months leading up to leukemia diagnosis may be reflective of early signs or manifestations of the disease before clinical diagnosis.

Side Effects of Medications:

- Cancer treatment medications may have effects on various blood parameters.
- Positive correlations emerge if specific medications impact multiple indicators in a consistent manner.

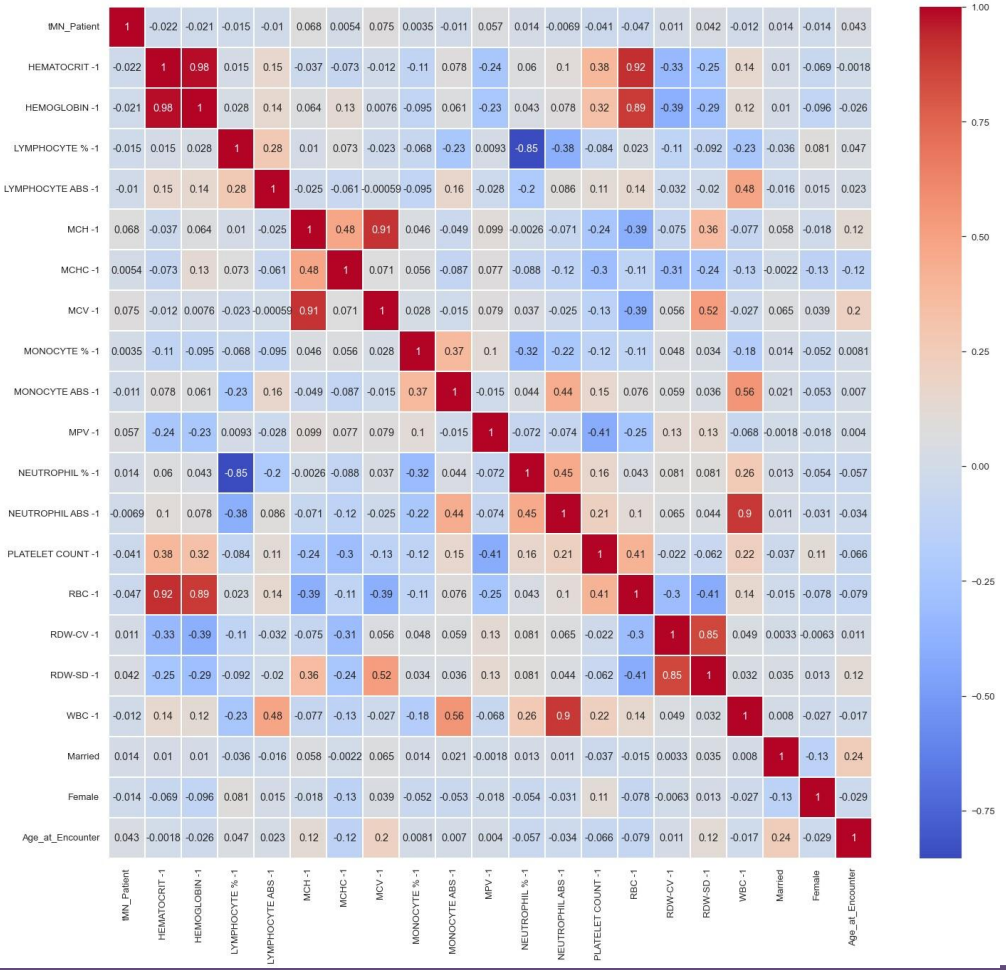
Seasonal Variations:

- Capturing partial pattern of seasons (Summer with higher temperature and different environment factors)

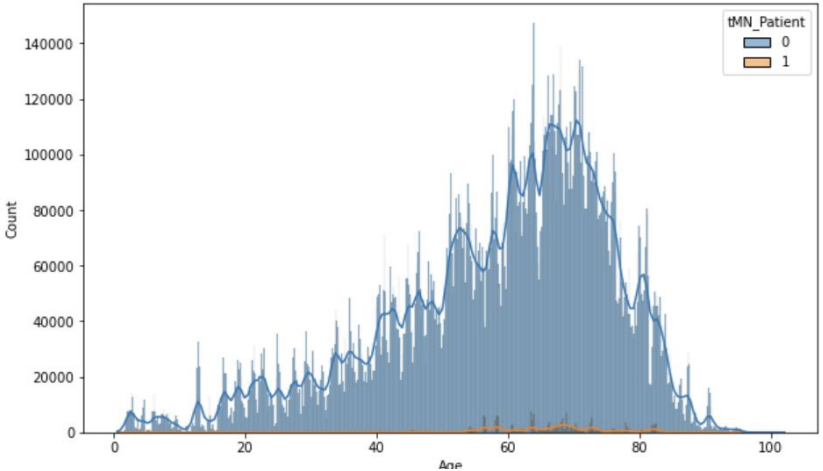


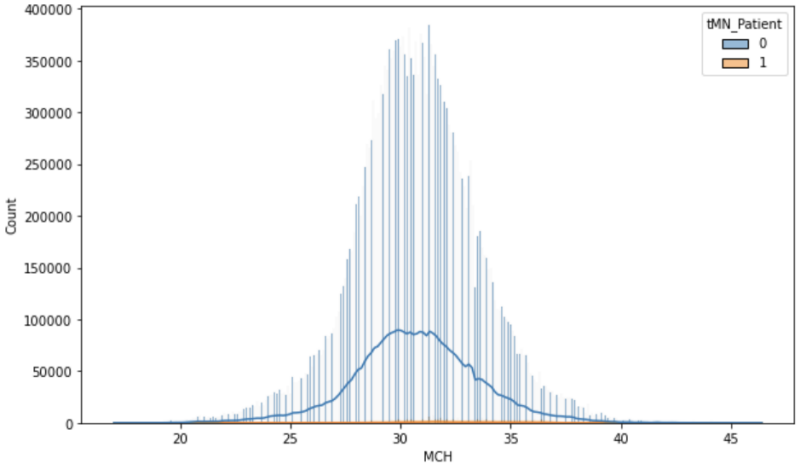
Overall Trend

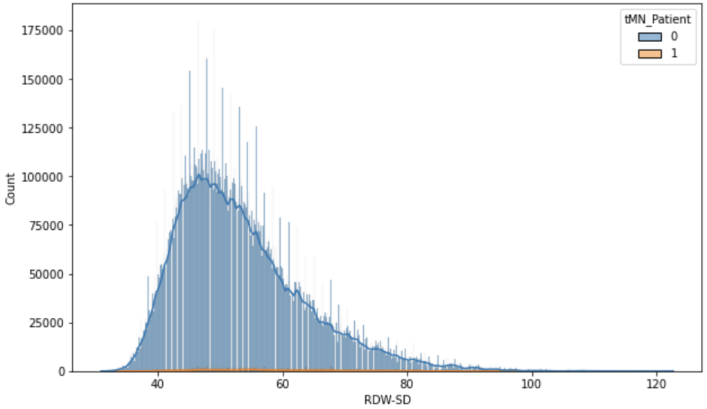
- There are positive and negative changes at the same time
- But slightly variations

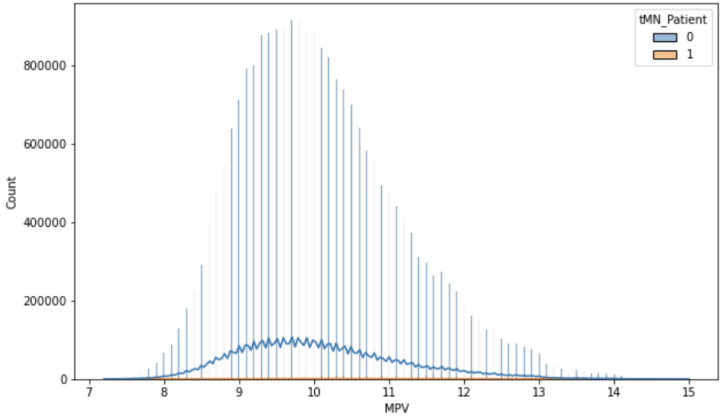


Age 50-83, peak around 70 years old









Model Result - Performance

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.6022	0.009	-69.155	0.000	-0.619	-0.585
HEMATOCRIT1	-1.2158	0.119	-10.224	0.000	-1.449	-0.983
HEMOGLOBIN1	2.3557	0.112	21.025	0.000	2.136	2.575
LYMPHOCYTEPCT1	0.0457	0.015	2.952	0.003	0.015	0.076
LYMPHOCYTEABS1	0.2163	0.028	7.784	0.000	0.162	0.271
MCH1	-2.2633	0.075	-30.274	0.000	-2.410	-2.117
MCHC1	0.6039	0.038	16.031	0.000	0.530	0.678
MCV1	2.3362	0.068	34.289	0.000	2.203	2.470
MONOCYTEPCT1	0.0796	0.009	9.341	0.000	0.063	0.096
MONOCYTEABS1	-0.1063	0.013	-8.317	0.000	-0.131	-0.081
MPV1	0.4637	0.004	113.775	0.000	0.456	0.472
NEUTROPHILPCT1	0.2174	0.017	12.761	0.000	0.184	0.251
NEUTROPHILABS1	0.4609	0.056	8.158	0.000	0.350	0.572
PLATELETCOUNT1	-0.1591	0.005	-29.207	0.000	-0.170	-0.148
RBC1	-1.3680	0.054	-25.311	0.000	-1.474	-1.262
RDWCV1	0.2414	0.016	15.569	0.000	0.211	0.272
RDWSD1	-0.3271	0.017	-18.693	0.000	-0.361	-0.293
WBC1	-0.5287	0.070	-7.509	0.000	-0.667	-0.391
Married	0.0491	0.008	5.829	0.000	0.033	0.066
Female	-0.2587	0.008	-34.315	0.000	-0.273	-0.244
AgeatEncounter	0.6762	0.005	131.054	0.000	0.666	0.686

No. Observations:	434116
Df Residuals:	434095
Df Model:	20
Scale:	1.0000
Log-Likelihood:	-2.4025e+05
Deviance:	4.8049e+05
Pearson chi2:	4.31e+05
Pseudo R-squ. (CS):	0.2438

Deviance and Pearson Chi-Square:

- The Deviance is 4.8049e+05, and the Pearson chi-square is 4.31e+05.
- These values are indicators of the goodness-of-fit of the model. Lower values suggest a better fit.

Feature Interpretation - Part 1

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.6022	0.009	-69.155	0.000	-0.619	-0.585
HEMATOCRIT1	-1.2158	0.119	-10.224	0.000	-1.449	-0.983
HEMOGLOBIN1	2.3557	0.112	21.025	0.000	2.136	2.575
LYMPHOCYTEPCT1	0.0457	0.015	2.952	0.003	0.015	0.076
LYMPHOCYTEABS1	0.2163	0.028	7.784	0.000	0.162	0.271
MCH1	-2.2633	0.075	-30.274	0.000	-2.410	-2.117
MCHC1	0.6039	0.038	16.031	0.000	0.530	0.678
MCV1	2.3362	0.068	34.289	0.000	2.203	2.470
MONOCYTEPCT1	0.0796	0.009	9.341	0.000	0.063	0.096
MONOCYTEABS1	-0.1063	0.013	-8.317	0.000	-0.131	-0.081
MPV1	0.4637	0.004	113.775	0.000	0.456	0.472
NEUTROPHILPCT1	0.2174	0.017	12.761	0.000	0.184	0.251
NEUTROPHILABS1	0.4609	0.056	8.158	0.000	0.350	0.572
PLATELETCOUNT1	-0.1591	0.005	-29.207	0.000	-0.170	-0.148
RBC1	-1.3680	0.054	-25.311	0.000	-1.474	-1.262
RDWCV1	0.2414	0.016	15.569	0.000	0.211	0.272
RDWSD1	-0.3271	0.017	-18.693	0.000	-0.361	-0.293
WBC1	-0.5287	0.070	-7.509	0.000	-0.667	-0.391
Married	0.0491	0.008	5.829	0.000	0.033	0.066
Female	-0.2587	0.008	-34.315	0.000	-0.273	-0.244
AgeatEncounter	0.6762	0.005	131.054	0.000	0.666	0.686

Strong Influence:

HEMOGLOBIN1 (2.3557):

- Positive coefficient indicates a strong positive impact on the log-odds.
- Individuals with higher hemoglobin levels are more likely to belong to the positive class.

MCV1 (2.3362): increase risks if positive

- Positive coefficient suggests a strong positive impact.
- Higher mean corpuscular volume increases the odds of being in the positive class.

MPV1 (0.4637):

- Positive coefficient indicates a strong positive influence.
- Increased mean platelet volume correlates with higher odds of the positive class.

Feature Interpretation - Part 2

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.6022	0.009	-69.155	0.000	-0.619	-0.585
HEMATOCRIT1	-1.2158	0.119	-10.224	0.000	-1.449	-0.983
HEMOGLOBIN1	2.3557	0.112	21.025	0.000	2.136	2.575
LYMPHOCYTEPCT1	0.0457	0.015	2.952	0.003	0.015	0.076
LYMPHOCYTEABS1	0.2163	0.028	7.784	0.000	0.162	0.271
MCH1	-2.2633	0.075	-30.274	0.000	-2.410	-2.117
MCHC1	0.6039	0.038	16.031	0.000	0.530	0.678
MCV1	2.3362	0.068	34.289	0.000	2.203	2.470
MONOCYTEPCT1	0.0796	0.009	9.341	0.000	0.063	0.096
MONOCYTEABS1	-0.1063	0.013	-8.317	0.000	-0.131	-0.081
MPV1	0.4637	0.004	113.775	0.000	0.456	0.472
NEUTROPHILPCT1	0.2174	0.017	12.761	0.000	0.184	0.251
NEUTROPHILABS1	0.4609	0.056	8.158	0.000	0.350	0.572
PLATELETCOUNT1	-0.1591	0.005	-29.207	0.000	-0.170	-0.148
RBC1	-1.3680	0.054	-25.311	0.000	-1.474	-1.262
RDWCV1	0.2414	0.016	15.569	0.000	0.211	0.272
RDWSD1	-0.3271	0.017	-18.693	0.000	-0.361	-0.293
WBC1	-0.5287	0.070	-7.509	0.000	-0.667	-0.391
Married	0.0491	0.008	5.829	0.000	0.033	0.066
Female	-0.2587	0.008	-34.315	0.000	-0.273	-0.244
AgeatEncounter	0.6762	0.005	131.054	0.000	0.666	0.686

Moderate Influence:

NEUTROPHILPCT1 (0.2174):

- Positive coefficient suggests a moderate positive impact.
- Higher neutrophil percentage moderately increases the odds of the positive class.

NEUTROPHILABS1 (0.4609):

- Positive coefficient indicates a moderate positive influence.
- Increased absolute neutrophil count moderately raises the odds of the positive class.

LYMPHOCYTEABS1 (0.2163):

- Positive coefficient suggests a moderate positive impact.
- Higher absolute lymphocyte count moderately increases the odds of the positive class.

Feature Interpretation - Part 3

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.6022	0.009	-69.155	0.000	-0.619	-0.585
HEMATOCRIT1	-1.2158	0.119	-10.224	0.000	-1.449	-0.983
HEMOGLOBIN1	2.3557	0.112	21.025	0.000	2.136	2.575
LYMPHOCYTEPCT1	0.0457	0.015	2.952	0.003	0.015	0.076
LYMPHOCYTEABS1	0.2163	0.028	7.784	0.000	0.162	0.271
MCH1	-2.2633	0.075	-30.274	0.000	-2.410	-2.117
MCHC1	0.6039	0.038	16.031	0.000	0.530	0.678
MCV1	2.3362	0.068	34.289	0.000	2.203	2.470
MONOCYTEPCT1	0.0796	0.009	9.341	0.000	0.063	0.096
MONOCYTEABS1	-0.1063	0.013	-8.317	0.000	-0.131	-0.081
MPV1	0.4637	0.004	113.775	0.000	0.456	0.472
NEUTROPHILPCT1	0.2174	0.017	12.761	0.000	0.184	0.251
NEUTROPHILABS1	0.4609	0.056	8.158	0.000	0.350	0.572
PLATELETCOUNT1	-0.1591	0.005	-29.207	0.000	-0.170	-0.148
RBC1	-1.3680	0.054	-25.311	0.000	-1.474	-1.262
RDWCV1	0.2414	0.016	15.569	0.000	0.211	0.272
RDWSD1	-0.3271	0.017	-18.693	0.000	-0.361	-0.293
WBC1	-0.5287	0.070	-7.509	0.000	-0.667	-0.391
Married	0.0491	0.008	5.829	0.000	0.033	0.066
Female	-0.2587	0.008	-34.315	0.000	-0.273	-0.244
AgeatEncounter	0.6762	0.005	131.054	0.000	0.666	0.686

Small Influence: lower, increase the risks of tMNs

WBC1 (-0.5287):

- Negative coefficient suggests a small negative impact.
- Higher white blood cell count decreases the odds of the positive class.

RBC1 (-1.3680):

- Negative coefficient indicates a small negative influence.
- Higher red blood cell count decreases the odds of the positive class.

HEMATOCRIT1 (-1.2158):

- Negative coefficient suggests a small negative impact.
- Higher hematocrit levels decrease the odds of the positive class

Feature Interpretation - Part 4

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.6022	0.009	-69.155	0.000	-0.619	-0.585
HEMATOCRIT1	-1.2158	0.119	-10.224	0.000	-1.449	-0.983
HEMOGLOBIN1	2.3557	0.112	21.025	0.000	2.136	2.575
LYMPHOCYTEPCT1	0.0457	0.015	2.952	0.003	0.015	0.076
LYMPHOCYTEABS1	0.2163	0.028	7.784	0.000	0.162	0.271
MCH1	-2.2633	0.075	-30.274	0.000	-2.410	-2.117
MCHC1	0.6039	0.038	16.031	0.000	0.530	0.678
MCV1	2.3362	0.068	34.289	0.000	2.203	2.470
MONOCYTEPCT1	0.0796	0.009	9.341	0.000	0.063	0.096
MONOCYTEABS1	-0.1063	0.013	-8.317	0.000	-0.131	-0.081
MPV1	0.4637	0.004	113.775	0.000	0.456	0.472
NEUTROPHILPCT1	0.2174	0.017	12.761	0.000	0.184	0.251
NEUTROPHILABS1	0.4609	0.056	8.158	0.000	0.350	0.572
PLATELETCOUNT1	-0.1591	0.005	-29.207	0.000	-0.170	-0.148
RBC1	-1.3680	0.054	-25.311	0.000	-1.474	-1.262
RDWCV1	0.2414	0.016	15.569	0.000	0.211	0.272
RDWSD1	-0.3271	0.017	-18.693	0.000	-0.361	-0.293
WBC1	-0.5287	0.070	-7.509	0.000	-0.667	-0.391
Married	0.0491	0.008	5.829	0.000	0.033	0.066
Female	-0.2587	0.008	-34.315	0.000	-0.273	-0.244
AgeatEncounter	0.6762	0.005	131.054	0.000	0.666	0.686

Other Influences:

RDWSD1 (-0.3271):

- Negative coefficient indicates an influence, but it might be of smaller magnitude.

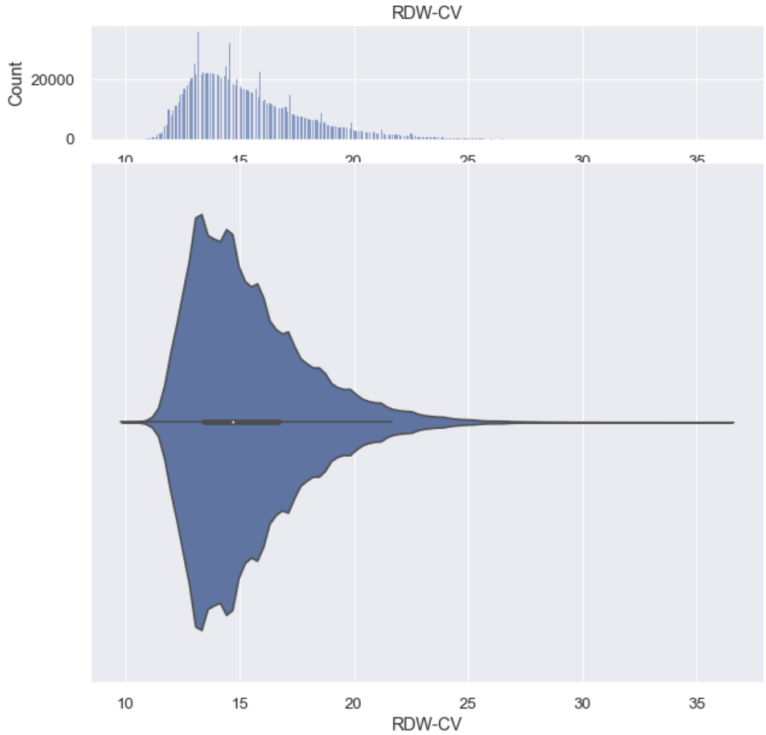
Binary Features (Married, Female):

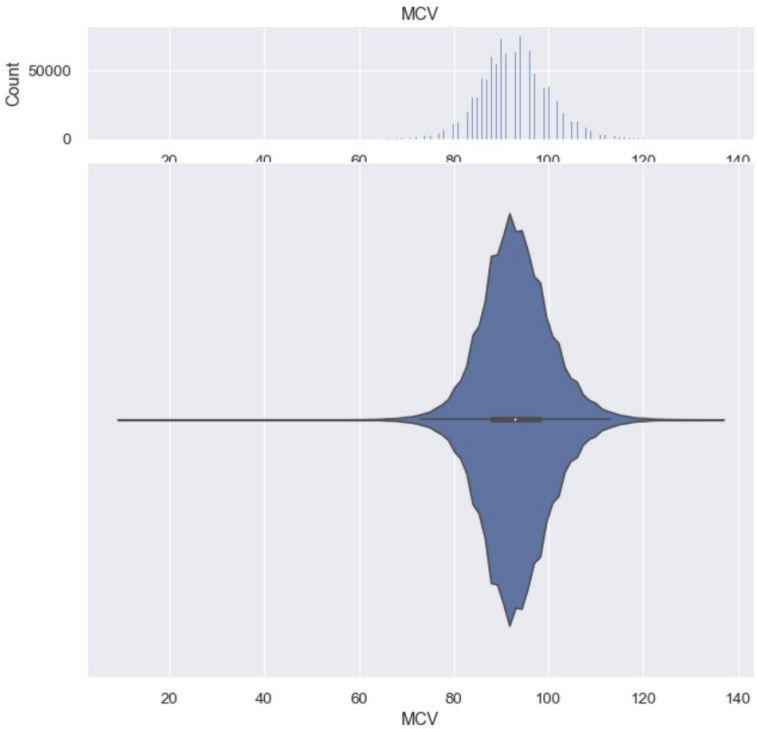
- Positive (Married) and negative (Female) coefficients suggest influences, but the magnitudes are smaller compared to continuous variables.

AgeatEncounter (0.6762):

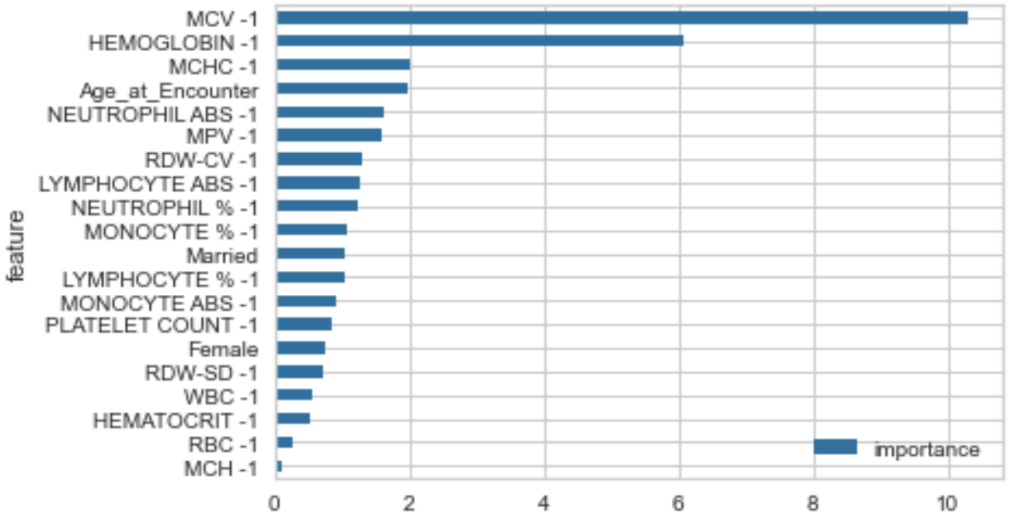
- Positive coefficient indicates a strong positive impact.
- Higher age at encounter significantly increases the odds of the positive class.

Abnormal Pattern Elements





Feature Importance

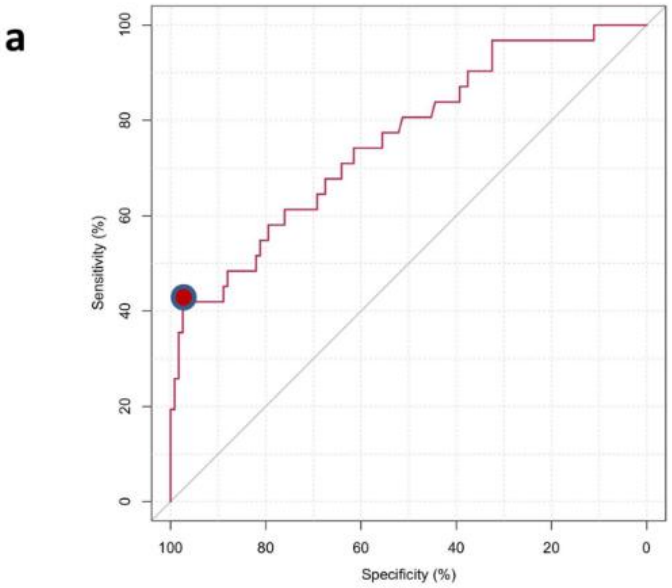


Most significant factors:

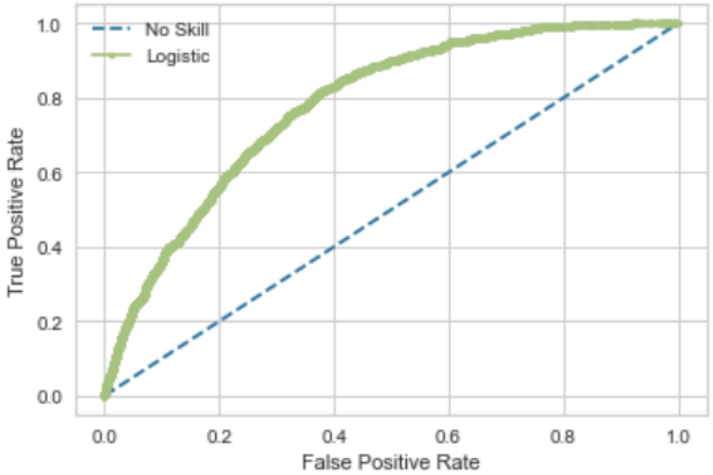
- The foremost factor, MCV-1, bears a weight five times greater than the third most important factor. It has the most substantial influence on the final outcome.

- HGB -1 is the second-best influencing factor

Model Comparison



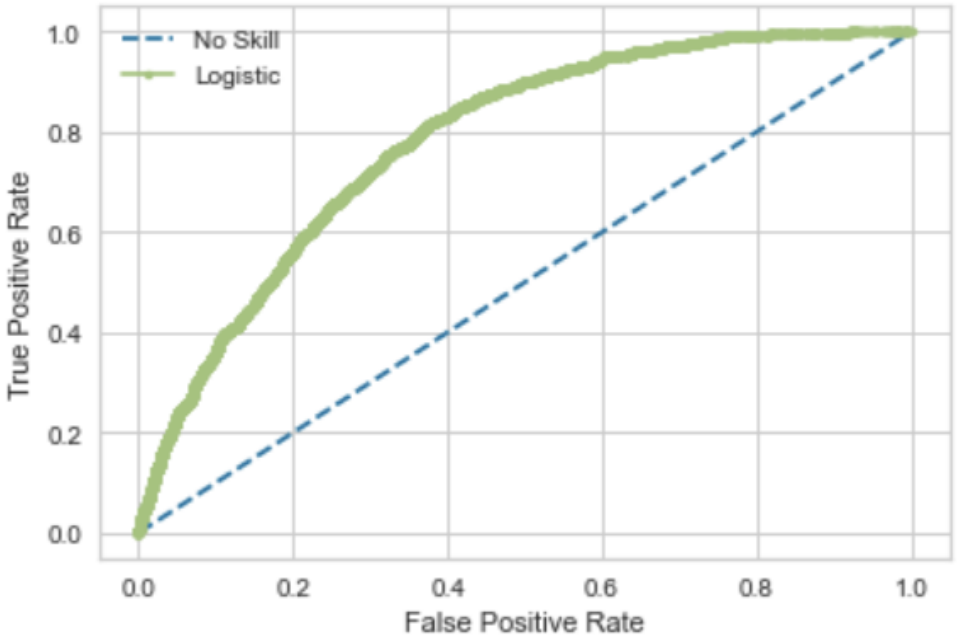
Extended Data Fig. 3 | Performance of the combined model in predicting progression to AML. a, Receiver operating characteristic curve for prediction of AML development using model 1 (see Methods). The red dot indicates the point on the curve with the highest positive predictive value with sensitivity of 41.9% and specificity of 95.7%. **b**, **c**, Kaplan–Meier



- Model 1 (from article): The ROC curve indicates a point with sensitivity of 41.9% and specificity of 95.7%.
- Model 2 (Our model): The ROC AUC is 0.783, and additional metrics are provided (precision, recall, and F1-score).
- Because True positive more important in this case, our model is out-performing the model

End

		Actual	
		Positive	Negative
Predicted	Positive	**True Positive** *Predicted has cancer* *Has Cancer*	**False Positive** *Predicted has cancer/Does not have cancer*
	Negative	**False Negative** *Predicted not cancer* *Has cancer*	**True Negative** *Predicted not cancer* *Does not have cancer*



- Interpretation of AUC:.
  - A value of 0.783 suggests that the model has a reasonably good ability to discriminate between positive and negative instances.
- Performance Benchmarks:
  - While there's no universal rule, you can use the following general benchmarks for interpreting AUC:
    - 0.5 to 0.6: Poor
    - 0.6 to 0.7: Fair
    - 0.7 to 0.8: Good
    - 0.8 to 0.9: Very good
    - 0.9 to 1.0: Excellent

Predictor Weight Importance

Capture non-linear relationships:

- Non-linear relationships between features and y that are not fully capture by correlation alone

Multicollinearity:

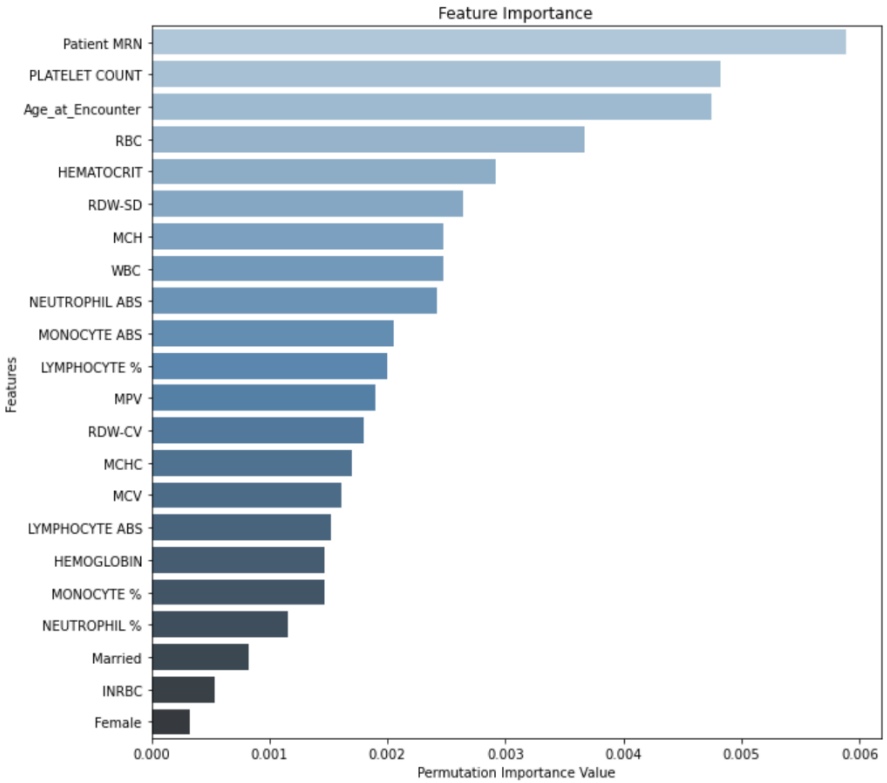
- Affect model differently from correlation which merely use pairwise relationships

Features interactions

- Capture interactions between features, where the combined effect of two or more features affects the target differently than the individual effects. combined effects on y

Higher noise sensitivity

- Correlation might be sensitive to noise or outliers in the data, while permutation importance tends to be more robust.
- If a feature has a strong impact on model performance, permutation importance would be likely to reflect the weight more reliably.



GLM Model Selection

GLMs are particularly useful when the assumptions of classical linear regression are not met, such as when dealing with non-normally distributed and heteroscedastic data or when modeling binary outcomes.

The generalized linear model (GLM) relates the linear predictor to the mean of the response variable through a link function. For binary logistic regression, a common type of GLM used for classification problems, the model can be expressed as follows:

$$\text{logit}(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

- π is the probability of the positive class (success) for a given set of predictor values.
- $\text{logit}(\pi)$ is the natural logarithm of the odds ratio $\frac{\pi}{1-\pi}$.
- β_0 is the intercept.
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients associated with the predictor variables $x_1, x_2, \ldots, x_n$.

Interpretability: The coefficients in GLMs provide interpretable information about the impact of predictors on the probability of the outcome.

Probabilistic Predictions: GLMs naturally output probabilities, allowing users to make probabilistic interpretations of classification results.

Flexibility: GLMs can be adapted to handle different types of response variable distributions, making them versatile for various data types.

GLM Model Result (Original)

```
No. Observations:                 841717
Df Residuals:                     841695
Df Model:                             21
Scale:                            1.0000
Log-Likelihood:                  -33404.
Deviance:                         66807.
Pearson chi2:                   5.36e+08
Pseudo R-squ. (CS):             0.007601
```

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-5.6828	0.035	-163.421	0.000	-5.751	-5.615
HEMATOCRIT	-0.2890	0.349	-0.827	0.408	-0.974	0.396
HEMOGLOBIN	0.6368	0.339	1.881	0.060	-0.027	1.300
INRBC	-0.0009	0.005	-0.188	0.851	-0.010	0.009
LYMPHOCYTEPCT	0.3771	0.034	10.949	0.000	0.310	0.445
LYMPHOCYTEABS	-0.4737	0.068	-6.974	0.000	-0.607	-0.341
MCH	-1.6901	0.240	-7.028	0.000	-2.161	-1.219
MCHC	0.7094	0.122	5.820	0.000	0.471	0.948
MCV	1.4898	0.215	6.943	0.000	1.069	1.910
MONOCYTEPCT	0.0774	0.017	4.687	0.000	0.045	0.110
MONOCYTEABS	0.0239	0.011	2.240	0.025	0.003	0.045
MPV	0.1075	0.012	8.636	0.000	0.083	0.132
NEUTROPHILPCT	0.1881	0.038	4.996	0.000	0.114	0.262
NEUTROPHILABS	-0.0126	0.051	-0.246	0.806	-0.113	0.088
PLATELETCOUNT	-0.4321	0.020	-21.291	0.000	-0.472	-0.392
RBC	-0.7316	0.155	-4.710	0.000	-1.036	-0.427
RDWCV	0.2609	0.040	6.483	0.000	0.182	0.340
RDWSD	-0.1410	0.044	-3.206	0.001	-0.227	-0.055
WBC	0.0257	0.068	0.381	0.703	-0.107	0.158
Female	0.1534	0.027	5.720	0.000	0.101	0.206
Married	0.2695	0.031	8.569	0.000	0.208	0.331
AgeatEncounter	0.3540	0.015	23.517	0.000	0.324	0.383

So, for logistic regression, the log-odds of the probability of the positive class is modeled as a linear combination of the predictor variables.

If we have a specific set of coefficients like β0=−5.6828 and β1(MCH) = -1.6901, it means the logistic regression model for the data is:

$$\text{logit}(\pi) = -5.6828 - 1.6901x_1$$

Intercept (β0): When all predictor variables are zero, the log-odds of the positive class is −5.6828.

Coefficient for x1(β1): A one-unit increase in X1 is associated with a −1.6901 decrease in the log-odds of the positive class.

To interpret the coefficients, we can exponentiate them to obtain odds ratios. For example, e^(-1.6901) would give us the odds ratio associated with a one-unit increase in X1

GLM Model Result (Improved)

```
No. Observations:                841717
Df Residuals:                    841699
Df Model:                            17
Scale:                           1.0000
Log-Likelihood:                 -33404.
Deviance:                        66808.
Pearson chi2:                  5.27e+08
Pseudo R-squ. (CS):            0.007600
```

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-5.6827	0.035	-163.458	0.000	-5.751	-5.615
HEMOGLOBIN	0.3785	0.130	2.902	0.004	0.123	0.634
LYMPHOCYTEPCT	0.3714	0.032	11.645	0.000	0.309	0.434
LYMPHOCYTEABS	-0.4530	0.055	-8.252	0.000	-0.561	-0.345
MCH	-1.6813	0.240	-7.001	0.000	-2.152	-1.211
MCHC	0.7563	0.108	6.999	0.000	0.545	0.968
MCV	1.4688	0.213	6.890	0.000	1.051	1.887
MONOCYTEPCT	0.0748	0.015	4.861	0.000	0.045	0.105
MONOCYTEABS	0.0277	0.007	4.033	0.000	0.014	0.041
MPV	0.1073	0.012	8.627	0.000	0.083	0.132
NEUTROPHILPCT	0.1851	0.035	5.339	0.000	0.117	0.253
PLATELETCOUNT	-0.4321	0.020	-21.291	0.000	-0.472	-0.392
RBC	-0.7753	0.146	-5.323	0.000	-1.061	-0.490
RDWCV	0.2612	0.040	6.507	0.000	0.183	0.340
RDWSD	-0.1404	0.044	-3.197	0.001	-0.227	-0.054
Female	0.1523	0.027	5.687	0.000	0.100	0.205
Married	0.2696	0.031	8.573	0.000	0.208	0.331
AgeatEncounter	0.3536	0.015	23.510	0.000	0.324	0.383

Remove features that shades a negative impact on model performance

- Remove INRBC
- Remove NEUTROPHILABS
- Remove WBC
- Remove HCT

Model interpretation
- Coefficients and factor influences weight

Positive X1 Effect:

- If X1 represents a variable that is measured in units (e.g., MCV), then for each additional unit increase in X1, the odds of the positive class occurring are e^1.4898 times higher.

Strength of Association:

- The magnitude of the odds ratio can also indicate the strength of the association between X1 and the positive class.
- An odds ratio greater than 1 suggests a positive association, while an odds ratio less than 1 suggests a negative association.

Direction of Impact:

- The sign of the coefficient (β1) indicates the direction of the impact. A positive coefficient means that an increase in X1 is associated with an increase in the log-odds of the positive class.

Result (Continue)

1. Precision:

The precision is the ratio of correctly predicted positive observations to the total predicted positives. A high precision indicates that when the model predicts a positive class, it is likely to be correct.

2. Recall (Sensitivity or True Positive Rate):

Recall is the ratio of correctly predicted positive observations to the total actual positives. A high recall for class 0 indicates that the model is good at identifying true positives for that class, while a low recall for class 1 suggests that the model misses many actual positives for that class.

3. F1-Score:

The F1-score is the weighted average of precision and recall. It is a balance between precision and recall. A high F1-score indicates a good balance between precision and recall.

4. Accuracy:

Overall accuracy is the ratio of correctly predicted observations to the total observations. In this case, the model accuracy indicating that it correctly predicts the class for of the result percent of the samples.

5. Macro Avg:

The macro average calculates the metric independently for each class and then takes the unweighted average. In this case, it provides an average precision, recall, and F1-score across both classes.

6. Weighted Avg:

weighted average calculates the metric for each class and then takes the average, weighted by the number of samples in each class. It provides an overall average that gives more importance to the class with more samples.

Result

11/9

	precision	recall	f1-score	support
0	0.73	0.69	0.71	584238
1	0.69	0.12	0.2	983
accuracy			0.71	585221
macro avg	0.71	0.22	0.46	585221
weighted avg	0.73	0.68	0.71	585221

11/21

	precision	recall	f1-score	support
0	0.81	0.83	0.82	274756
1	0.79	0.08	0.15	983
accuracy			0.82	275739
macro avg	0.8	0.44	0.49	275739
weighted avg	0.81	0.81	0.82	275739

11/16

	precision	recall	f1-score	support
0	0.75	0.86	0.81	461965
1	0.72	0.09	0.16	983
accuracy			0.81	462948
macro avg	0.73	0.48	0.49	462948
weighted avg	0.75	0.85	0.81	462948

11/26

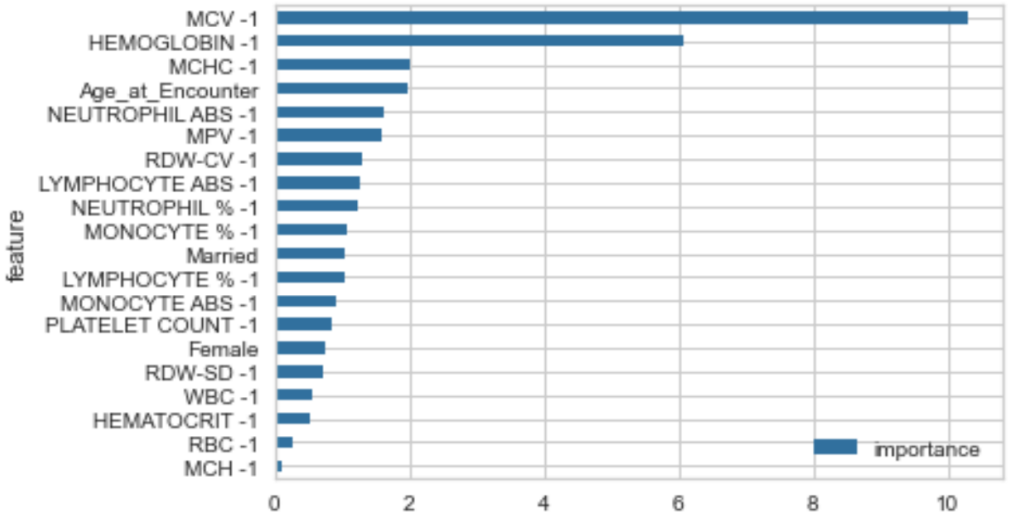
	precision	recall	f1-score	support
0	0.84	0.78	0.81	167104
1	0.82	0.14	0.24	983
accuracy			0.81	168087
macro avg	0.83	0.46	0.53	168087
weighted avg	0.84	0.76	0.81	168087

LR

	precision	recall	f1-score
0	0.68	0.75	0.72
1	0.72	0.65	0.68
accuracy			0.70
macro avg	0.70	0.70	0.70
weighted avg	0.70	0.70	0.70

~70%

lr - Feature Importance

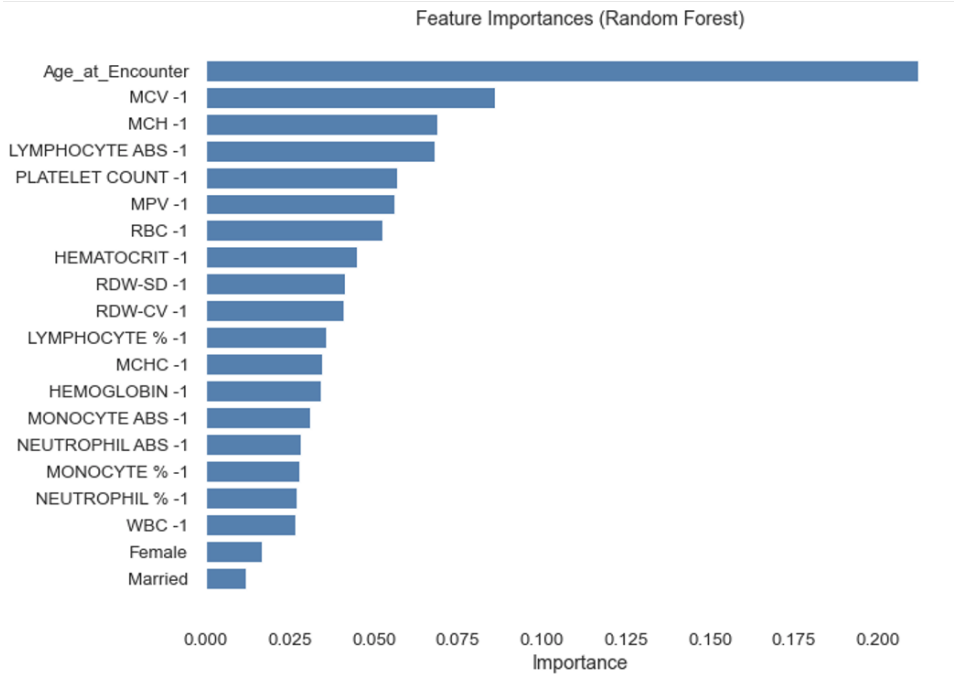


Most significant factors:

- The foremost factor, MCV-1, bears a weight five times greater than the third most important factor. It has the most substantial influence on the final outcome.

- HGB -1 is the second-best influencing factor

RF



features make more sense than rf

MCV and MCH, more meaningful features taking the weights.

HGB going down

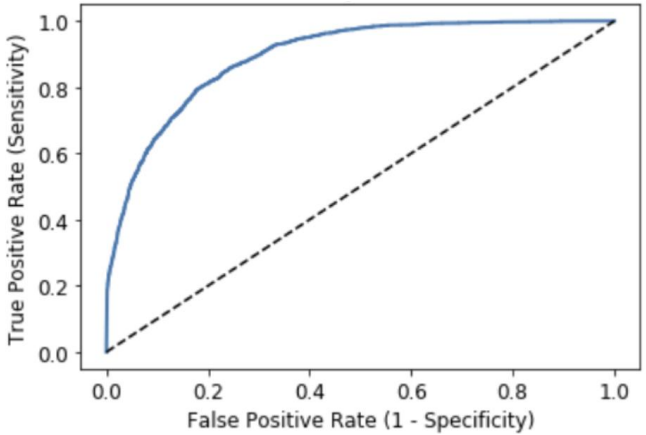
RF

	precision	recall	f1-score
0	0.89	0.82	0.85
1	0.83	0.81	0.82
accuracy			0.85
macro avg	0.86	0.82	0.84
weighted avg	0.89	0.82	0.85

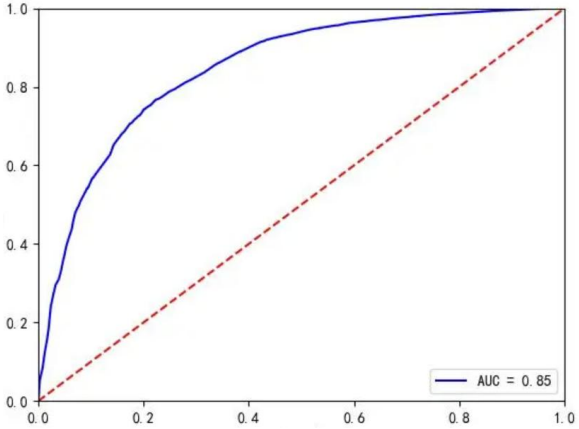
~80-90%

rf

sensitivity-specificity



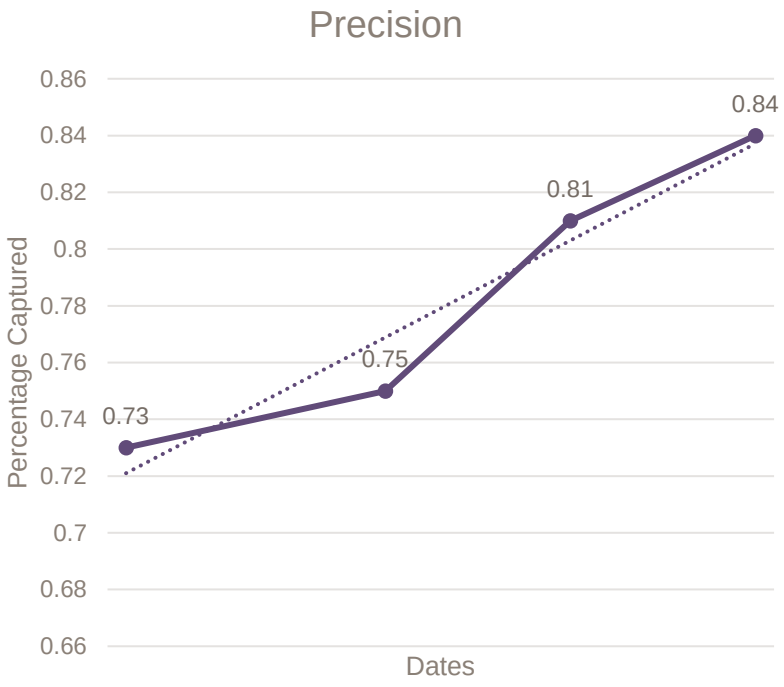
auc-roc



add on background for article comparisons

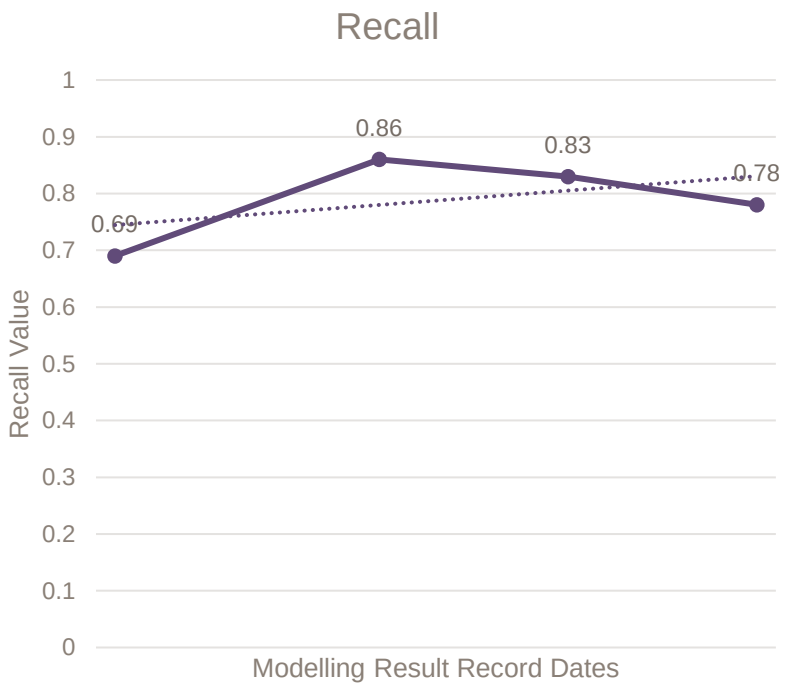
expense decrease

Precision



- The precision represents the ratio of correctly predicted positive observations to the total predicted positives. A high precision indicates that when the model predicts a positive class, it is likely to be correct.

Recall



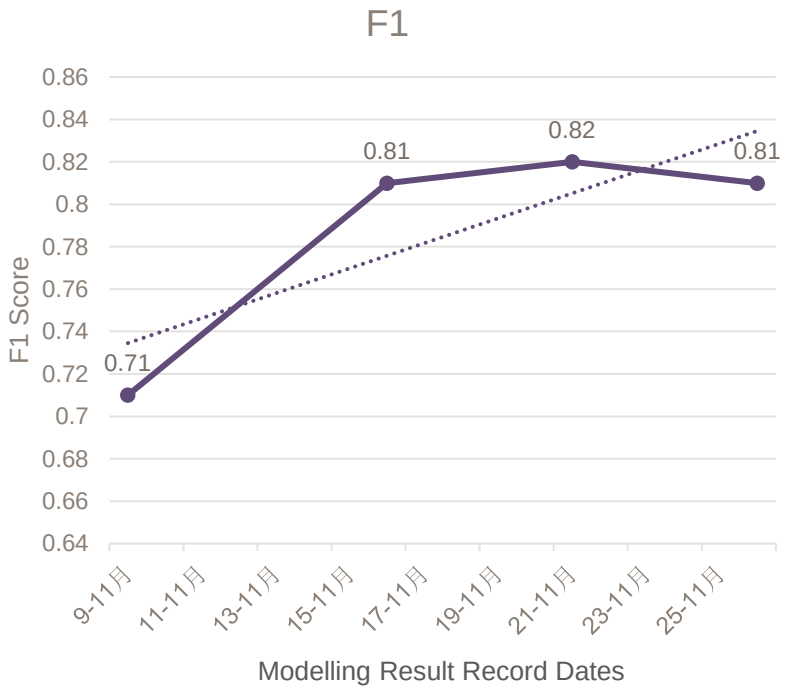
Recall (Sensitivity or True Positive Rate):

- Recall is the ratio of correctly predicted positive observations to the total actual positives. A higher recall indicates that the model is better at identifying true positives for that class. It measures the ability of the model to capture all the positive instances.

Precision-Recall Tradeoff

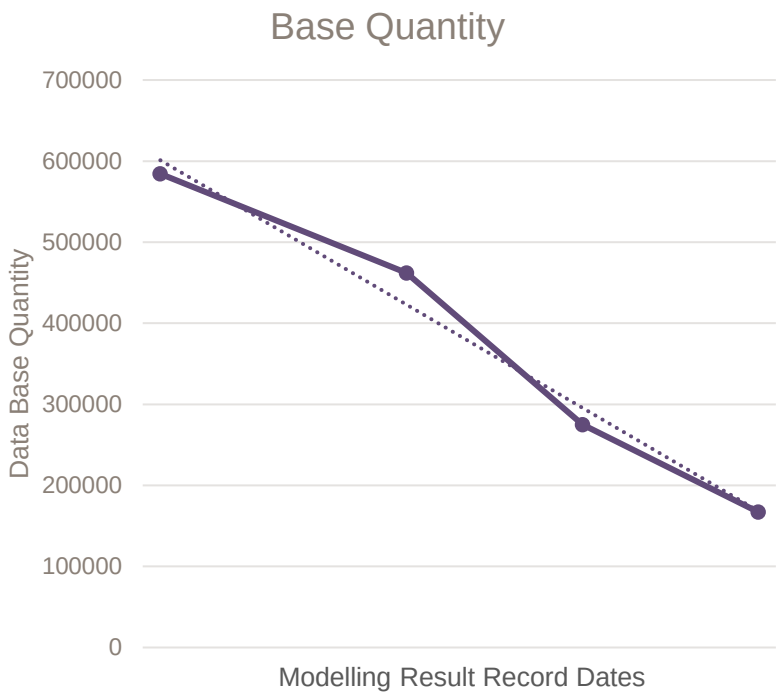
- The precision-recall tradeoff arises because increasing one of these metrics often leads to a decrease in the other.
- Lowering the threshold increases the number of instances predicted as positive, which can increase recall but may also introduce more false positives, reducing precision.
- Raising the threshold leads to fewer positive predictions, potentially increasing precision but at the cost of missing some positive instances, reducing recall.

F1 Score



- F1 score is a useful metric when there is an uneven distribution between the classes, and both false positives and false negatives need to be considered. It helps strike a balance between precision and recall, providing a single value that summarizes the performance of a classification model.

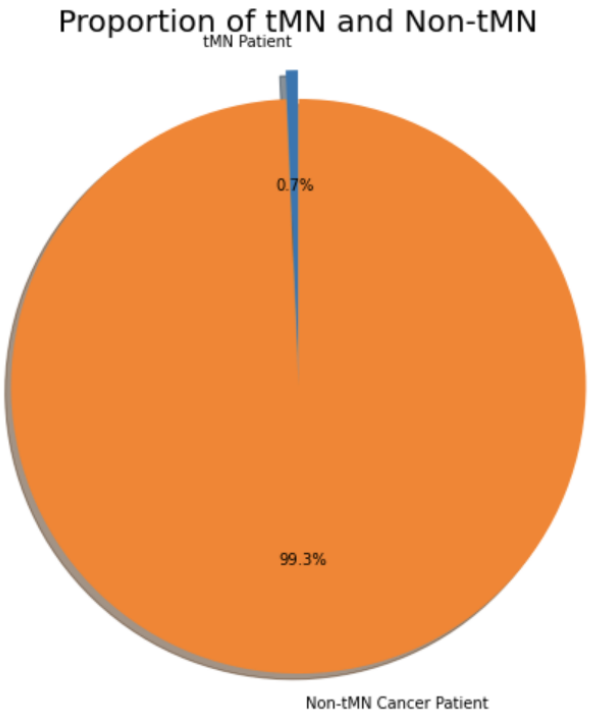
Support



Cutting Out

- Broken records
- Null values
- Falsely calculated results
- Non-informative features
- Noise

Proportion of Target Patient Group (Latest)



Reason for group 1 performance vs group 0 performance differences

- Same feature, same database, same patient source, but too small proportion for model to learn

Future Plans

Because you suggest we only use the available data in EPIC, following with your instruction and advice, I plan to adapt the following methods to improve the modeling result, including:

1. Algorithm (4 alternatives)
2. Change factors/variables
3. Change the weighting of factors/variables
4. Punishment of false prediction of case 1
5. After improving the overall result of base model, going in more in depth for other models

12.4 Report

	Study	PDID	WHOcat	TypeAML	Type2	C_Risk	M_Risk	NK	AOD	Center	ERDate	Intergroup	Study1
309	AMLSG0704	PD10789a	yes	AML	AML	Favorable	Favorable	0	43	39	0 days	0	
1372	AMLSG0704	PD10790a	no	AML	AML	Adverse	Adverse	0	56	29	0 days	0	
628	AMLSG0704	PD10792a	NA	AML	AML	Intermediate	Inter-1	1	48	36	0 days	0	
358	AMLSG0704	PD10793a	yes	AML	AML	Adverse	Adverse	0	54	2	0 days	0	
629	AMLSG0704	PD10794a	yes	AML	AML	Intermediate	Favorable	1	46	42	0 days	0	
630	AMLSG0704	PD10795a	yes	AML	AML	Intermediate	Favorable	1	49	2	0 days	0	
631	AMLSG0704	PD10796a	yes	AML	AML	Intermediate	Favorable	1	53	43	0 days	0	
1169	AMLSG0704	PD10798a	yes	AML	AML	Intermediate	Inter-2	0	58	43	0 days	0	
632	AMLSG0704	PD10800a	no	AML	AML	Intermediate	Inter-1	1	18	42	0 days	0	
1373	AMLSG0704	PD10801a	no	AML	AML	Intermediate	Inter-2	0	49	22	0 days	0	
633	AMLSG0704	PD10802a	no	AML	AML	Intermediate	Inter-1	1	50	44	0 days	0	
634	AMLSG0704	PD10803a	yes	AML	AML	Intermediate	Inter-1	1	44	14	0 days	0	
1170	AMLSG0704	PD10804a	no	AML	AML	Intermediate	Inter-2	0	57	43	0 days	0	
1374	AMLSG0704	PD10805a	yes	AML	AML	Adverse	Adverse	0	48	25	0 days	0	
1485	AMLSG0704	PD10806a	no	oAML	oAML	Adverse	Adverse	0	58	34	0 days	0	
635	AMLSG0704	PD10807a	yes	AML	AML	Intermediate	Inter-1	1	46	30	0 days	0	
1574	AMLSG0704	PD10808a	no	tAML	tAML	Adverse	Adverse	0	54	28	0 days	0	
636	AMLSG0704	PD10809a	yes	AML	AML	Intermediate	Favorable	1	44	14	0 days	0	

Showing 1 to 18 of 1,540 entries, 106 total columns

Data - Part 1 - non-genetic data

clinicalData

Filter Cols: « < 1 - 50 > »

	Study	PDID	WHOcat	TypeAML	Type2	C_Risk	M_Risk	NK	AOD	Center	ERDate	Intergroup	Study1	BaselineTherapy_	VPA	ATRA_arm	TPL_o	TPL_type	Date_LF	TPL_Art	CR_date	Rec
309	AMLSG0704	PD10789a	yes	AML	AML	Favorable	Favorable	0	43	39	0 days	0	1	1	0	1	0	NA	1085 days	NA	27 days	
1372	AMLSG0704	PD10790a	no	AML	AML	Adverse	Adverse	0	56	29	0 days	0	1	1	0	0	1	FREMD	277 days	FREMD	179 days	
628	AMLSG0704	PD10792a	NA	AML	AML	Intermediate	Inter-1	1	48	36	0 days	0	1	1	0	1	1	ALLO	667 days	ALLO	26 days	
358	AMLSG0704	PD10793a	yes	AML	AML	Adverse	Adverse	0	54	2	0 days	0	1	1	0	0	1	ALLO	1721 days	ALLO	64 days	
629	AMLSG0704	PD10794a	yes	AML	AML	Intermediate	Favorable	1	46	42	0 days	0	1	1	0	1	1	FREMD	2071 days	FREMD	72 days	
630	AMLSG0704	PD10795a	yes	AML	AML	Intermediate	Favorable	1	49	2	0 days	0	1	1	0	0	0	NA	1646 days	NA	29 days	
631	AMLSG0704	PD10796a	yes	AML	AML	Intermediate	Favorable	1	53	43	0 days	0	1	1	0	1	1	ALLO	1041 days	ALLO	62 days	
1169	AMLSG0704	PD10798a	yes	AML	AML	Intermediate	Inter-2	0	58	43	0 days	0	1	1	0	1	1	FREMD	1640 days	FREMD	35 days	
632	AMLSG0704	PD10800a	no	AML	AML	Intermediate	Inter-1	1	18	42	0 days	0	1	1	0	1	1	FREMD	437 days	FREMD	56 days	
1373	AMLSG0704	PD10801a	no	AML	AML	Intermediate	Inter-2	0	49	22	0 days	0	1	1	0	0	1	FREMD	539 days	FREMD	37 days	
633	AMLSG0704	PD10802a	no	AML	AML	Intermediate	Inter-1	1	50	44	0 days	0	1	1	0	1	0	NA	1496 days	NA	23 days	
634	AMLSG0704	PD10803a	yes	AML	AML	Intermediate	Inter-1	1	44	14	0 days	0	1	1	0	1	1	ALLO	1874 days	ALLO	111 days	
1170	AMLSG0704	PD10804a	no	AML	AML	Intermediate	Inter-2	0	57	43	0 days	0	1	1	0	0	1	FREMD	700 days	FREMD	155 days	
1374	AMLSG0704	PD10805a	yes	AML	AML	Adverse	Adverse	0	48	25	0 days	0	1	1	0	1	1	ALLO	462 days	ALLO	98 days	
1485	AMLSG0704	PD10806a	no	oAML	oAML	Adverse	Adverse	0	58	34	0 days	0	1	1	0	0	1	FREMD	208 days	FREMD	33 days	
635	AMLSG0704	PD10807a	yes	AML	AML	Intermediate	Inter-1	1	46	30	0 days	0	1	1	0	1	1	FREMD	244 days	FREMD	NA	
1574	AMLSG0704	PD10808a	no	tAML	tAML	Adverse	Adverse	0	54	28	0 days	0	1	1	0	0	1	ALLO	393 days	ALLO	204 days	
636	AMLSG0704	PD10809a	yes	AML	AML	Intermediate	Favorable	1	44	14	0 days	0	1	1	0	0	1	ALLO	1925 days	ALLO	49 days	
47	AMLSG0704	PD10810a	NA	AML	AML	NA	NA	NA	41	30	0 days	0	1	1	0	0	1	FREMD	2159 days	FREMD	124 days	
637	AMLSG0704	PD10811a	yes	AML	AML	Intermediate	Favorable	1	51	43	0 days	0	1	1	0	0	0	NA	1999 days	NA	83 days	
638	AMLSG0704	PD10812a	no	AML	AML	Intermediate	Inter-1	1	38	11	0 days	0	1	1	0	1	1	FREMD	2098 days	FREMD	33 days	
1171	AMLSG0704	PD10813a	yes	AML	AML	Favorable	Favorable	0	49	39	0 days	0	1	1	0	1	0	NA	1054 days	NA	29 days	
639	AMLSG0704	PD10814a	no	AML	AML	Intermediate	Inter-1	1	49	23	0 days	0	1	1	0	0	1	ALLO	1782 days	ALLO	187 days	
640	AMLSG0704	PD10815a	yes	AML	AML	Intermediate	Favorable	1	50	2	0 days	0	1	1	0	0	0	NA	1678 days	NA	32 days	
48	AMLSG0704	PD10816a	NA	AML	AML	NA	NA	NA	53	9	0 days	0	1	1	0	0	0	NA	46 days	NA	NA	
1508	AMLSG0704	PD10817a	no	AML	AML	Adverse	Adverse	0	51	29	0 days	0	1	1	0	0	1	FREMD	290 days	FREMD	83 days	
641	AMLSG0704	PD10818a	yes	AML	AML	Intermediate	Favorable	1	34	45	0 days	1	1	1	0	0	1	FREMD	1830 days	FREMD	15 days	
642	AMLSG0704	PD10819a	no	AML	AML	Intermediate	Inter-1	1	55	28	0 days	0	1	1	0	1	1	ALLO	363 days	ALLO	46 days	

Showing 1 to 28 of 1,540 entries, 106 total columns

Data - Part 2 - non-genetic data

clinicalData

Filter | Cols: « < 1 - 50 > »

Recurrence_date	TPL_date	TPL_Phase	Time_Diag_TPL	Time_1CR_TPL	Family_donnor	Post_Induction_Therapy_Remission	Status	OS	EFSSTAT	ereignart	efs	rfsstat	rfs	cuminc	rfsstat_allo	rfs_allo	cuminc_allo	HB	platelet
NA	NA	NA	NA	NA	1	CR	0	1085	0	NA	1085	0	1058	0	0	1058	0	6.2	20
NA	75 days	RD1	75	NA	0	RD	1	277	1	Death_in_CR	277	1	98	2	1	98	2	7.6	76
NA	226 days	CR1	226	200	1	CR	1	667	1	Death_in_CR	667	1	641	2	0	200	0	16.0	99
NA	86 days	CR1	86	22	1	CR	0	1721	1	RD_after_Ind1	28	0	1657	0	0	22	0	6.8	60
NA	98 days	CR1	98	26	0	CR	0	2071	1	RD_after_Ind1	27	0	1999	0	0	26	0	7.6	74
NA	NA	NA	NA	NA	0	CR	0	1646	0	NA	1646	0	1617	0	0	1617	0	8.0	58
344 days	134 days	CR1	134	72	1	CR	0	1041	1	Recurrence	344	1	282	1	0	72	0	7.4	50
865 days	996 days	CR2	996	961	0	CR	1	1640	1	Recurrence	865	1	830	1	1	830	1	6.4	25
168 days	85 days	CR1	85	29	0	CR	0	437	1	Recurrence	168	1	112	1	0	29	0	9.1	43
406 days	156 days	CR1	156	119	0	CR	1	539	1	Recurrence	406	1	369	1	0	119	0	11.9	42
NA	NA	NA	NA	NA	NA	CR	0	1496	0	NA	1496	0	1473	0	0	1473	0	8.9	130
NA	91 days	PR1	91	NA	1	RD	0	1874	1	RD_or_PR_after_Ind2	54	0	1763	0	0	1763	0	8.0	80
496 days	141 days	RD1	141	NA	0	RD	1	700	1	RD_after_Ind1	35	1	341	1	1	341	1	7.7	157
NA	68 days	RD1	68	NA	1	RD	1	462	1	RD_after_Ind1	20	1	364	2	1	364	2	7.7	119
152 days	161 days	Rez1	161	128	0	CR	1	208	1	Recurrence	137	1	119	1	1	119	1	11.2	449
NA	138 days	RD1	138	NA	0	RD	1	244	1	RD_or_PR_after_Ind2	66	NA	NA	NA	NA	NA	NA	7.7	55
375 days	174 days	PR1	174	NA	1	RD	1	393	1	RD_or_PR_after_Ind2	115	1	171	1	1	171	1	5.4	16
NA	104 days	CR1	104	55	1	CR	0	1925	0	NA	1925	0	1876	0	0	55	0	10.4	64
NA	96 days	RD1	96	NA	0	RD	0	2159	1	RD_after_Ind1	14	0	2035	0	0	2035	0	8.0	17
NA	NA	NA	NA	NA	0	CR	0	1999	0	NA	1999	0	1916	0	0	1916	0	12.8	46
NA	100 days	CR1	100	67	0	CR	0	2098	0	NA	2098	0	2065	0	0	67	0	9.7	47
NA	NA	NA	NA	NA	NA	CR	0	1054	0	NA	1054	0	1025	0	0	1025	0	9.4	33
NA	90 days	RD1	90	NA	1	RD	0	1782	1	RD_after_Ind1	35	0	1595	0	0	1595	0	9.8	50
NA	NA	NA	NA	NA	0	CR	0	1678	0	NA	1678	0	1646	0	0	1646	0	9.8	36
NA	NA	NA	NA	NA	0	RD	1	46	1	RD_after_Ind1	21	NA	NA	NA	NA	NA	NA	8.6	33
250 days	109 days	CR1	109	26	0	CR	1	290	1	RD_after_Ind1	25	1	167	1	0	26	0	13.6	178
344 days	501 days	CR2	501	486	0	CR	0	1830	1	Recurrence	344	1	329	1	1	329	1	5.0	40
306 days	151 days	CR1	151	105	1	CR	1	363	1	Recurrence	306	1	260	1	0	105	0	8.7	40

Data - Part 3 - non-genetic data

Status	OS	EFSSTAT	ereignart	efs	rfsstat	rfs	cuminc	rfsstat_allo	rfs_allo	cuminc_allo	HB	platelet	wbc	LDH	gender	BM_Blasts	Performance_ECOG	PB_Blasts	Splenomegaly	MLL_PTD	inv3_t3_3
0	1085	0	NA	1085	0	1058	0	0	1058	0	6.2	20	2.70	237	1	45.0	1	4	1	0	0
1	277	1	Death_in_CR	277	1	98	2	1	98	2	7.6	76	59.40	803	2	15.0	2	16	1	0	0
1	667	1	Death_in_CR	667	1	641	2	0	200	0	16.0	99	1.50	212	1	32.0	0	0	0	0	0
0	1721	1	RD_after_Ind1	28	0	1657	0	0	22	0	6.8	60	9.80	519	1	30.0	1	0	0	0	0
0	2071	1	RD_after_Ind1	27	0	1999	0	0	26	0	7.6	74	8.20	279	2	52.0	1	4	1	0	0
0	1646	0	NA	1646	0	1617	0	0	1617	0	8.0	58	1.60	200	2	50.0	1	3	0	0	0
0	1041	1	Recurrence	344	1	282	1	0	72	0	7.4	50	56.30	523	1	88.0	2	NA	0	0	0
1	1640	1	Recurrence	865	1	830	1	1	830	1	6.4	25	0.80	139	1	62.0	1	3	0	0	0
0	437	1	Recurrence	168	1	112	1	0	29	0	9.1	43	41.40	832	1	77.0	2	66	1	0	0
1	539	1	Recurrence	406	1	369	1	0	119	0	11.9	42	9.90	214	2	80.0	1	2	0	0	0
0	1496	0	NA	1496	0	1473	0	0	1473	0	8.9	130	57.30	419	2	80.0	1	9	0	1	0
0	1874	1	RD_or_PR_after_Ind2	54	0	1763	0	0	1763	0	8.0	80	50.00	1130	2	NA	0	90	0	0	0
1	700	1	RD_after_Ind1	35	1	341	1	1	341	1	7.7	157	124.60	739	1	60.0	0	85	0	1	0
1	462	1	RD_after_Ind1	20	1	364	2	1	364	2	7.7	119	108.00	986	2	90.0	1	91	1	0	1
1	208	1	Recurrence	137	1	119	1	1	119	1	11.2	449	2.60	202	1	58.0	0	13	0	0	0
1	244	1	RD_or_PR_after_Ind2	66	NA	NA	NA	NA	NA	NA	7.7	55	20.10	337	2	60.0	1	14	1	0	0
1	393	1	RD_or_PR_after_Ind2	115	1	171	1	1	171	1	5.4	16	5.00	587	2	25.0	1	1	0	0	0
0	1925	0	NA	1925	0	1876	0	0	55	0	10.4	64	91.10	224	2	80.0	0	92	0	0	0
0	2159	1	RD_after_Ind1	14	0	2035	0	0	2035	0	8.0	17	20.20	355	1	90.0	0	96	1	0	NA
0	1999	0	NA	1999	0	1916	0	0	1916	0	12.8	46	34.00	276	2	75.0	0	89	0	0	0
0	2098	0	NA	2098	0	2065	0	0	67	0	9.7	47	143.10	2410	2	NA	1	74	0	1	0
0	1054	0	NA	1054	0	1025	0	0	1025	0	9.4	33	9.30	418	1	75.0	1	34	1	0	0
0	1782	1	RD_after_Ind1	35	0	1595	0	0	1595	0	9.8	50	14.90	367	1	86.0	1	12	1	0	0
0	1678	0	NA	1678	0	1646	0	0	1646	0	9.8	36	52.70	359	1	85.0	2	92	0	0	0
1	46	1	RD_after_Ind1	21	NA	NA	NA	NA	NA	NA	8.6	33	13.60	330	1	81.0	1	45	0	0	NA
1	290	1	RD_after_Ind1	25	1	167	1	0	26	0	13.6	178	0.90	391	1	90.0	1	37	0	0	0
0	1830	1	Recurrence	344	1	329	1	1	329	1	5.0	40	56.20	570	1	96.0	1	79	0	0	0
1	363	1	Recurrence	306	1	260	1	0	105	0	8.7	40	71.60	648	1	90.0	1	75	0	0	0

Key columns related to Sediment Plot - Part 1

TPL_type	Date_LF	TPL_Art	CR_date	Recurrence_date	TPL_date	TPL_Phase	Time_Diag_TPL	Time_1CR_TPL	Family_donnor	Post_Induction_Therapy_Remission	Status	OS	EFSSTAT	ereignart	efs
	1085		27						1	CR	0	1085	0		1085
FREMD	277	FREMD	179		75	RD1	75		0	RD	1	277	1	Death_in_CR	277
ALLO	667	ALLO	26		226	CR1	226	200	1	CR	1	667	1	Death_in_CR	667
ALLO	1721	ALLO	64		86	CR1	86	22	1	CR	0	1721	1	RD_after_Ind1	28
FREMD	2071	FREMD	72		98	CR1	98	26	0	CR	0	2071	1	RD_after_Ind1	27
	1646		29						0	CR	0	1646	0		1646
ALLO	1041	ALLO	62	344	134	CR1	134	72	1	CR	0	1041	1	Recurrence	344
FREMD	1640	FREMD	35	865	996	CR2	996	961	0	CR	1	1640	1	Recurrence	865
FREMD	437	FREMD	56	168	85	CR1	85	29	0	CR	0	437	1	Recurrence	168
FREMD	539	FREMD	37	406	156	CR1	156	119	0	CR	1	539	1	Recurrence	406
	1496		23							CR	0	1496	0		1496
ALLO	1874	ALLO	111		91	PR1	91		1	RD	0	1874	1	RD_or_PR_after_Ind2	54
FREMD	700	FREMD	155	496	141	RD1	141		0	RD	1	700	1	RD_after_Ind1	35
ALLO	462	ALLO	98		68	RD1	68		1	RD	1	462	1	RD_after_Ind1	20
FREMD	208	FREMD	33	152	161	Rez1	161	128	0	CR	1	208	1	Recurrence	137
FREMD	244	FREMD			138	RD1	138		0	RD	1	244	1	RD_or_PR_after_Ind2	66
ALLO	393	ALLO	204	375	174	PR1	174		1	RD	1	393	1	RD_or_PR_after_Ind2	115
ALLO	1925	ALLO	49		104	CR1	104	55	1	CR	0	1925	0		1925
FREMD	2159	FREMD	124		96	RD1	96		0	RD	0	2159	1	RD_after_Ind1	14
	1999		83						0	CR	0	1999	0		1999
FREMD	2098	FREMD	33		100	CR1	100	67	0	CR	0	2098	0		2098
	1054		29							CR	0	1054	0		1054
ALLO	1782	ALLO	187		90	RD1	90		1	RD	0	1782	1	RD_after_Ind1	35
	1678		32						0	CR	0	1678	0		1678
	46								0	RD	1	46	1	RD_after_Ind1	21
FREMD	290	FREMD	83	250	109	CR1	109	26	0	CR	1	290	1	RD_after_Ind1	25
FREMD	1830	FREMD	15	344	501	CR2	501	486	0	CR	0	1830	1	Recurrence	344
ALLO	363	ALLO	46	306	151	CR1	151	105	1	CR	1	363	1	Recurrence	306
TPL_(Spende	240	TPL_(Spende	141	209	110	RD1	110		0	CR	1	240	1	RD_after_Ind1	29
ALLO	408	ALLO	130		75	PR1	75		1	RD	1	408	1	RD_or_PR_after_Ind2	56
AUTO	696	AUTO	64	279	157	CR1	157	93	0	CR	1	696	1	Recurrence	306
FREMD	143	FREMD			135	RD1	135		0	RD	1	143	1	RD_after_Ind1	120

Key columns related to Sediment Plot - Part 2

rfsstat	rfs	cuminc	rfsstat_allo	rfs_allo	cuminc_allo
0	1058	0	0	1058	0
1	98	2	1	98	2
1	641	2	0	200	0
0	1657	0	0	22	0
0	1999	0	0	26	0
0	1617	0	0	1617	0
1	282	1	0	72	0
1	830	1	1	830	1
1	112	1	0	29	0
1	369	1	0	119	0
0	1473	0	0	1473	0
0	1763	0	0	1763	0
1	341	1	1	341	1
1	364	2	1	364	2
1	119	1	1	119	1
1	171	1	1	171	1
0	1876	0	0	55	0
0	2035	0	0	2035	0
0	1916	0	0	1916	0
0	2065	0	0	67	0
0	1025	0	0	1025	0
0	1595	0	0	1595	0
0	1646	0	0	1646	0
1	167	1	0	26	0
1	329	1	1	329	1
1	260	1	0	105	0
1	68	1	1	68	1
1	278	2	1	278	2
1	215	1	1	215	1

C_Risk	M_Risk
Favorable	Favorable
Adverse	Adverse
Intermediate	Inter-1
Adverse	Adverse
Intermediate	Favorable
Intermediate	Favorable
Intermediate	Favorable
Intermediate	Inter-2
Intermediate	Inter-1
Intermediate	Inter-2
Intermediate	Inter-1
Intermediate	Inter-1
Intermediate	Inter-2
Adverse	Adverse
Adverse	Adverse
Intermediate	Inter-1
Adverse	Adverse
Intermediate	Favorable
Intermediate	Favorable
Intermediate	Inter-1
Favorable	Favorable
Intermediate	Inter-1
Intermediate	Favorable

- We have no risk scores to references in our database

Key Columns - Part 1:

- TPL_type: Indicates the type of treatment, possibly referring to a specific therapy or protocol.
- Date_LF: Represents the date of the last follow-up for the patient, providing the end point for the observation period.
- TPL_Art: Indicates the type of treatment related to art (possibly an abbreviation), likely associated with the treatment protocol.
- TPL_date: Date associated with the TPL treatment.
- TPL_Phase: Indicates the phase of TPL treatment.
- Time_Diag_TPL: Time from diagnosis to the start of TPL treatment.
- Time_1CR_TPL: Time from the first complete remission (CR) to the start of TPL treatment.
- cuminc: Cumulative incidence of a specific event.
- rfsstat_allo: Recurrence-free survival status specific to allo (possibly allogeneic treatment).
- rfs_allo: Recurrence-free survival time specific to allo.
- cuminc_allo: Cumulative incidence specific to allo. （bone marrow transplant）

Key Columns - Part 2:

- CR_date: Date of achieving complete remission (CR).
- Recurrence_date: Date of recurrence after initial treatment.
- Post_Induction_Therapy_Remission: Indicates whether the patient achieved remission after post-induction therapy.
- Status: Overall clinical status or outcome.
- OS (Overall Survival): Survival time from a specific time point, often diagnosis.
- EFSSTAT: Event-free survival status.
- ereignart: Type of event (e.g., death, recurrence) leading to an event-free survival.
- efs: Event-free survival time.
- rfsstat: Recurrence-free survival status.
- rfs: Recurrence-free survival time.

Questions related to TPL columns

What are the TPL_type?

- FREMD meaning?
- Why separate as FREMD and ALLO treatment?

TPL_type	Date_LF	TPL_Art
	1085	
FREMD	277	FREMD
ALLO	667	ALLO
ALLO	1721	ALLO
FREMD	2071	FREMD
	1646	
ALLO	1041	ALLO
FREMD	1640	FREMD
FREMD	437	FREMD
FREMD	539	FREMD
	1496	
ALLO	1874	ALLO
FREMD	700	FREMD
ALLO	462	ALLO
FREMD	208	FREMD
FREMD	244	FREMD
ALLO	393	ALLO
ALLO	1925	ALLO
FREMD	2159	FREMD

We need to change these part as tMN related disease or tMN itself

Data Problems:

Relapse:

- Too many categories and variety we could not capture one by one
- Around 550,000 records
- Unless we target the exact same disease name, since it would be different for every cancer patient.

In the article, it is always related to AML, and set AML categories so they can target the exact keyword

- We have no risk scores to references in our database

multi-stage prediction model

Replicate Design - Part 1

1. **TPL_type:**
   - This is a categorical variable indicating the type of treatment. There isn't a specific formula for this; it's a classification based on the type of treatment received.
2. **Date_LF:**
   - This is the date of the last follow-up and is recorded as a specific date.
3. **TPL_Art:**
   - Similar to TPL_type, it is a categorical variable indicating the type of treatment related to art.
4. **TPL_date:**
   - Date associated with the TPL treatment. It is recorded as a specific date.
5. **TPL_Phase:**
   - Indicates the phase of TPL treatment. This is typically a categorical variable.
6. **Time_Diag_TPL:**
   - Time from diagnosis to the start of TPL treatment
   - *Time_Diag_TPL=TPL_date−Diagnosis_date*
7. **Time_1CR_TPL:**
   - Time from the first complete remission (CR) to the start of TPL treatment.
   - *Time_1CR_TPL=TPL_date−CR_date*

Replicate Design - Part 2

8. **CR_date:**
   - Date of achieving complete remission (CR)
   - (If resolved date is not null).
9. **Recurrence_date:**
   - Date of recurrence after the initial treatment. It is recorded as a specific date.
   - (if the ICD or diagnosis ever occur, more than once)
10. **Post_Induction_Therapy_Remission:**
   - Indicates whether the patient achieved remission after post-induction therapy. (Remission - Resolved Date)
11. **Status:**
   - Overall clinical status or outcome (no specific matched explanations).
12. **OS (Overall Survival):**
   - Survival time from a specific time point, often diagnosis.
   - *OS=Date(alive)−First Cancer Diagnosis_date (first)*

Replicate Design - Part 3

13. **EFSSTAT:**
    - Event-free survival status. This is typically a categorical variable.
    - (Survive after cancer, no tMN development)
14. **ereignart:**
    - Type of event (e.g., death, recurrence) leading to an event-free survival.
    - (Death, remission or survive after cancer without remission)
15. **efs:**
    - Event-free survival time.
    - *efs* = *Date* (today) − *First Cancer Diagnosis_date*
16. **rfsstat:**
    - Recurrence-free survival status.
    - (No relapse, no same diagnosis occurs)

Replicate Design - Part 4

17. **rfs:**
    - Recurrence-free survival time.
    - *rfs=Recurrence_date（if recurrence exist）−Diagnosis_date*
18. **cuminc:**
    - Cumulative incidence of a specific event. Calculated using Kaplan-Meier or cumulative incidence functions.
19. **rfsstat_allo:**
    - Recurrence-free survival status specific to allo (allogeneic hematopoietic cell transplant (allograft))
20. **rfs_allo:**
    - Recurrence-free survival time specific to allo.
    - *rfs_allo=Recurrence_date−Allo_Treatment_date*
21. **cuminc_allo:**
    - Cumulative incidence specific to allo. Calculated using Kaplan-Meier or cumulative incidence functions for events related to allo treatment.

Model Design Process

1. Start with what data we have
2. According to the data traits and distributions, we pick the model
3. Run the appropriate model and optimize model results

Choose a Model:

- Consider the nature of the problem (classification etc.).
- Consider the size of the dataset.
- Choose models based on the characteristics of your data (linear models for linear relationships, tree-based models for non-linear relationships, etc.)

Data Problems

Since we have none of the 23 key columns in the article's data set, we need to replicate and create all the key columns by ourselves.

We have very messy raw data and have none of the key categories they are using in our database.

Specifically:

- We can see a patient dead, but no mark reason of dead. So we need to make assumption that patient dead from the disease before the death date
- We assume relapse will use the exact same disease name from before
- We assume the closest date we can get represent the true status of the corresponding cancer and disease.
- No clear grouping or differentiation showed in the article.
- No specific time recorded for relapse and remission.
- No risk factors we can evaluate our patients for.

Data Traits

Dataset in the article:

- time specific columns features
- classifications for different disease and capture all the situations

Anderson data:

- not much time columns to replicate multiple disease classifications probability variation results.
- Cannot be captured for all cancer cases and would not be a unified result

Additional Features Design

Inherited, behavioral/lifestyle, environmental risk factors and comorbid conditions:

- tMN several blood indicators do not change
- Age, background disease, gender, allergy (Inherited)
- Smoking, germline variation, obesity (BMI) (disease related), blood pressure (Unhealthy lifestyle)
- Chemotherapy type, frequency and time, and the relevant medication usage. (Chemical factor/inducement)
- Relevant diseases related to Hematopoietic stem cells (HSCs) change or CHIP early stage before Leukemia (Propagating inflammation, accelerating cardiovascular disease, co-morbidities, hematologic cancers, atherosclerotic cardiovascular diseases, inflammatory conditions, premature menopause, HIV)

1. Non-remission Death:
    - Death without remission from their own cancer
    - all cancer patients
    - can first dead from their own cancers

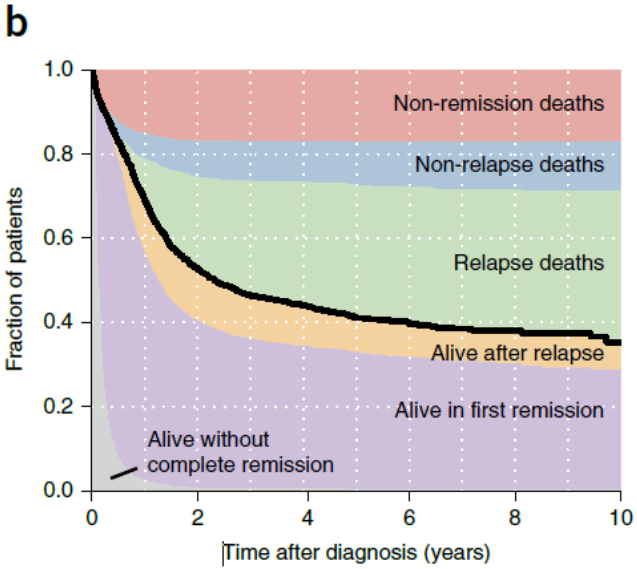
2. Non-relapse deaths:
    - remission from their own cancer but die from other cancer or diseases (Non-tMN diseases)
    - cancers can be in remission, but dead from other possibilities

3. Relapse death
    - develop Tmn LATER
    - REPLAP IS TMN, and dead from tMN

what is their chances of diagnosis
when would they die from tmn

4. did not develop and alive (could be middle process)

5. alive without complete remission
    - cancer still active but did not dead,
    - and no tMN development



time component

> Nat Genet. 2017 Mar;49(3):332-340. doi: 10.1038/ng.3756. Epub 2017 Jan 16.

Precision oncology for acute myeloid leukemia using a knowledge bank approach

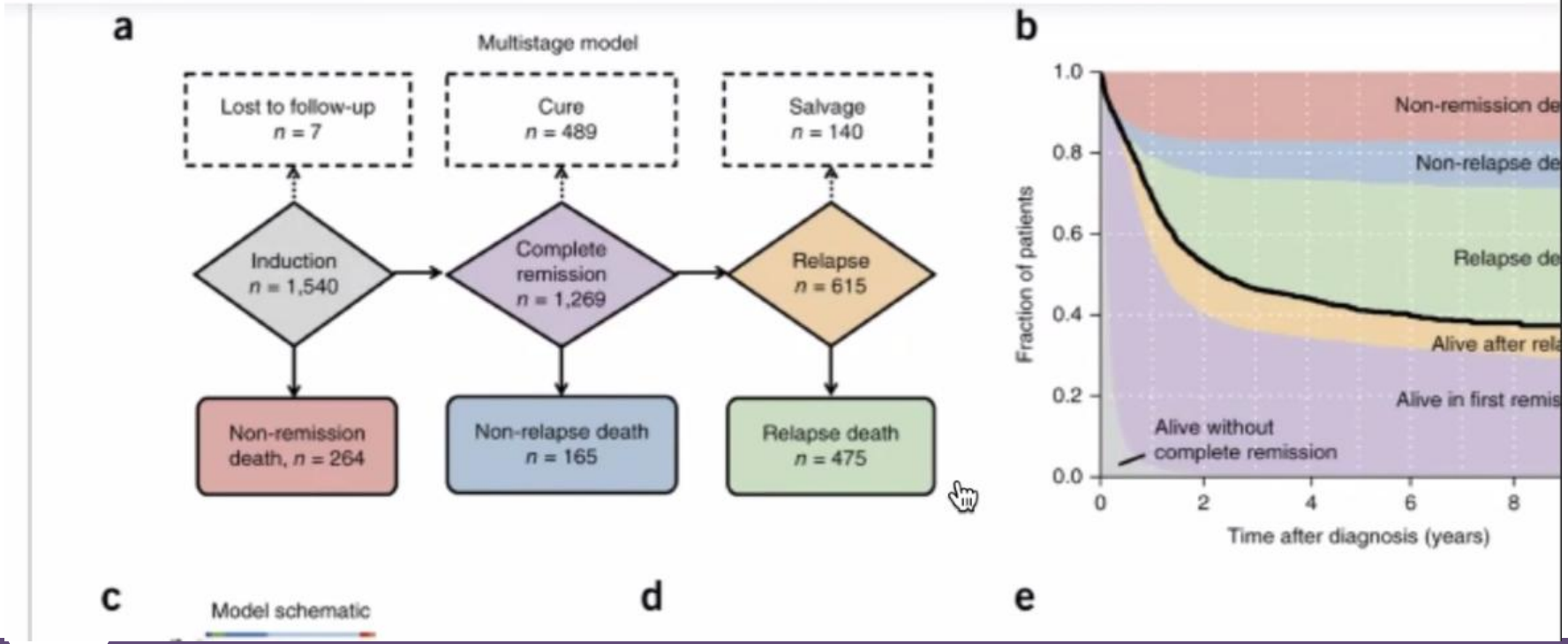
Moritz Gerstung [1 2], Elli Papaemmanuil [1 3], Inigo Martincorena [1], Lars Bullinger [4], Verena I Gaidzik [4], Peter Paschka [4], Michael Heuser [5], Felicitas Thol [5], Niccolo Bolli [1 6], Peter Ganly [7], Arnold Ganser [5], Ultan McDermott [1], Konstanze Döhner [4], Richard F Schlenk [4], Hartmut Döhner [4], Peter J Campbell [1 8]

Affiliations + expand

PMID: 28092685 PMCID: PMC5764082 DOI: 10.1038/ng.3756

Free PMC article

Precision oncology for acute myeloid leukemia using a knowledge ba



Problems

dead, not chance of develop the tMNs

competing risk

don’t develop
dead too early
or really don’t have a tMN

multi-stage model

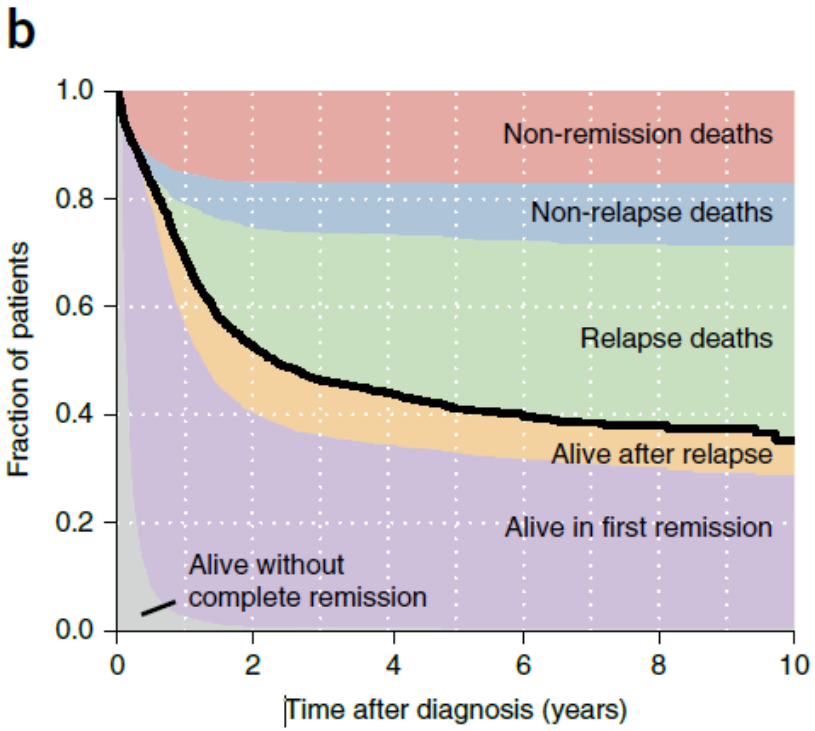
get a first chemotherapy, proabbility of achieve remission or not and dead
once remission, relapse in the future. Dead from other disease.
dead from relapse
two years later, the chance of alive that disease leukemia comes abck as a replapse
chance from leukemia coming back

did not develop and alive

alive without complete remission

cancer still active but did not dead,

and no tMN development



Our data set is not appropriate for direct clinical use, as the algorithm has not yet been prospectively validated and sequencing was performed using a research pipeline. Nonetheless, as a research tool, we have created a prototype portal within our website[21] (http://cancer.sanger.ac.uk/aml-multistage) that allows outcome predictions to be generated based on this data set for user-defined constellations of genomic features, clinical variables and treatment strategies (Supplementary Fig. 5). The underlying algorithm is capable of imputing missing variables and computing confidence intervals for each prediction.

The knowledge bank

cancer.sanger.ac.uk/aml-multistage/

AML multistage predictions research only

1. Select sample (or enter new patient's variables below)

Please select

Data may be rounded for privacy reasons.

2. Enter/amend variables

Clinical variables

Gender

male female N/A

Clinical

Lactic Acid Dehydrogenase (units/l) [83-7627]

Performance (ECOG)

0 1 2 3 4 N/A

White cell count (1e-9/l) [0.2-532.7]

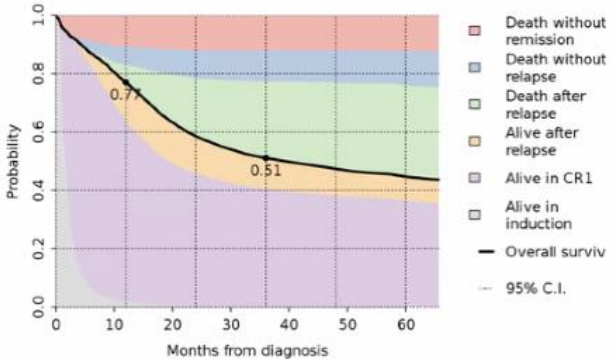
Platelet count (1e-9/l) [2-916]

Results | Log hazard | Coefficients | Help

Patient summary

Patient:
Driver mutations:
Blood counts:
Treatment:

Outcome after diagnosis



Outcome 3 years after diagnosis

Death without remission	12%
Death without relapse	11%
Death after relapse	26%
Alive after relapse	10%
Alive in CR1	41%
Alive without CR	0%

1. Non-remission Death:
    - Death without remission from their own cancer
    - all cancer patients
    - can first dead from their own cancers

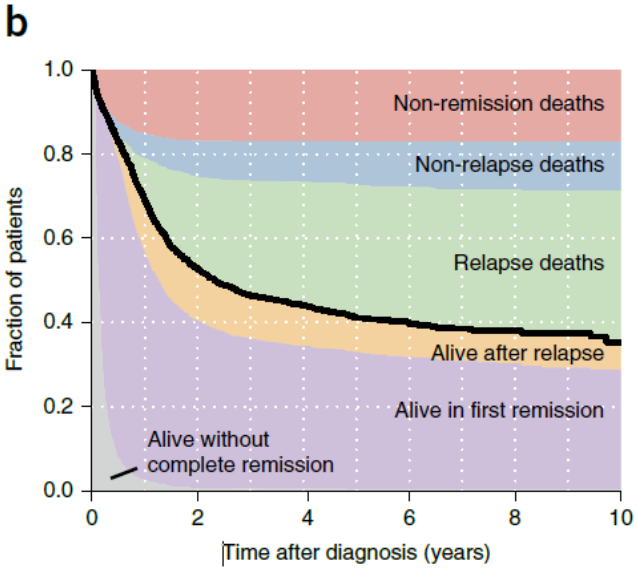
2. Non-relapse deaths:
    - remission from their own cancer but die from other cancer or diseases (Non-tMN diseases)
    - cancers can be in remission, but dead from other possibilities

3. Relapse death
    - develop Tmn LATER
    - REPLAP IS TMN, and dead from tMN

what is their chances of diagnosis
when would they die from tmn

4. did not develop and alive (could be middle process)

5. alive without complete remission
    - cancer still active but did not dead,
    - and no tMN development



would have had their treatments changed using a knowledge bank in comparison to the current recommendations. It is therefore natural to assess how many patients would have had t nt altered

On av prognos at 3 year

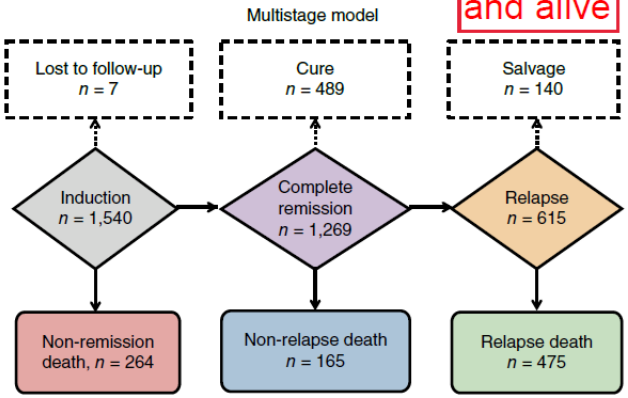
1119 records

Relapse and alive

529889 records

94155 patients

a



death date before recover/NA

death after same disease

c

Model schematic

d

tree diagram

as a track
grouping different

Source

Precision oncology for acute myeloid leukemia using a knowledge bank approach

Article: https://www.nature.com/articles/ng.3756#Sec15
Calculator: https://www.nature.com/articles/ng.3756#Sec15
Github: https://github.com/mg14/aml-multistage