Introduction to Project 3: Colorectal Cancer and Serum Metabolic Profiling

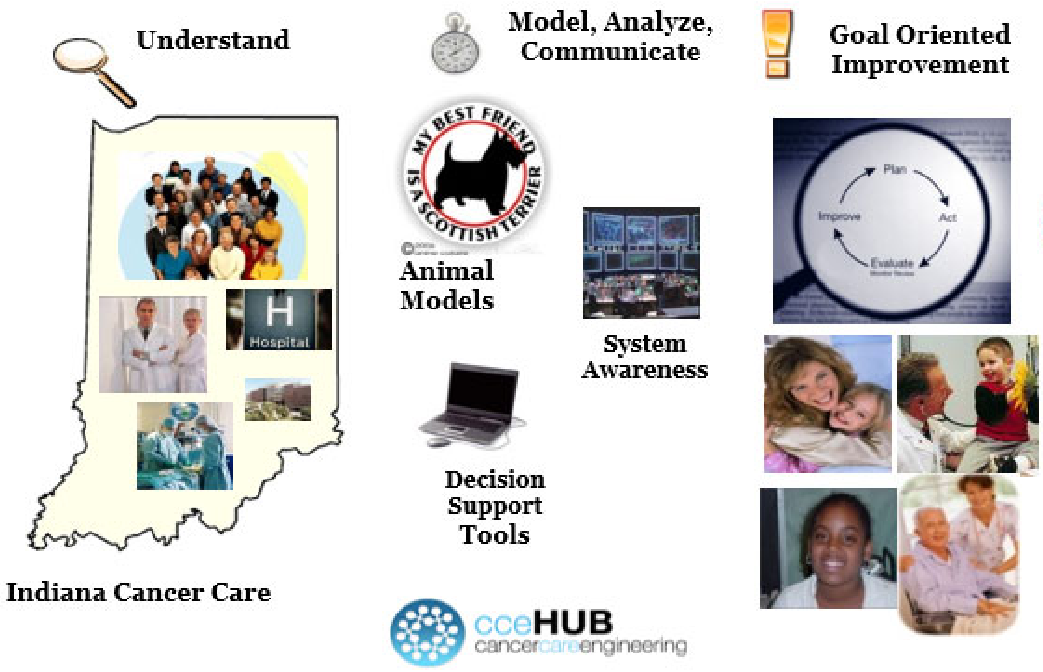
AUTHOR
Min Zhang and Zhaoxia Yu

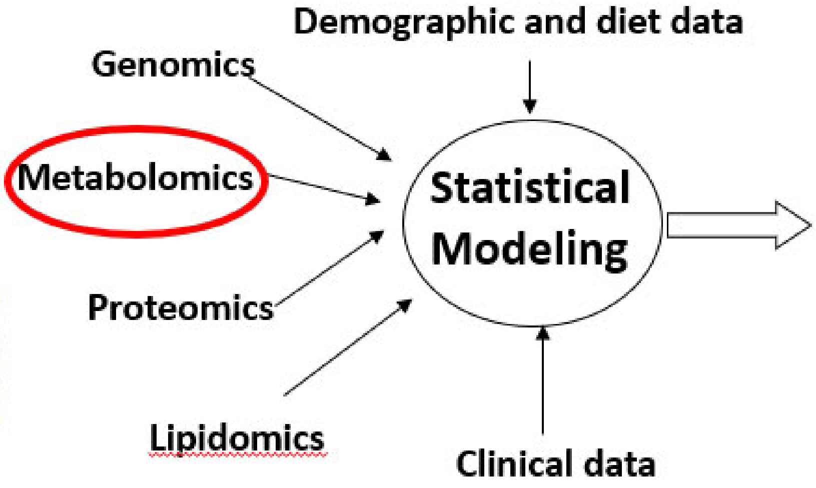
Summer Program ISI-BUDS 2023 Project 3

Instructors and TA

- Min Zhang: Department of Epidemiology and Biostatistics, University of California, Irvine
- Zhaoxia Yu: Department of Statistics, University of California, Irvine
- Thanasi Bakis: Department of Statistics, University of California, Irvine

Metabolomics Biomarker Discovery




Demographic and diet data
Genomics
Metabolomics
Proteomics
Lipidomics
Statistical Modeling
Clinical data

REGENSTRIEF INSTITUTE, INC.

PURDUE UNIVERSITY Discovery Park

UNITED STATES DEPARTMENT OF VETERANS AFFAIRS

WALTHER CANCER INSTITUTE
Collaborating for a cure...

Cancer Care Engineering Project (CCE)

Background

- Colorectal cancer (CRC) is among the most common (2nd in women and 3rd in men) and deadly cancers worldwide.
- Despite advances, there’s a pressing need for robust biomarkers to improve CRC screening, surveillance, and therapy monitoring.
- Cancer patients often show altered / abnormal metabolism.

- A study monitored 158 **targeted** metabolites from 25 potentially significant metabolic pathways in 234 serum samples. These samples were collected from three patient groups: 66 CRC patients, 76 polyp patients (polyp is a benign status), and 92 healthy controls (Zhu et al. 2014).

Background (Zhu et al. 2014)

- Partial least-squares-discriminant analysis (PLS-DA) models were used to distinguish CRC patients from both healthy controls and polyp patients. The Receiver Operating Characteristic (ROC) curves based on these PLS-DA models indicated
    - high sensitivity: 0.96 for differentiating CRC patients from healthy controls; and 0.89 for from poly patients
    - good specificity: 0.80 and 0.88
    - excellent areas under the curve: 0.93 and 0.95

Data Exploration

Read data

A useful website about reading data from excel

http://www.sthda.com/english/wiki/reading-data-from-excel-files-xls-xlsx-into-r

```
#install.packages("readxl")
# Load the library
library("readxl")
```

```
Warning: package 'readxl' was built under R version 4.2.3
```

```
library("tidyr")
library("ggplot2")
crc <- read_excel("CRC raw data_ updated with patient ID.xlsx", sheet = "Sheet1", na=c("NA", "-"))
```

```
attributes(crc)
crc <- as.data.frame(crc)
```

Explore the Data

1. Examine the data dimension and names

   ▶ Code

2. Check the first column

   ▶ Code

```
 C  H  P 
66 92 76 
```

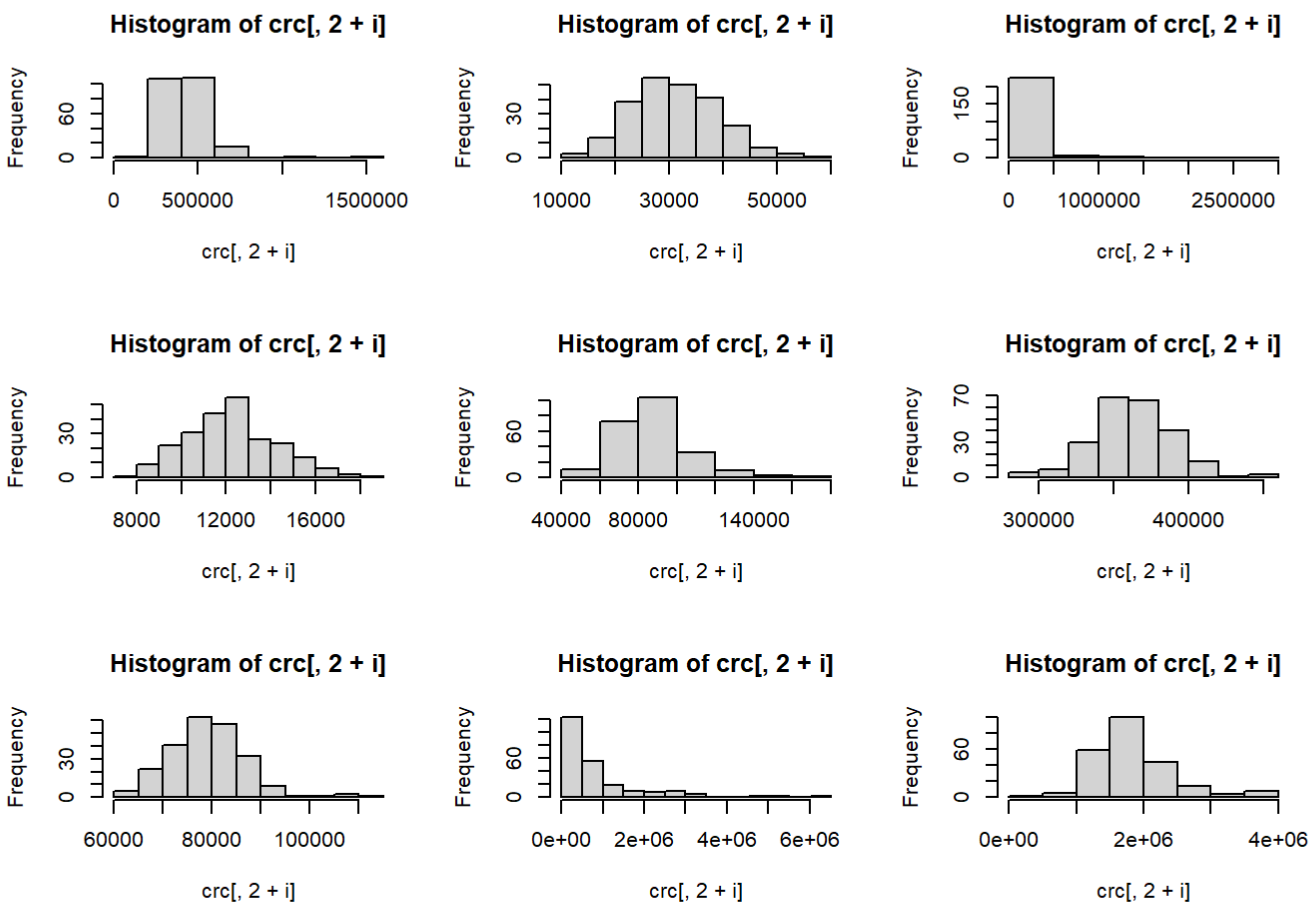
3. Check the second column

   ▶ Code

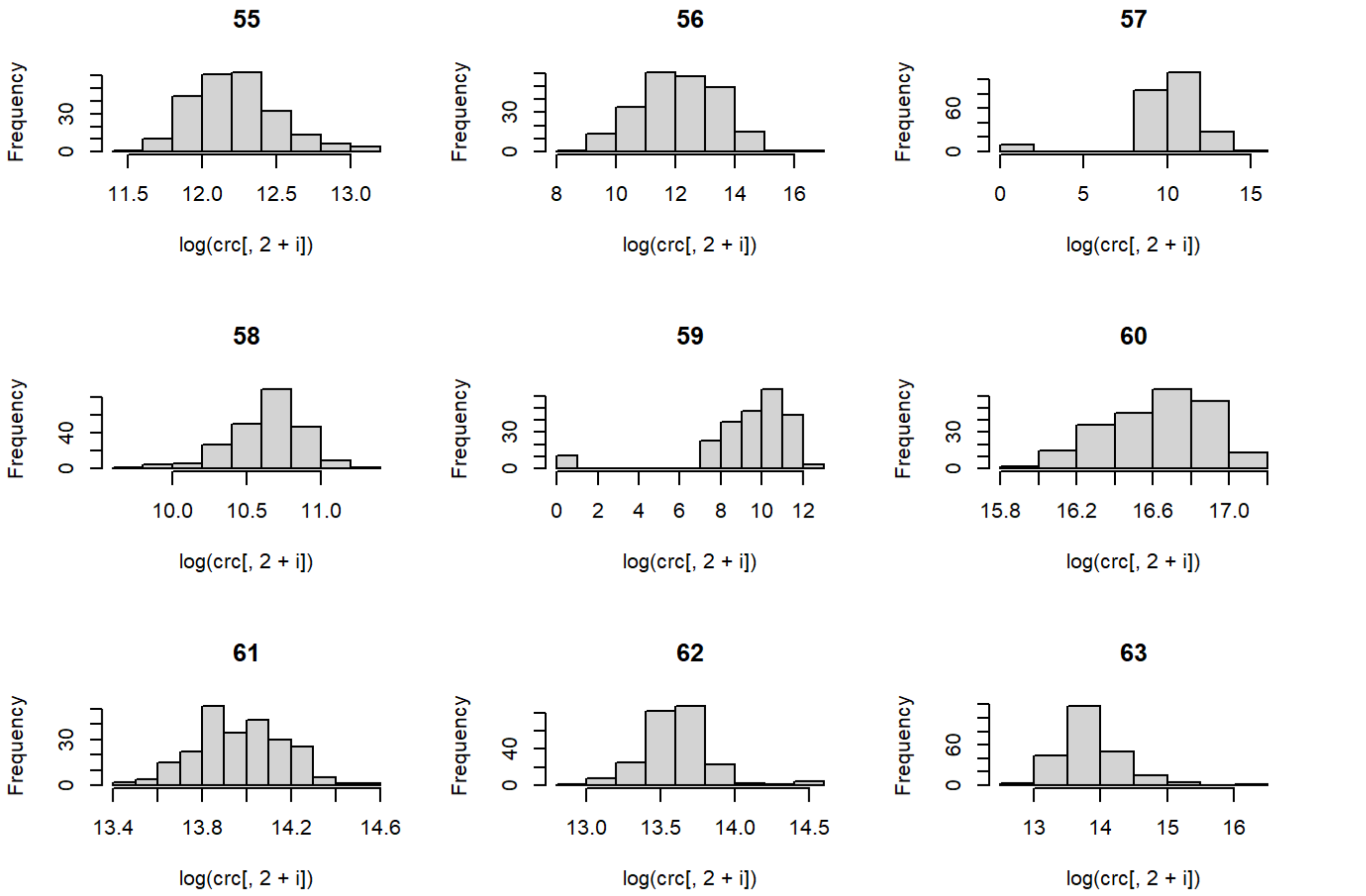
```
[1] 224
```

4. Columns 3-115 are metabolism variables. We will check the first nine by drawing their histograms

   ▶ Code


Histogram of crc[, 2 + i]
Frequency
0
60
0
500000
1500000
crc[, 2 + i]
Histogram of crc[, 2 + i]
Frequency
0
30
10000
30000
50000
crc[, 2 + i]
Histogram of crc[, 2 + i]
Frequency
0
150
0
1000000
2500000
crc[, 2 + i]
Histogram of crc[, 2 + i]
Frequency
0
30
8000
12000
16000
crc[, 2 + i]
Histogram of crc[, 2 + i]
Frequency
0
60
40000
80000
140000
crc[, 2 + i]
Histogram of crc[, 2 + i]
Frequency
0
30
70
300000
400000
crc[, 2 + i]
Histogram of crc[, 2 + i]
Frequency
0
30
60000
80000
100000
crc[, 2 + i]
Histogram of crc[, 2 + i]
Frequency
0
60
0e+00
2e+06
4e+06
6e+06
crc[, 2 + i]
Histogram of crc[, 2 + i]
Frequency
0
60
0e+00
2e+06
4e+06
crc[, 2 + i]

▶ Code



5. Characteristics of a metabolism

▶ Code

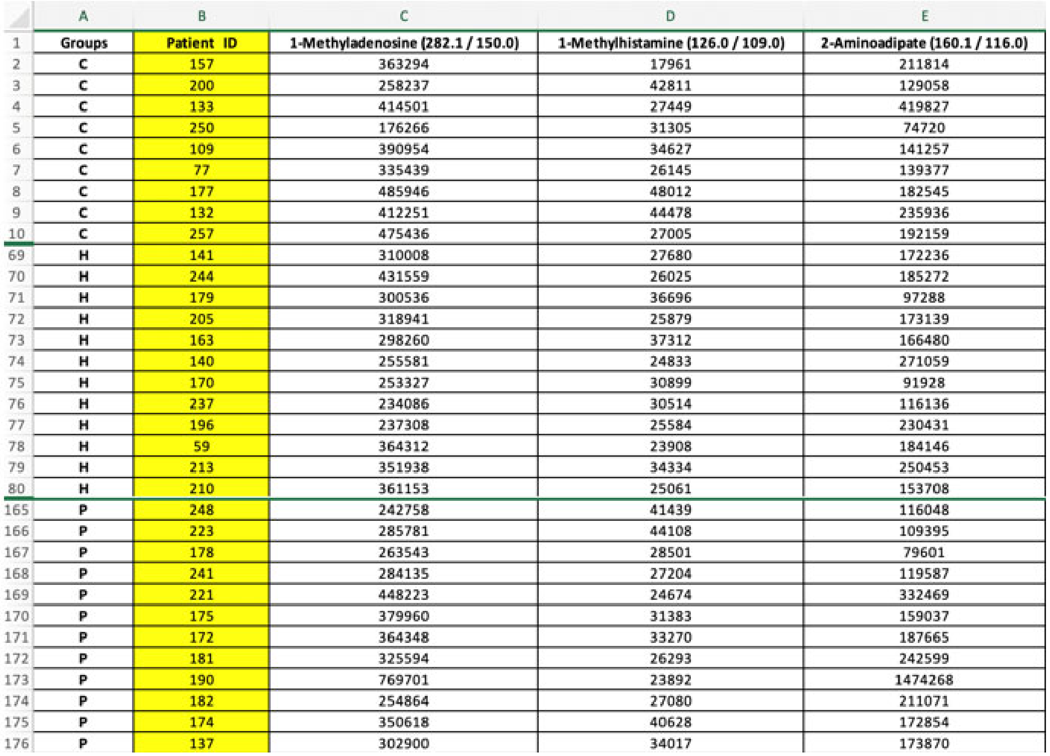
6. Columns 116-124 (clinical variables)

```
names(crc[,116:124])
```

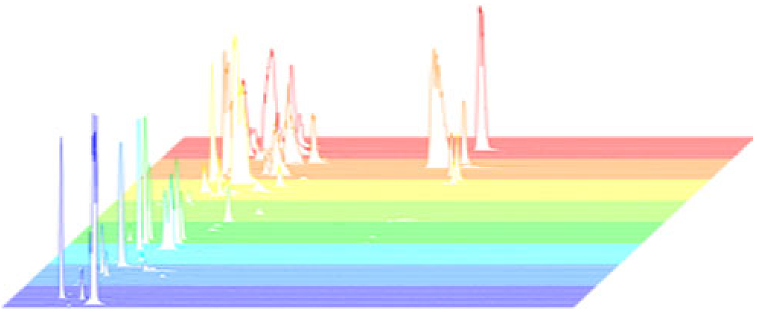
```
[1] "Age at  Consent"   "Gender"            "Smoking condition"
[4] "Drink   Alcohol?"  "Diagnosis"         "Stage"
[7] "Height [cm]"       "Weight [kg]"       "BMI [kg/mÂ²]"
```

```
colnames(crc)[116:124]=c('age', 'gender', 'smoking', 'alcohol', 'diagnosis',
                         'stage', 'height', 'weight', 'bmi')
```

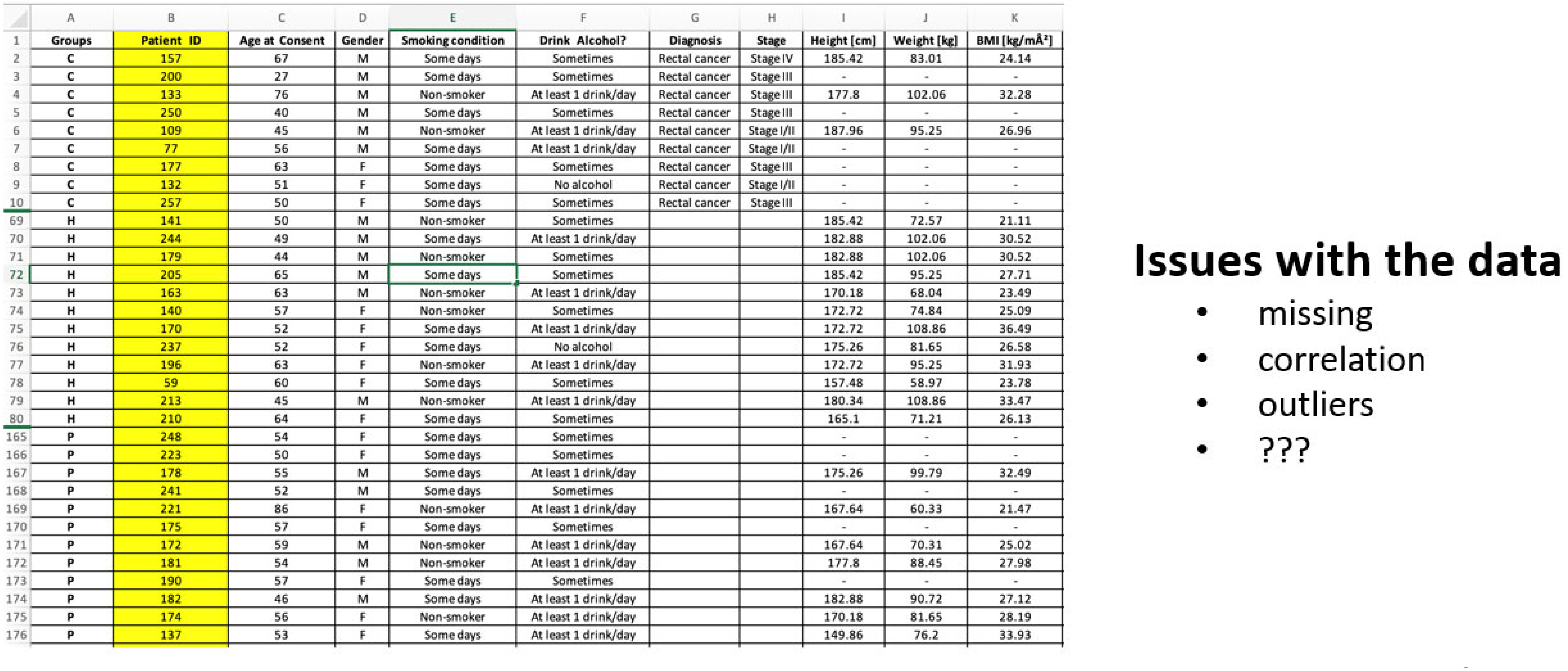
Metabolomics Data



	A	B	C	D	E
1	Groups	Patient ID	1-Methyladenosine (282.1 / 150.0)	1-Methylhistamine (126.0 / 109.0)	2-Aminoadipate (160.1 / 116.0)
2	C	157	363294	17961	211814
3	C	200	258237	42811	129058
4	C	133	414501	27449	419827
5	C	250	176266	31305	74720
6	C	109	390954	34627	141257
7	C	77	335439	26145	139377
8	C	177	485946	48012	182545
9	C	132	412251	44478	235936
10	C	257	475436	27005	192159
69	H	141	310008	27680	172236
70	H	244	431559	26025	185272
71	H	179	300536	36696	97288
72	H	205	318941	25879	173139
73	H	163	298260	37312	166480
74	H	140	255581	24833	271059
75	H	170	253327	30899	91928
76	H	237	234086	30514	116136
77	H	196	237308	25584	230431
78	H	59	364312	23908	184146
79	H	213	351938	34334	250453
80	H	210	361153	25061	153708
165	P	248	242758	41439	116048
166	P	223	285781	44108	109395
167	P	178	263543	28501	79601
168	P	241	284135	27204	119587
169	P	221	448223	24674	332469
170	P	175	379960	31383	159037
171	P	172	364348	33270	187665
172	P	181	325594	26293	242599
173	P	190	769701	23892	1474268
174	P	182	254864	27080	211071
175	P	174	350618	40628	172854
176	P	137	302900	34017	173870



Metabolomics Data



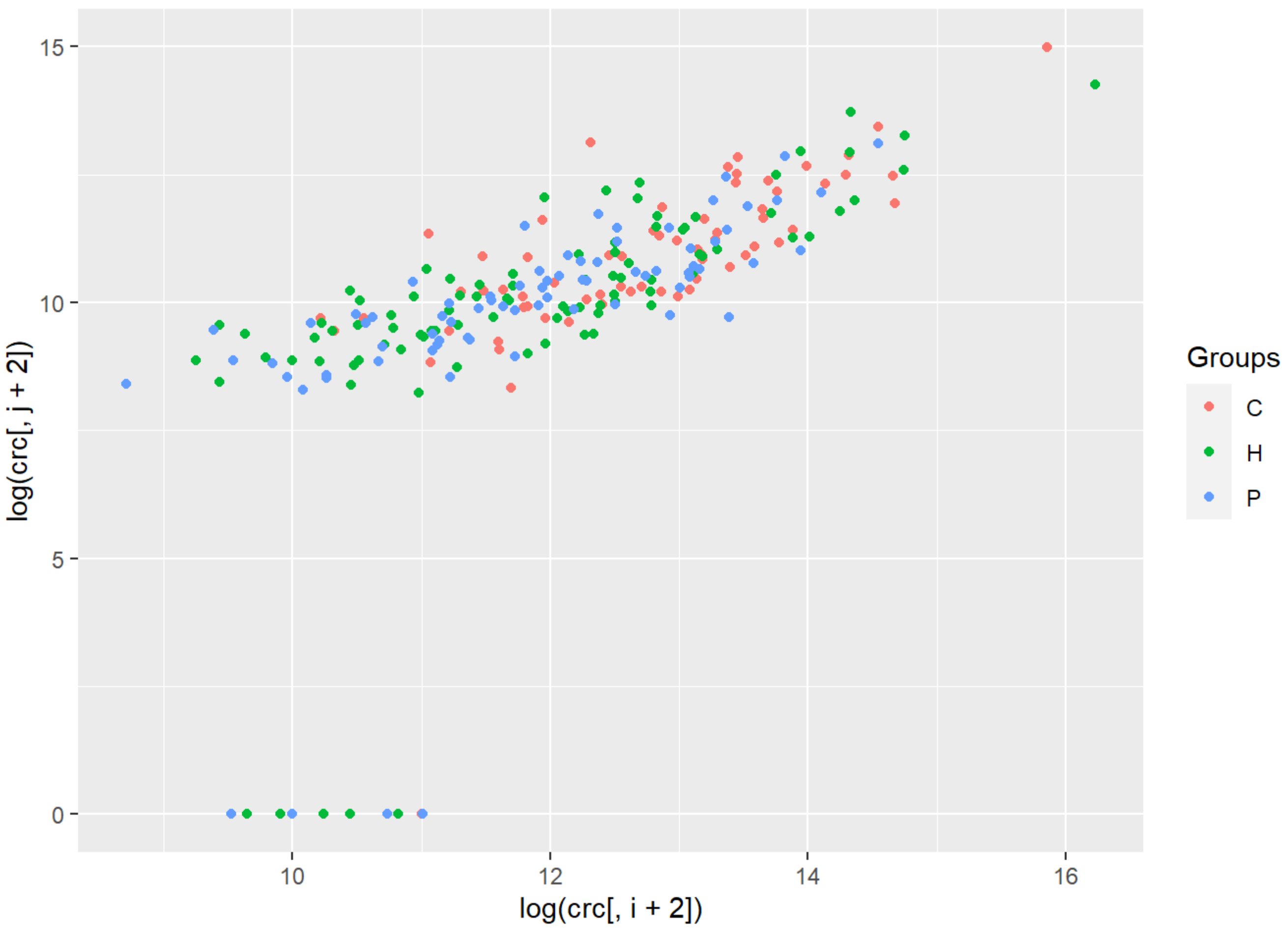
	A	B	C	D	E	F	G	H	I	J	K
1	**Groups**	**Patient ID**	**Age at Consent**	**Gender**	**Smoking condition**	**Drink Alcohol?**	**Diagnosis**	**Stage**	**Height [cm]**	**Weight [kg]**	**BMI [kg/mÅ²]**
2	**C**	157	67	M	Some days	Sometimes	Rectal cancer	Stage IV	185.42	83.01	24.14
3	**C**	200	27	M	Some days	Sometimes	Rectal cancer	Stage III	-	-	-
4	**C**	133	76	M	Non-smoker	At least 1 drink/day	Rectal cancer	Stage III	177.8	102.06	32.28
5	**C**	250	40	M	Some days	Sometimes	Rectal cancer	Stage III	-	-	-
6	**C**	109	45	M	Non-smoker	At least 1 drink/day	Rectal cancer	Stage I/II	187.96	95.25	26.96
7	**C**	77	56	M	Some days	At least 1 drink/day	Rectal cancer	Stage I/II	-	-	-
8	**C**	177	63	F	Some days	Sometimes	Rectal cancer	Stage III	-	-	-
9	**C**	132	51	F	Some days	No alcohol	Rectal cancer	Stage I/II	-	-	-
10	**C**	257	50	F	Some days	Sometimes	Rectal cancer	Stage III	-	-	-
69	**H**	141	50	M	Non-smoker	Sometimes			185.42	72.57	21.11
70	**H**	244	49	M	Some days	At least 1 drink/day			182.88	102.06	30.52
71	**H**	179	44	M	Non-smoker	Sometimes			182.88	102.06	30.52
72	**H**	205	65	M	Some days	Sometimes			185.42	95.25	27.71
73	**H**	163	63	M	Non-smoker	At least 1 drink/day			170.18	68.04	23.49
74	**H**	140	57	F	Non-smoker	Sometimes			172.72	74.84	25.09
75	**H**	170	52	F	Some days	At least 1 drink/day			172.72	108.86	36.49
76	**H**	237	52	F	Some days	No alcohol			175.26	81.65	26.58
77	**H**	196	63	F	Non-smoker	At least 1 drink/day			172.72	95.25	31.93
78	**H**	59	60	F	Some days	Sometimes			157.48	58.97	23.78
79	**H**	213	45	M	Non-smoker	At least 1 drink/day			180.34	108.86	33.47
80	**H**	210	64	F	Some days	Sometimes			165.1	71.21	26.13
165	**P**	248	54	F	Some days	Sometimes			-	-	-
166	**P**	223	50	F	Some days	Sometimes			-	-	-
167	**P**	178	55	M	Some days	At least 1 drink/day			175.26	99.79	32.49
168	**P**	241	52	M	Some days	Sometimes			-	-	-
169	**P**	221	86	F	Non-smoker	At least 1 drink/day			167.64	60.33	21.47
170	**P**	175	57	F	Some days	Sometimes			-	-	-
171	**P**	172	59	M	Non-smoker	At least 1 drink/day			167.64	70.31	25.02
172	**P**	181	54	M	Non-smoker	At least 1 drink/day			177.8	88.45	27.98
173	**P**	190	57	F	Some days	Sometimes			-	-	-
174	**P**	182	46	M	Some days	At least 1 drink/day			182.88	90.72	27.12
175	**P**	174	56	F	Non-smoker	At least 1 drink/day			170.18	81.65	28.19
176	**P**	137	53	F	Some days	At least 1 drink/day			149.86	76.2	33.93

Issues with the data

- missing
- correlation
- outliers
- ???

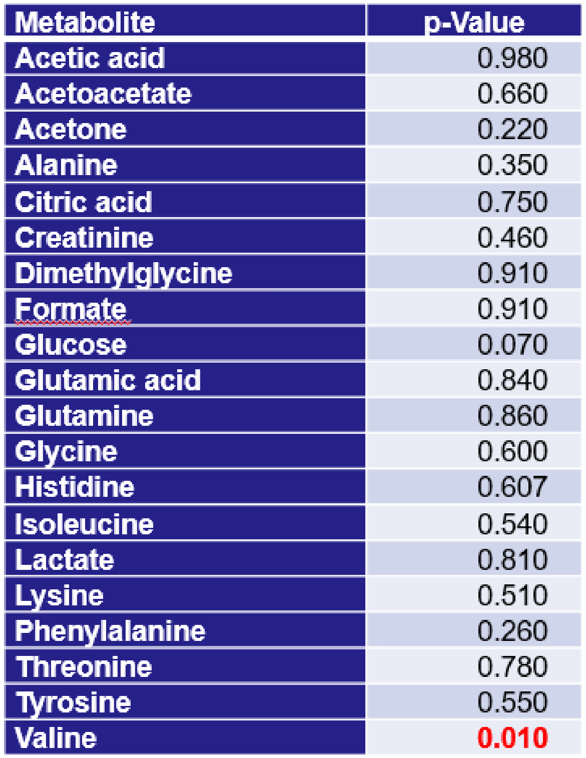
Examine missing data. What is the missing mechanism? Are the undetected values really undetected or missing values? What way makes more sense?

```
i=56; j=57
ggplot(crc, aes(x=log(crc[,i+2]), y=log(crc[,j+2]), color=Groups))+
  geom_point()
```



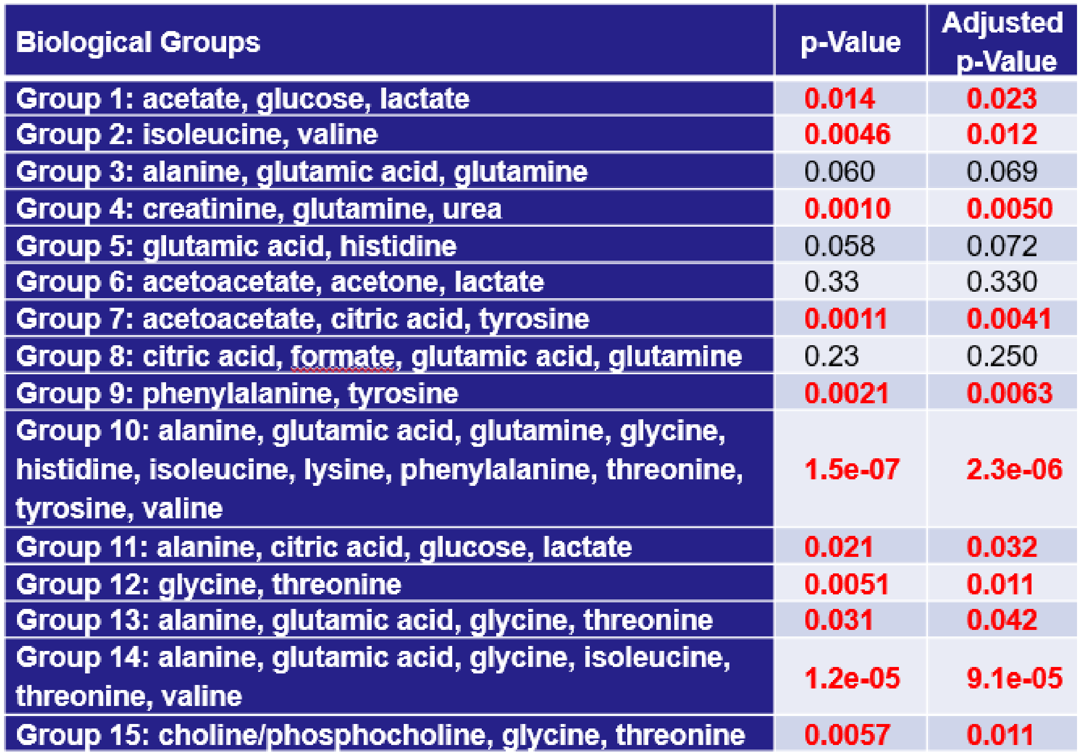
Metabolomics Data analysis

Individual Metabolite Analysis



Metabolite	p-Value
Acetic acid	0.980
Acetoacetate	0.660
Acetone	0.220
Alanine	0.350
Citric acid	0.750
Creatinine	0.460
Dimethylglycine	0.910
Formate	0.910
Glucose	0.070
Glutamic acid	0.840
Glutamine	0.860
Glycine	0.600
Histidine	0.607
Isoleucine	0.540
Lactate	0.810
Lysine	0.510
Phenylalanine	0.260
Threonine	0.780
Tyrosine	0.550
Valine	0.010

Multiple Metabolite Analysis



Biological Groups	p-Value	Adjusted p-Value
Group 1: acetate, glucose, lactate	0.014	0.023
Group 2: isoleucine, valine	0.0046	0.012
Group 3: alanine, glutamic acid, glutamine	0.060	0.069
Group 4: creatinine, glutamine, urea	0.0010	0.0050
Group 5: glutamic acid, histidine	0.058	0.072
Group 6: acetoacetate, acetone, lactate	0.33	0.330
Group 7: acetoacetate, citric acid, tyrosine	0.0011	0.0041
Group 8: citric acid, formate, glutamic acid, glutamine	0.23	0.250
Group 9: phenylalanine, tyrosine	0.0021	0.0063
Group 10: alanine, glutamic acid, glutamine, glycine, histidine, isoleucine, lysine, phenylalanine, threonine, tyrosine, valine	1.5e-07	2.3e-06
Group 11: alanine, citric acid, glucose, lactate	0.021	0.032
Group 12: glycine, threonine	0.0051	0.011
Group 13: alanine, glutamic acid, glycine, threonine	0.031	0.042
Group 14: alanine, glutamic acid, glycine, isoleucine, threonine, valine	1.2e-05	9.1e-05
Group 15: choline/phosphocholine, glycine, threonine	0.0057	0.011

Source: Chen et al. (2015), *Journal of Proteome Research*, 14(6): 2492-2499.

Scientific Questions

- In the healthy population,
  - the metabolisms are associated with which clinical / demographic variables?
- Which metabolisms are different between
  - healthy and cancer subjects?
  - healthy and polyp subjects?

  - polyp and cancer subjects?
  - between the two subtypes of cancer?
- Do different group have the same metabolite correlation structure?
- Missing data treatment: should the “1”s be replaced by “NA”? If we do so, can we improve “the results”?
- Use nested CV to correct the bias in predictive accuracy
- …

How to read a research paper?

https://www.elsevier.com/connect/infographic-how-to-read-a-scientific-paper

Main takeaways

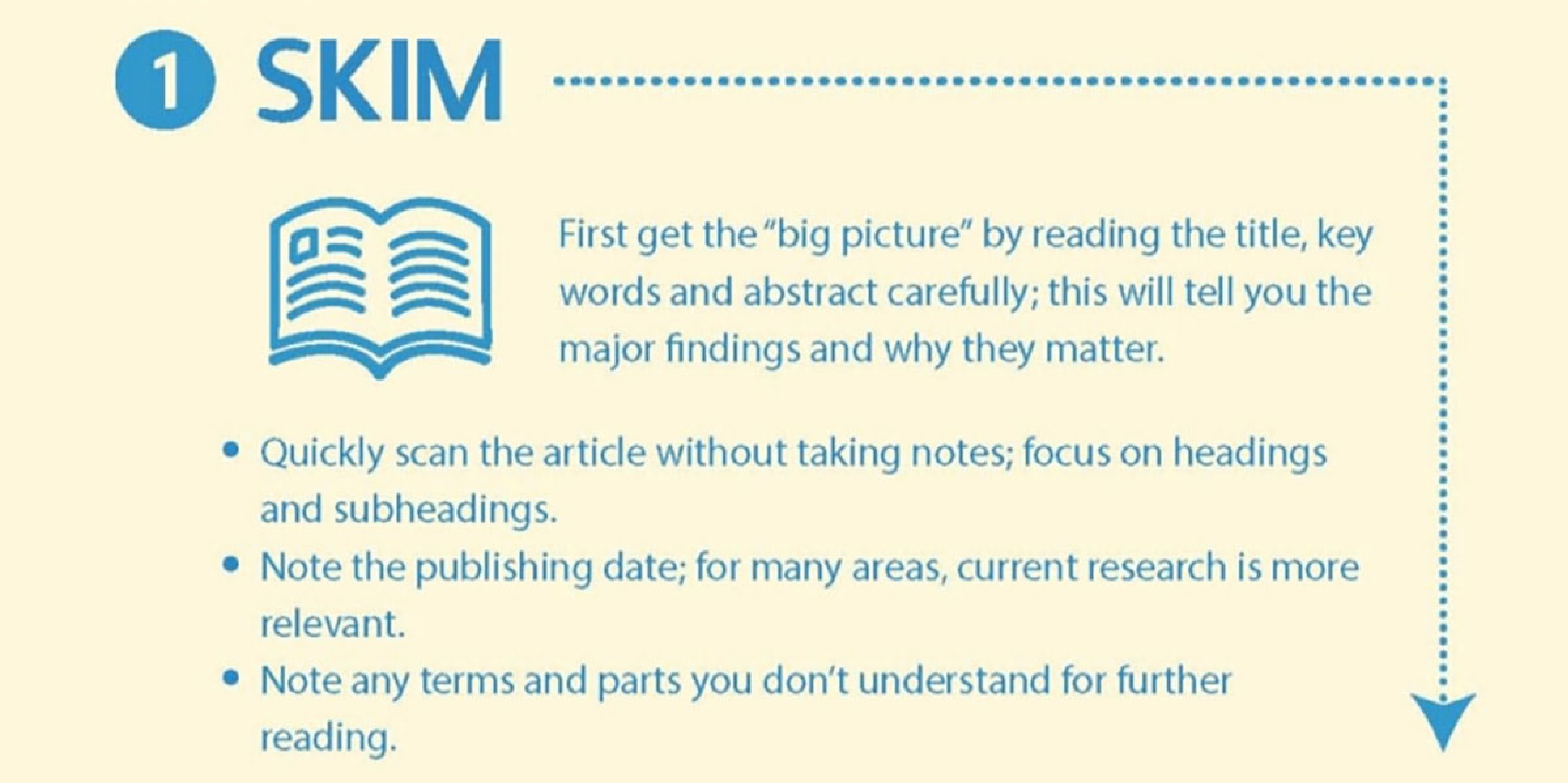
1. **Strategic Reading:** Reading a scientific paper should not be a linear process (from beginning to end). Instead, it requires a strategic approach that goes beyond a surface-level understanding.
2. **Critical Mindset:** Adopt a critical mindset while reading. Challenge the findings and question your understanding to deepen your comprehension of the content.
3. **Fluid Navigation:** It’s okay to navigate backwards and forwards through the paper. Scientific literature often requires revisiting sections to fully grasp complex ideas and information.

---

4. **Note-taking:** Keeping notes is a crucial part of the reading process. Notes will help synthesize information, identify important points, and assist in your comprehension and recall of key findings.
5. **Multi-tab Browsing:** To comprehend the full depth of a scientific paper, often you’ll need to have multiple tabs open in your browser for cross-referencing, fact-checking, and understanding contextual and background information.

Use (Zhu et al. 2014) as an example.

Step1: SKIM to find the big picture



1 SKIM

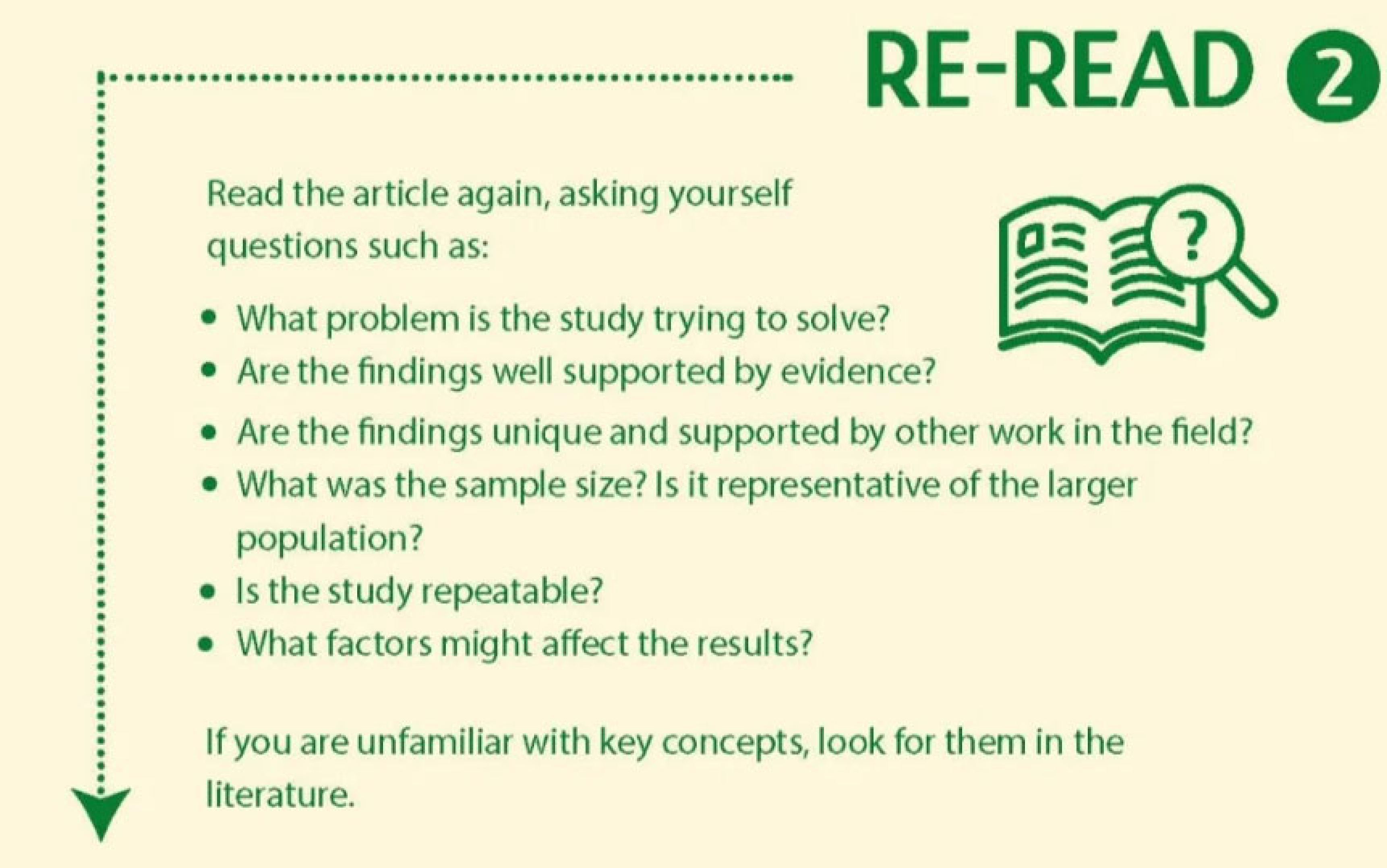
First get the "big picture" by reading the title, key words and abstract carefully; this will tell you the major findings and why they matter.

- Quickly scan the article without taking notes; focus on headings and subheadings.
- Note the publishing date; for many areas, current research is more relevant.
- Note any terms and parts you don't understand for further reading.

Big picture:

- Based on the title, the article offers a screening method for colorectal cancer by using targeted serum metabolic profiling.
- The abstract told us that colorectal cancer imposes serious public health burden and the screening method introduced by the article has very satisfactory results. The main statistical method seems to be the partial least-squares discriminant analysis (PLS-DA)

Step2: RE-READ



RE-READ 2

Read the article again, asking yourself questions such as:

- What problem is the study trying to solve?
- Are the findings well supported by evidence?
- Are the findings unique and supported by other work in the field?
- What was the sample size? Is it representative of the larger population?
- Is the study repeatable?
- What factors might affect the results?

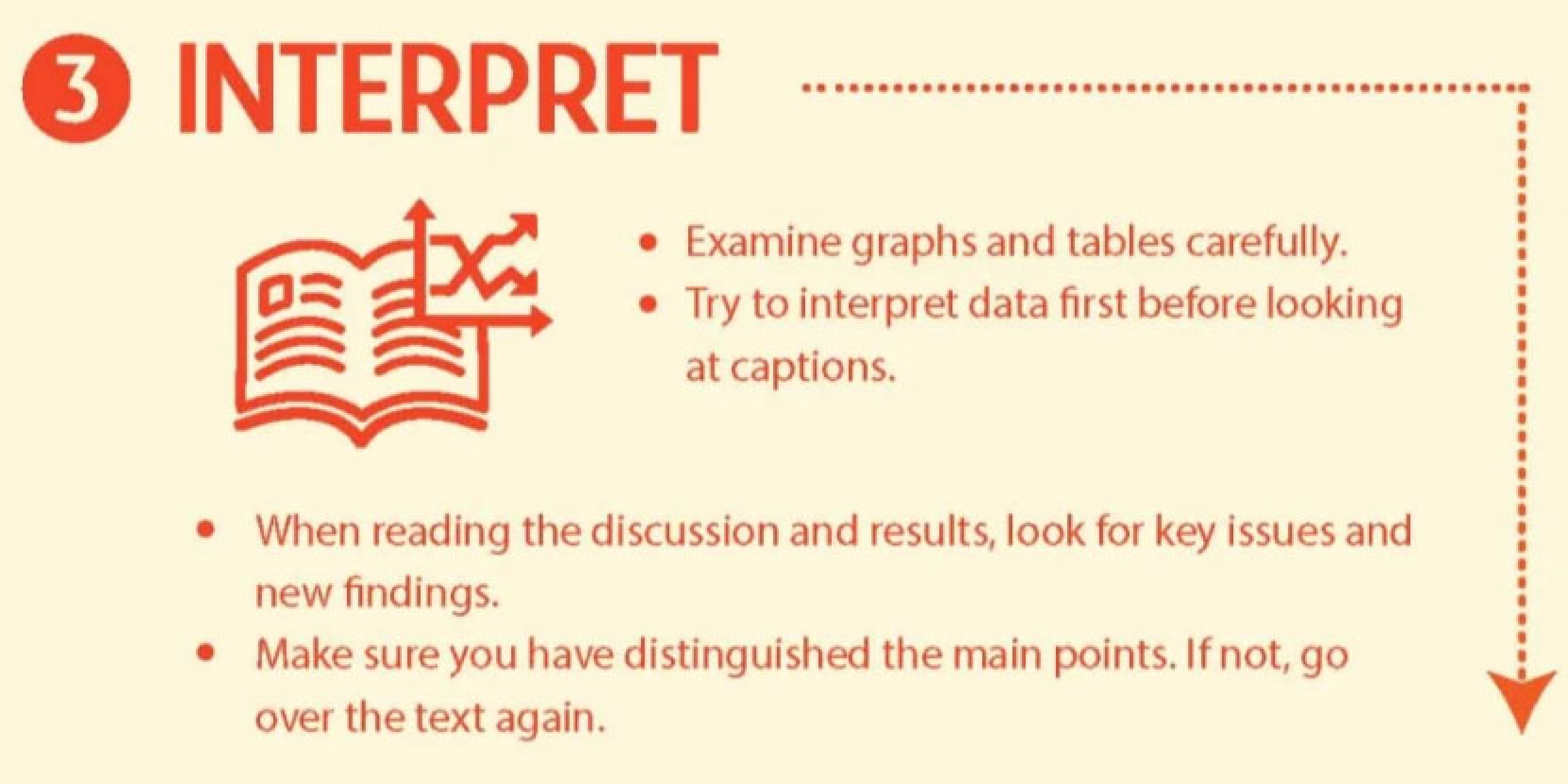
If you are unfamiliar with key concepts, look for them in the literature.

- What problem is the study trying to solve? **A**: providing a new screening test for colorectal cancer because existing ones are not accurate, invasive, or expensive.
- Are the findings well supported by evidence? **A**: yes. The new approach (metabolic profiling + PLS-DA) has high sensitivity, good specificity, and excellent areas under the curve.

- Are the findings unique and supported by other work in the field? **A**:

- What was the sample size? Is it representative of the larger population? **A**: 66 CRC patients, 76 polyp patients, and 92 healthy controls. Patients were age- and gender-matched in each group. The sample is not a representative sample of the larger population.
- Is the study repeatable? **A**: not sure about scientifically. Since the data is available, we can check whether similar results can be obtained.
- What factors might affect the results? **A**: The authors detected 112 metabolisms from 158 targeted ones due to QC filtering. QC might be an issue.

Step3: INTREPRET



3 INTERPRET

- Examine graphs and tables carefully.
- Try to interpret data first before looking at captions.
- When reading the discussion and results, look for key issues and new findings.
- Make sure you have distinguished the main points. If not, go over the text again.

Step4: SUMMARIZE

SUMMARIZE 4

- Take notes; it improves reading comprehension and helps you remember key points.
- If you have a printed version, highlight key points and write on the article. If it's on screen, make use of markers and comments.

References

Zhu, Jiangjiang, Danijel Djukovic, Lingli Deng, Haiwei Gu, Farhan Himmati, E. Gabriela Chiorean, and Daniel Raftery. 2014. "Colorectal Cancer Detection Using Targeted Serum Metabolic Profiling." *Journal of Proteome Research* 13 (9): 4120–30. https://doi.org/10.1021/pr500494u.