

Exploratory and Predictive Analytics for Precision Medicine

Phu Dang (PACE Scholar) & **Leena Kang** (PATHS Scholar)
Data Science, Applied Mathematics
Mentor: Dr. Niema Moshiri
University of California San Diego

Outline - FOR WORK/TIME DISTRIBUTION, WILL BE HIDDEN

- Background (Phu) & Data + RQ (Leena)
- Exploratory Analysis
 - Uni-, Bi-, Multivariate Analyses (Phu)
 - Data Collection Bias Analysis (Phu)
- Predictive Analysis
 - Machine Learning Models (Leena)
 - Solutions to Imbalanced Data (Leena)
 - Performance Metrics & Assessment (Phu)
- Limitations (Leena) & Potential Applications (Phu)
- Conclusion & Closing

Outline

- **Background & Data + Research Question**
- Exploratory Analysis
 - Uni-, Bi-, Multivariate Analyses
- Predictive Analysis
 - Machine Learning Models
 - Solutions to Imbalanced Data
 - Performance Metrics & Assessment
- Limitations
- Closing & Acknowledgements

Background

Background

<https://library.ucsd.edu/dc/object/bb2493244b>

LIBRARY DIGITAL COLLECTIONS

Search Digital Collections

Component 1 of 51

Women's Healthy Eating and Living (WHEL) Study

Complete Set of WHEL Data

File Size 13.1 MB

File Format ZIP Format

Scope And Content This file was modified on 2022-08-23 to include mealtime data and again on 2022-12-21 to include a properly formatted version of the same file.

Download file View file

Collection

- Women's Healthy Eating and Living (WHEL) Study

Cite This Work

Pierce, John P.; Faerber, Susan; Wright, Fred A.; Rock, Cheryl L.; Newman, Vicky; Flatt, Shirley W.; Kealey, Sheila; Jones, Vicki E.; Caan, Bette J.; Gold, Ellen B.; Haan, Mary; Hollenbach, Kathryn A.; Jones, Lovell A.; Marshall, James R.; Ritenbaugh, Cheryl; Stefanick, Marcia L.; Thomson, Cynthia; Wasserman, Linda; Natarajan, Loki; Thomas, Ronald G.; Gilpin, Elizabeth; Parker, Barbara A.; Greenberg, E. Robert; Al-Delaimy, Wael K.; Bardwell, Wayne A.; Carlson, Robert W.; Emond, Jennifer A.; Hajek, Richard A.; Karanja, Njeri; Madlensky, Lisa (2016). Women's Healthy Eating and Living (WHEL) Study. UC San Diego Library Digital Collections. <https://doi.org/10.6075/J0H12ZWX>

Components

- Complete Set of WHEL Data
- Bibliography, updated January 2016
- Data - Baseline
 - Blood Carotenoid
 - Demographics
 - Family History
 - Food from FFQ
 - Food from NDS
 - Health Status
 - Lifestyle Questionnaire
 - Medical/Clinical Measurements
 - Personal Habits
 - Reproductive History
 - Supplement Count
 - Supplement Intake
 - Thoughts and Feelings
 - WHEL Baseline Data Documentation, updated January 2016
- Data - Year 1
 - Blood Carotenoid
 - Clinical Measurements
 - Food from FFQ
 - Food from NDS
 - Health Status
 - Personal Habits
 - Supplement Count

Background

<https://library.ucsd.edu/dc/object/bb2493244b>

“Precision medicine...is an innovative approach to tailoring disease prevention and treatment that takes into account differences in people's genes, environments, and lifestyles.” - FDA

LIBRARY DIGITAL COLLECTIONS ≡

Search Digital Collections 🔍

Women's Healthy Eating and Living (WHEL) Study

Component 1 of 51 ◀ ▶

Complete Set of WHEL Data

File Size 13.1 MB

File Format ZIP Format

Scope And Content This file was modified on 2022-08-23 to include mealtime data and again on 2022-12-21 to include a properly formatted version of the same file.

Download file View file

+

Collection

- Women's Healthy Eating and Living (WHEL) Study

Cite This Work

Pierce, John P.; Faerber, Susan; Wright, Fred A.; Rock, Cheryl L.; Newman, Vicky; Flatt, Shirley W.; Kealey, Sheila; Jones, Vicki E.; Caan, Bette J.; Gold, Ellen B.; Haan, Mary; Hollenbach, Kathryn A.; Jones, Lovell A.; Marshall, James R.; Ritenbaugh, Cheryl; Stefanick, Marcia L.; Thomson, Cynthia; Wasserman, Linda; Natarajan, Loki; Thomas, Ronald G.; Gilpin, Elizabeth; Parker, Barbara A.; Greenberg, E. Robert; Al-Delaimy, Wael K.; Bardwell, Wayne A.; Carlson, Robert W.; Emond, Jennifer A.; Hajek, Richard A.; Karanja, Njeri; Madlensky, Lisa (2016). Women's Healthy Eating and Living (WHEL) Study. UC San Diego Library Digital Collections. <https://doi.org/10.6075/J0H12ZWX>

Components

- Complete Set of WHEL Data
- Bibliography, updated January 2016
- Data - Baseline
 - Blood Carotenoid
 - Demographics
 - Family History
 - Food from FFQ
 - Food from NDS
 - Health Status
 - Lifestyle Questionnaire
 - Medical/Clinical Measurements
 - Personal Habits
 - Reproductive History
 - Supplement Count
 - Supplement Intake
 - Thoughts and Feelings
 - WHEL Baseline Data Documentation, updated January 2016
- Data - Year 1
 - Blood Carotenoid
 - Clinical Measurements
 - Food from FFQ
 - Food from NDS
 - Health Status
 - Personal Habits
 - Supplement Count

Background

<https://library.ucsd.edu/dc/object/bb2493244b>

“Precision medicine...is an innovative approach to tailoring disease prevention and treatment that takes into account differences in people's genes, environments, and lifestyles.” - FDA

Surveyed women w/ varying stages of breast cancer

LIBRARY DIGITAL COLLECTIONS ≡

Search Digital Collections 🔍

Women's Healthy Eating and Living (WHEL) Study

Component 1 of 51 ◀ ▶

Complete Set of WHEL Data

File Size 13.1 MB

File Format ZIP Format

Scope And Content This file was modified on 2022-08-23 to include mealtime data and again on 2022-12-21 to include a properly formatted version of the same file.

Download file View file

+ +

Collection

- Women's Healthy Eating and Living (WHEL) Study

Cite This Work

Pierce, John P.; Faerber, Susan; Wright, Fred A.; Rock, Cheryl L.; Newman, Vicky; Flatt, Shirley W.; Kealey, Sheila; Jones, Vicki E.; Caan, Bette J.; Gold, Ellen B.; Haan, Mary; Hollenbach, Kathryn A.; Jones, Lovell A.; Marshall, James R.; Ritenbaugh, Cheryl; Stefanick, Marcia L.; Thomson, Cynthia; Wasserman, Linda; Natarajan, Loki; Thomas, Ronald G.; Gilpin, Elizabeth; Parker, Barbara A.; Greenberg, E. Robert; Al-Delaimy, Wael K.; Bardwell, Wayne A.; Carlson, Robert W.; Emond, Jennifer A.; Hajek, Richard A.; Karanja, Njeri; Madlensky, Lisa (2016). Women's Healthy Eating and Living (WHEL) Study. UC San Diego Library Digital Collections. <https://doi.org/10.6075/J0H12ZWX>

Components

- Complete Set of WHEL Data
- Bibliography, updated January 2016
- Data - Baseline
 - Blood Carotenoid
 - Demographics
 - Family History
 - Food from FFQ
 - Food from NDS
 - Health Status
 - Lifestyle Questionnaire
 - Medical/Clinical Measurements
 - Personal Habits
 - Reproductive History
 - Supplement Count
 - Supplement Intake
 - Thoughts and Feelings
 - WHEL Baseline Data Documentation, updated January 2016
- Data - Year 1
 - Blood Carotenoid
 - Clinical Measurements
 - Food from FFQ
 - Food from NDS
 - Health Status
 - Personal Habits
 - Supplement Count

Background

<https://library.ucsd.edu/dc/object/bb2493244b>

“Precision medicine...is an innovative approach to tailoring disease prevention and treatment that takes into account differences in people's genes, environments, and lifestyles.” - FDA

Surveyed women w/ varying stages of breast cancer



Dietary intervention (high vegetable & low fat)

LIBRARY DIGITAL COLLECTIONS

Search Digital Collections

Components

- Complete Set of WHEL Data
- Bibliography, updated January 2016
- Data - Baseline
 - Blood Carotenoid
 - Demographics
 - Family History
 - Food from FFQ
 - Food from NDS
 - Health Status
 - Lifestyle Questionnaire
 - Medical/Clinical Measurements
 - Personal Habits
 - Reproductive History
 - Supplement Count
 - Supplement Intake
 - Thoughts and Feelings
 - WHEL Baseline Data Documentation, updated January 2016
- Data - Year 1
 - Blood Carotenoid
 - Clinical Measurements
 - Food from FFQ
 - Food from NDS
 - Health Status
 - Personal Habits
 - Supplement Count

Component 1 of 51

Women's Healthy Eating and Living (WHEL) Study

Complete Set of WHEL Data

File Size 13.1 MB

File Format ZIP Format

Scope And Content This file was modified on 2022-08-23 to include mealtime data and again on 2022-12-21 to include a properly formatted version of the same file.

Download file View file

Collection

- Women's Healthy Eating and Living (WHEL) Study

Cite This Work

Pierce, John P.; Faerber, Susan; Wright, Fred A.; Rock, Cheryl L.; Newman, Vicky; Flatt, Shirley W.; Kealey, Sheila; Jones, Vicki E.; Caan, Bette J.; Gold, Ellen B.; Haan, Mary; Hollenbach, Kathryn A.; Jones, Lovell A.; Marshall, James R.; Ritenbaugh, Cheryl; Stefanick, Marcia L.; Thomson, Cynthia; Wasserman, Linda; Natarajan, Loki; Thomas, Ronald G.; Gilpin, Elizabeth; Parker, Barbara A.; Greenberg, E. Robert; Al-Delaimy, Wael K.; Bardwell, Wayne A.; Carlson, Robert W.; Emond, Jennifer A.; Hajek, Richard A.; Karanja, Njeri; Madlensky, Lisa (2016). Women's Healthy Eating and Living (WHEL) Study. UC San Diego Library Digital Collections. <https://doi.org/10.6075/J0H12ZXW>

Background

<https://library.ucsd.edu/dc/object/bb2493244b>

“Precision medicine...is an innovative approach to tailoring disease prevention and treatment that takes into account differences in people's genes, environments, and lifestyles.” - FDA

Surveyed women w/ varying stages of breast cancer



Dietary intervention (high vegetable & low fat)



Assess impact on the likelihood of additional breast cancer events

LIBRARY DIGITAL COLLECTIONS

Search Digital Collections

Components

- Complete Set of WHEL Data
- Bibliography, updated January 2016
- Data - Baseline**
 - Blood Carotenoid
 - Demographics
 - Family History
 - Food from FFQ
 - Food from NDS
 - Health Status
 - Lifestyle Questionnaire
 - Medical/Clinical Measurements
 - Personal Habits
 - Reproductive History
 - Supplement Count
 - Supplement Intake
 - Thoughts and Feelings
 - WHEL Baseline Data Documentation, updated January 2016
- Data - Year 1**
 - Blood Carotenoid
 - Clinical Measurements
 - Food from FFQ
 - Food from NDS
 - Health Status
 - Personal Habits
 - Supplement Count

Women's Healthy Eating and Living (WHEL) Study

Component 1 of 51

Complete Set of WHEL Data

File Size 13.1 MB

File Format ZIP Format

Scope And Content This file was modified on 2022-08-23 to include mealtime data and again on 2022-12-21 to include a properly formatted version of the same file.

[Download file](#) [View file](#)

Collection

- Women's Healthy Eating and Living (WHEL) Study

Cite This Work

Pierce, John P.; Faerber, Susan; Wright, Fred A.; Rock, Cheryl L.; Newman, Vicky; Flatt, Shirley W.; Kealey, Sheila; Jones, Vicki E.; Caan, Bette J.; Gold, Ellen B.; Haan, Mary; Hollenbach, Kathryn A.; Jones, Lovell A.; Marshall, James R.; Ritenbaugh, Cheryl; Stefanick, Marcia L.; Thomson, Cynthia; Wasserman, Linda; Natarajan, Loki; Thomas, Ronald G.; Gilpin, Elizabeth; Parker, Barbara A.; Greenberg, E. Robert; Al-Delaimy, Wael K.; Bardwell, Wayne A.; Carlson, Robert W.; Emond, Jennifer A.; Hajek, Richard A.; Karanja, Njeri; Madlensky, Lisa (2016). Women's Healthy Eating and Living (WHEL) Study. UC San Diego Library Digital Collections. <https://doi.org/10.6075/J0H12ZXW>

Outline

- **Background & Data + Research Question**
- Exploratory Analysis
 - Uni-, Bi-, Multivariate Analyses
- Predictive Analysis
 - Machine Learning Models
 - Solutions to Imbalanced Data
 - Performance Metrics & Assessment
- Limitations
- Closing & Acknowledgements

The Dataset: 4 subparts

The Dataset: 4 subparts

Baseline

- ‘Starting point’
 - 12 Datasets
 - 3088 patients
-

The Dataset: 4 subparts

Baseline

- ‘Starting point’
- 12 Datasets
- 3088 patients

Year 1 & Year 4

- After 1 & 4 years
 - 8 Datasets each
 - ~2000-3000 patients
-

The Dataset: 4 subparts

- | Baseline | Year 1 & Year 4 | Final |
|--|---|--|
| <ul style="list-style-type: none">○ ‘Starting point’○ 12 Datasets○ 3088 patients | <ul style="list-style-type: none">○ After 1 & 4 years○ 8 Datasets each○ ~2000-3000 patients | <ul style="list-style-type: none">○ Condition after the study○ 2 Datasets○ 3088 patients |
-

The Dataset: 4 subparts

Baseline	Year 1 & Year 4	Final
○ 'Starting point'	○ After 1 & 4 years	○ Condition after the study
○ 12 Datasets	○ 8 Datasets each	○ 2 Datasets
○ 3088 patients	○ ~2000-3000 patients	○ 3088 patients

Includes health status, dietary patterns, medical measurements, blood carotenoids, emotional health, etc.

The Dataset: 4 subparts

Baseline	Year 1 & Year 4	Final
○ 'Starting point'	○ After 1 & 4 years	○ Condition after the study
○ 12 Datasets	○ 8 Datasets each	○ 2 Datasets
○ 3088 patients	○ ~2000-3000 patients	○ 3088 patients

Includes health status, dietary patterns, medical measurements, blood carotenoids, emotional health, etc.

30 Datasets with ~3000 observations (rows) and ~27 variables (columns) each

The Dataset: Baseline Medical/Clinical Measurements

In [21]: `base_measure.head(10)`

Out[21]:

	ID	Lump/Mast	Radiation	Chemo	Tumor Type	Estr Recep	Prog Recep	Node Exam	Node Pos	Stage	...	Weight	BMI	Waist	Hip	Pulse	Bld Pres -Sys	Bld Pres -Dias	Menopause	An Es U
0	1002	1	2	2	3	1	1	18	1.0	II	...	65.50	23.0697	70.6	94.5	-9	-9	-9	1	
1	1003	1	2	2	3	1	1	14	0.0	I	...	60.00	23.4375	87.0	94.0	-9	-9	-9	2	
2	1005	2	1	1	3	3	3	29	0.0	II	...	68.63	32.4181	92.0	110.0	-9	-9	-9	2	
3	1007	2	1	2	4	0	0	21	1.0	II	...	55.45	23.3808	70.0	99.0	-9	-9	-9	2	
4	1008	2	1	1	3	1	1	16	0.0	I	...	64.40	25.4737	75.6	100.0	-9	-9	-9	2	
5	1009	2	1	1	3	1	1	6	1.0	II	...	81.90	31.5568	98.7	118.8	-9	-9	-9	3	
6	1010	2	1	2	3	1	1	23	4.0	II	...	72.27	31.4871	86.0	111.5	-9	-9	-9	2	
7	1011	1	2	2	1	0	1	8	6.0	II	...	59.00	23.0469	73.5	95.0	-9	-9	-9	3	
8	1012	1	2	1	3	1	1	9	0.0	I	...	62.50	27.2304	83.0	108.0	-9	-9	-9	2	
9	1015	2	1	2	3	1	1	8	4.0	II	...	82.00	27.9121	86.0	106.0	-9	-9	-9	2	

10 rows × 22 columns

The Dataset: Baseline Medical/Clinical Measurements

```
In [21]: base_measure.head(10)
```

Out[21]:

ID	Lump/Mast	Radiation	Chemo	Tumor Type	Estr Recep	Prog Recep	Node Exam	Node Pos	Stage	...	Weight	BMI	Waist	Hip	Pulse	Bld Pres-Sys	Bld Pres-Dias	Menopause	An Es U
0	1002	1	2	2	3	1	1	18	1.0	II	...	65.50	23.0697	70.6	94.5	-9	-9	-9	1
1	1003	1	2	2	3	1	1	14	0.0	I	...	60.00	23.4375	87.0	94.0	-9	-9	-9	2
2	1005	2	1	1	3	3	3	29	0.0	II	...	68.63	32.4181	92.0	110.0	-9	-9	-9	2
3	1007	2	1	2	4	0	0	21	1.0	II	...	55.45	23.3808	70.0	99.0	-9	-9	-9	2
4	1008	2	1	1	3	1	1	16	0.0	I	...	64.40	25.4737	75.6	100.0	-9	-9	-9	2
5	1009	2	1	1	3	1	1	6	1.0	II	...	81.90	31.5568	98.7	118.8	-9	-9	-9	3
6	1010	2	1	2	3	1	1	23	4.0	II	...	72.27	31.4871	86.0	111.5	-9	-9	-9	2
7	1011	1	2	2	1	0	1	8	6.0	II	...	59.00	23.0469	73.5	95.0	-9	-9	-9	3
8	1012	1	2	1	3	1	1	9	0.0	I	...	62.50	27.2304	83.0	108.0	-9	-9	-9	2
9	1015	2	1	2	3	1	1	8	4.0	II	...	82.00	27.9121	86.0	106.0	-9	-9	-9	2

10 rows × 22 columns

The Dataset: Baseline Medical/Clinical Measurements

In [21]: `base_measure.head(10)`

Out[21]:

	ID	Lum /Mast	Radiation	Chemo	Tumor Type	Estr Recep	Prog Recep	Node Exam	Node Pos	Stage	...	Weight	BMI	Waist	Hip	Pulse	Bld Pres -Sys	Bld Pres -Dias	Menopause	An Es U
0	1002	1	2	2	3	1	1	18	1.0	II	...	65.50	23.0697	70.6	94.5	-9	-9	-9	1	
1	1003	1	2	2	3	1	1	14	0.0	I	...	60.00	23.4375	87.0	94.0	-9	-9	-9	2	
2	1005	2	1	1	3	3	3	29	0.0	II	...	68.63	32.4181	92.0	110.0	-9	-9	-9	2	
3	1007	2	1	2	4	0	0	21	1.0	II	...	55.45	23.3808	70.0	99.0	-9	-9	-9	2	
4	1008	2	1	1	3	1	1	16	0.0	I	...	64.40	25.4737	75.6	100.0	-9	-9	-9	2	
5	1009	2	1	1	3	1	1	6	1.0	II	...	81.90	31.5568	98.7	118.8	-9	-9	-9	3	
6	1010	2	1	2	3	1	1	23	4.0	II	...	72.27	31.4871	86.0	111.5	-9	-9	-9	2	
7	1011	1	2	2	1	0	1	8	6.0	II	...	59.00	23.0469	73.5	95.0	-9	-9	-9	3	
8	1012	1	2	1	3	1	1	9	0.0	I	...	62.50	27.2304	83.0	108.0	-9	-9	-9	2	
9	1015	2	1	2	3	1	1	8	4.0	II	...	82.00	27.9121	86.0	106.0	-9	-9	-9	2	

10 rows × 22 columns

Patient ID's - connected and included in ALL
30 datasets

Based on the WHEL's Study Dataset...

Based on the WHEL's Study Dataset...

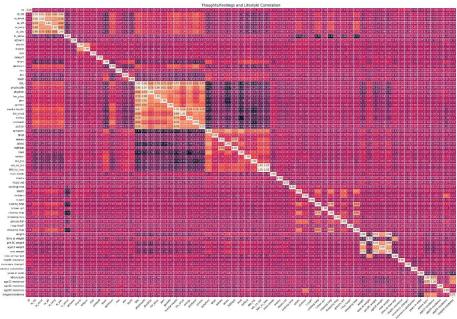
Is there some **predictive power** in a patient's diet,
clinical measurements, and lifestyle (over time) with
their recurrence in their breast cancer condition?

Outline

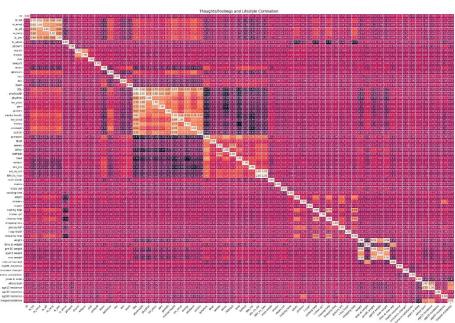
- Background & Data + Research Question
- **Exploratory Analysis**
 - Uni-, Bi-, Multivariate Analyses
- Predictive Analysis
 - Machine Learning Models
 - Solutions to Imbalanced Data
 - Performance Metrics & Assessment
- Limitations
- Closing & Acknowledgements

Uni-, Bi-, Multivariate Analyses

Uni-, Bi-, Multivariate Analyses

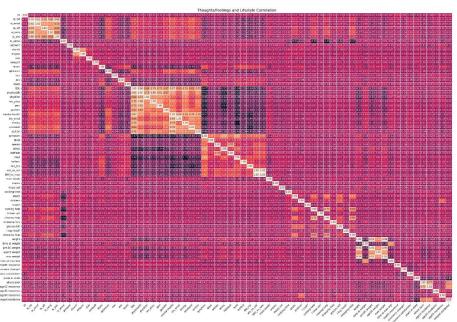


Uni-, Bi-, Multivariate Analyses



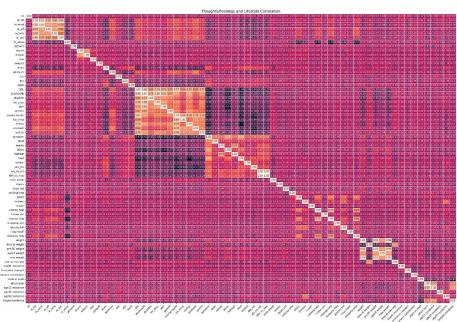
weight -	1.00	-0.36	0.91	0.54	0.94
time at weight -	-0.36	1.00	-0.30	-0.10	-0.35
pre-BC weight -	0.91	-0.30	1.00	0.56	0.92
age18 weight -	0.54	-0.10	0.56	1.00	0.59
	weight	time at weight	pre-BC weight	age18 weight	max weight
	time at				

Uni-, Bi-, Multivariate Analyses



weight -	1.00	-0.36	0.91	0.54	0.94
time at weight -	-0.36	1.00	-0.30	-0.10	-0.35
pre-BC weight -	0.91	-0.30	1.00	0.56	0.92
age18 weight -	0.54	-0.10	0.56	1.00	0.59
weight					
time at weight					
pre-BC weight					
age18 weight					
max weight					
time at					

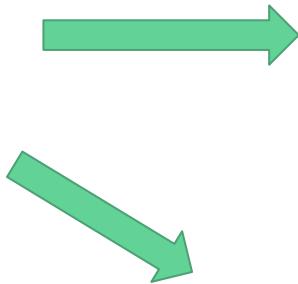
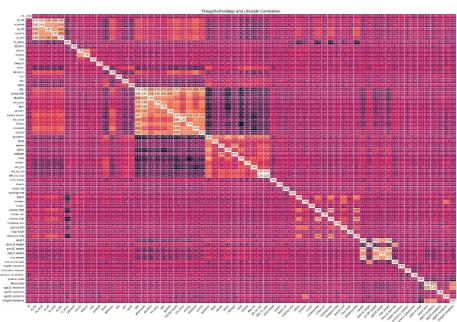
Uni-, Bi-, Multivariate Analyses



	weight	-0.36	0.91	0.54	0.94
time at weight	-0.36	1.00	-0.30	-0.10	-0.35
pre-BC weight	0.91	-0.30	1.00	0.56	0.92
age18 weight	0.54	-0.10	0.56	1.00	0.59
weight					
time at weight					
pre-BC weight					
age18 weight					
max weight					
time at					

	QOL	0.90	0.63	0.79	0.70	0.67	0.87	0.64
physhealth	1.00							
physfctn	0.90	1.00						
lim_phys	0.63	0.75	1.00					
lim_emot	0.79	0.88	0.51	1.00				
energy	0.70	0.80	0.52	0.59	1.00			
genhlth	0.67	0.68	0.46	0.43	0.42	1.00		
mental health	0.87	0.58	0.36	0.51	0.43	0.51	1.00	
emotwell	0.64	0.33	0.18	0.32	0.22	0.28	0.84	1.00
socfctn	0.75	0.60	0.44	0.49	0.45	0.55	0.74	0.40
QOL								
physhealth								
physfctn								
lim_phys								
pain								
genhlth								
mental health								
lim_emot								

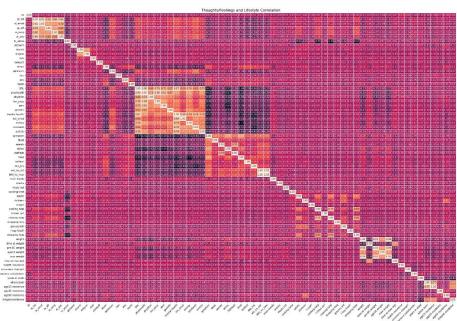
Uni-, Bi-, Multivariate Analyses



	weight	-1.00	-0.36	0.91	0.54	0.94	
time at weight	-0.36	1.00	-0.30	-0.10	-0.35		
pre-BC weight	0.91	-0.30	1.00	0.56	0.92		
age18 weight	0.54	-0.10	0.56	1.00	0.59		
max weight						1.00	
time at							1.00

	QOL	physhealth	physfctn	lim_phys	pain	genhlth	mental health	lim_emot
QOL	1.00	0.90	0.63	0.79	0.70	0.67	0.87	0.64
physhealth	0.90	1.00	0.75	0.88	0.80	0.68	0.58	0.33
physfctn	0.63	0.75	1.00	0.51	0.52	0.46	0.36	0.18
lim_phys	0.79	0.88	0.51	1.00	0.59	0.43	0.51	0.32
pain	0.70	0.80	0.52	0.59	1.00	0.42	0.43	0.22
genhlth	0.67	0.68	0.46	0.43	0.42	1.00	0.51	0.28
mental health	0.87	0.58	0.36	0.51	0.43	0.51	1.00	0.84
lim_emot	0.64	0.33	0.18	0.32	0.22	0.28	0.84	1.00
energy	0.75	0.60	0.44	0.49	0.45	0.55	0.74	0.40
emotwell	0.62	0.35	0.19	0.27	0.26	0.41	0.78	0.55
socfctn	0.75	0.58	0.35	0.54	0.46	0.41	0.77	0.49

Uni-, Bi-, Multivariate Analyses

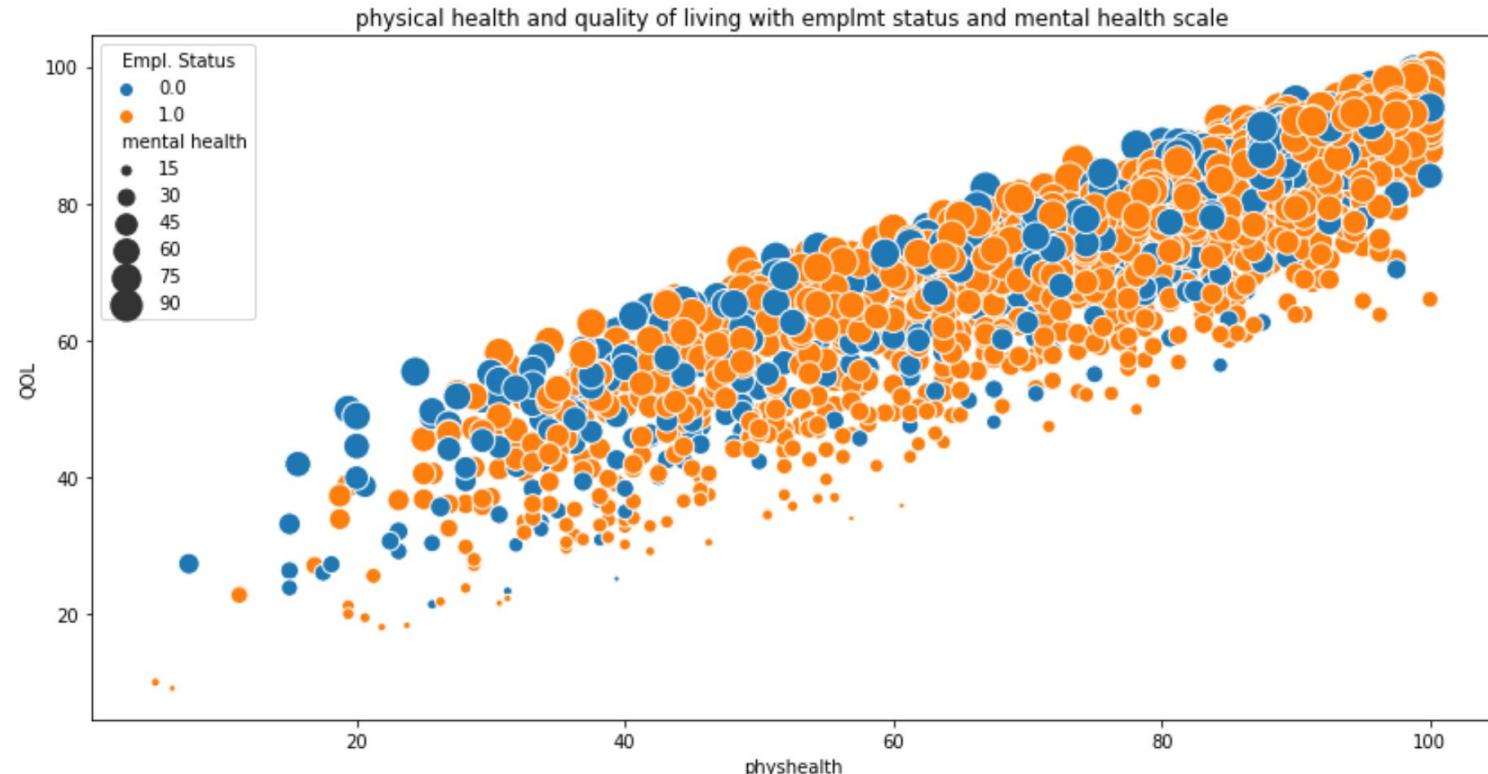


	QOL	physhealth	physfctn	lim_phys	pain	genhlth	mental health	lim_emot
QOL	1.00	0.90	0.63	0.79	0.70	0.67	0.87	0.64
physhealth	0.90	1.00	0.75	0.88	0.80	0.68	0.58	0.33
physfctn	0.63	0.75	1.00	0.51	0.52	0.46	0.36	0.18
lim_phys	0.79	0.88	0.51	1.00	0.59	0.43	0.51	0.32
pain	0.70	0.80	0.52	0.59	1.00	0.42	0.43	0.22
genhlth	0.67	0.68	0.46	0.43	0.42	1.00	0.51	0.28
mental health	0.87	0.58	0.36	0.51	0.43	0.51	1.00	0.84
lim_emot	0.64	0.33	0.18	0.32	0.22	0.28	0.84	1.00
energy	0.75	0.60	0.44	0.49	0.45	0.55	0.74	0.40
emotwell	0.62	0.35	0.19	0.27	0.26	0.41	0.78	0.55
socfctn	0.75	0.58	0.35	0.54	0.46	0.41	0.77	0.49

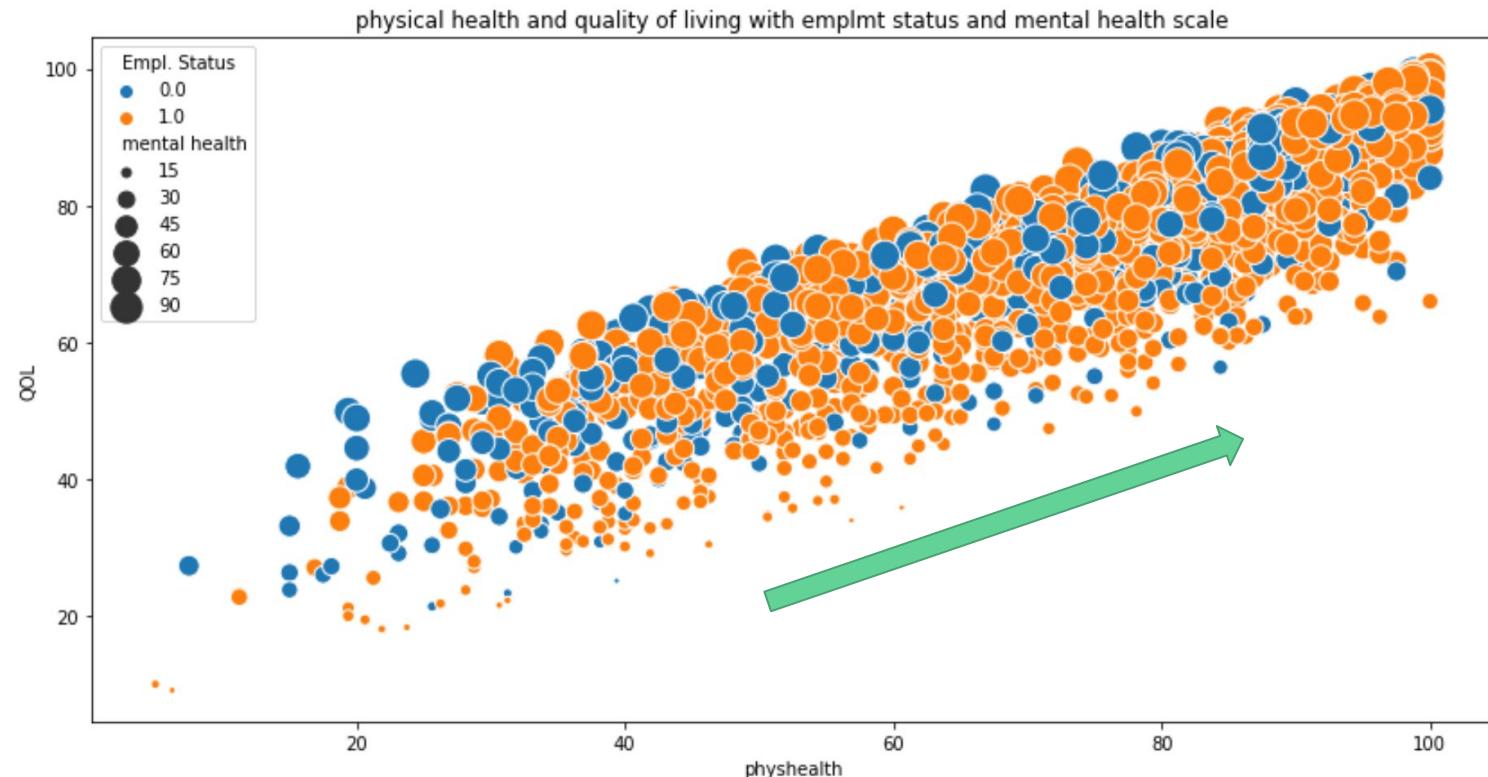
weight	1.00	-0.36	0.91	0.54	0.94
time at weight	-0.36	1.00	-0.30	-0.10	-0.35
pre-BC weight	0.91	-0.30	1.00	0.56	0.92
age18 weight	0.54	-0.10	0.56	1.00	0.59

sanity checks

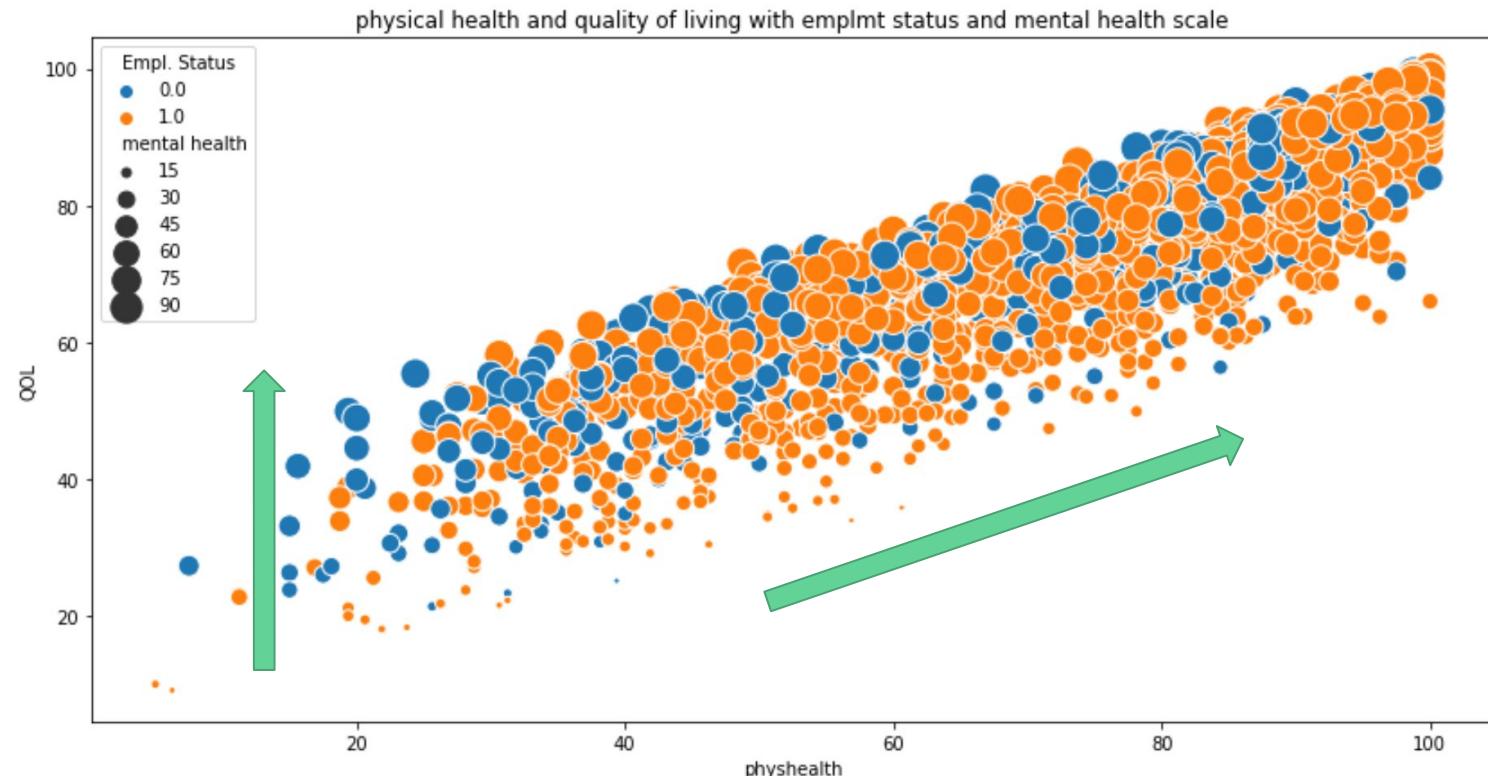
Uni-, Bi-, Multivariate Analyses



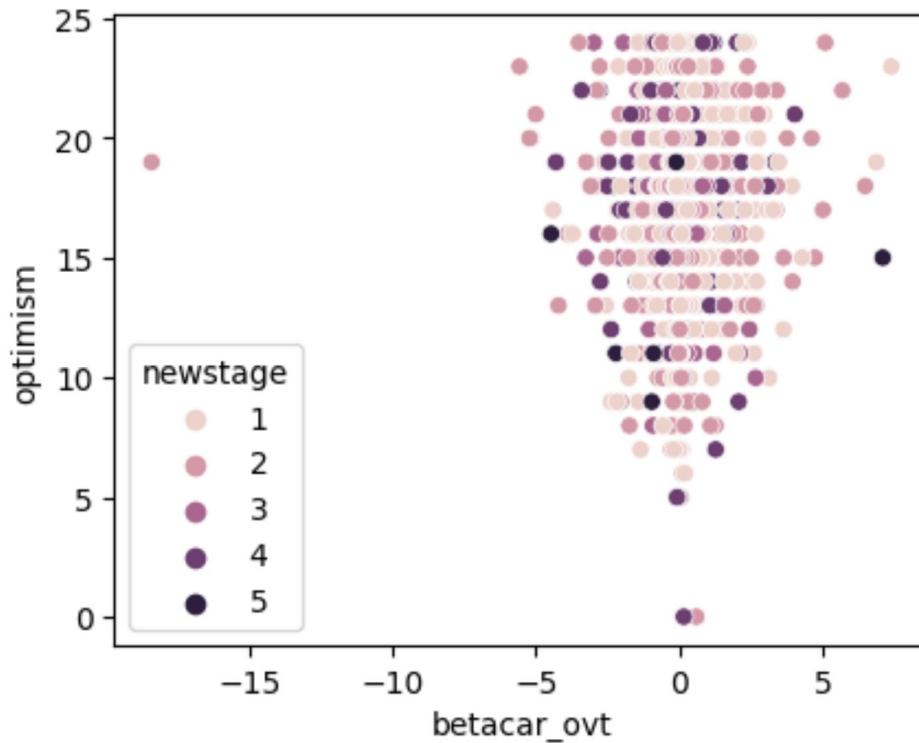
Uni-, Bi-, Multivariate Analyses



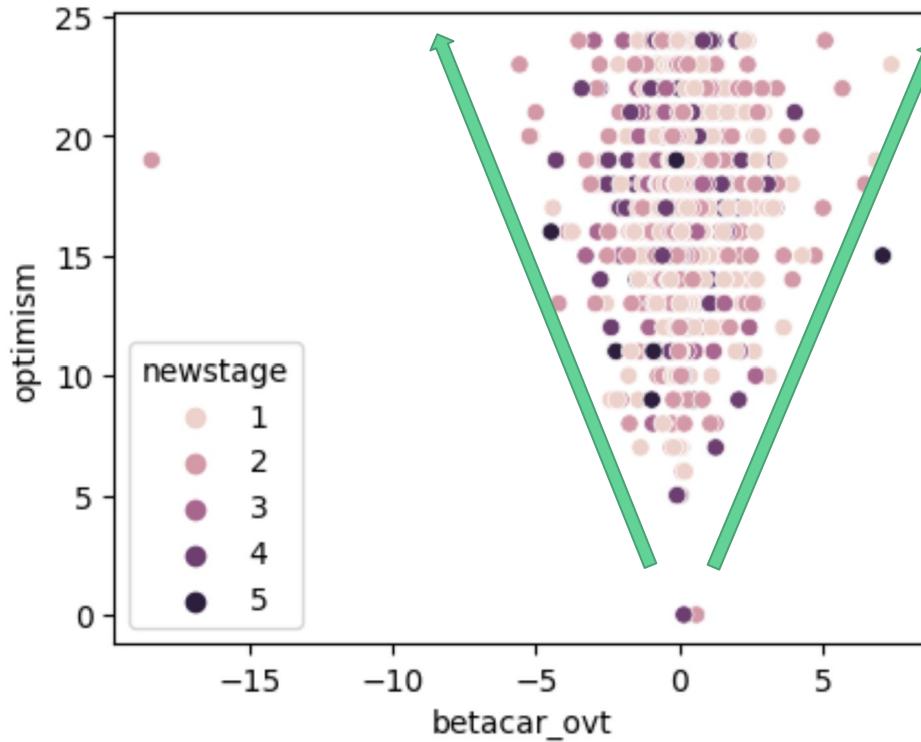
Uni-, Bi-, Multivariate Analyses



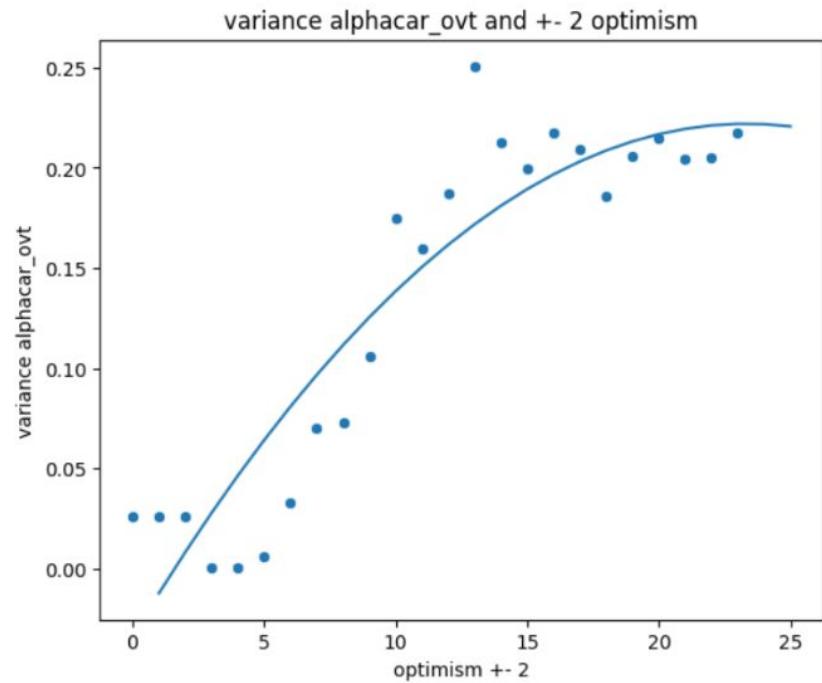
Uni-, Bi-, Multivariate Analyses



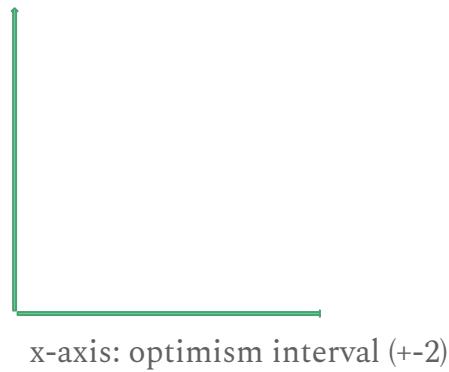
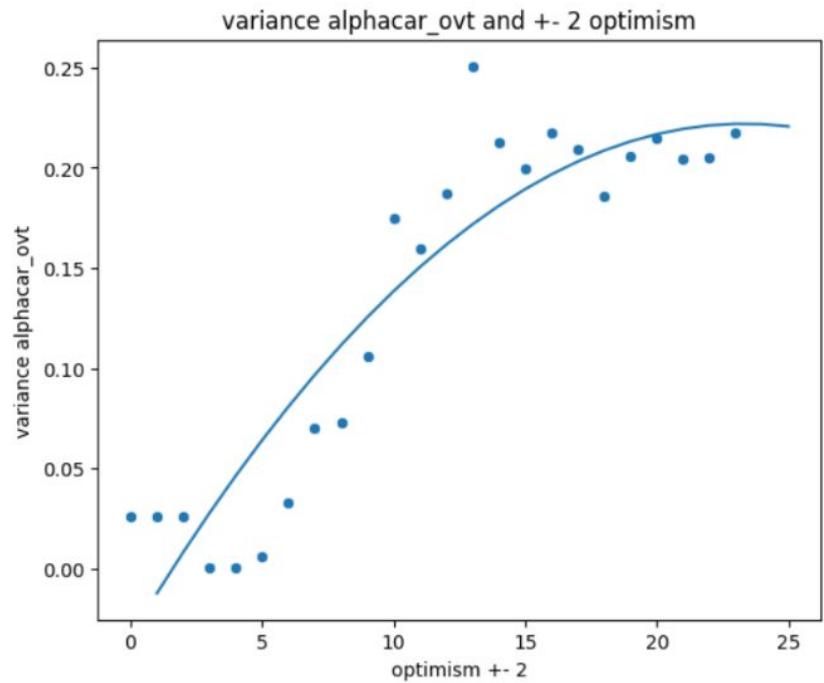
Uni-, Bi-, Multivariate Analyses



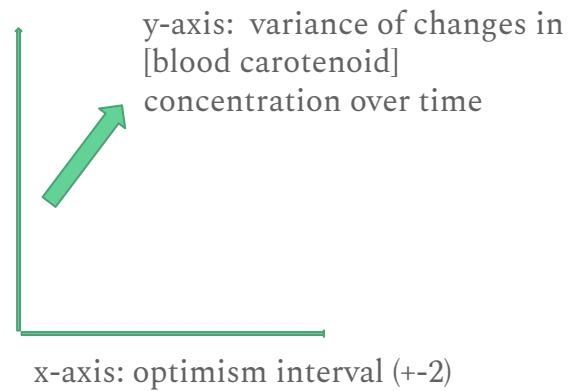
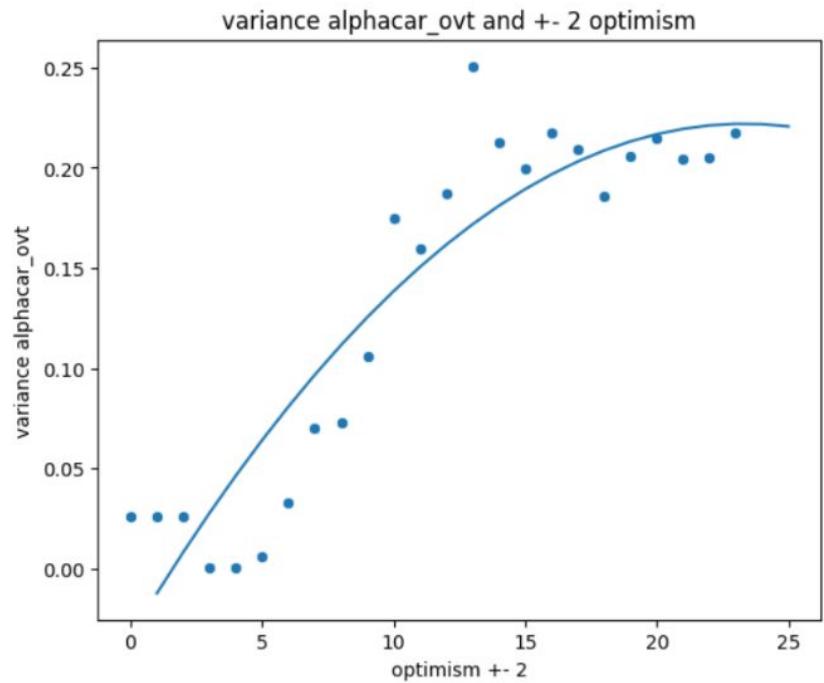
Uni-, Bi-, Multivariate Analyses



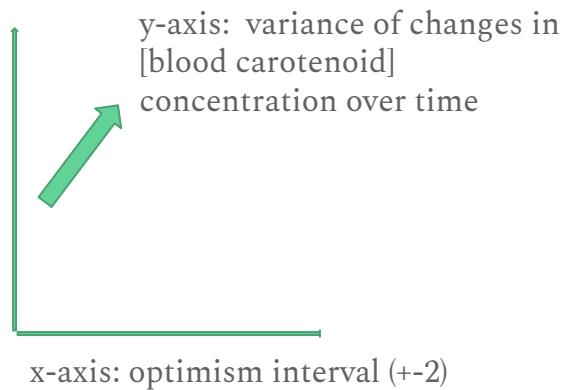
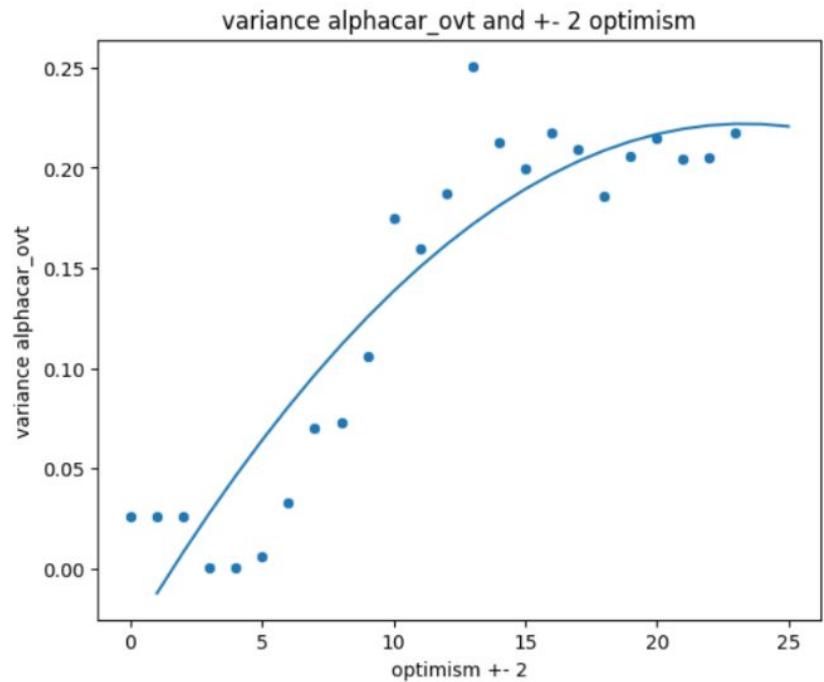
Uni-, Bi-, Multivariate Analyses



Uni-, Bi-, Multivariate Analyses



Uni-, Bi-, Multivariate Analyses



→ Variances of changes in alpha-carotene concentrations increases over increasing optimism intervals (+-2), then flatten out / slightly decrease

Outline

- Background & Data + Research Question
- Exploratory Analysis
 - Uni-, Bi-, Multivariate Analyses
- Predictive Analysis
 - Machine Learning Models
 - Solutions to Imbalanced Data
 - Performance Metrics & Assessment
- Limitations
- Closing & Acknowledgements

Binary Classification Model: Preprocessing



Binary Classification Model: Preprocessing

Target variable: breast cancer status

0 : No evidence of recurrence

> 0 : Local, Regional, Distant recurrence

```
[84]: ┌─ endpoints_df.head(10)
```

Out[84]:

	id	intgrp	vitality	brcastatus	othcstatus
0	1002	3	1	0	0.0
1	1003	3	1	0	0.0
2	1005	4	1	0	0.0
3	1007	3	1	0	0.0
4	1008	4	1	0	0.0
5	1009	4	1	3	0.0
6	1010	3	0	4	0.0
7	1011	4	0	4	0.0
8	1012	4	1	0	0.0
9	1015	3	0	2	0.0



Binary Classification Model: Preprocessing

Target variable: breast cancer status

0 : No evidence of recurrence

> 0 : Local, Regional, Distant recurrence

```
[84]: ┌─ endpoints_df.head(10)
```

Out[84]:

	id	irtgpr	vitality	brcaststatus	othcstatus
0	1002	3	1	0	0.0
1	1003	3	1	0	0.0
2	1005	4	1	0	0.0
3	1007	3	1	0	0.0
4	1008	4	1	0	0.0
5	1009	4	1	3	0.0
6	1010	3	0	4	0.0
7	1011	4	0	4	0.0
8	1012	4	1	0	0.0
9	1015	3	0	2	0.0

Binary Classification Model: Preprocessing

Target variable: breast cancer status

0 : No evidence of recurrence

> 0 : Local, Regional, Distant recurrence

[84]: endpoints_df.head(10)

Out[84]:

	id	irtgpr	vitality	brcaststatus	othcstatus
0	1002	3	1	0	0.0
1	1003	3	1	0	0.0
2	1005	4	1	0	0.0
3	1007	3	1	0	0.0
4	1008	4	1	0	0.0
5	1009	4	1	3	0.0
6	1010	3	0	4	0.0
7	1011	4	0	4	0.0
8	1012	4	1	0	0.0
9	1015	3	0	2	0.0

Features: Year 4 - Baseline

Binary Classification Model: Preprocessing

Target variable: breast cancer status

0 : No evidence of recurrence

> 0 : Local, Regional, Distant recurrence

```
[84]: endpoints_df.head(10)
```

Out[84]:

	id	intgrp	vitality	brcaststatus	othcstatus
0	1002	3	1	0	0.0
1	1003	3	1	0	0.0
2	1005	4	1	0	0.0
3	1007	3	1	0	0.0
4	1008	4	1	0	0.0
5	1009	4	1	3	0.0
6	1010	3	0	4	0.0
7	1011	4	0	4	0.0
8	1012	4	1	0	0.0
9	1015	3	0	2	0.0

Features: Year 4 - Baseline

- Metrics: clinical measurements, blood carotenoids
- Diet: Food From NDS
- Lifestyle: personal habits, emotional health,
- quality of life (QOL)

Binary Classification Model: Preprocessing (Merging)

Binary Classification Model: Preprocessing (Merging)

	id	recurrence
0	1002	0
1	1003	0
2	1005	0
3	1007	0
4	1008	0
...
3083	13241	0
3084	13242	0
3085	13243	0
3086	13244	0
3087	13245	0

3088 rows × 2 columns

Binary Classification Model: Preprocessing (Merging)

	id	recurrence
0	1002	0
1	1003	0
2	1005	0
3	1007	0
4	1008	0
...
3083	13241	0
3084	13242	0
3085	13243	0
3086	13244	0
3087	13245	0

3088 rows × 2 columns

Target variable

Binary Classification Model: Preprocessing (Merging)

	id	recurrence
0	1002	0
1	1003	0
2	1005	0
3	1007	0
4	1008	0
...
3083	13241	0
3084	13242	0
3085	13243	0
3086	13244	0
3087	13245	0

3088 rows × 2 columns

Baseline (*Clinical measurements*)

ID	B_Height	B_Weight	B_BMI	B_Waist	B_Hip	B_Pulse
1245	159.00	61.6	24.3661	74.3	99.5	32
1248	171.46	61.2	20.8174	69.2	96.0	33
1249	161.50	60.4	23.1575	71.5	95.1	33
1250	168.40	45.3	15.9740	61.0	86.5	30
1253	161.20	78.6	30.2477	90.5	109.0	32



Year 4 (*Clinical measurements*)

ID	y4_Heightcm	y4_Weightkg	y4_BMI	y4_Waistcm	y4_Hipcm	y4_Pulse30
1002	166.50000	65.000000	23.446871	71.400002	97.900002	28
1007	153.39999	64.900002	27.579985	83.400002	105.600000	37
1010	154.30000	72.800003	30.577330	84.300003	104.000000	28
1012	151.50000	68.500000	29.844570	96.199997	110.000000	33
1018	162.89999	51.500000	19.407299	70.800003	88.599998	28

Binary Classification Model: Preprocessing (Merging)

	id	recurrence
0	1002	0
1	1003	0
2	1005	0
3	1007	0
4	1008	0
...
3083	13241	0
3084	13242	0
3085	13243	0
3086	13244	0
3087	13245	0

3088 rows × 2 columns

Baseline (*Clinical measurements*)

ID	B_Height	B_Weight	B_BMI	B_Waist	B_Hip	B_Pulse
1245	159.00	61.6	24.3661	74.3	99.5	32
1248	171.46	61.2	20.8174	69.2	96.0	33
1249	161.50	60.4	23.1575	71.5	95.1	33
1250	168.40	45.3	15.9740	61.0	86.5	30
1253	161.20	78.6	30.2477	90.5	109.0	32



Year 4 (*Clinical measurements*)

ID	y4_Heightcm	y4_Weightkg	y4_BMI	y4_Waistcm	y4_Hipcm	y4_Pulse30
1002	166.50000	65.000000	23.446871	71.400002	97.900002	28
1007	153.39999	64.900002	27.579985	83.400002	105.600000	37
1010	154.30000	72.800003	30.577330	84.300003	104.000000	28
1012	151.50000	68.500000	29.844570	96.199997	110.000000	33
1018	162.89999	51.500000	19.407299	70.800003	88.599998	28

id recurrence y4_Heightcm y4_Weightkg

0	1245	0	-1.10001	2.800002
1	1249	0	-0.30000	7.600000
2	1250	0	-1.90000	-1.399998
3	1253	0	-0.20000	1.900000
4	1254	0	0.39999	-3.000002

Binary Classification Model: Preprocessing (Merging)

	id	recurrence
0	1002	0
1	1003	0
2	1005	0
3	1007	0
4	1008	0
...
3083	13241	0
3084	13242	0
3085	13243	0
3086	13244	0
3087	13245	0

3088 rows × 2 columns

Baseline (*Clinical measurements*)

ID	B_Height	B_Weight	B_BMI	B_Waist	B_Hip	B_Pulse
1245	159.00	61.6	24.3661	74.3	99.5	32
1248	171.46	61.2	20.8174	69.2	96.0	33
1249	161.50	60.4	23.1575	71.5	95.1	33
1250	168.40	45.3	15.9740	61.0	86.5	30
1253	161.20	78.6	30.2477	90.5	109.0	32

Year 4 (*Clinical measurements*)

ID	y4_Heightcm	y4_Weightkg	y4_BMI	y4_Waistcm	y4_Hipcm	y4_Pulse30
1002	166.50000	65.000000	23.446871	71.400002	97.900002	28
1007	153.39999	64.900002	27.579985	83.400002	105.600000	37
1010	154.30000	72.800003	30.577330	84.300003	104.000000	28
1012	151.50000	68.500000	29.844570	96.199997	110.000000	33
1018	162.89999	51.500000	19.407299	70.800003	88.599998	28

Merged based on 'ID'

	id	recurrence	y4_Heightcm	y4_Weightkg
0	1245	0	-1.10001	2.800002
1	1249	0	-0.30000	7.600000
2	1250	0	-1.90000	-1.399998
3	1253	0	-0.20000	1.900000
4	1254	0	0.39999	-3.000002

(Year 4) - (Baseline)

Binary Classification Model: Preprocessing (Merging)

Repeat merging for
rest of dataframes ...

Binary Classification Model: Preprocessing (Merging)

Repeat merging for
rest of dataframes ...

Shape: (655, 68)



2433 rows lost !!

	ID	QOL_ovt	energy_ovt	genhlth_ovt	optimism_ovt	pain_ovt	socfctn_ovt	lutein_ovt
0	1245	-32.312500	-20.0	-40.0	-2.0	-12.5	-50.0	-0.1697
1	1256	-36.687500	-5.0	10.0	-3.0	-50.0	-37.5	0.0980
2	1265	-6.312500	-5.0	-10.0	2.0	-12.5	0.0	0.0349
3	1266	0.437500	-10.0	0.0	2.0	25.0	-12.5	-0.1125
4	1276	30.646484	30.0	5.0	-1.0	25.0	37.5	0.0766

Binary Classification Model: Preprocessing (Merging)

Repeat merging for
rest of dataframes ...

Shape: (655, 68)

→ 2433 rows lost !!

	ID	QOL_ovt	energy_ovt	genhlth_ovt	optimism_ovt	pain_ovt	socfctn_ovt	lutein_ovt
0	1245	-32.312500	-20.0	-40.0	-2.0	-12.5	-50.0	-0.1697
1	1256	-36.687500	-5.0	10.0	-3.0	-50.0	-37.5	0.0980
2	1265	-6.312500	-5.0	-10.0	2.0	-12.5	0.0	0.0349
3	1266	0.437500	-10.0	0.0	2.0	25.0	-12.5	-0.1125
4	1276	30.646484	30.0	5.0	-1.0	25.0	37.5	0.0766

In [18]: ovt['recurrence_y'].value_counts()

Out[18]: 0 629
1 26
Name: recurrence_y, dtype: int64

629 (96%) had no recurrence
26 (4%) had recurrence

Outline

- Background & Data + Research Question
- Exploratory Analysis
 - Uni-, Bi-, Multivariate Analyses
- **Predictive Analysis**
 - Machine Learning Models
 - **Solutions to Imbalanced Data**
 - Performance Metrics & Assessment
- Limitations
- Closing & Acknowledgements

Handling Imbalanced Data: Procedures

Handling Imbalanced Data: Procedures

Original, Preprocessed Data

	ID	QOL_ovt	energy_ovt	genhlth_ovt	optimism_ovt	pain_ovt	socfctn_ovt
0	1245	-32.312500	-20.0	-40.0	-2.0	-12.5	-50.0
1	1256	-36.687500	-5.0	10.0	-3.0	-50.0	-37.5
2	1265	-6.312500	-5.0	-10.0	2.0	-12.5	0.0
3	1266	0.437500	-10.0	0.0	2.0	25.0	-12.5
4	1276	30.646484	30.0	5.0	-1.0	25.0	37.5

Handling Imbalanced Data: Procedures

Original, Preprocessed Data

	ID	QOL_ovt	energy_ovt	genhlth_ovt	optimism_ovt	pain_ovt	socfctn_ovt
0	1245	-32.312500	-20.0	-40.0	-2.0	-12.5	-50.0
1	1256	-36.687500	-5.0	10.0	-3.0	-50.0	-37.5
2	1265	-6.312500	-5.0	-10.0	2.0	-12.5	0.0
3	1266	0.437500	-10.0	0.0	2.0	25.0	-12.5
4	1276	30.646484	30.0	5.0	-1.0	25.0	37.5

*Apply Resample
Method*



Handling Imbalanced Data: Procedures

Original, Preprocessed Data

	ID	QOL_ovt	energy_ovt	genhlth_ovt	optimism_ovt	pain_ovt	socfctn_ovt
0	1245	-32.312500	-20.0	-40.0	-2.0	-12.5	-50.0
1	1256	-36.687500	-5.0	10.0	-3.0	-50.0	-37.5
2	1265	-6.312500	-5.0	-10.0	2.0	-12.5	0.0
3	1266	0.437500	-10.0	0.0	2.0	25.0	-12.5
4	1276	30.646484	30.0	5.0	-1.0	25.0	37.5

*Apply Resample
Method*



New, Resampled Data

	ID	QOL_ovt	energy_ovt	genhlth_ovt	optimism_ovt	pain_ovt	socfctn_ovt	target
90	1563	-33.000000		-15.0	-20.0	-2.0	-50.0	-50.0
280	5569	23.187500		15.0	25.0	1.0	37.5	0.0
301	7249	13.312500		20.0	0.0	1.0	12.5	25.0
118	3299	7.292969		0.0	0.0	2.0	0.0	0.0
323	7305	-1.062500		0.0	5.0	4.0	-12.5	0.0

Handling Imbalanced Data: Procedures

Original, Preprocessed Data

	ID	QOL_ovt	energy_ovt	genhlth_ovt	optimism_ovt	pain_ovt	socfctn_ovt
0	1245	-32.312500	-20.0	-40.0	-2.0	-12.5	-50.0
1	1256	-36.687500	-5.0	10.0	-3.0	-50.0	-37.5
2	1265	-6.312500	-5.0	-10.0	2.0	-12.5	0.0
3	1266	0.437500	-10.0	0.0	2.0	25.0	-12.5
4	1276	30.646484	30.0	5.0	-1.0	25.0	37.5

*Apply Resample
Method*



New, Resampled Data

	ID	QOL_ovt	energy_ovt	genhlth_ovt	optimism_ovt	pain_ovt	socfctn_ovt	target
90	1563	-33.000000		-15.0	-20.0	-2.0	-50.0	-50.0
280	5569	23.187500		15.0	25.0	1.0	37.5	0.0
301	7249	13.312500		20.0	0.0	1.0	12.5	25.0
118	3299	7.292969		0.0	0.0	2.0	0.0	0.0
323	7305	-1.062500		0.0	5.0	4.0	-12.5	0.0

Train/Test Data



Handling Imbalanced Data: Procedures

Original, Preprocessed Data

	ID	QOL_ovt	energy_ovt	genhlth_ovt	optimism_ovt	pain_ovt	socfctn_ovt
0	1245	-32.312500	-20.0	-40.0	-2.0	-12.5	-50.0
1	1256	-36.687500	-5.0	10.0	-3.0	-50.0	-37.5
2	1265	-6.312500	-5.0	-10.0	2.0	-12.5	0.0
3	1266	0.437500	-10.0	0.0	2.0	25.0	-12.5
4	1276	30.646484	30.0	5.0	-1.0	25.0	37.5

Apply Resample
Method



New, Resampled Data

	ID	QOL_ovt	energy_ovt	genhlth_ovt	optimism_ovt	pain_ovt	socfctn_ovt	target
90	1563	-33.000000		-15.0	-20.0	-2.0	-50.0	-50.0
280	5569	23.187500		15.0	25.0	1.0	37.5	0.0
301	7249	13.312500		20.0	0.0	1.0	12.5	25.0
118	3299	7.292969		0.0	0.0	2.0	0.0	0.0
323	7305	-1.062500		0.0	5.0	4.0	-12.5	0.0

Train/Test Data



Classification Models



Linear regression
KNN
SVM (linear kernel)
Decision tree
Random forest

Handling Imbalanced Data: Procedures

Original, Preprocessed Data

ID	QOL_ovt	energy_ovt	genhlth_ovt	optimism_ovt	pain_ovt	socfctn_ovt	
0	1245	-32.312500	-20.0	-40.0	-2.0	-12.5	-50.0
1	1256	-36.687500	-5.0	10.0	-3.0	-50.0	-37.5
2	1265	-6.312500	-5.0	-10.0	2.0	-12.5	0.0
3	1266	0.437500	-10.0	0.0	2.0	25.0	-12.5
4	1276	30.646484	30.0	5.0	-1.0	25.0	37.5

Apply Resample
Method



New, Resampled Data

ID	QOL_ovt	energy_ovt	genhlth_ovt	optimism_ovt	pain_ovt	socfctn_ovt	target
90	1563	-33.000000	-15.0	-20.0	-2.0	-50.0	-50.0
280	5569	23.187500	15.0	25.0	1.0	37.5	0.0
301	7249	13.312500	20.0	0.0	1.0	12.5	25.0
118	3299	7.292969	0.0	0.0	2.0	0.0	0.0
323	7305	-1.062500	0.0	5.0	4.0	-12.5	0.0

Train/Test Data



Classification Models

Compute Scores



Linear regression
KNN
SVM (linear kernel)
Decision tree
Random forest

Handling Imbalanced Data: Procedures

Original, Preprocessed Data

ID	QOL_ovt	energy_ovt	genhlth_ovt	optimism_ovt	pain_ovt	socfctn_ovt	
0	1245	-32.312500	-20.0	-40.0	-2.0	-12.5	-50.0
1	1256	-36.687500	-5.0	10.0	-3.0	-50.0	-37.5
2	1265	-6.312500	-5.0	-10.0	2.0	-12.5	0.0
3	1266	0.437500	-10.0	0.0	2.0	25.0	-12.5
4	1276	30.646484	30.0	5.0	-1.0	25.0	37.5

Apply Resample
Method



New, Resampled Data

ID	QOL_ovt	energy_ovt	genhlth_ovt	optimism_ovt	pain_ovt	socfctn_ovt	...
90	1563	-33.000000	-15.0	-20.0	-2.0	-50.0	-50.0
280	5569	23.187500	15.0	25.0	1.0	37.5	0.0
301	7249	13.312500	20.0	0.0	1.0	12.5	25.0
118	3299	7.292969	0.0	0.0	2.0	0.0	0.0
323	7305	-1.062500	0.0	5.0	4.0	-12.5	0.0

Train/Test Data

Evaluate, Performance

True Class	Predicted Class	
	True Positive (TP)	False Negative (FN)
False Positive (FP)	True Negative (TN)	

TPR (recall)
TNR (specificity)
PPV (precision)
PNV
Accuracy
F1

Compute Scores



Classification Models



Linear regression
KNN
SVM (linear kernel)
Decision tree
Random forest

Handling Imbalanced Data: Procedures

Original, Preprocessed Data

ID	QOL_ovt	energy_ovt	genhlth_ovt	optimism_ovt	pain_ovt	socfctn_ovt	
0	1245	-32.312500	-20.0	-40.0	-2.0	-12.5	-50.0
1	1256	-36.687500	-5.0	10.0	-3.0	-50.0	-37.5
2	1265	-6.312500	-5.0	-10.0	2.0	-12.5	0.0
3	1266	0.437500	-10.0	0.0	2.0	25.0	-12.5
4	1276	30.646484	30.0	5.0	-1.0	25.0	37.5

Apply Resample
Method



New, Resampled Data

ID	QOL_ovt	energy_ovt	genhlth_ovt	optimism_ovt	pain_ovt	socfctn_ovt	...
90	1563	-33.000000	-15.0	-20.0	-2.0	-50.0	-50.0
280	5569	23.187500	15.0	25.0	1.0	37.5	0.0
301	7249	13.312500	20.0	0.0	1.0	12.5	25.0
118	3299	7.292969	0.0	0.0	2.0	0.0	0.0
323	7305	-1.062500	0.0	5.0	4.0	-12.5	0.0

Train/Test Data

REPEAT PROCESS
FOR EACH
RESAMPLE METHOD

Evaluate, Performance

True Class	Predicted Class	
	True Positive (TP)	False Negative (FN)
False Positive (FP)	True Negative (TN)	

TPR (recall)
TNR (specificity)
PPV (precision)
PNV
Accuracy
 F_1

Compute Scores



Classification Models

Linear regression
KNN
SVM (linear kernel)
Decision tree
Random forest

Outline

- Background & Data + Research Question
- Exploratory Analysis
 - Uni-, Bi-, Multivariate Analyses
- **Predictive Analysis**
 - Machine Learning Models
 - Solutions to Imbalanced Data
 - **Performance Metrics & Assessment**
- Limitations
- Closing & Acknowledgements

Our Best Model & Performance: Up AND Downsampling on a Random Forest Classifier

Our Best Model & Performance: Up AND Downsampling on a Random Forest Classifier

ROC AUC: 0.9436

Classification Report

	precision	recall	f1-score	support
no recurrence	0.87	0.85	0.86	46
recurrence	0.87	0.89	0.88	54

Confusion Matrix: True Negative: 39 False Positive: 7
False Negative: 6 True Positive: 48

Accuracy: 0.87

Precision: 0.873

Specificity: 0.848

Recall: 0.889

F1: 0.881

Our Best Model & Performance: Up AND Downsampling on a Random Forest Classifier

ROC AUC: 0.9436

Classification Report

	precision	recall	f1-score	support
no recurrence	0.87	0.85	0.86	46
recurrence	0.87	0.89	0.88	54

Confusion Matrix: True Negative: 39 False Positive: 7
False Negative: 6 True Positive: 48

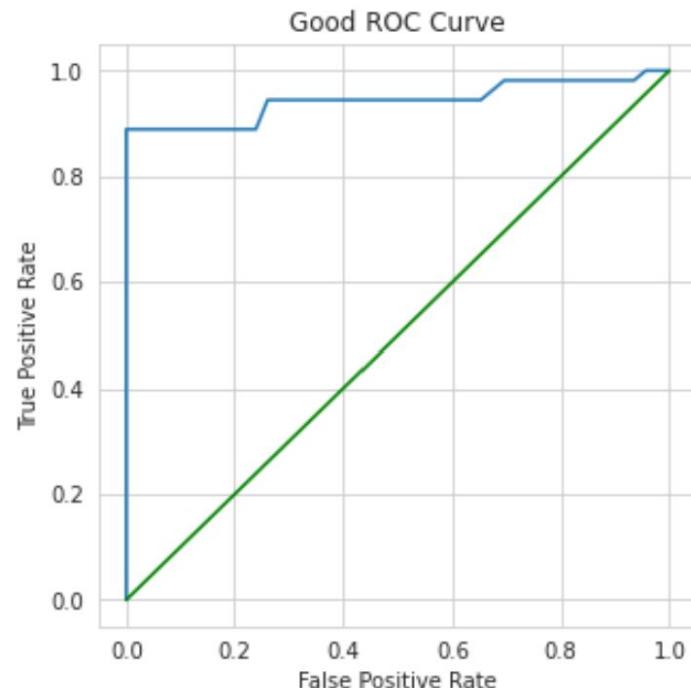
Accuracy: 0.87

Precision: 0.873

Specificity: 0.848

Recall: 0.889

F1: 0.881



Outline

- Background & Data + Research Question
- Exploratory Analysis
 - Uni-, Bi-, Multivariate Analyses
- Predictive Analysis
 - Machine Learning Models
 - Solutions to Imbalanced Data
 - Performance Metrics & Assessment
- **Limitations**
- Closing & Acknowledgements

Binary Classification Model: Preprocessing (Merging)

Repeat merging for
rest of dataframes ...

Shape: (655, 68)

→ 2433 rows lost !!

	ID	QOL_ovt	energy_ovt	genhlth_ovt	optimism_ovt	pain_ovt	socfctn_ovt	lutein_ovt
0	1245	-32.312500	-20.0	-40.0	-2.0	-12.5	-50.0	-0.1697
1	1256	-36.687500	-5.0	10.0	-3.0	-50.0	-37.5	0.0980
2	1265	-6.312500	-5.0	-10.0	2.0	-12.5	0.0	0.0349
3	1266	0.437500	-10.0	0.0	2.0	25.0	-12.5	-0.1125
4	1276	30.646484	30.0	5.0	-1.0	25.0	37.5	0.0766

In [18]: ovt['recurrence_y'].value_counts()

Out[18]: 0 629
1 26
Name: recurrence_y, dtype: int64

629 (96%) had no recurrence
26 (4%) had recurrence

Binary Classification Model: Preprocessing (Merging)

Repeat merging for
rest of dataframes ...

Shape: (655, 68)

2433 rows lost !!

	ID	QOL_ovt	energy_ovt	genhlth_ovt	optimism_ovt	pain_ovt	socfctn_ovt	lutein_ovt
0	1245	-32.312500	-20.0	-40.0	-2.0	-12.5	-50.0	-0.1697
1	1256	-36.687500	-5.0	10.0	-3.0	-50.0	-37.5	0.0980
2	1265	-6.312500	-5.0	-10.0	2.0	-12.5	0.0	0.0349
3	1266	0.437500	-10.0	0.0	2.0	25.0	-12.5	-0.1125
4	1276	30.646484	30.0	5.0	-1.0	25.0	37.5	0.0766

In [18]: ovt['recurrence_y'].value_counts()

Out[18]: 0 629
1 26
Name: recurrence_y, dtype: int64

629 (96%) had no recurrence
26 (4%) had recurrence

Limitations

- Lack/loss of data

Limitations

- **Lack/loss of data**
 - *Columns dropped* →

Limitations

- **Lack/loss of data**
 - *Columns dropped → Binary/Multiclass data*

Limitations

- **Lack/loss of data**
 - *Columns dropped* → Binary/Multiclass data
 - *Rows dropped* →

Limitations

- **Lack/loss of data**
 - *Columns dropped* → Binary/Multiclass data
 - *Rows dropped* → Patients not in all 3 checkpoints →

Limitations

- **Lack/loss of data**
 - *Columns dropped* → Binary/Multiclass data
 - *Rows dropped* → Patients not in all 3 checkpoints → potential biases in data collection

Limitations: Data Collection Bias Analysis

Limitations: Data Collection Bias Analysis

Operation: calculate changes over time
for thoughts & feelings

Before: 2861 rows

After: 1989 rows

Loss: 872 rows

Limitations: Data Collection Bias Analysis

Operation: calculate changes over time
for thoughts & feelings

Before: 2861 rows

After: 1989 rows

Loss: 872 rows

Education

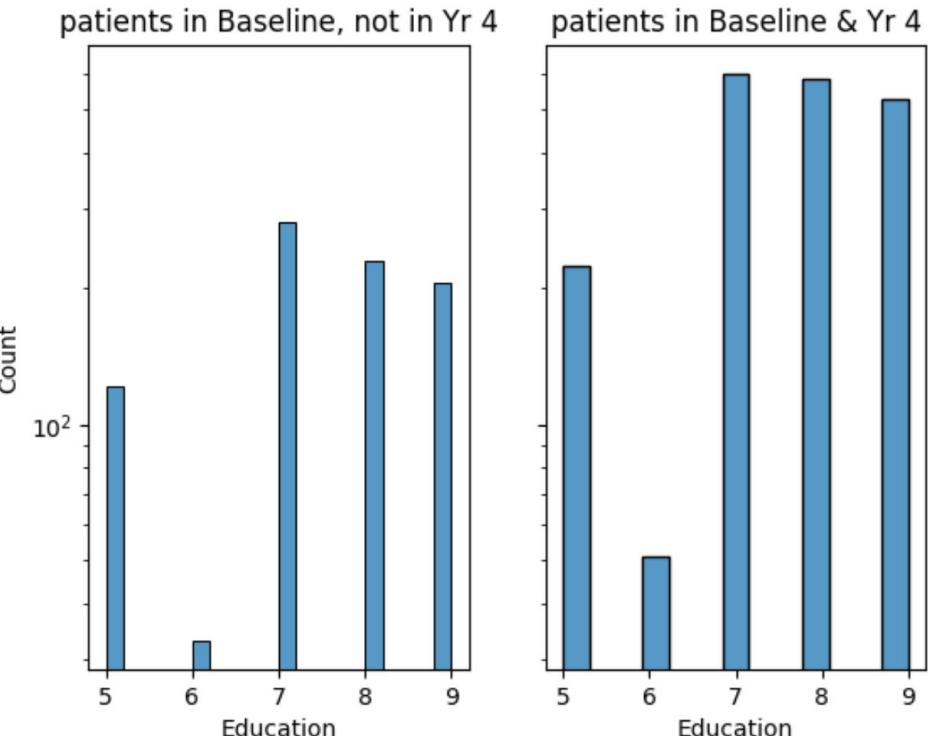
5 - High School Graduate or less

6 - Post High School Training

7 - Some College Education

8 - College/University Graduate

9 - Post College/University Education



Limitations: Data Collection Bias Analysis

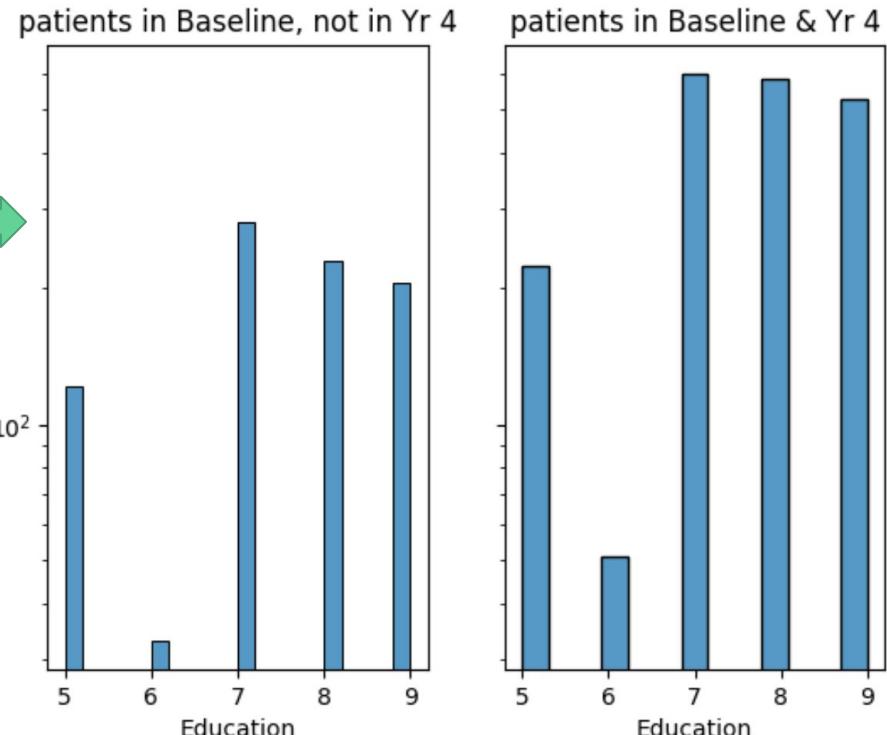
Operation: calculate changes over time
for thoughts & feelings

Before: 2861 rows

After: 1989 rows

Loss: 872 rows

y-axis in log scale



Education

5 - High School Graduate or less

6 - Post High School Training

7 - Some College Education

8 - College/University Graduate

9 - Post College/University Education

Limitations: Data Collection Bias Analysis

Operation: calculate changes over time
for thoughts & feelings

Before: 2861 rows

After: 1989 rows

Loss: 872 rows

y-axis in log scale

Education

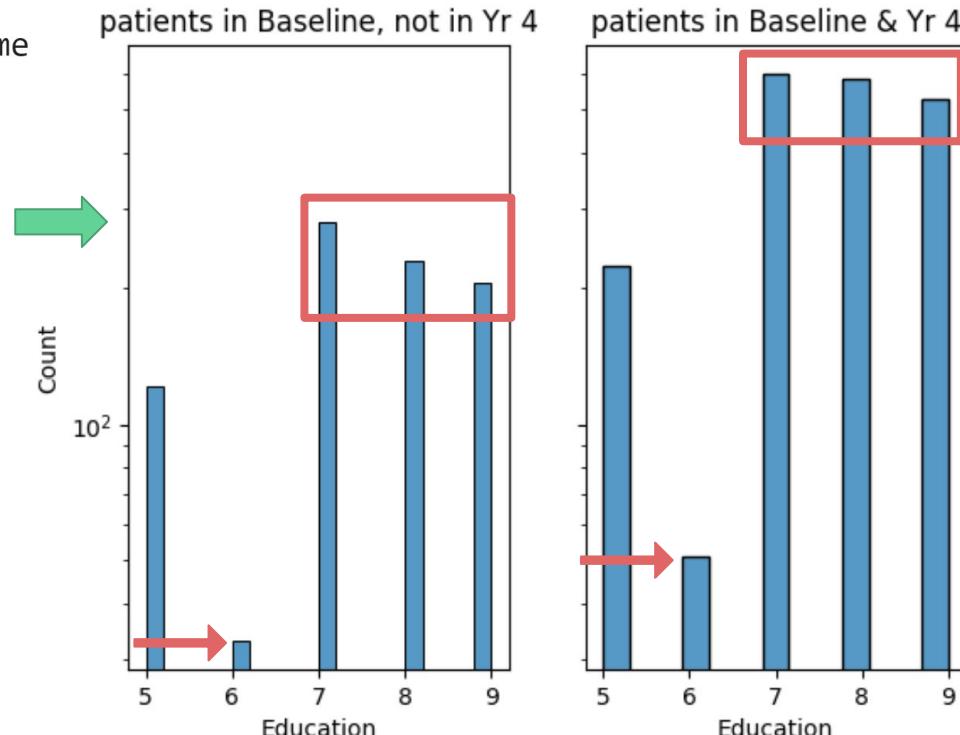
5 - High School Graduate or less

6 - Post High School Training

7 - Some College Education

8 - College/University Graduate

9 - Post College/University Education



Outline

- Background & Data + Research Question
- Exploratory Analysis
 - Uni-, Bi-, Multivariate Analyses
- Predictive Analysis
 - Machine Learning Models
 - Solutions to Imbalanced Data
 - Performance Metrics & Assessment
- Limitations
- **Closing & Acknowledgements**

Closing & Acknowledgements

Closing & Acknowledgements

- Our mentor - Dr. Niema Moshiri

Closing & Acknowledgements

- Our mentor - Dr. Niema Moshiri
- UCSD Library's Research Data Curation program

https://github.com/pndang/Project_WHEL

https://github.com/leena-kang/WHEL_Study

**Thank you!
Questions?**

l1kang@ucsd.edu

pndang@ucsd.edu

Handling Imbalanced Data: Resample Methods

Original Data

```
In [18]: ovt['recurrence_y'].value_counts()
```

```
Out[18]: 0    629  
1     26  
Name: recurrence_y, dtype: int64
```

Upsample Minority Class (1:1)

```
print(upsampled['recurrence_y'].value_counts())  
upsampled
```

```
1    472  
0    472  
Name: recurrence_y, dtype: int64
```

Downsample Majority Class (4:1)

```
downsampled = pd.concat([no_recur_downsampled, recur])  
downsampled.recurrence_y.value_counts()
```

```
Out[51]: 0    76  
1     19  
Name: recurrence_y, dtype: int64
```

SMOTE (imblearn)

```
In [72]: X_train_smote.shape, y_train_smote.shape
```

```
Out[72]: ((944, 67), (944,))
```

Limitations

- **Lack/loss of data**
 - *Columns dropped* → Binary/Multiclass data
 - *Rows dropped* → Patients not in Year 4 → potential biases in data collection
- **Accuracy score**

Limitations

- **Lack/loss of data**
 - *Columns dropped* → Binary/Multiclass data
 - *Rows dropped* → Patients not in Year 4 → potential biases in data collection
- **Accuracy score**

Limitations

- **Lack/loss of data**
 - *Columns dropped* → Binary/Multiclass data
 - *Rows dropped* → Patients not in Year 4 → potential biases in data collection
- **Accuracy score** → may not perform well with other tests

Limitations

- **Lack/loss of data**
 - *Columns dropped* → Binary/Multiclass data
 - *Rows dropped* → Patients not in Year 4 → potential biases in data collection
- **Accuracy score** → may not perform well with other tests
 - To improve our evaluation: ***cross validation***

Potential Applications