Analysis of some features for prognosis of Alzheimer onset:

Probability theory & Information theory

Luca, Alexandra, Ingrid

MMIV-ML group meeting, 13 January 2022

Mild Cognitive Impairment

time

Mild **C**ognitive **I**mpairment

Alzheimer **D**isease

Mild **C**ognitive **I**mpairment
(stable)

time

Mild Cognitive Impairment

Alzheimer Disease

Mild Cognitive Impairment
(stable)






Gender

AGE

RAVLT

ANARTERR

GDTOTAL

TRABSCOR

CATANIMSC

TRAASCOR

AVDELTOT

AVDEL30MIN

LRHHC

LRLV

FAQ

Apoe4

Gender

AGE

RAVLT

ANARTERR

GDTOTAL

TRABSCOR

CATANIMSC

TRAASCOR

AVDELTOT

AVDEL30MIN

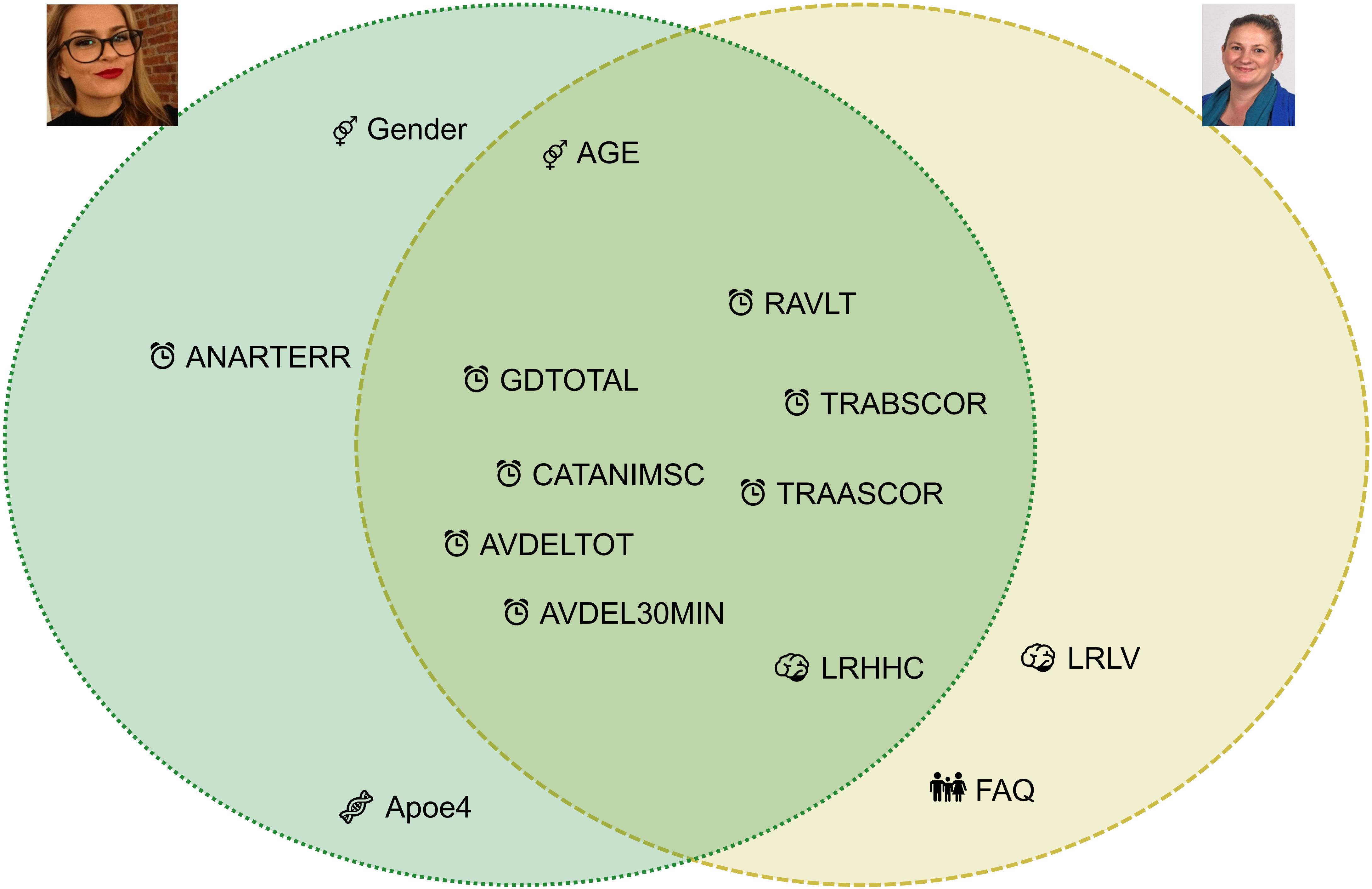
LRHHC

LRLV

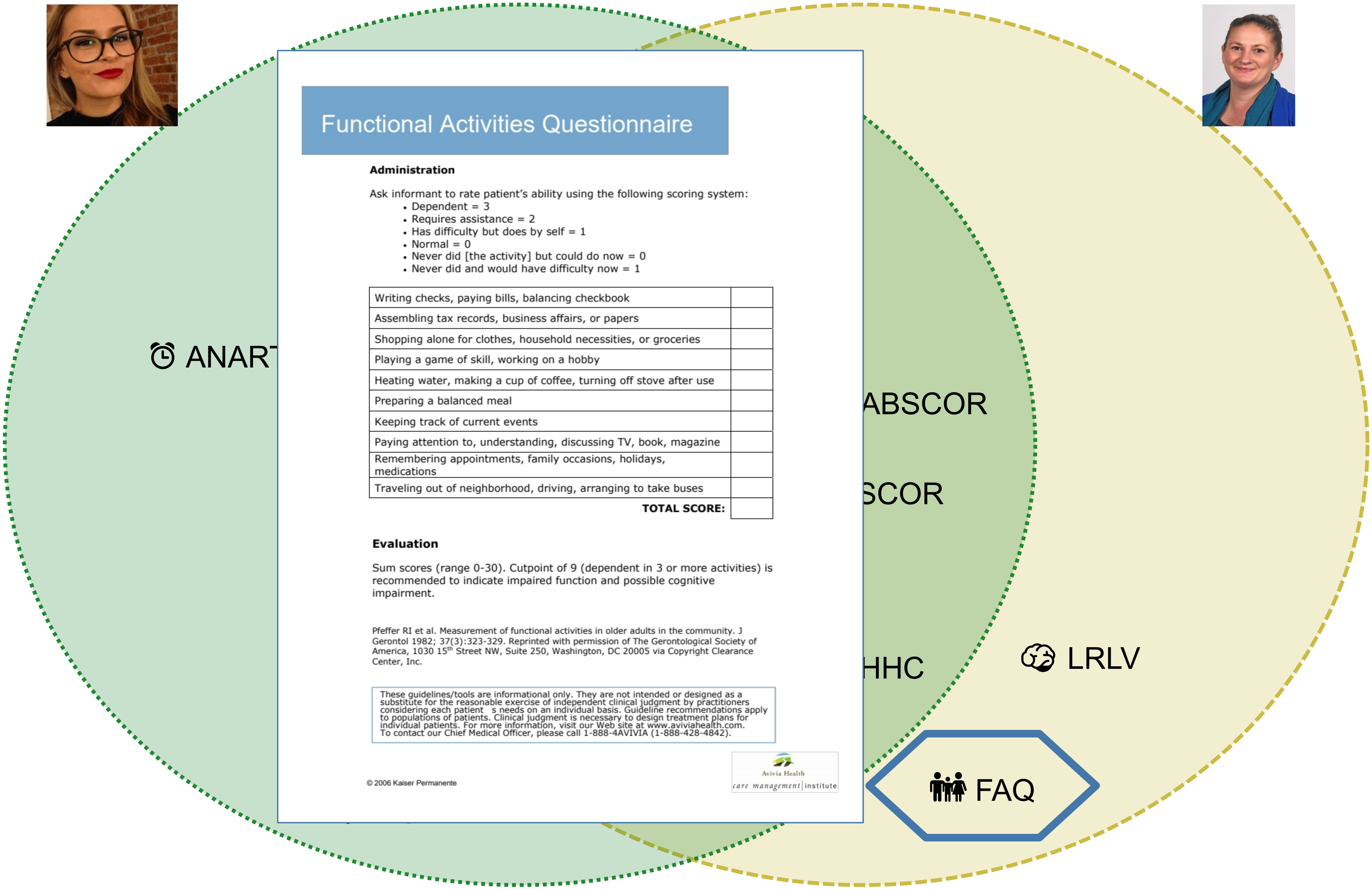
FAQ

Apoe4

How ‘good’ are these features at prognosing the later onset of Alzheimer?



How 'good' are these features at prognosing the later onset of Alzheimer?



Functional Activities Questionnaire

Administration

Ask informant to rate patient's ability using the following scoring system:

- Dependent = 3
- Requires assistance = 2
- Has difficulty but does by self = 1
- Normal = 0
- Never did [the activity] but could do now = 0
- Never did and would have difficulty now = 1

Writing checks, paying bills, balancing checkbook	
Assembling tax records, business affairs, or papers	
Shopping alone for clothes, household necessities, or groceries	
Playing a game of skill, working on a hobby	
Heating water, making a cup of coffee, turning off stove after use	
Preparing a balanced meal	
Keeping track of current events	
Paying attention to, understanding, discussing TV, book, magazine	
Remembering appointments, family occasions, holidays, medications	
Traveling out of neighborhood, driving, arranging to take buses	
TOTAL SCORE:	

Evaluation

Sum scores (range 0-30). Cutpoint of 9 (dependent in 3 or more activities) is recommended to indicate impaired function and possible cognitive impairment.

Pfeffer RI et al. Measurement of functional activities in older adults in the community. J Gerontol 1982; 37(3):323-329. Reprinted with permission of The Gerontological Society of America, 1030 15th Street NW, Suite 250, Washington, DC 20005 via Copyright Clearance Center, Inc.

Avivia Health care management | institute

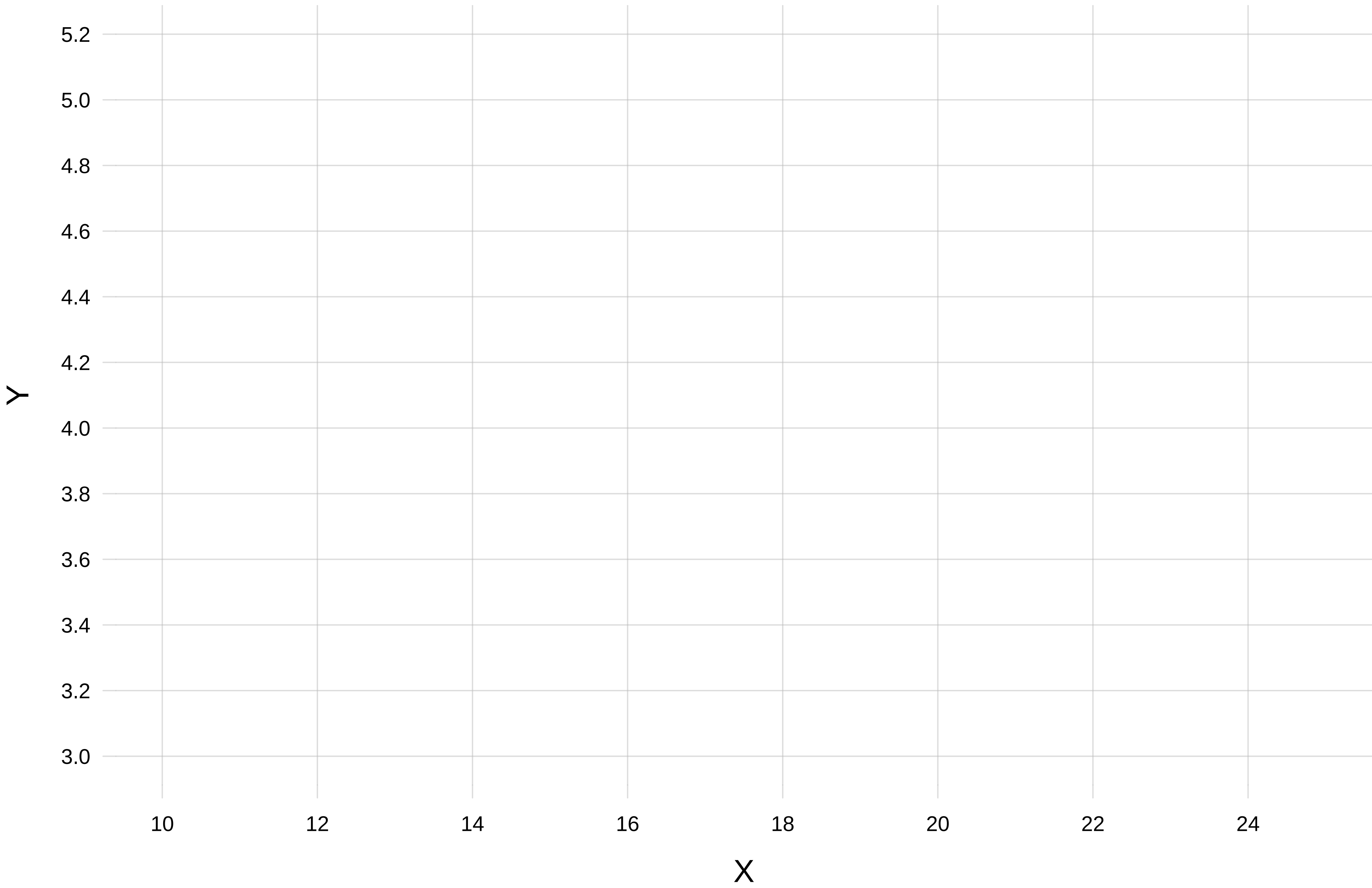
Data source: ADNI Alzheimer's Disease Neuroimaging Initiative

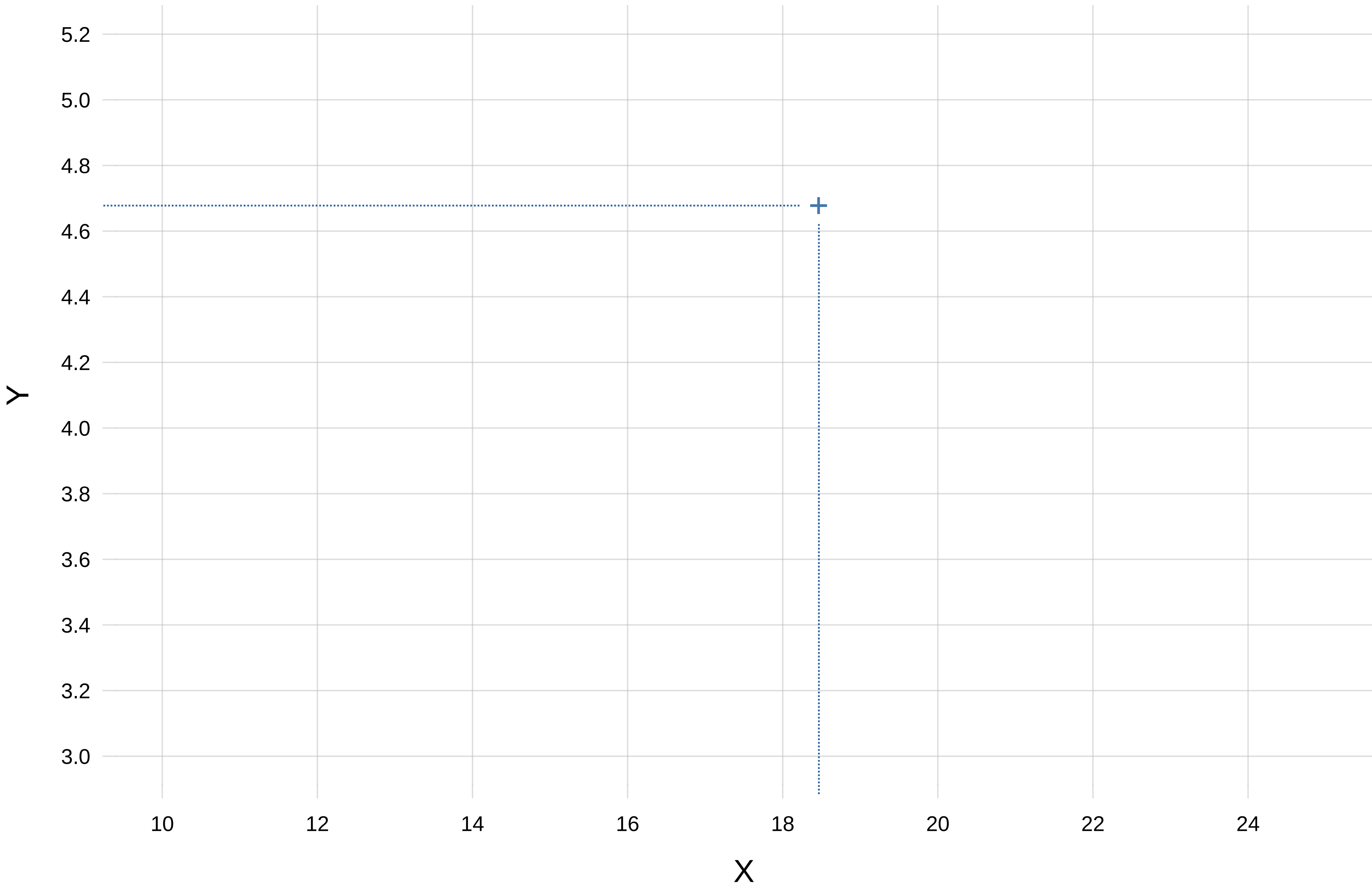
Ingrid's study: 12 + 1 variates, 678 datapoints

Alexandra's study: 11 + 1 variates, 708 datapoints (43 missing values)

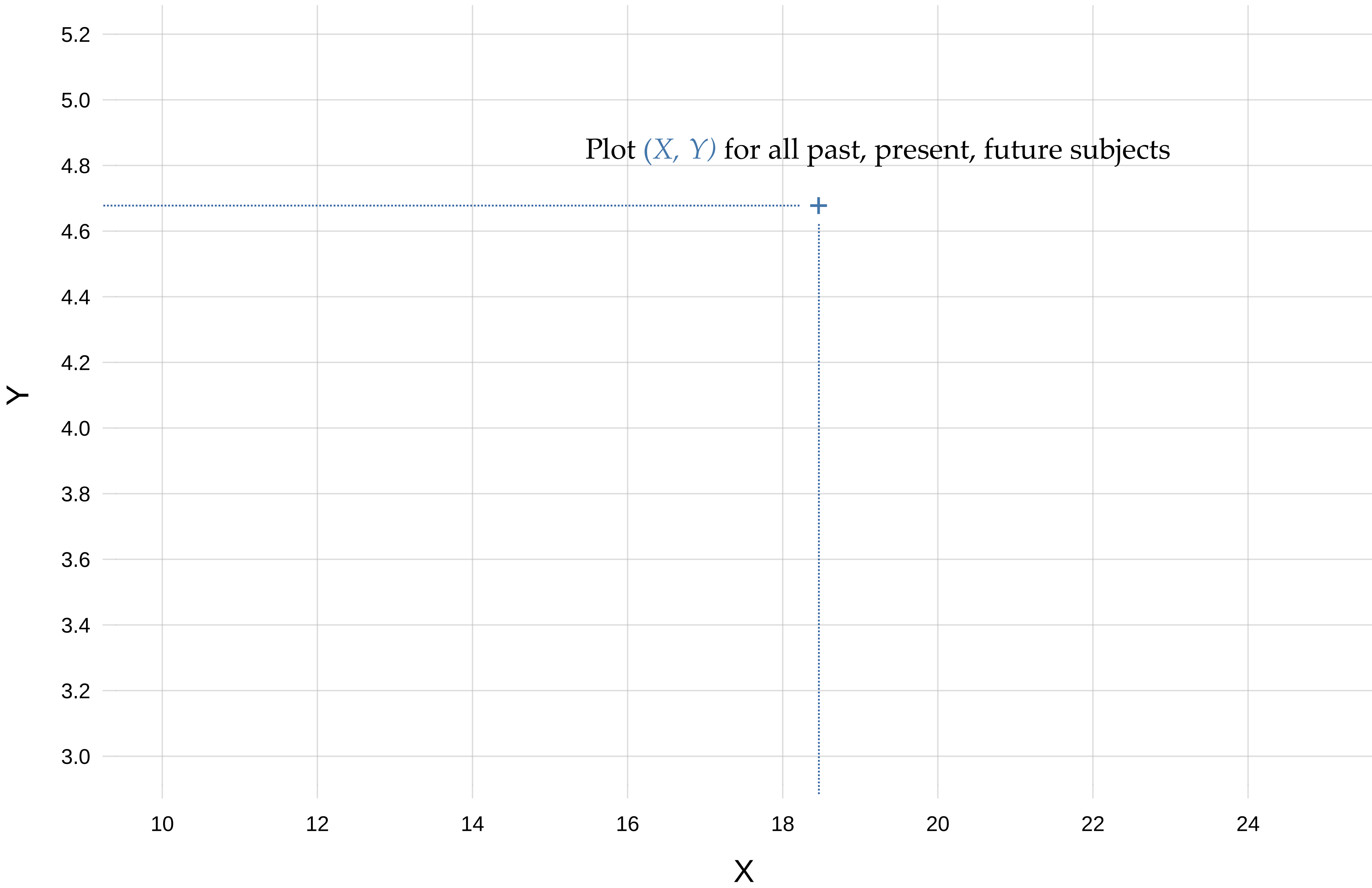
Computation time: ~65 h/study (3 parallel sessions to assess numeric convergence)

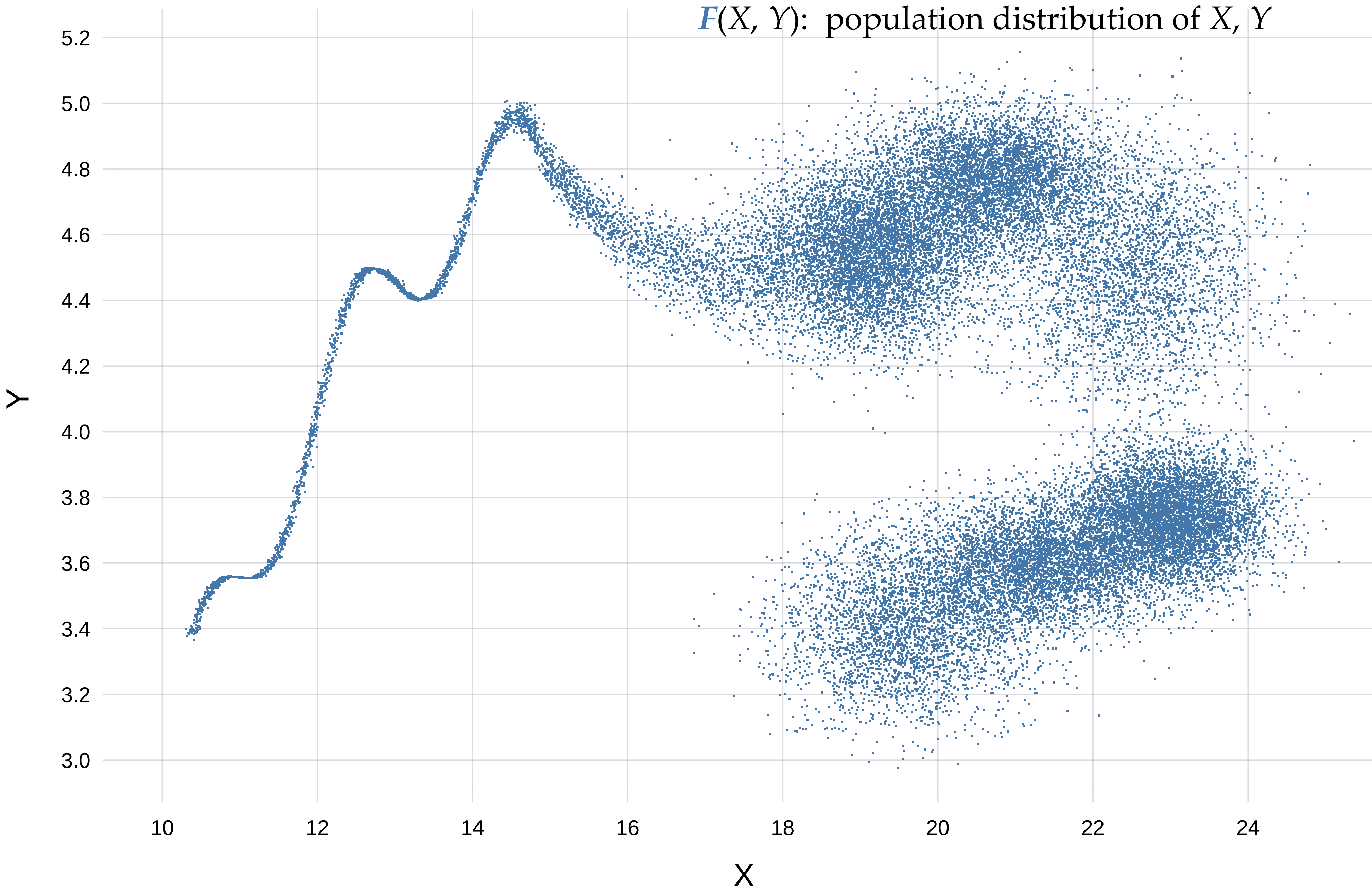
HPC UNINETT ʃigma2


5.2
5.0
4.8
4.6
4.4
4.2
4.0
3.8
3.6
3.4
3.2
3.0
Y
10
12
14
16
18
20
22
24
X

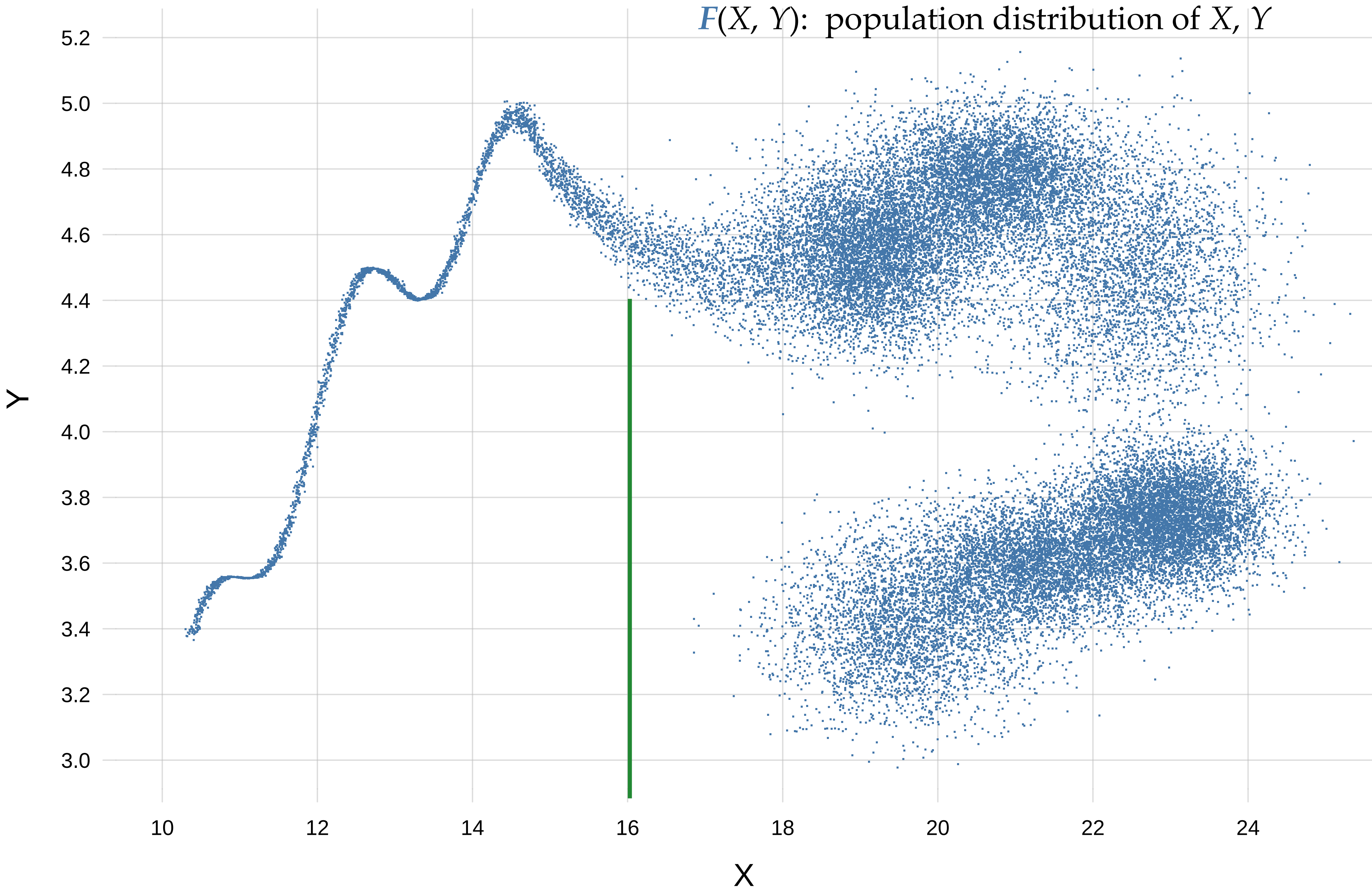


5.2
5.0
4.8
4.6
4.4
4.2
4.0
3.8
3.6
3.4
3.2
3.0
Y
10
12
14
16
18
20
22
24
X


Plot (X, Y) for all past, present, future subjects
5.2
5.0
4.8
4.6
4.4
4.2
4.0
3.8
3.6
3.4
3.2
3.0
Y
10
12
14
16
18
20
22
24
X


$F(X, Y)$: population distribution of X, Y
5.2
5.0
4.8
4.6
4.4
4.2
4.0
3.8
3.6
3.4
3.2
3.0
Y
10
12
14
16
18
20
22
24
X

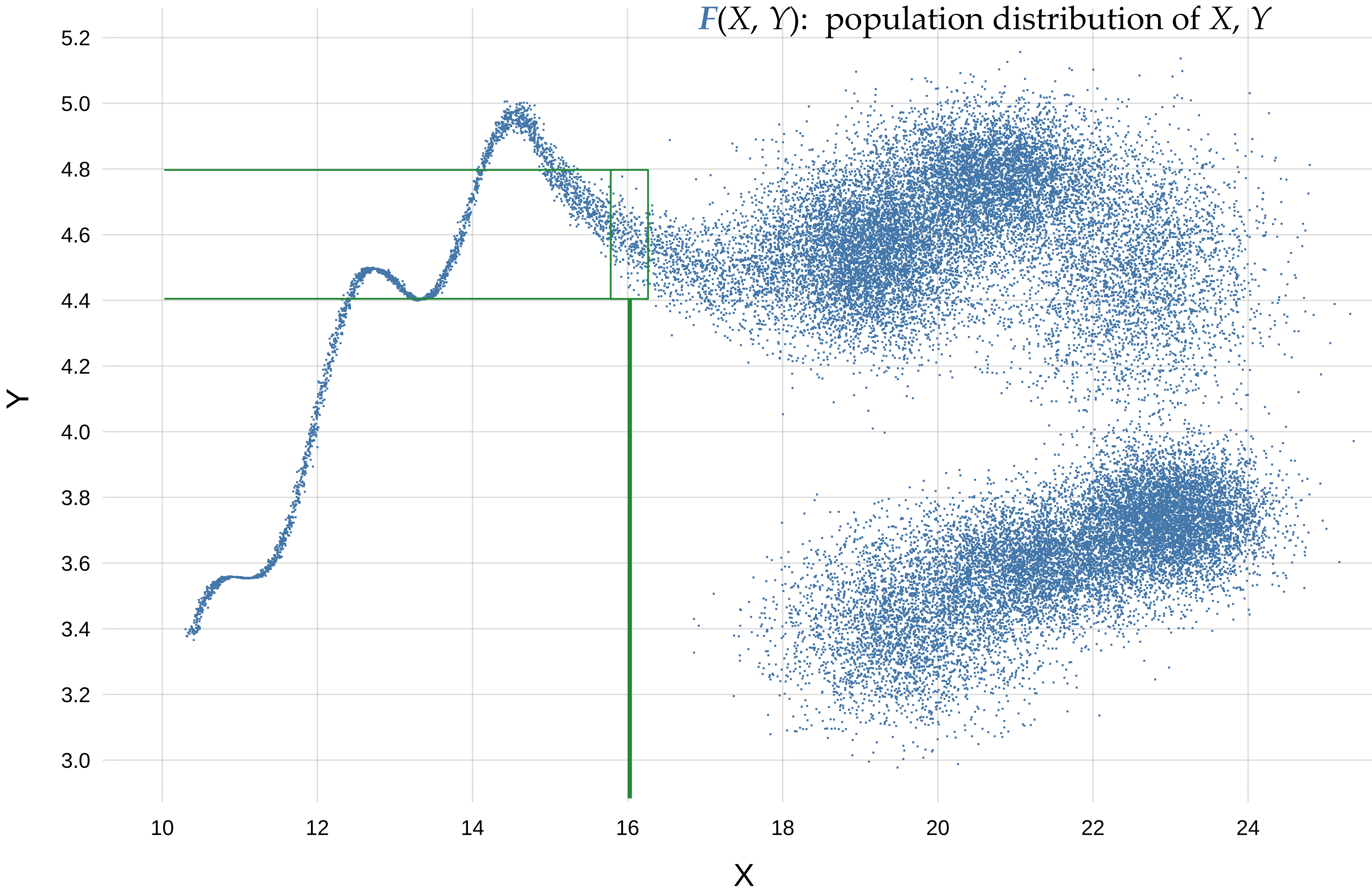
patient: $X = 16$



$F(X, Y)$: population distribution of X, Y
5.2
5.0
4.8
4.6
4.4
4.2
4.0
3.8
3.6
3.4
3.2
3.0
Y
10
12
14
16
18
20
22
24
X

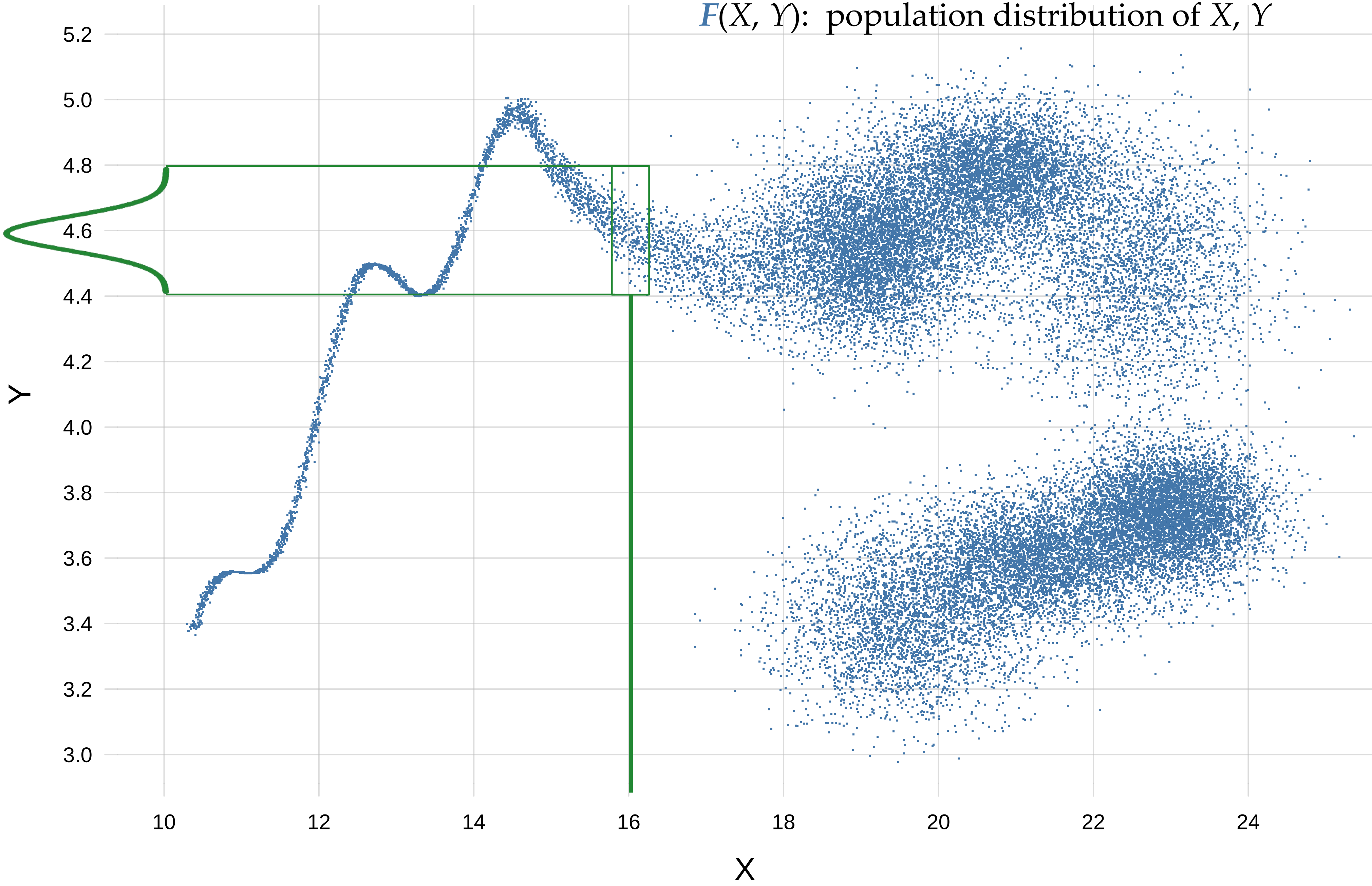
patient: $X = 16$

$\Rightarrow \quad Y \approx 4.5\text{–}4.7$

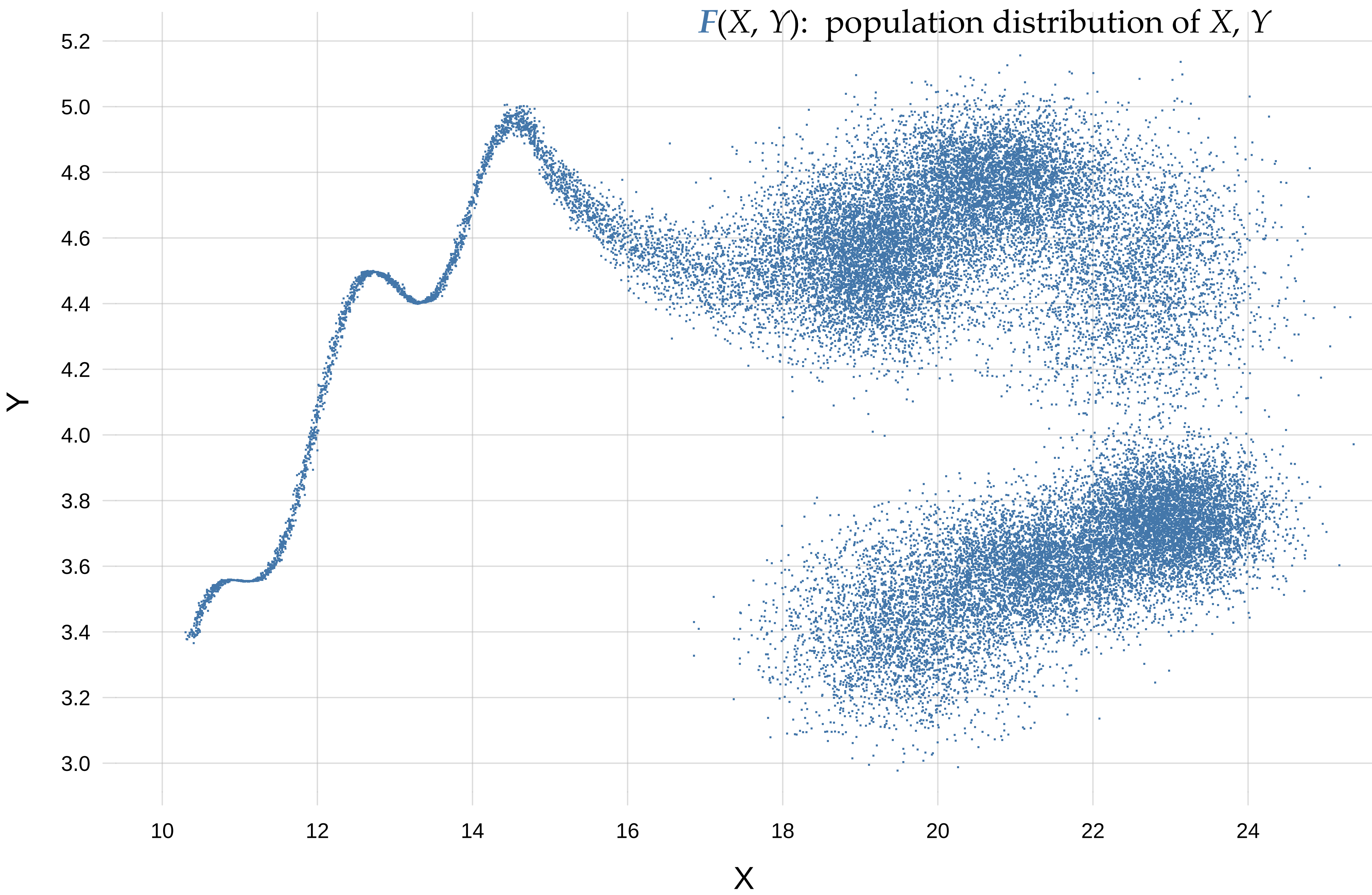


patient: $X = 16$

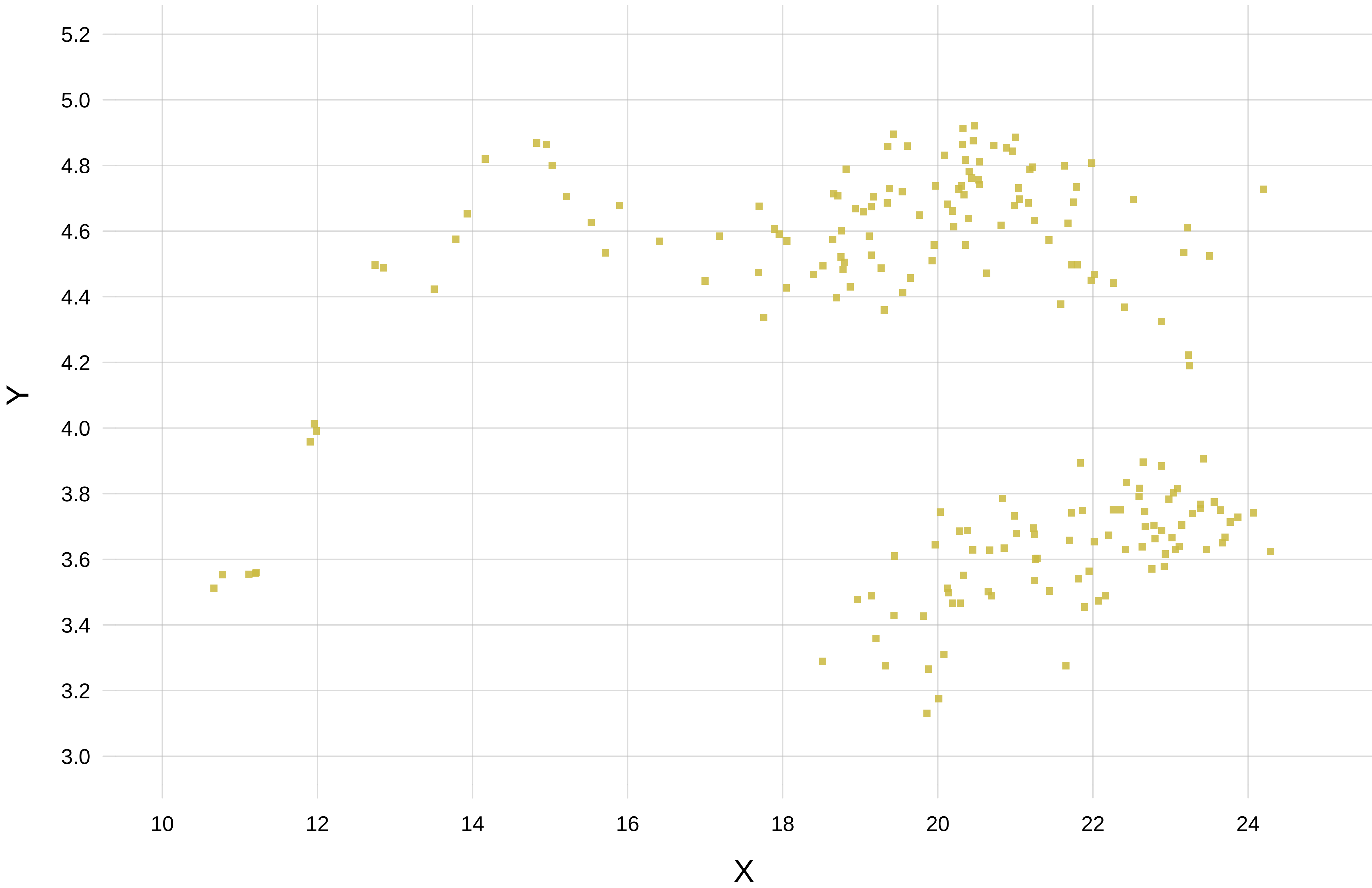
$\Rightarrow \quad Y \approx 4.5\text{–}4.7$



$$\mathrm{P}(y \mid x) = \boldsymbol{F}(y \mid x)$$



$F(X, Y)$: population distribution of X, Y
5.2
5.0
4.8
4.6
4.4
4.2
4.0
3.8
3.6
3.4
3.2
3.0
Y
10
12
14
16
18
20
22
24
X


5.2
5.0
4.8
4.6
4.4
4.2
4.0
3.8
3.6
3.4
3.2
3.0
Y
10
12
14
16
18
20
22
24
X

$$\mathrm{P}(y \mid x) = \boldsymbol{F}(y \mid x)$$

$$P(y \mid x) = \int F(y \mid x)\ p(F \mid \text{data})\ dF$$

probability = average over all possible population distributions

$$P(y \mid x) = \int F(y \mid x)\ p(F \mid \text{data})\ dF$$

probability = average over all possible population distributions

$$p(F \mid \text{data})$$

$$P(y \mid x) = \int F(y \mid x)\ p(F \mid \text{data})\ dF$$

probability = average over all possible population distributions

$$p(F \mid \text{data}) \propto \underbrace{F(y_1, x_1) \times F(y_2, x_2) \times F(y_3, x_3) \times \ldots}_{\text{how well the 'candidate' distribution fits the data}}$$

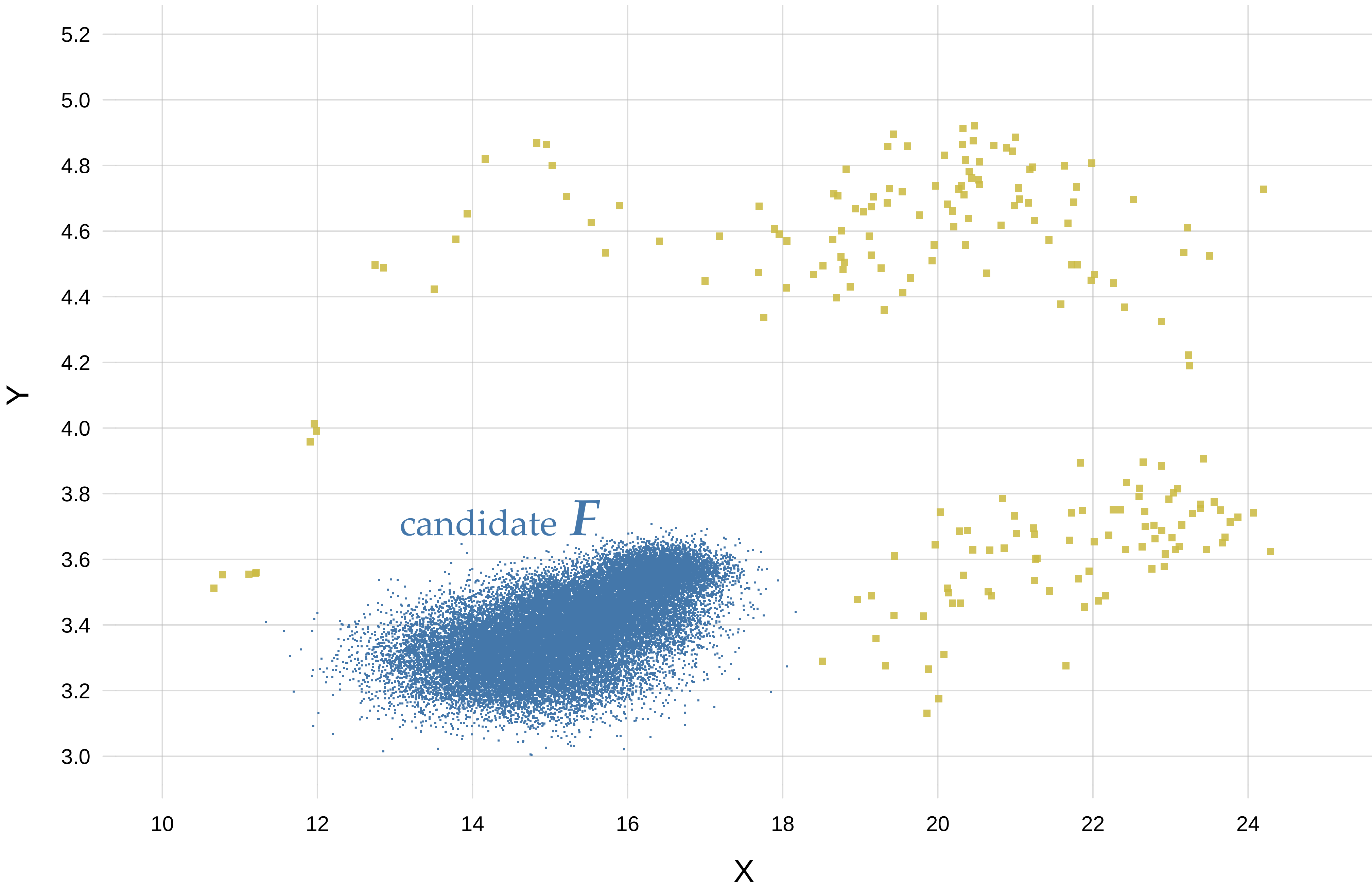
$$\mathrm{P}(y \mid x) = \int F(y \mid x)\ \mathrm{p}(F \mid \text{data})\ \mathrm{d}F$$

probability = average over all possible population distributions

$$\mathrm{p}(F \mid \text{data}) \propto \underbrace{F(y_1, x_1) \times F(y_2, x_2) \times F(y_3, x_3) \times \ldots}_{\text{how well the 'candidate' distribution fits the data}} \times \underbrace{\mathrm{p}(F \mid \text{prior info})}_{\text{extra-data knowledge}}$$

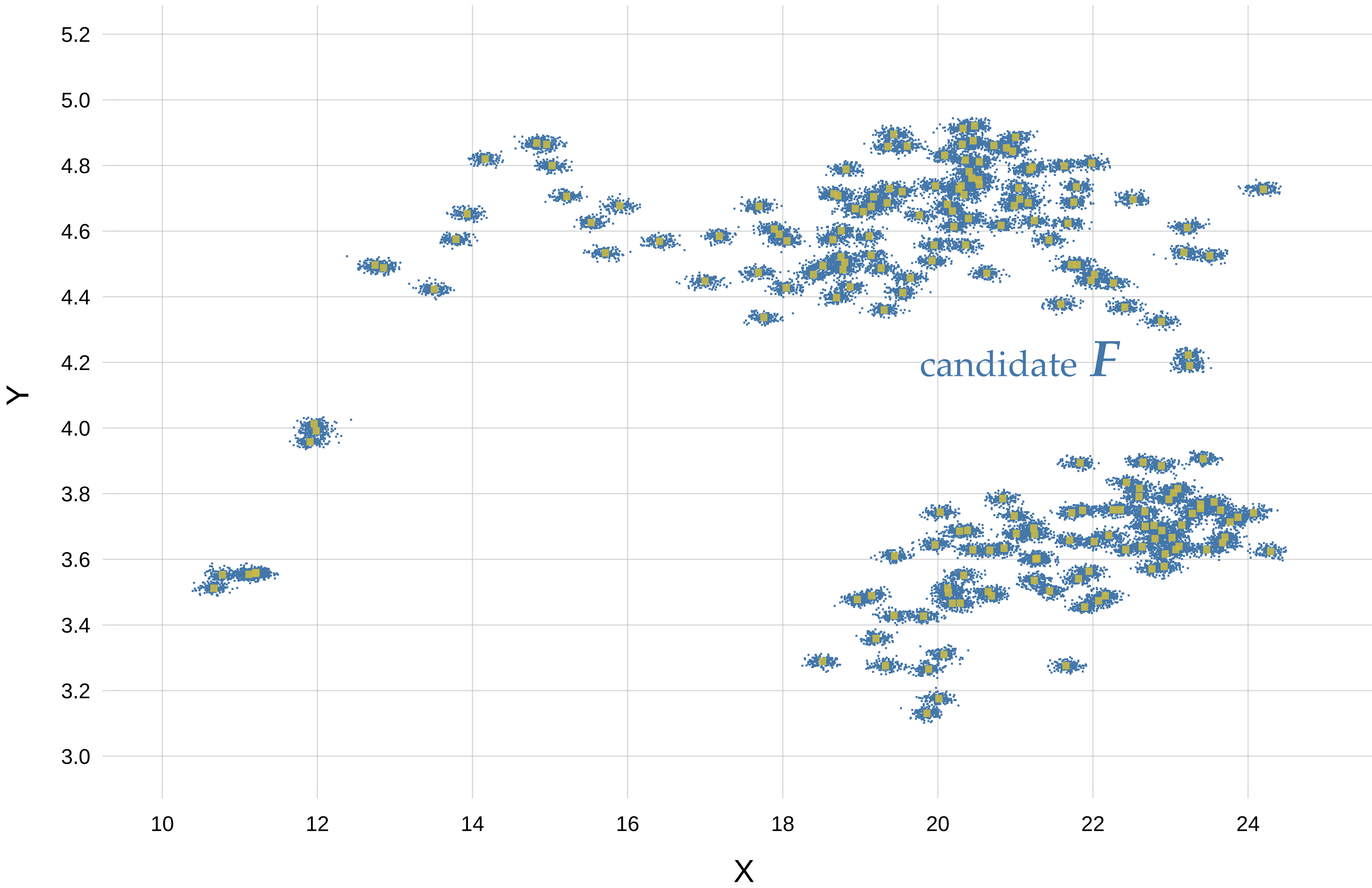
poor candidate: doesn't fit the data

$$\underbrace{F(y_1, x_1) \times F(y_2, x_2) \times F(y_3, x_3) \times \ldots}_{\text{low}} \times \underbrace{p(F \mid \text{prior info})}_{\text{high}}$$



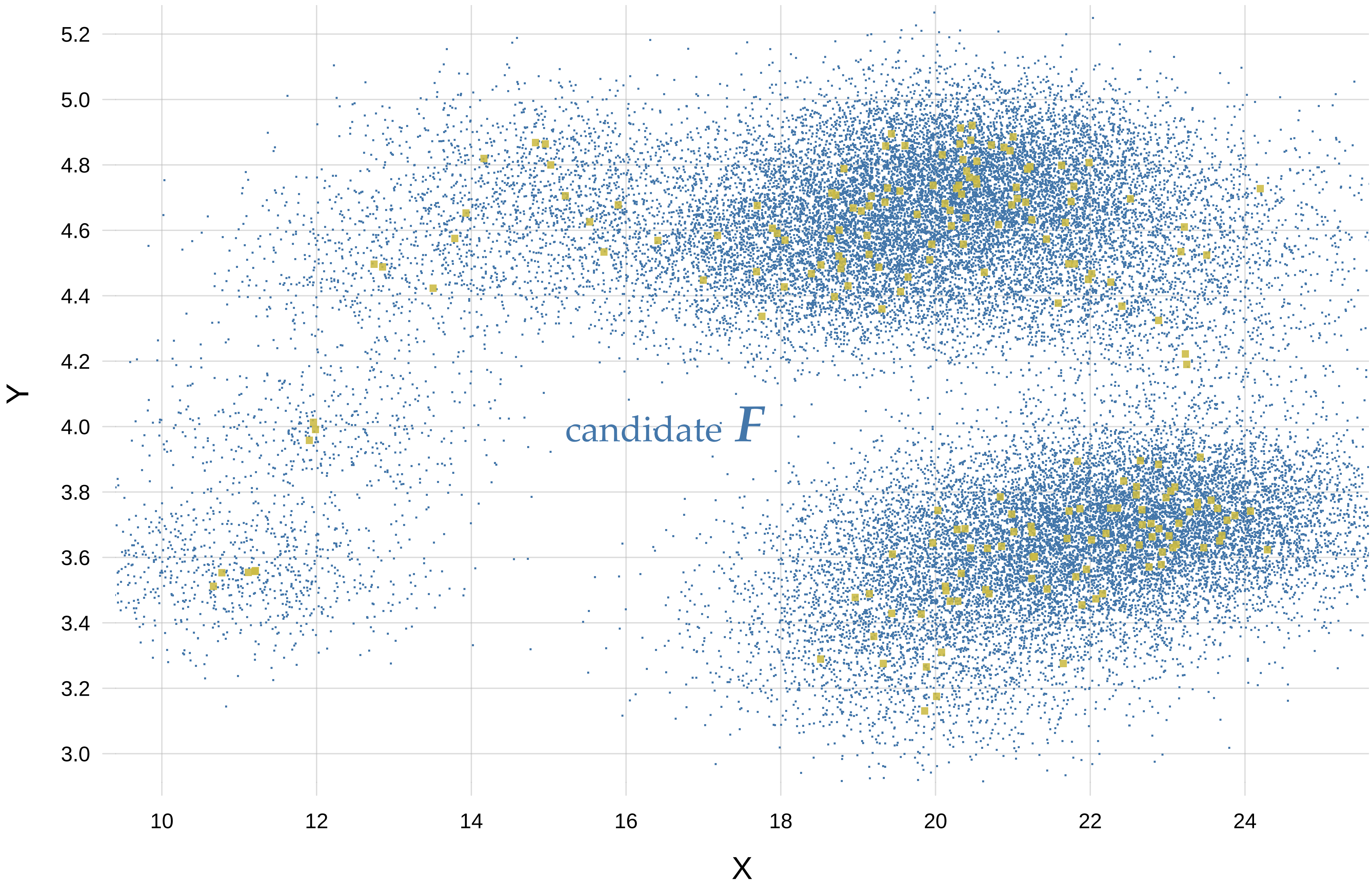
poor candidate: biologically implausible

$$\underbrace{F(y_1, x_1) \times F(y_2, x_2) \times F(y_3, x_3) \times \ldots}_{\text{high}} \times \underbrace{p(F \mid \text{prior info})}_{\text{low}}$$



reasonable candidate

$$\underbrace{F(y_1, x_1) \times F(y_2, x_2) \times F(y_3, x_3) \times \ldots}_{\text{high}} \times p(\underbrace{F \mid \text{prior info}}_{\text{high}})$$

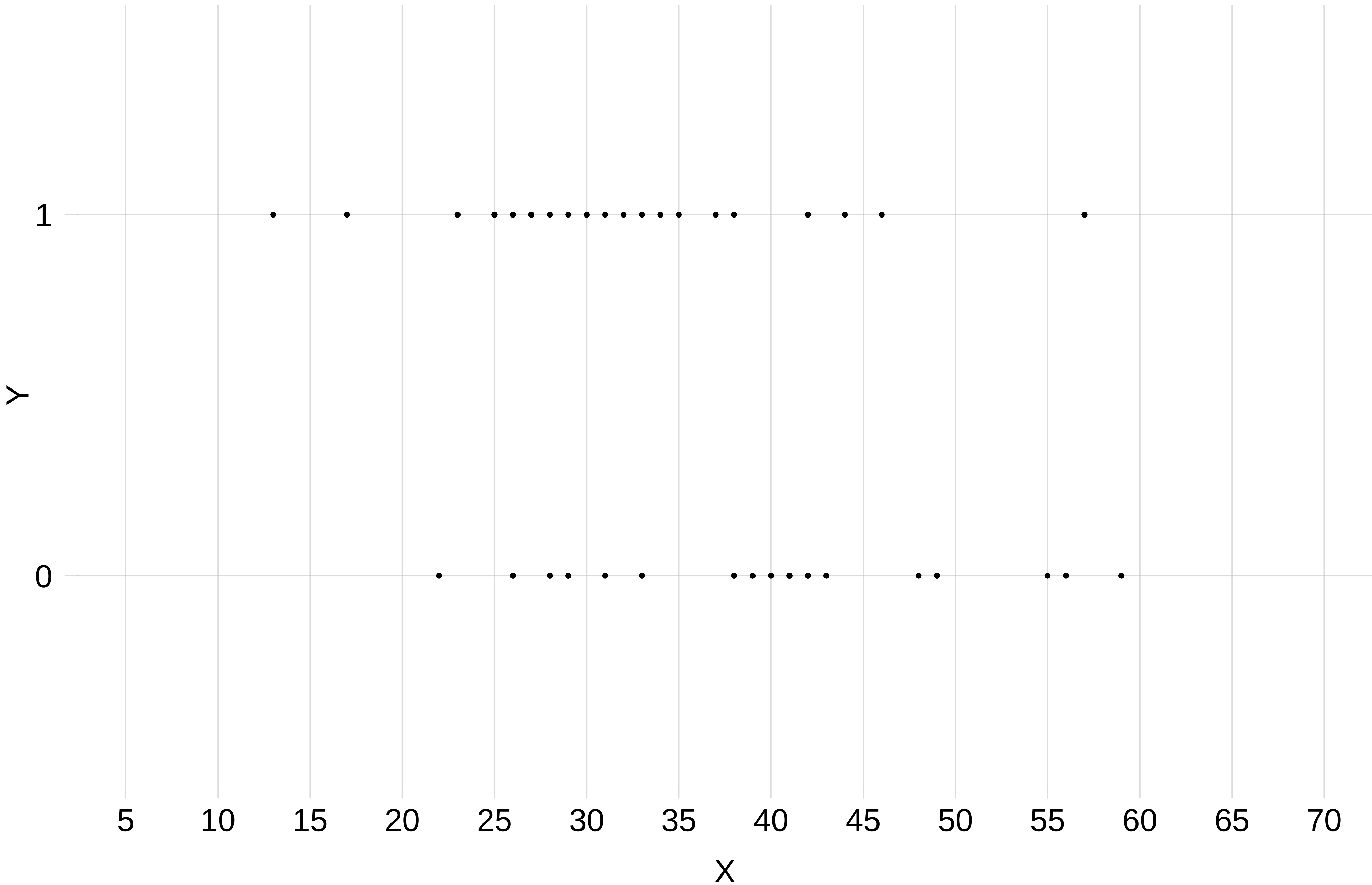


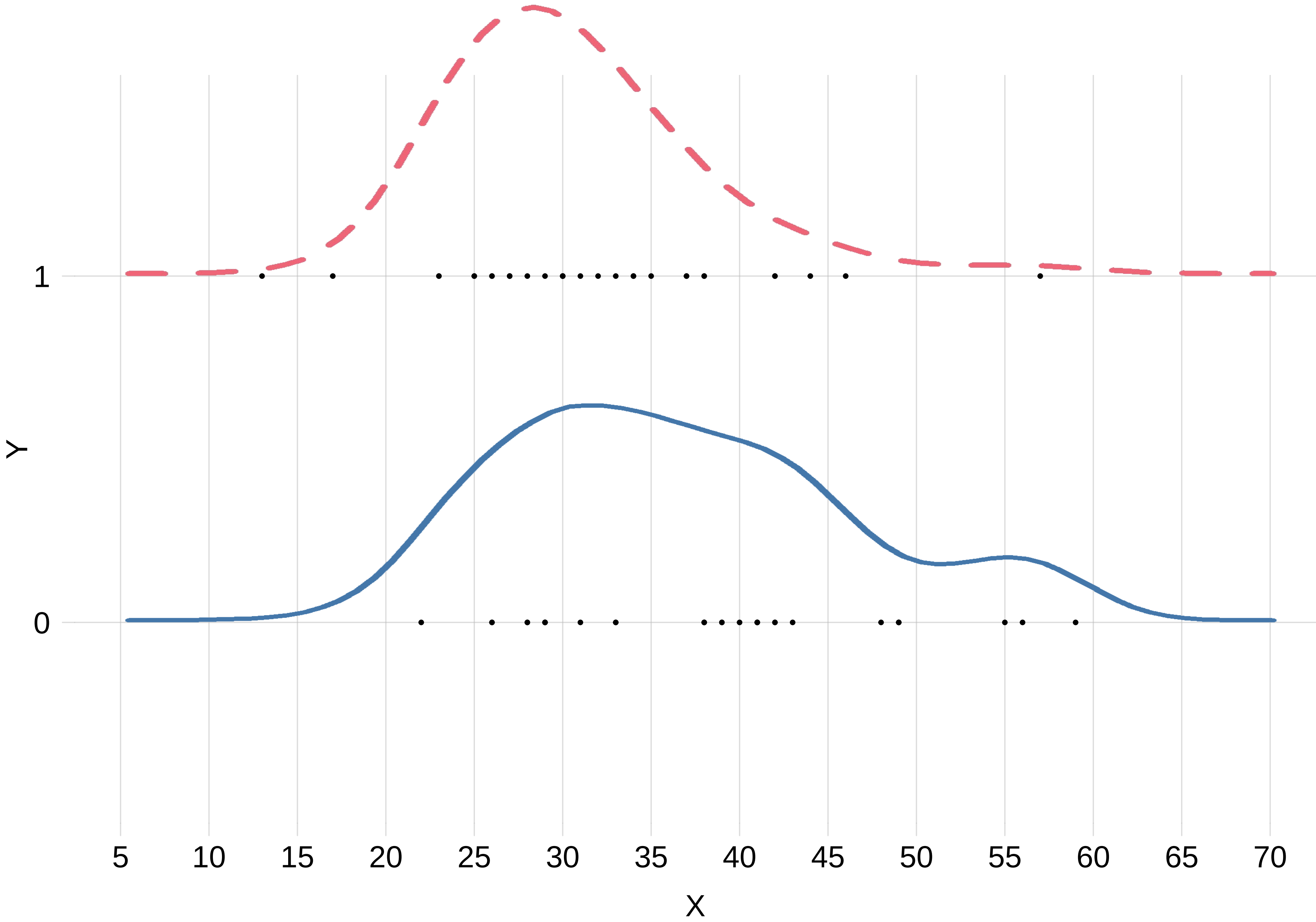
$$intuition \quad \rightarrow \quad mathematics$$

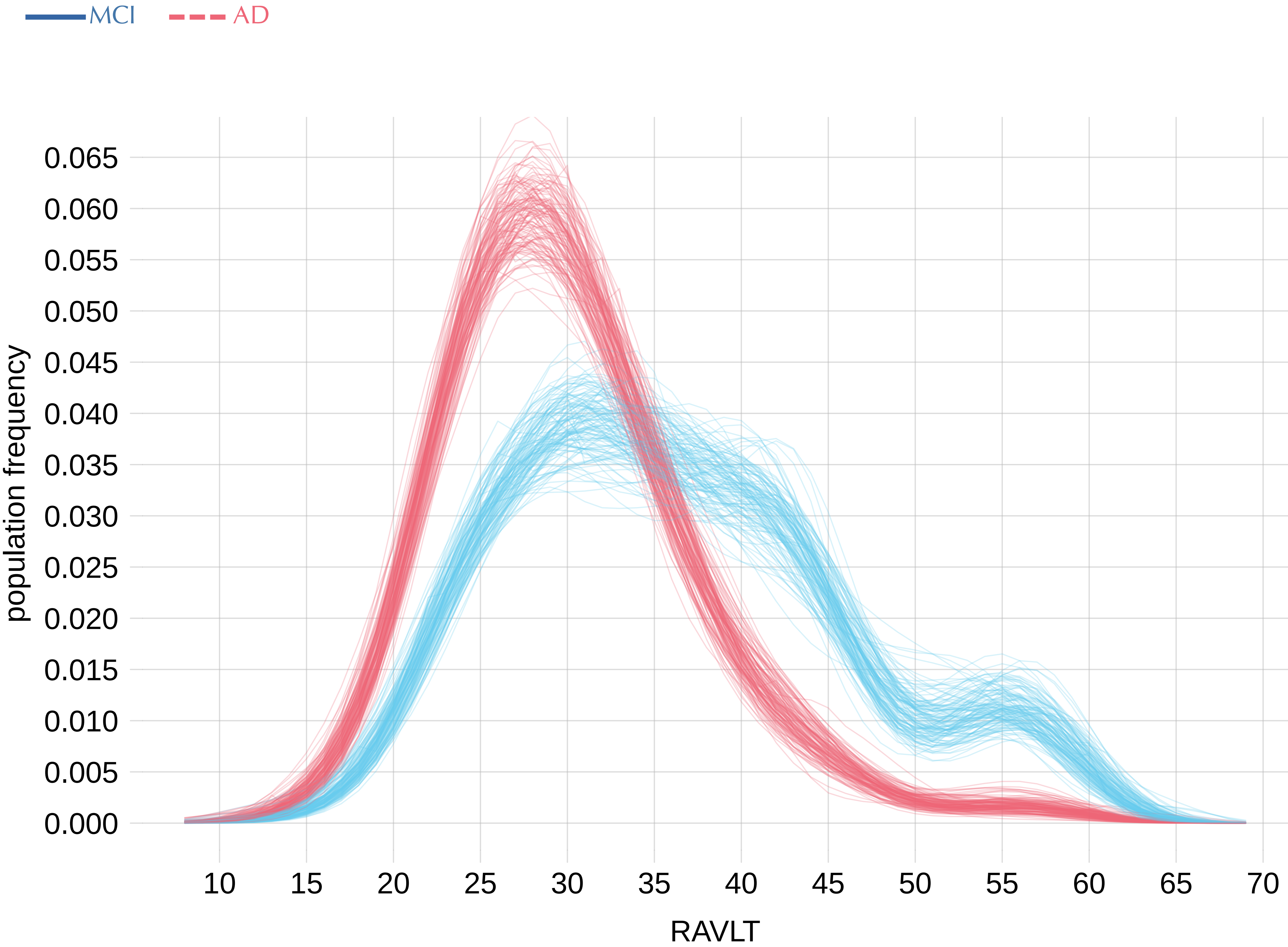
~~*intuition* $\longrightarrow$ *mathematics*~~

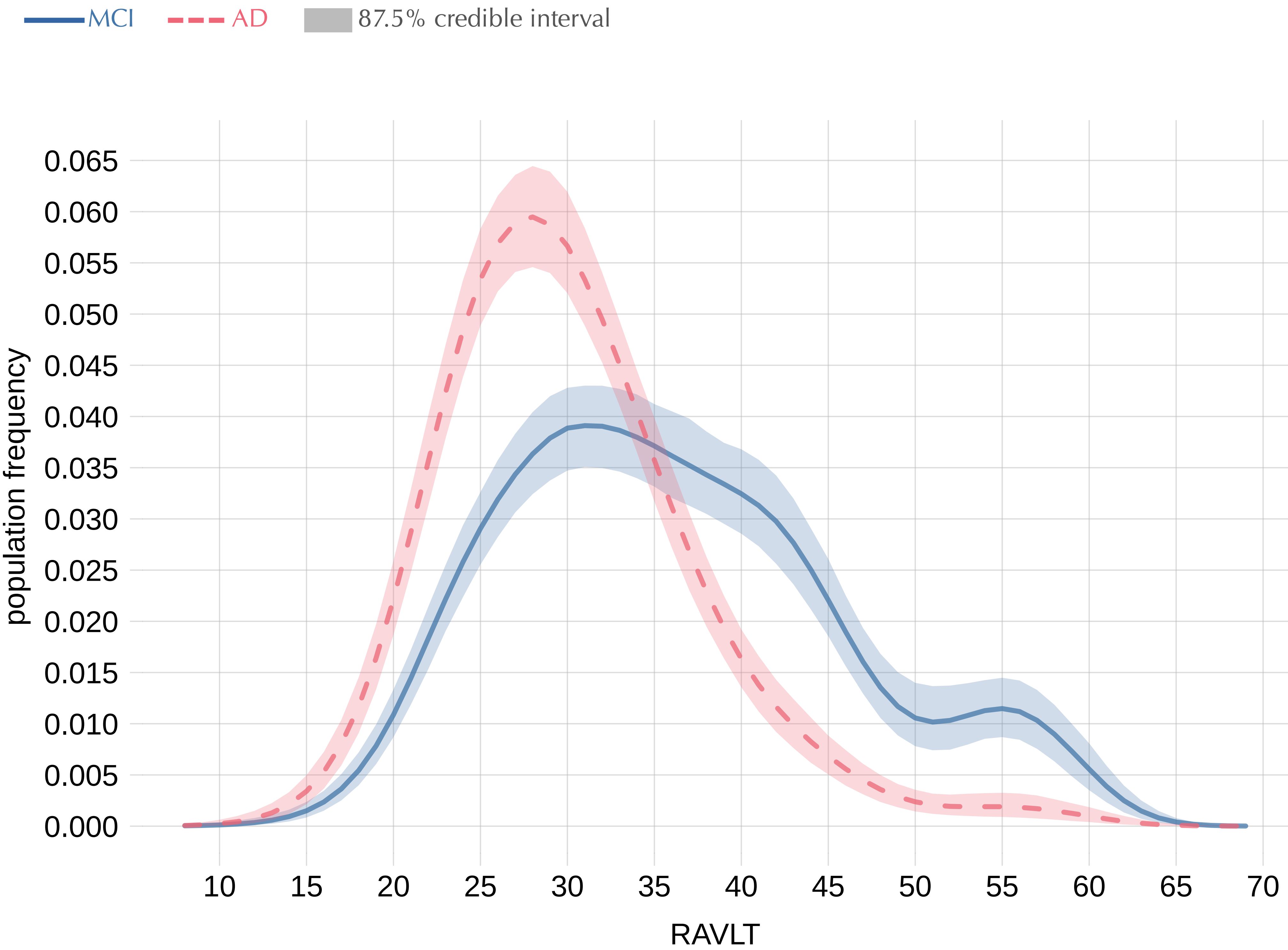
first principles $\rightarrow$ *mathematics* $\rightarrow$ *intuition*

('Bayesian')

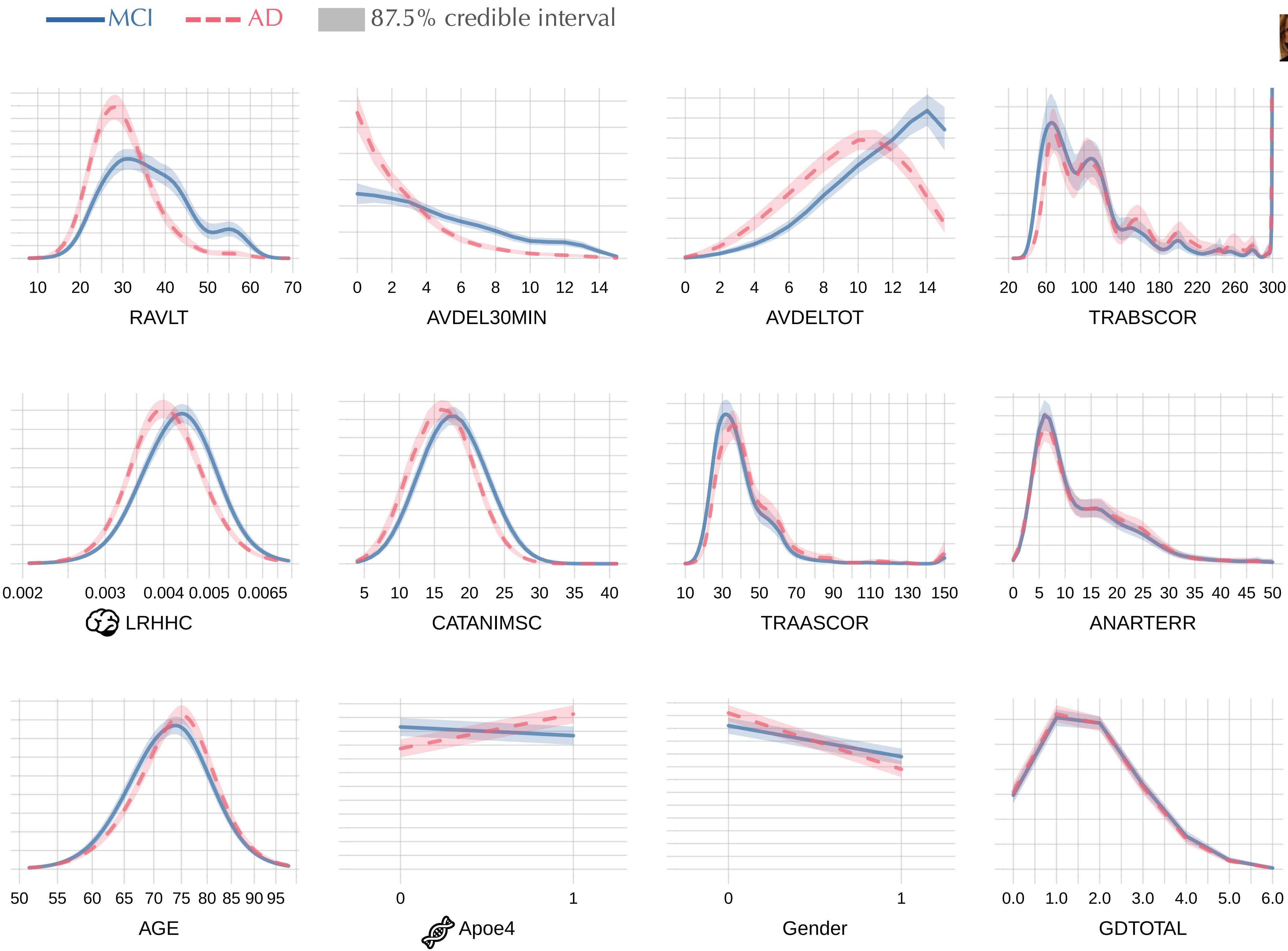

1
0
Y
5
10
15
20
25
30
35
40
45
50
55
60
65
70
X

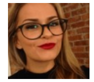

1
0
Y
5
10
15
20
25
30
35
40
45
50
55
60
65
70
X

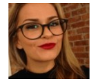

MCI
AD
population frequency
0.065
0.060
0.055
0.050
0.045
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
10
15
20
25
30
35
40
45
50
55
60
65
70
RAVLT

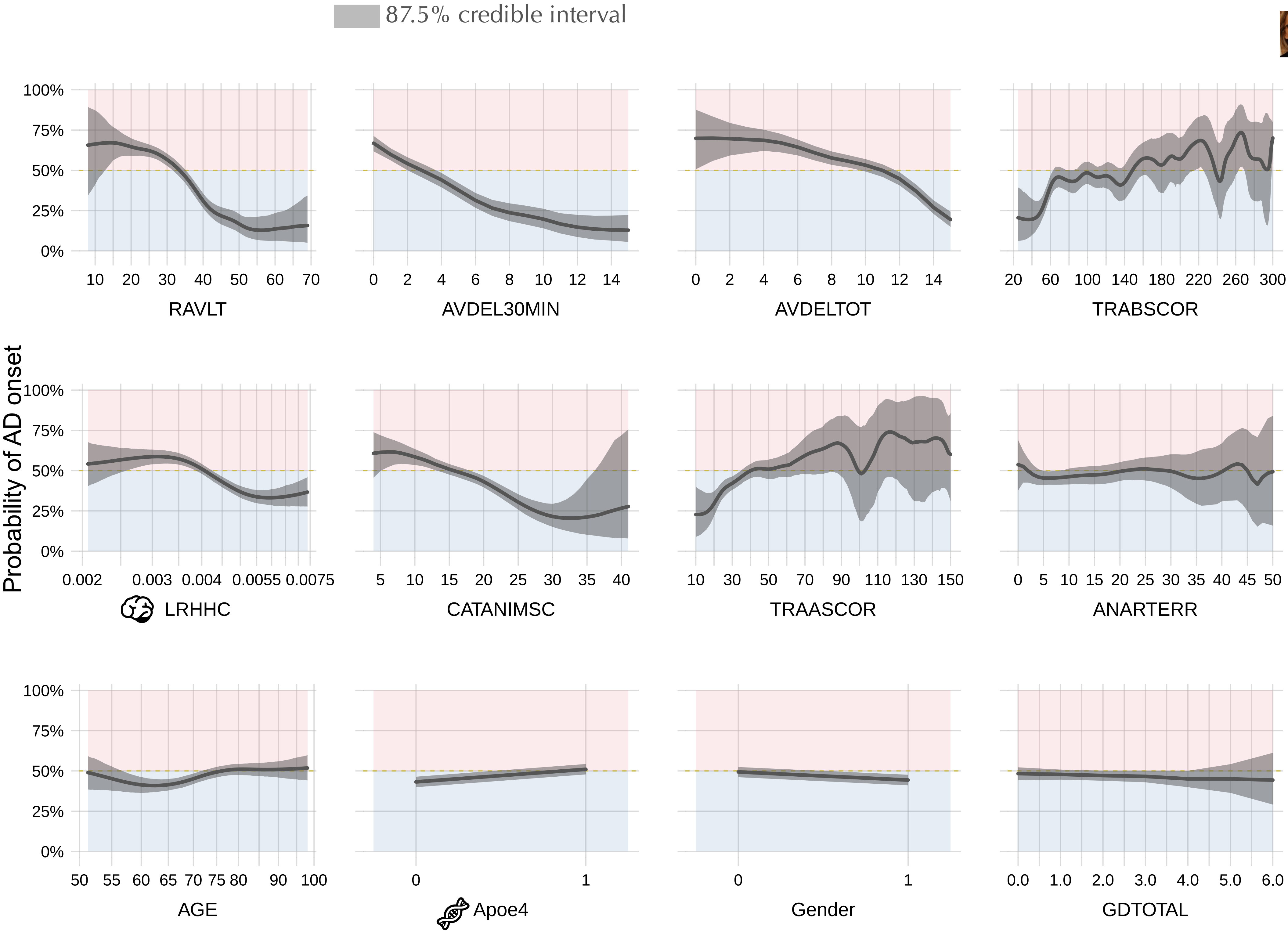


MCI
AD
87.5% credible interval
population frequency
0.065
0.060
0.055
0.050
0.045
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
10
15
20
25
30
35
40
45
50
55
60
65
70
RAVLT

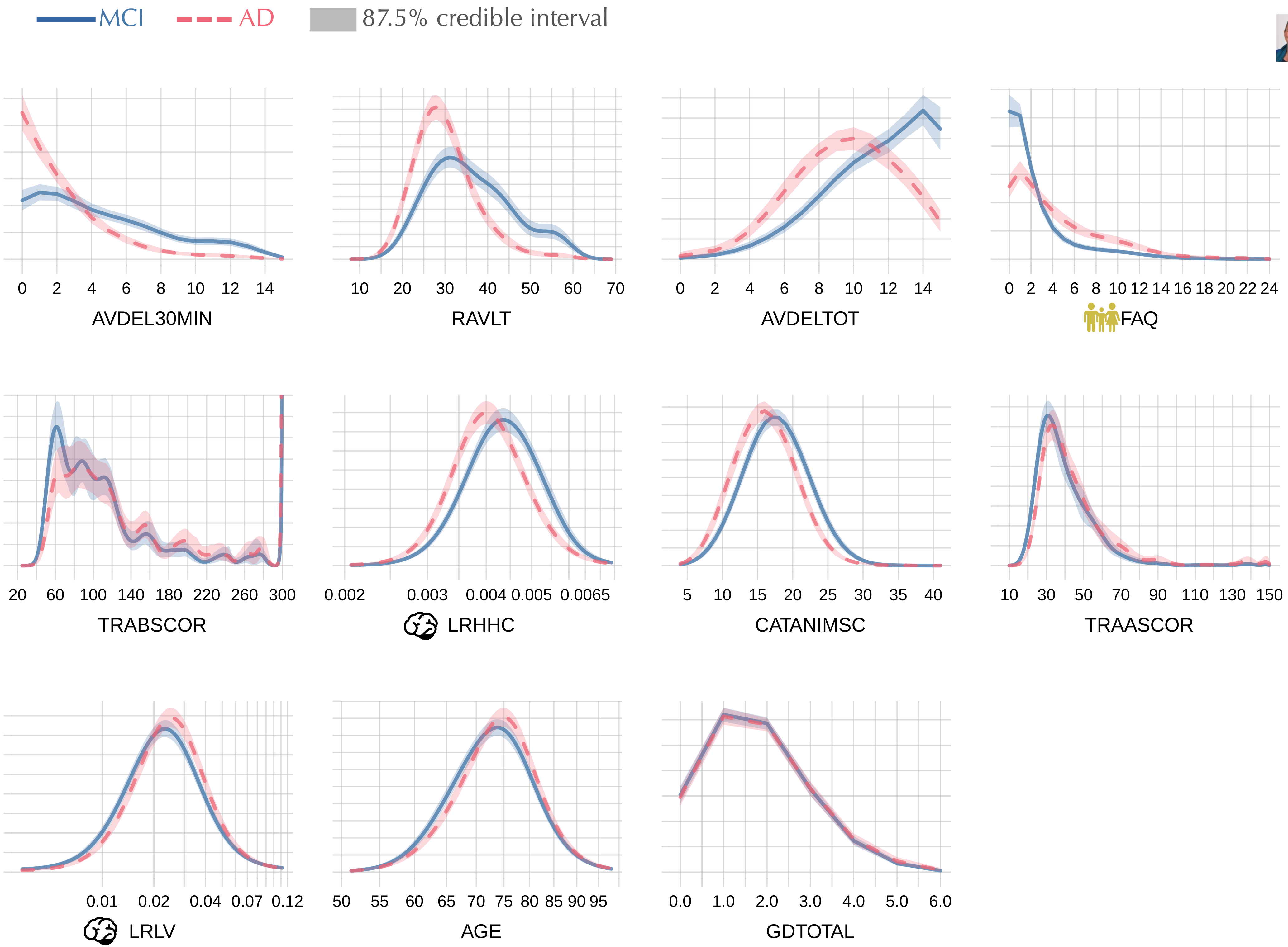

MCI
AD
87.5% credible interval
10 20 30 40 50 60 70
RAVLT
0 2 4 6 8 10 12 14
AVDEL30MIN
0 2 4 6 8 10 12 14
AVDELTOT
20 60 100 140 180 220 260 300
TRABSCOR
0.002 0.003 0.004 0.005 0.0065
LRHHC
5 10 15 20 25 30 35 40
CATANIMSC
10 30 50 70 90 110 130 150
TRAASCOR
0 5 10 15 20 25 30 35 40 45 50
ANARTERR
50 55 60 65 70 75 80 85 90 95
AGE
0 1
Apoe4
0 1
Gender
0.0 1.0 2.0 3.0 4.0 5.0 6.0
GDTOTAL

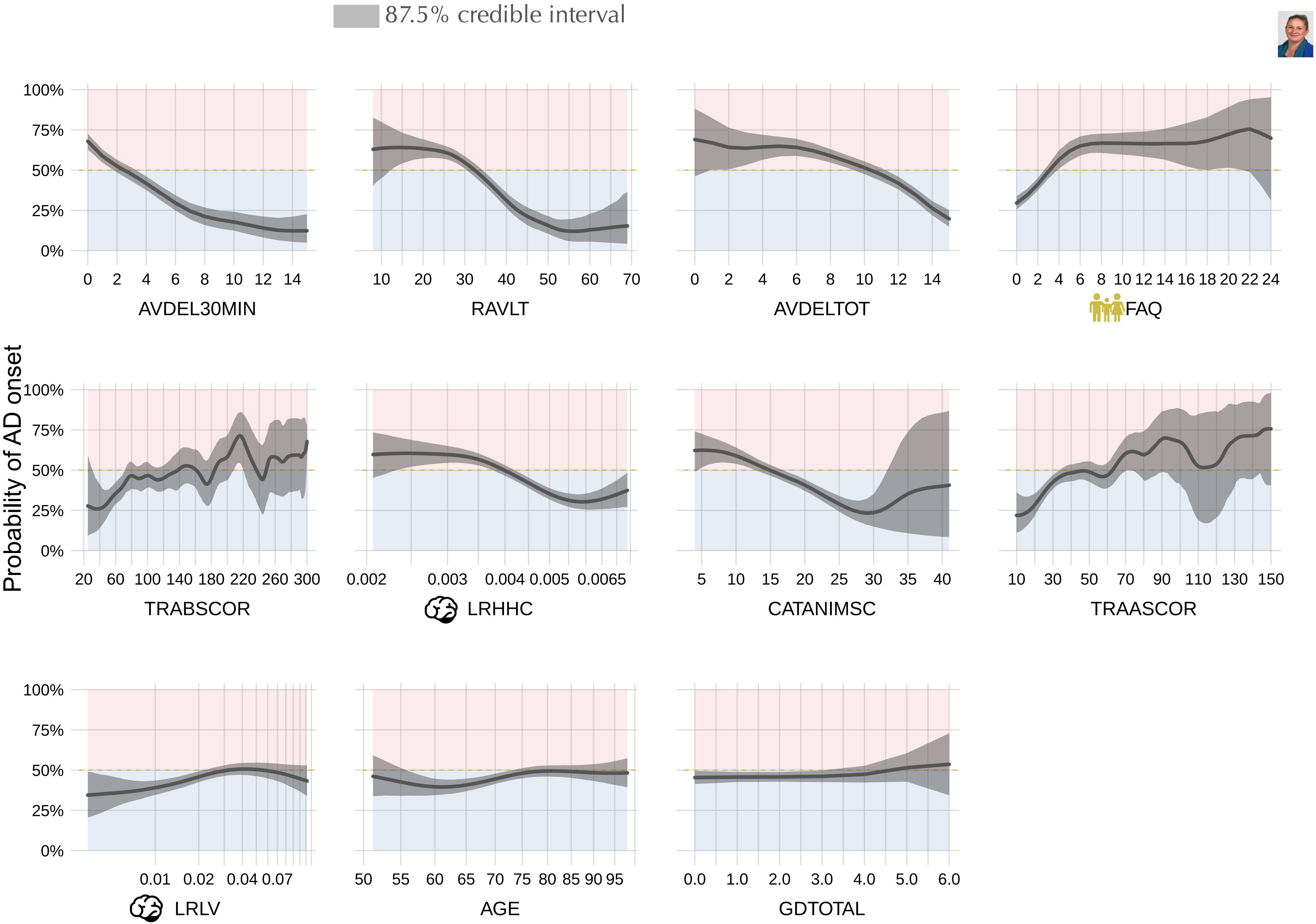






87.5% credible interval
Probability of AD onset
100%
75%
50%
25%
0%
RAVLT
10 20 30 40 50 60 70
AVDEL30MIN
0 2 4 6 8 10 12 14
AVDELTOT
0 2 4 6 8 10 12 14
TRABSCOR
20 60 100 140 180 220 260 300
LRHHC
0.002 0.003 0.004 0.0055 0.0075
CATANIMSC
5 10 15 20 25 30 35 40
TRAASCOR
10 30 50 70 90 110 130 150
ANARTERR
0 5 10 15 20 25 30 35 40 45 50
AGE
50 55 60 65 70 75 80 90 100
Apoe4
0 1
Gender
0 1
GDTOTAL
0.0 1.0 2.0 3.0 4.0 5.0 6.0


MCI
AD
87.5% credible interval
0 2 4 6 8 10 12 14
AVDEL30MIN
10 20 30 40 50 60 70
RAVLT
0 2 4 6 8 10 12 14
AVDELTOT
0 2 4 6 8 10 12 14 16 18 20 22 24
FAQ
20 60 100 140 180 220 260 300
TRABSCOR
0.002 0.003 0.004 0.005 0.0065
LRHHC
5 10 15 20 25 30 35 40
CATANIMSC
10 30 50 70 90 110 130 150
TRAASCOR
0.01 0.02 0.04 0.07 0.12
LRLV
50 55 60 65 70 75 80 85 90 95
AGE
0.0 1.0 2.0 3.0 4.0 5.0 6.0
GDTOTAL



87.5% credible interval
Probability of AD onset
100%
75%
50%
25%
0%
AVDEL30MIN
RAVLT
AVDELTOT
FAQ
TRABSCOR
LRHHC
CATANIMSC
TRAASCOR
LRLV
AGE
GDTOTAL

Interesting characteristics of $\boldsymbol{F}(Y, X)$ in Ingrid's & Alexandra's studies:

- Several high-density regions in the 12D space

- Some features seem more robust if used in a 'discriminative' way: $\mathrm{P}(Y \mid X)$, others in a 'generative' way: $\mathrm{P}(X \mid Y)$

$$\mathrm{P}(Y| X_{\mathrm{d}}, X_{\mathrm{g}}) \propto \mathrm{P}(X_{\mathrm{g}}| Y)\, \mathrm{P}(Y| X_{\mathrm{d}})$$

How to quantify the ‘importance’ or ‘prognostic power’ of a set of features?

“Language is a product of, and reflects, thinking.
Sloppy usage reflects sloppy thinking, a kind of thinking
incompatible with good scientific habits of mind”

(D. J. Helfand)

Prediction problem:
guess the six digits of the winning lottery ticket ??????

Clue A: ✔✔✔✔??

Clue B: ✔✔✔?✔?

Clue C: ???✔✔✔

What is the 'importance' or 'predictive power' of each clue?

Scenario 1: we can use **only one** clue

Clue **A**: ✔✔✔✔??

Clue **B**: ✔✔✔?✔?

Clue **C**: ???✔✔✔

increasing importance

Best: **A** or **B** (each gives 1/81 winning chance)

Worst: **C** (gives 1/729 winning chance)

Scenario 2: we can use **all** clues

Clue **A**: ✔✔✔✔??
Clue **B**: ✔✔✔?✔?
Clue **C**: ???✔✔✔

→ We fully know the winning number! 💰

Scenario 2: what happens if we **discard** clues?

Clue **A**: ✔✔✔✔??

Clue **B**: ✔✔✔?✔?

Clue **C**: ???✔✔✔

Scenario 2: what happens if we **discard** clues?

Clue **A**: ✔✔✔✔??
Clue **B**: ✔✔✔?✔?
Clue **C**: ???✔✔✔

- Discard **A**: still 100% win $\Rightarrow$ **A** has ‘importance **= 0**’

Scenario 2: what happens if we **discard** clues?

Clue **A**: ✔✔✔✔??
Clue **B**: ✔✔✔?✔?
Clue **C**: ???✔✔✔

- Discard **A**: still 100% win $\Rightarrow$ **A** has 'importance = **0**'
- Discard **B**: still 100% win $\Rightarrow$ **B** has 'importance = **0**'

Scenario 2: what happens if we **discard** clues?

Clue **A**: ✔✔✔✔??

Clue **B**: ✔✔✔?✔?

Clue **C**: ???✔✔✔

- Discard **A**: still 100% win $\Rightarrow$ **A** has 'importance = **0**'
- Discard **B**: still 100% win $\Rightarrow$ **B** has 'importance = **0**'

- Discard **A** *and* **B**: 1/9 winning chance

Scenario 2: what happens if we **discard** clues?

Clue **A**: ✔✔✔✔??
Clue **B**: ✔✔✔?✔?
Clue **C**: ???✔✔✔

- Discard **A**: still 100% win ⇒ **A** has 'importance **= 0**'
- Discard **B**: still 100% win ⇒ **B** has 'importance **= 0**'

- Discard **A** *and* **B**: 1/9 winning chance

⇒ **A** and **B** together have 'importance **> 0**'

Scenario 2: what happens if we **discard** clues?

Clue **A**: ✔✔✔✔??
Clue **B**: ✔✔✔?✔?
Clue **C**: ???✔✔✔

- Discard **A**: still 100% win $\Rightarrow$ **A** has ‘importance = **0**’
- Discard **B**: still 100% win $\Rightarrow$ **B** has ‘importance = **0**’

- Discard **A** *and* **B**: 1/9 winning chance

$\Rightarrow$ **A** and **B** together have ‘importance > **0**’

$$‘\,0 + 0 \neq 0\,’$$

‘Importance’ or ‘predictive power’ is *not* an *additive* property

Scenario 3: we have to **discard one** clue. Which?

Clue **A**: ✓✓✓✓??
Clue **B**: ✓✓✓?✓?
Clue **C**: ???✓✓✓

increasing importance

discard **A**: still 100% win

discard **B**: still 100% win

discard **C**: 1/9 winning chance

→ If we have to discard one clue, it's most important that we keep **C**

increasing importance

Scenario 1:
choose one clue

C **A** **B**

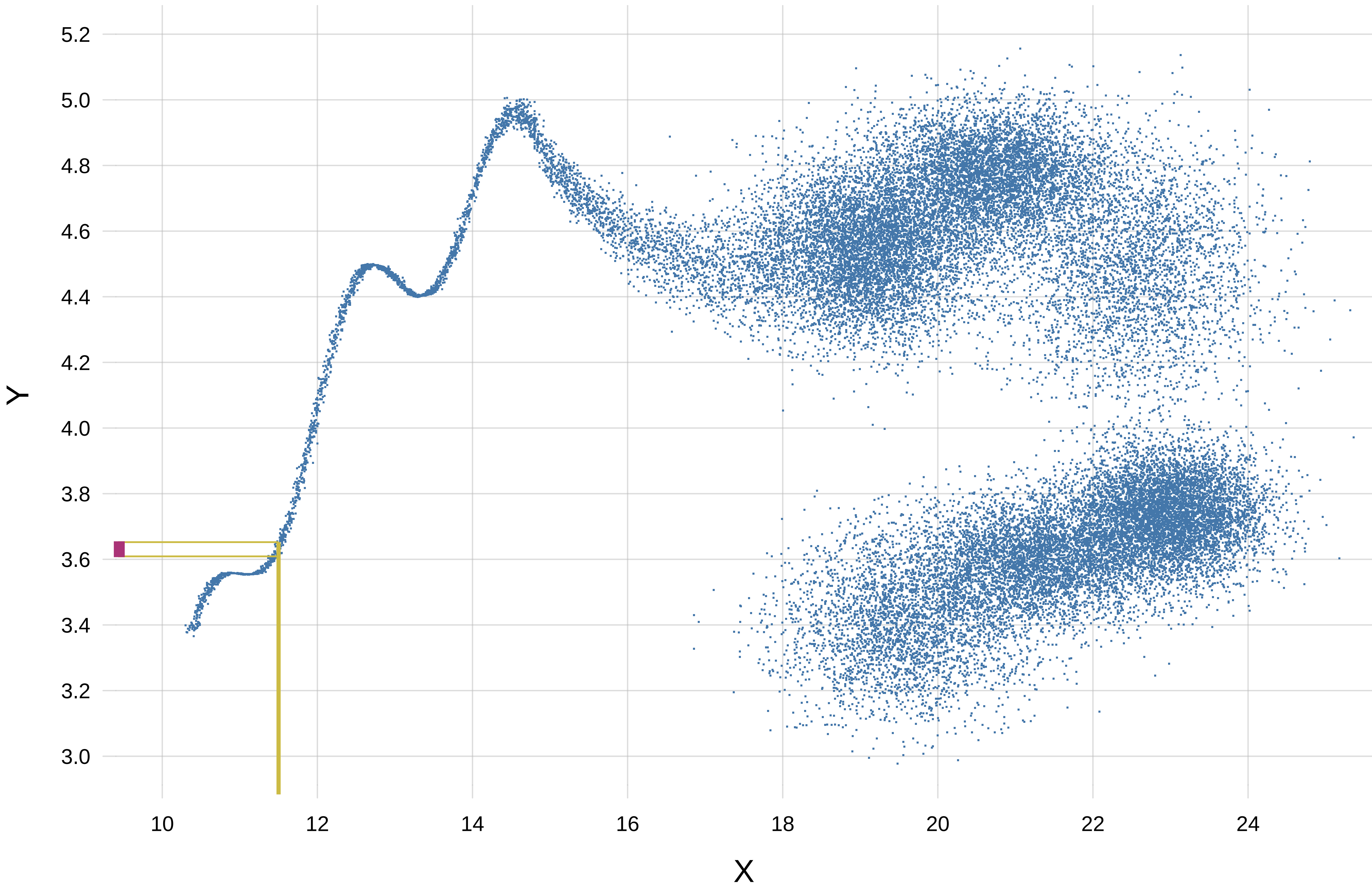
Scenario 3:
discard one clue

A **B** **C**

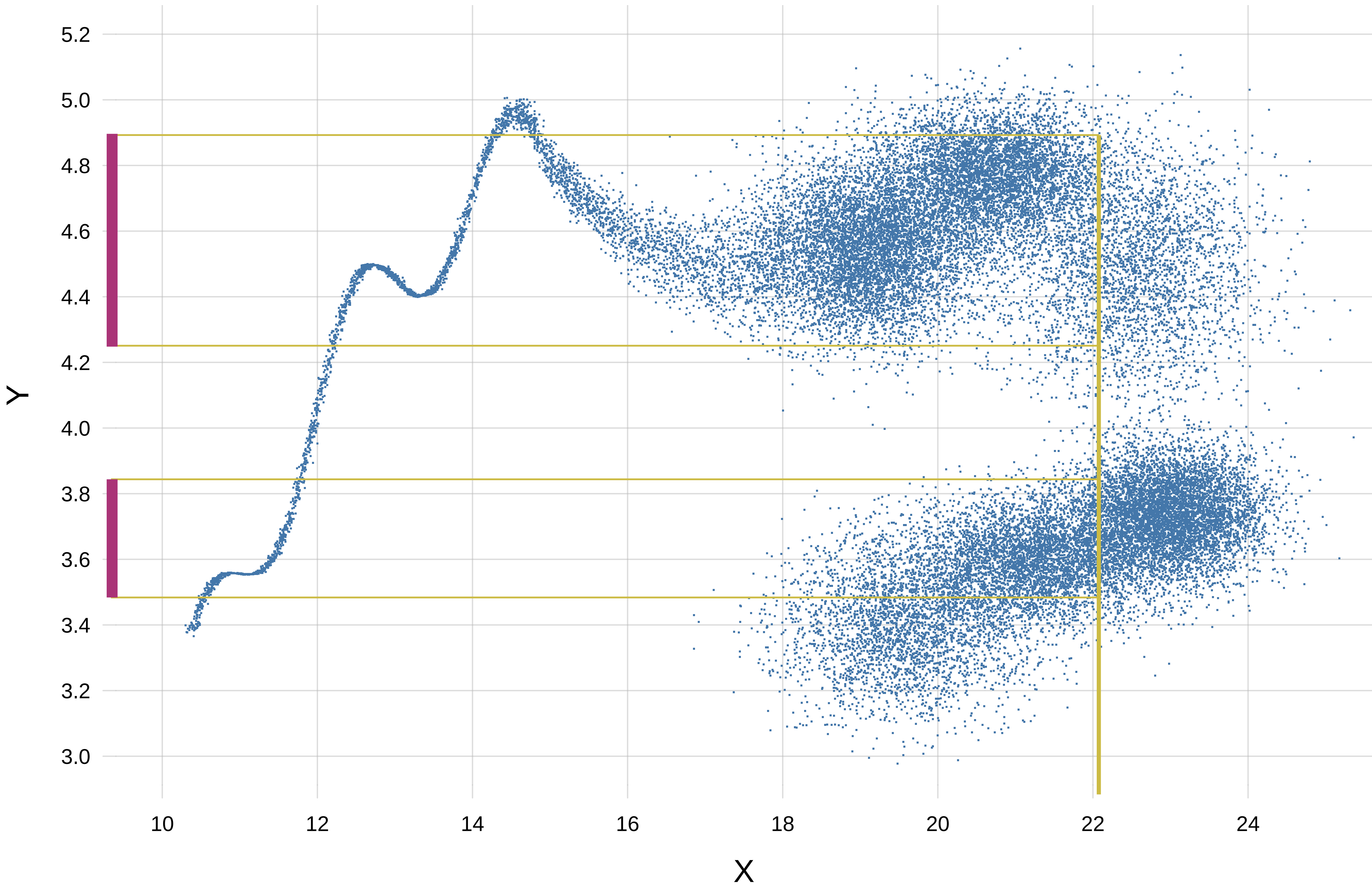
increasing importance

‘Importance’ or ‘predictive power’ of X is *context-dependent*
(which other features are we considering?)

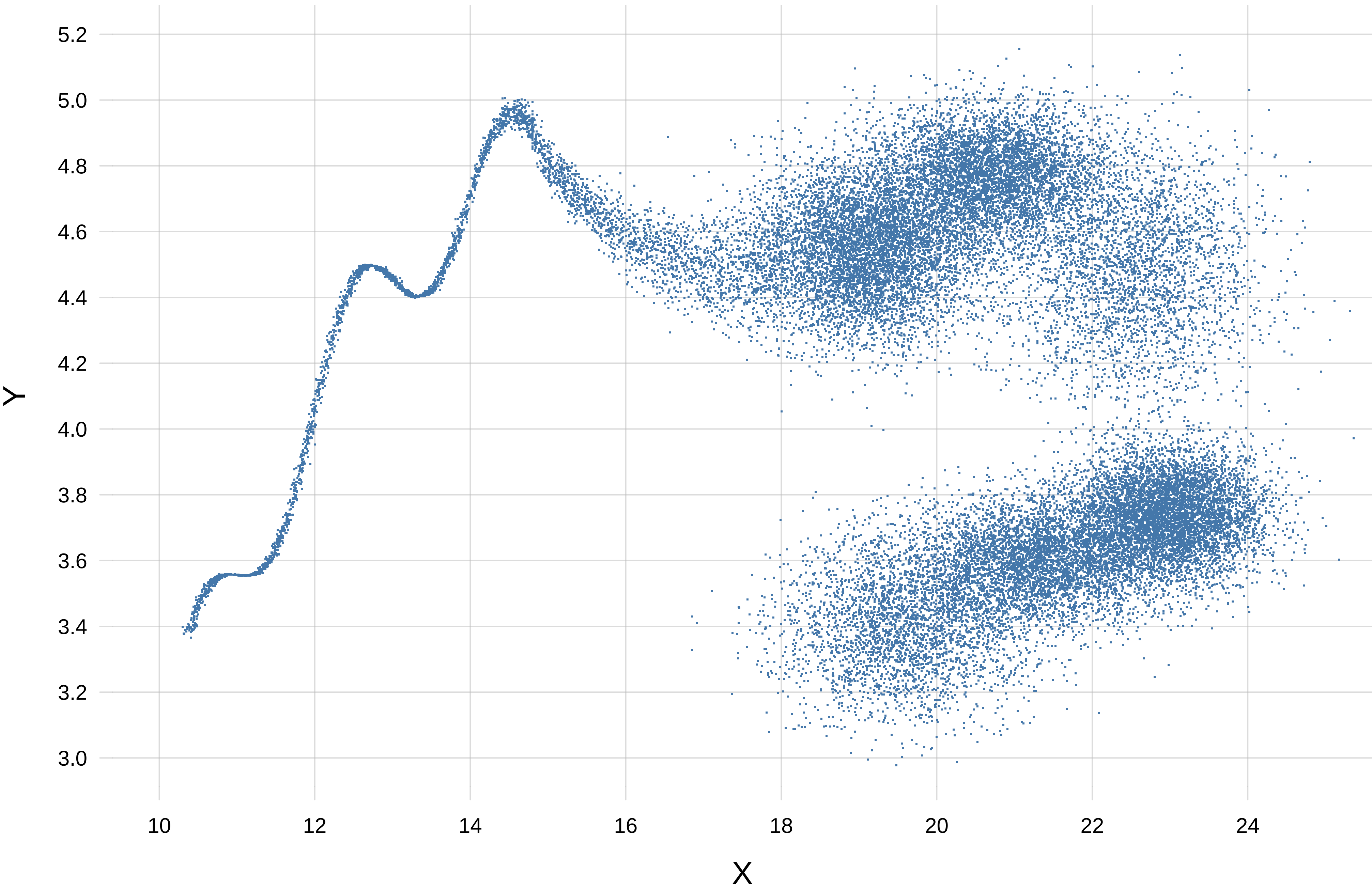
$x = 11.5 \;\Rightarrow\; y \approx 3.60\text{–}3.65$

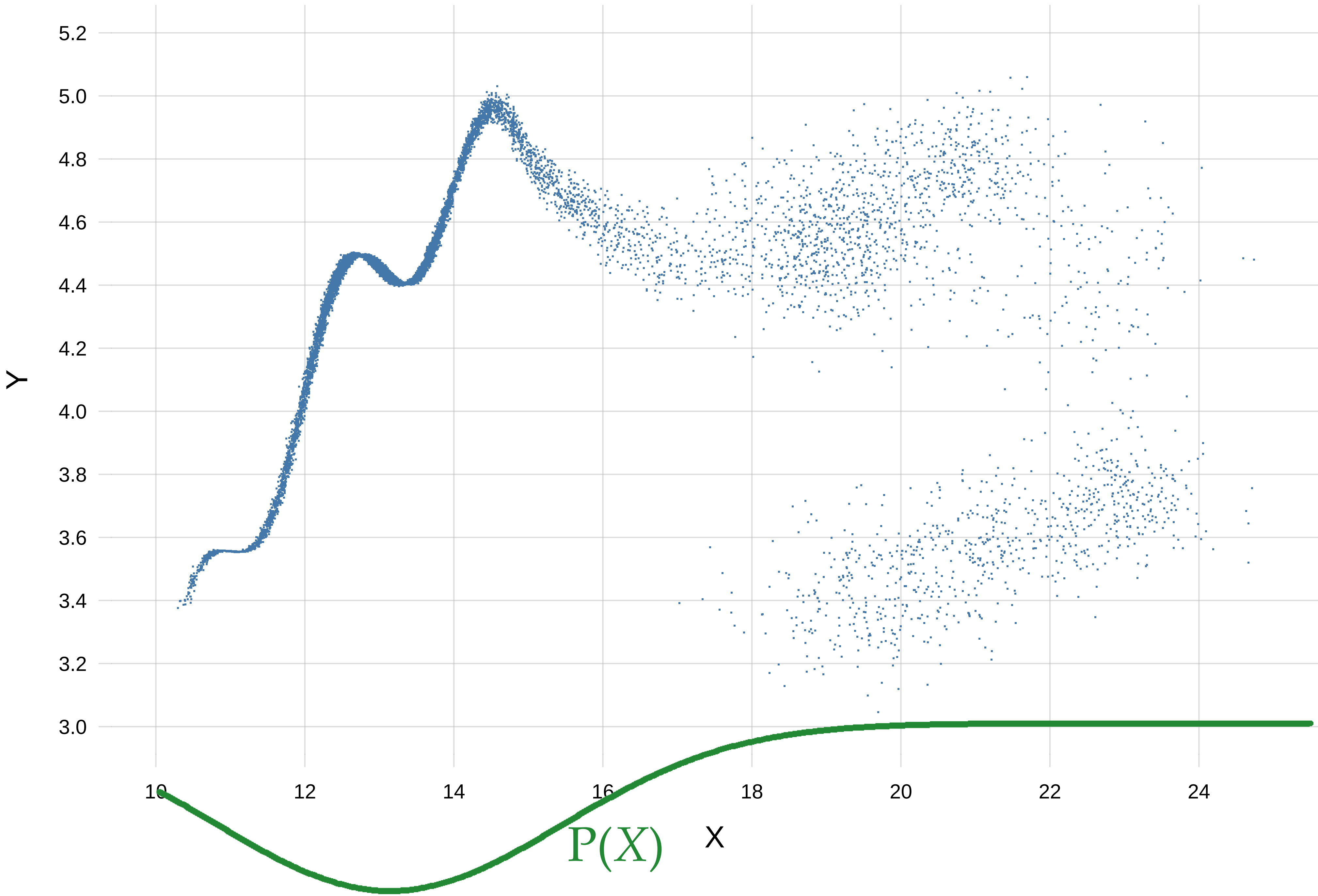


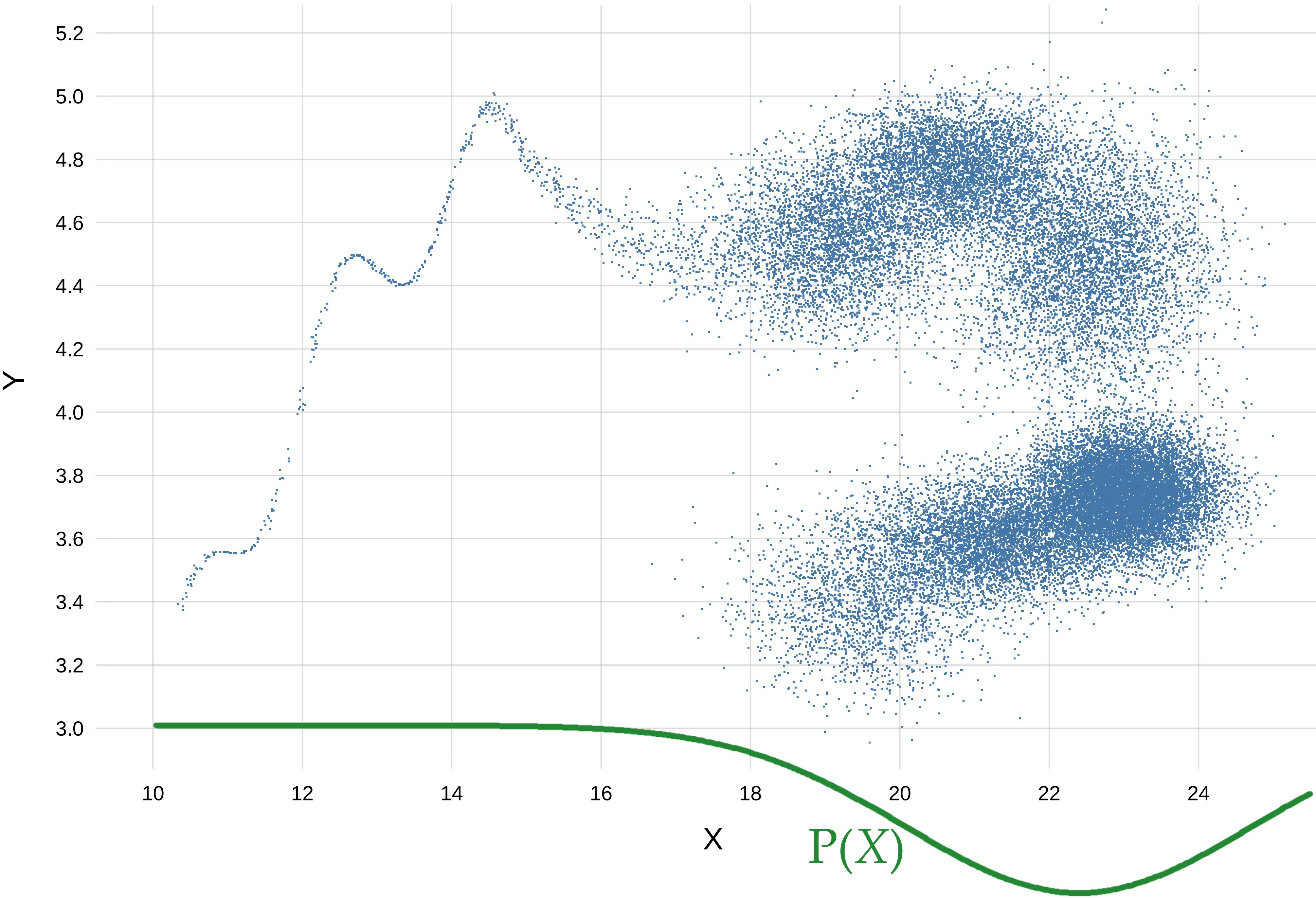
$x = 22 \quad \Rightarrow \quad y \approx 3.50–3.85$ or $4.25–4.90$



What is the 'overall predictive power' of X?



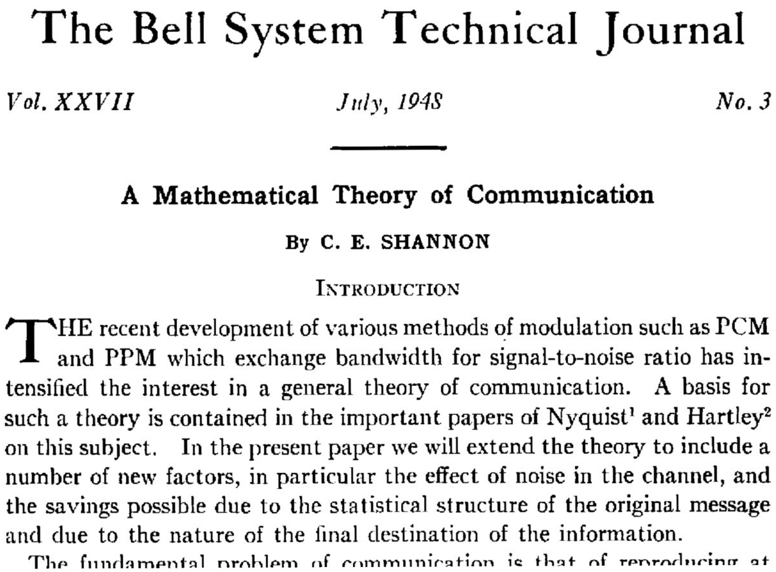

5.2
5.0
4.8
4.6
4.4
4.2
4.0
3.8
3.6
3.4
3.2
3.0
Y
10
12
14
16
18
20
22
24
X
P(X)


5.2
5.0
4.8
4.6
4.4
4.2
4.0
3.8
3.6
3.4
3.2
3.0
Y
10
12
14
16
18
20
22
24
X
P(X)

The ‘importance’ or ‘predictive power’ of X depends on $\mathrm{P}(X)$

The ‘importance’ or ‘predictive power’ of X depends on $\mathrm{P}(X)$

⚠ Careful with ‘data balancing’! ⚠



The Bell System Technical Journal

Vol. XXVII July, 1948 No. 3

A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

The fundamental problem of communication is that of reproducing at

Information Theory

The Bell System Technical Journal

Vol. XXVII July, 1948 No. 3

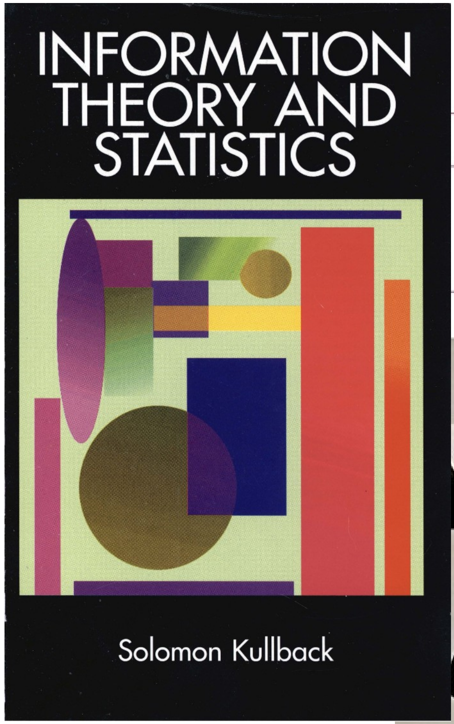
A Mathematical Theory of Communication

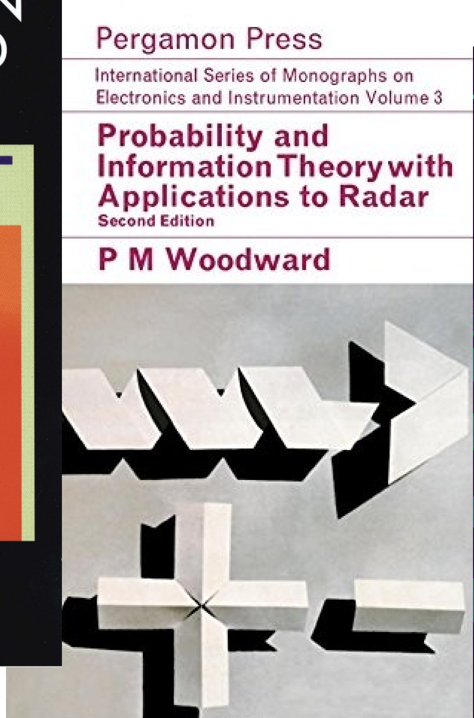
By C. E. SHANNON

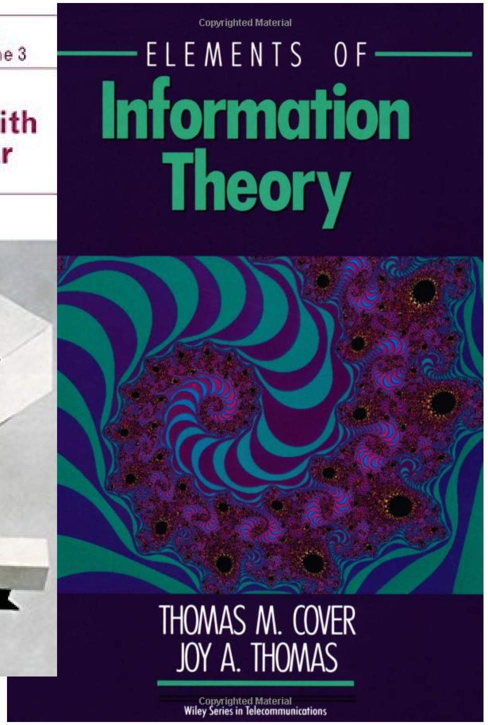
INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

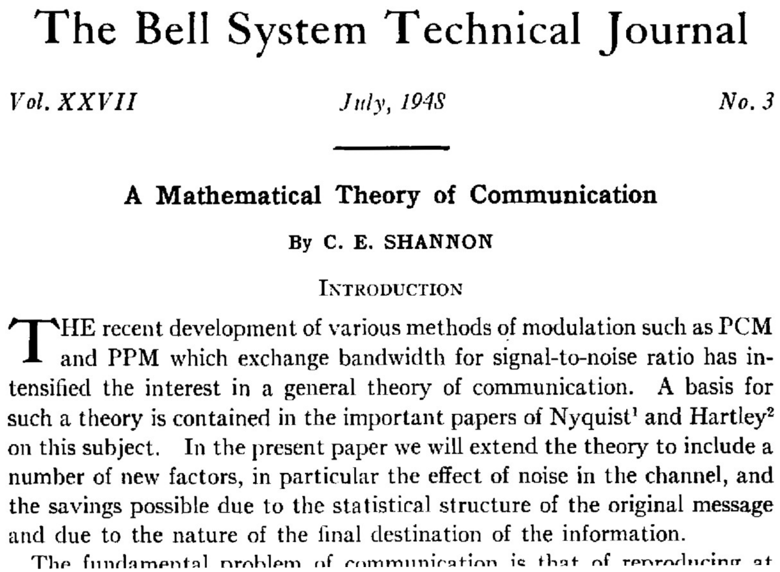
The fundamental problem of communication is that of reproducing at







Information Theory


The Bell System Technical Journal

Vol. XXVII — July, 1948 — No. 3

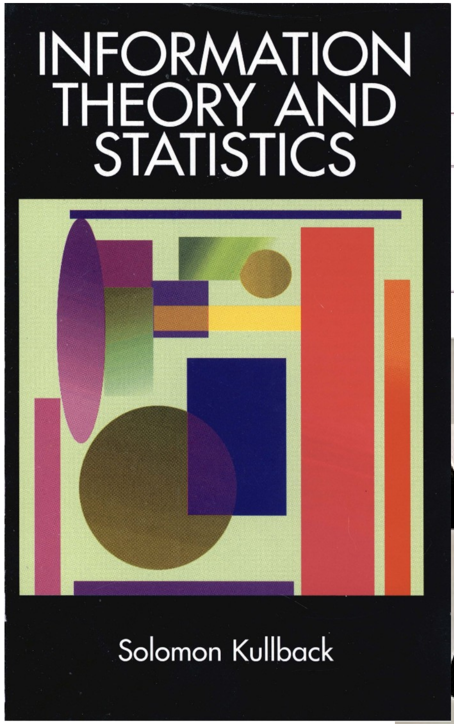
A Mathematical Theory of Communication

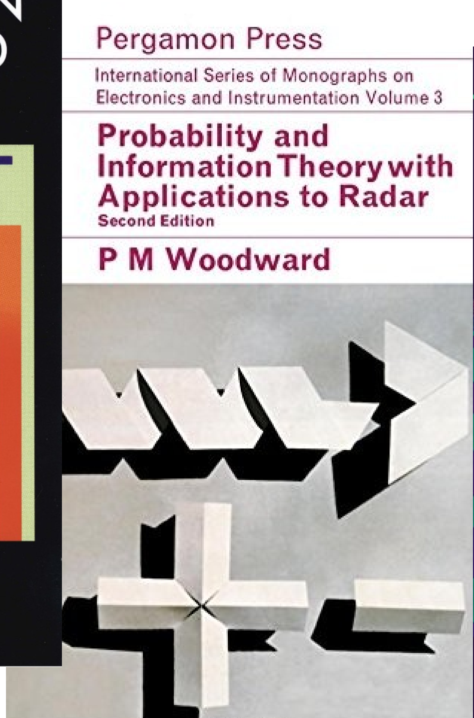
By C. E. SHANNON

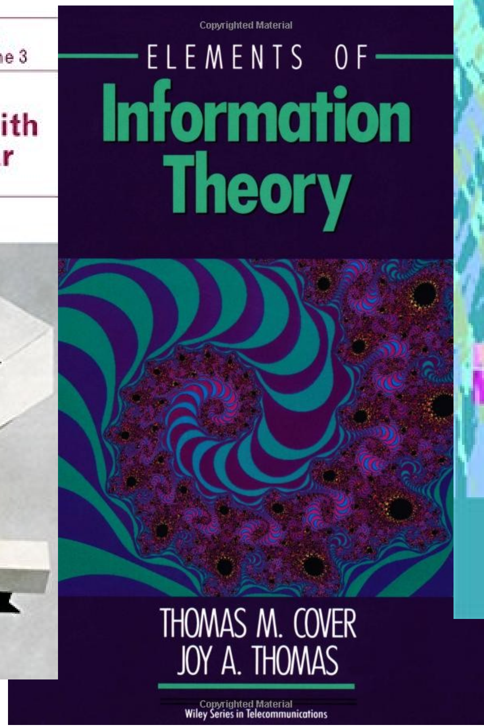
INTRODUCTION

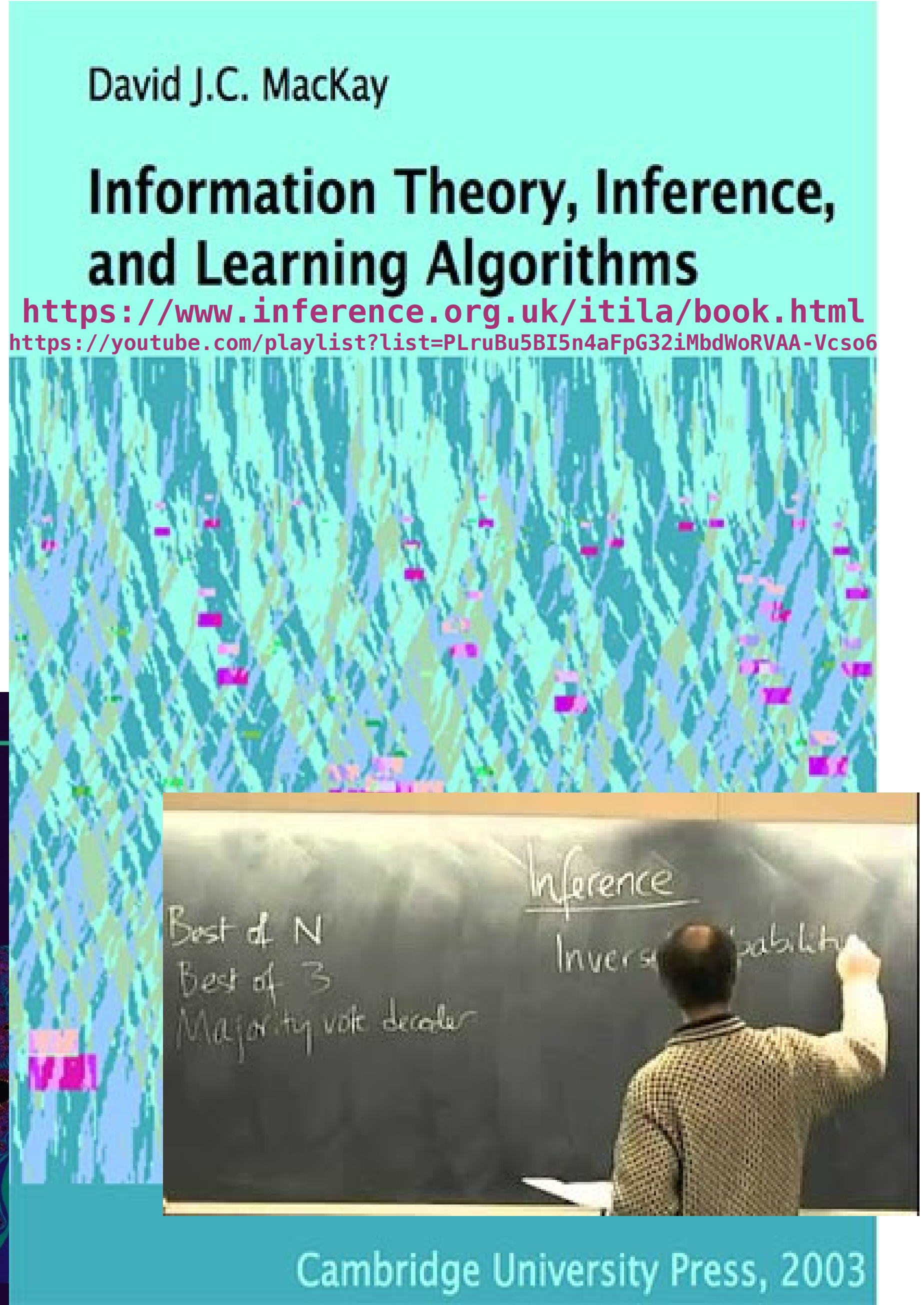
THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

The fundamental problem of communication is that of reproducing at









https://www.inference.org.uk/itila/book.html

https://youtube.com/playlist?list=PLruBu5BI5n4aFpG32iMbdWoRVAA-Vcso6

'predictive power' of X for Y $\coloneqq$ **Mutual information** between Y and X

(mean transinformation content)

$$I(X;Y) := \int \mathrm{p}(y|\,x)\,\mathrm{p}(x)\ \log\left[\frac{\mathrm{p}(y|\,x)}{\mathrm{p}(y)}\right] \mathrm{d}y\,\mathrm{d}x$$

'predictive power' of X for Y $\coloneqq$ **Mutual information** between Y and X

(mean transinformation content)

$$I(X;Y) := \int \mathrm{p}(y \mid x)\, \mathrm{p}(x) \log\left[\frac{\mathrm{p}(y \mid x)}{\mathrm{p}(y)}\right] \mathrm{d}y\, \mathrm{d}x$$

'predictive power' of X for Y $\coloneqq$ **Mutual information** between Y and X

(mean transinformation content)

$$I(X;Y) := \int \mathrm{p}(y|\,x)\,\mathrm{p}(x)\log\left[\frac{\mathrm{p}(y|\,x)}{\mathrm{p}(y)}\right]\mathrm{d}y\,\mathrm{d}x$$

$$I(Y;\, X_1, X_2) \geq I(Y;\, X_1)$$

$$I(Y;\, X_1, X_2) \geq I(Y;\, X_2)$$

but $\quad I(Y;\, X_1, X_2) \neq I(Y;\, X_1) + I(Y;\, X_2)$

IEC 80000-13

Edition 1.0 2008-03

INTERNATIONAL STANDARD

NORME INTERNATIONALE

Quantities and units –
Part 13: Information science and technology

Grandeurs et unités –
Partie 13: Science et technologies de l'information

IEC ISO

IEC 80000-13

Edition 1.0 2008-03

INTERNATIONAL STANDARD

INFORMATION SCIENCE AND TECHNOLOGY				QUANTITIES
Item No.	**Name**	**Symbol**	**Definition**	**Remarks**
13-24 (*902*)	information content *fr* *quantité* (f) *d'information*	$I(x)$	$I(x) = \mathrm{lb}\,\frac{1}{p(x)}\ \mathrm{Sh} = \lg \frac{1}{p(x)}\ \mathrm{Hart} = \ln \frac{1}{p(x)}\ \mathrm{nat}$ where $p(x)$ is the probability of event x	See ISO/IEC 2382-16, item 16.03.02. See also IEC 60027-3.
13-25 (*903*)	entropy *fr* *entropie* (f)	H	$H(X) = \sum_{i=1}^{n} p(x_i) I(x_i)$ for the set $X = \{x_1, \dots, x_n\}$ where $p(x_i)$ is the probability and $I(x_i)$ is the information content of event x_i	See ISO/IEC 2382-16, item 16.03.03.
13-30 (*908*)	joint information content *fr* *quantité* (f) *d'information conjointe*	$I(x, y)$	$I(x, y) = \mathrm{lb}\,\frac{1}{p(x, y)}\ \mathrm{Sh} = \lg \frac{1}{p(x, y)}\ \mathrm{Hart} = \ln \frac{1}{p(x, y)}\ \mathrm{nat}$ where $p(x, y)$ is the joint probability of events x and y	
13-35 (*912*)	transinformation content *fr* *transinformation* (f)	$T(x, y)$	$T(x, y) = I(x) + I(y) - I(x, y)$ where $I(x)$ and $I(y)$ are the information contents (13-24) of events x and y, respectively, and $I(x, y)$ is their joint information content (13-30)	See ISO/IEC 2382-16, item 16.04.07.
13-36 (*913*)	mean transinformation content *fr* *transinformation* (f) *moyenne*	T	$T(X, Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) T(x_i, y_j)$ for the sets $X = \{x_1, \dots, x_n\}$, $Y = \{y_1, \dots, y_m\}$, where $p(x_i, y_j)$ is the joint probability of events x_i and y_j, and $T(x_i, y_j)$ is their transinformation content (item 13-35)	See ISO/IEC 2382-16, item 16.04.08.

UNITS				INFORMATION SCIENCE AND TECHNOLOGY
Item No.	**Name**	**Symbol**	**Definition**	**Conversion factors and remarks**
13-24.a	shannon	Sh	value of the quantity when the argument is equal to 2	1 Sh ≈ 0,693 nat ≈ 0,301 Hart
13-24.b	hartley	Hart	value of the quantity when the argument is equal to 10	1 Hart ≈ 3,322 Sh ≈ 2,303 nat
13-24.c	natural unit of information	nat	value of the quantity when the argument is equal to e	1 nat ≈ 1,433 Sh ≈ 0,434 Hart
13-25.a	shannon	Sh		
13-25.b	hartley	Hart		
13-25.c	natural unit of information	nat		
13-30.a	shannon	Sh		
13-30.b	hartley	Hart		
13-30.c	natural unit of information	nat		
13-35.a	shannon	Sh		
13-35.b	hartley	Hart		
13-35.c	natural unit of information	nat		
13-36.a	shannon	Sh		In practice, the unit "shannon per character" is generally used, and sometimes the units "hartley per character" and "natural unit per character".
13-36.b	hartley	Hart		
13-36.c	natural unit of information	nat		

If Y is binary:

$$0 \text{ Sh} \leq I(Y;X) \leq 1 \text{ Sh}$$

If Y is binary:

$$0 \text{ Sh} \leq I(Y;X) \leq 1 \text{ Sh}$$

X and Y are independent
Using X is no better than flipping a coin

If Y is binary:

$$0\ \mathrm{Sh} \leq I(Y;X) \leq 1\ \mathrm{Sh}$$

X and Y are independent
Using X is no better than flipping a coin

Y is a deterministic function of X
X always yields perfect predictions

If Y is binary:

$$0 \text{ Sh} \leq I(Y;X) \leq 1 \text{ Sh}$$

X and Y are independent
Using X is no better than flipping a coin

Y is a deterministic function of X
X always yields perfect predictions

$$I(Y;X) = 0.22 \text{ Sh}$$

If Y is binary:

$$0 \text{ Sh} \le I(Y;X) \le 1 \text{ Sh}$$

X and Y are independent
Using X is no better than flipping a coin

Y is a deterministic function of X
X always yields perfect predictions

$$I(Y;X) = 0.22 \text{ Sh}$$

In $N=100$ new prognoses:

- we are **completely certain** about 22
- we are **completely uncertain** about $100-22=78$

→ approx $22+78/2=\mathbf{61}$ correct prognoses (TP+TN)

If Y is binary:

$$0\ \text{Sh} \le I(Y;X) \le 1\ \text{Sh}$$

X and Y are independent
Using X is no better than flipping a coin

Y is a deterministic function of X
X always yields perfect predictions

$$I(Y;X) = 0.22\ \text{Sh}$$

In $N = 100$ new prognoses:

- we are **completely certain** about 22
- we are **completely uncertain** about $100 - 22 = 78$

→ approx $22 + 78/2 = \mathbf{61}$ correct prognoses (TP+TN)

Any prognostic algorithm which does less than this is performing poorly

If Y is binary:

$$0\ \mathrm{Sh} \;\le\; I(Y;X) \;\le\; 1\ \mathrm{Sh}$$

X and Y are independent
Using X is no better than flipping a coin

Y is a deterministic function of X
X always yields perfect predictions

$$I(Y;X) = 0.22\ \mathrm{Sh}$$

In $N = 100$ new prognoses:

- we are **completely certain** about 22
- we are **completely uncertain** about $100 - 22 = 78$

→ approx $22 + 78/2 = \mathbf{61}$ correct prognoses (TP+TN)

Any prognostic algorithm which does less than this is performing poorly

Upper bound ≈ $$\frac{1}{2} + \frac{1}{2}\sqrt{1 - (1 - 0.22)^{\frac{4}{3}}} \approx 77\,\% \quad \pm 0.8\sqrt{N}\,\%$$ correct prognoses

If Y is binary:

$$0\ \mathrm{Sh} \leq I(Y;X) \leq 1\ \mathrm{Sh}$$

X and Y are independent
Using X is no better than flipping a coin

Y is a deterministic function of X
X always yields perfect predictions

$$I(Y;X) = 0.22\ \mathrm{Sh}$$

In $N = 100$ new prognoses:

- we are **completely certain** about 22
- we are **completely uncertain** about $100 - 22 = 78$

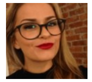
→ approx $22 + 78/2 =$ **61** correct prognoses (TP+TN)

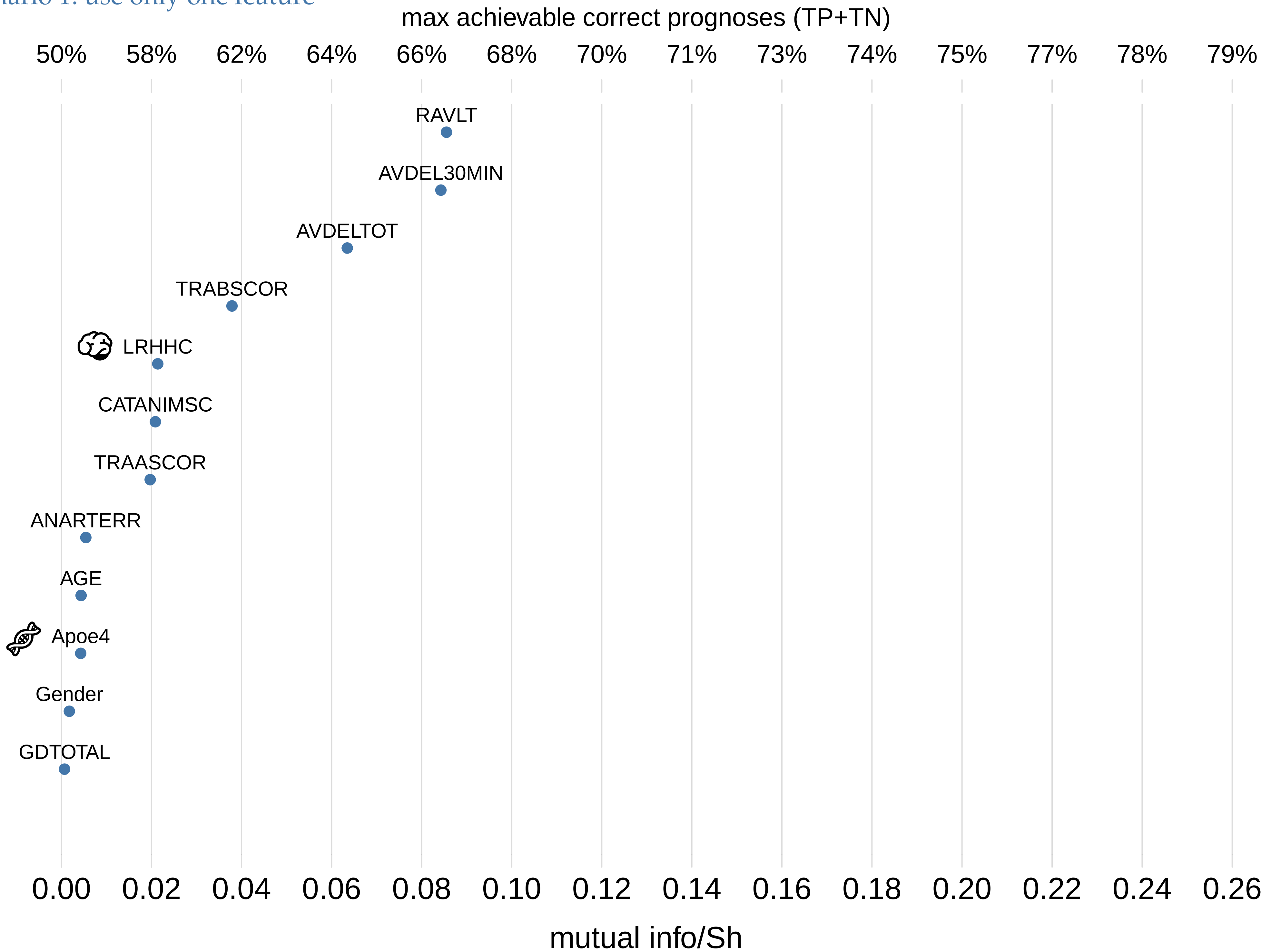
Any prognostic algorithm which does less than this is performing poorly

Upper bound ≈ $$\frac{1}{2} + \frac{1}{2}\sqrt{1 - (1 - 0.22)^{\frac{4}{3}}} \approx 77\,\% \quad \pm 0.8\sqrt{N}\,\%$$ correct prognoses

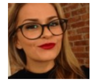
Maximum accuracy attainable
by *any* algorithm which uses only feature set X

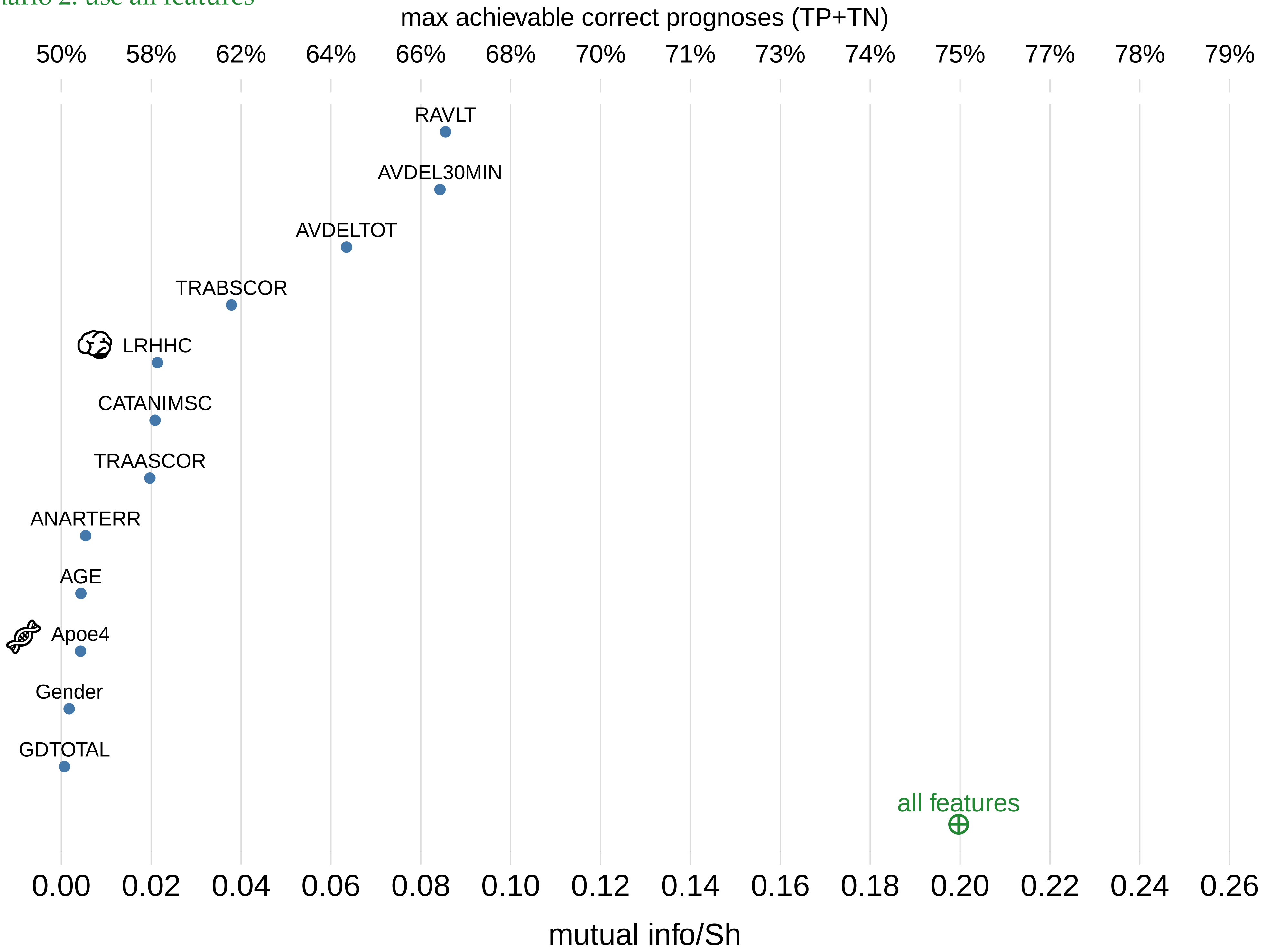
Scenario 1: use only one feature



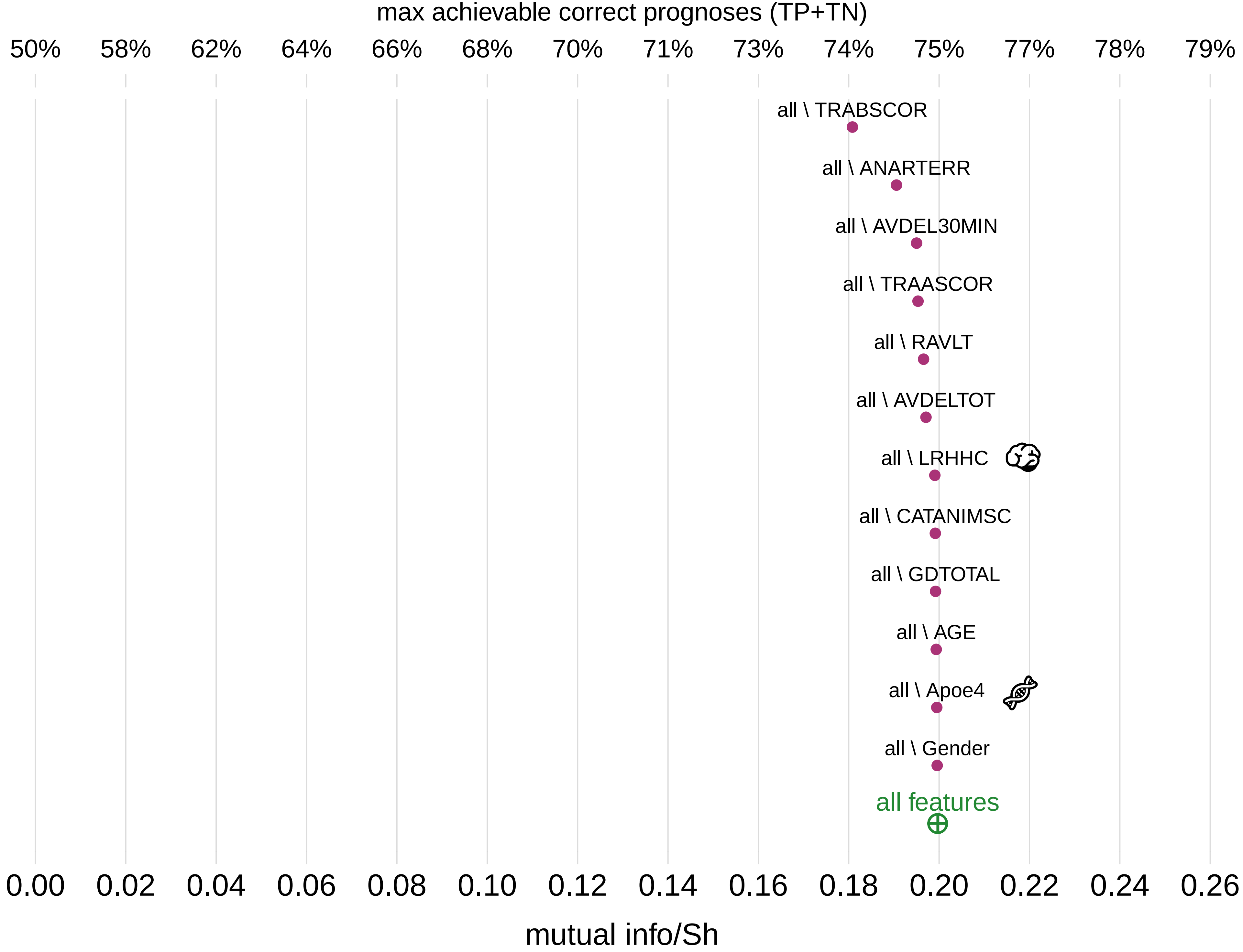


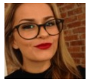
Scenario 2: use all features

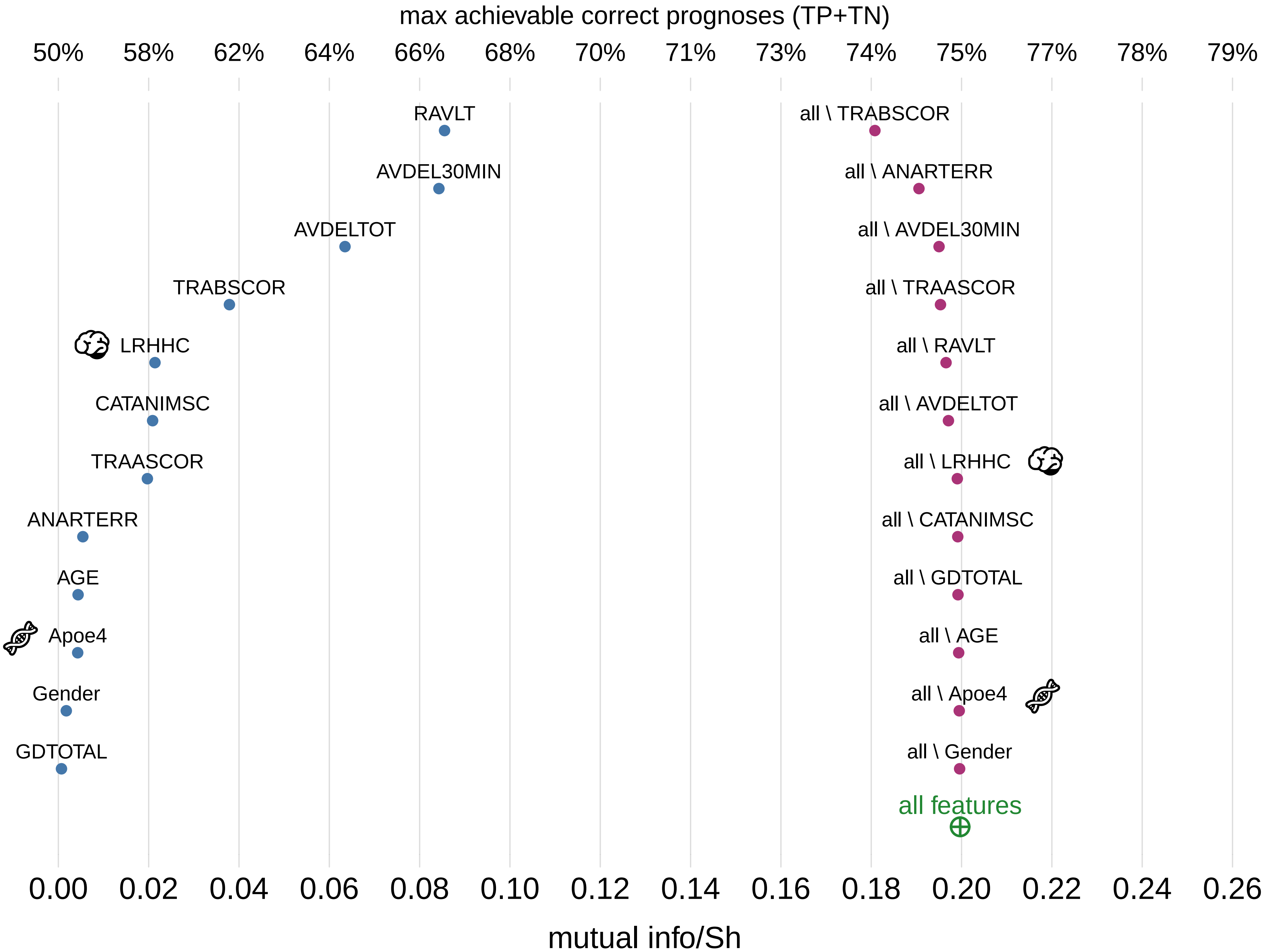


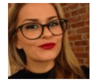


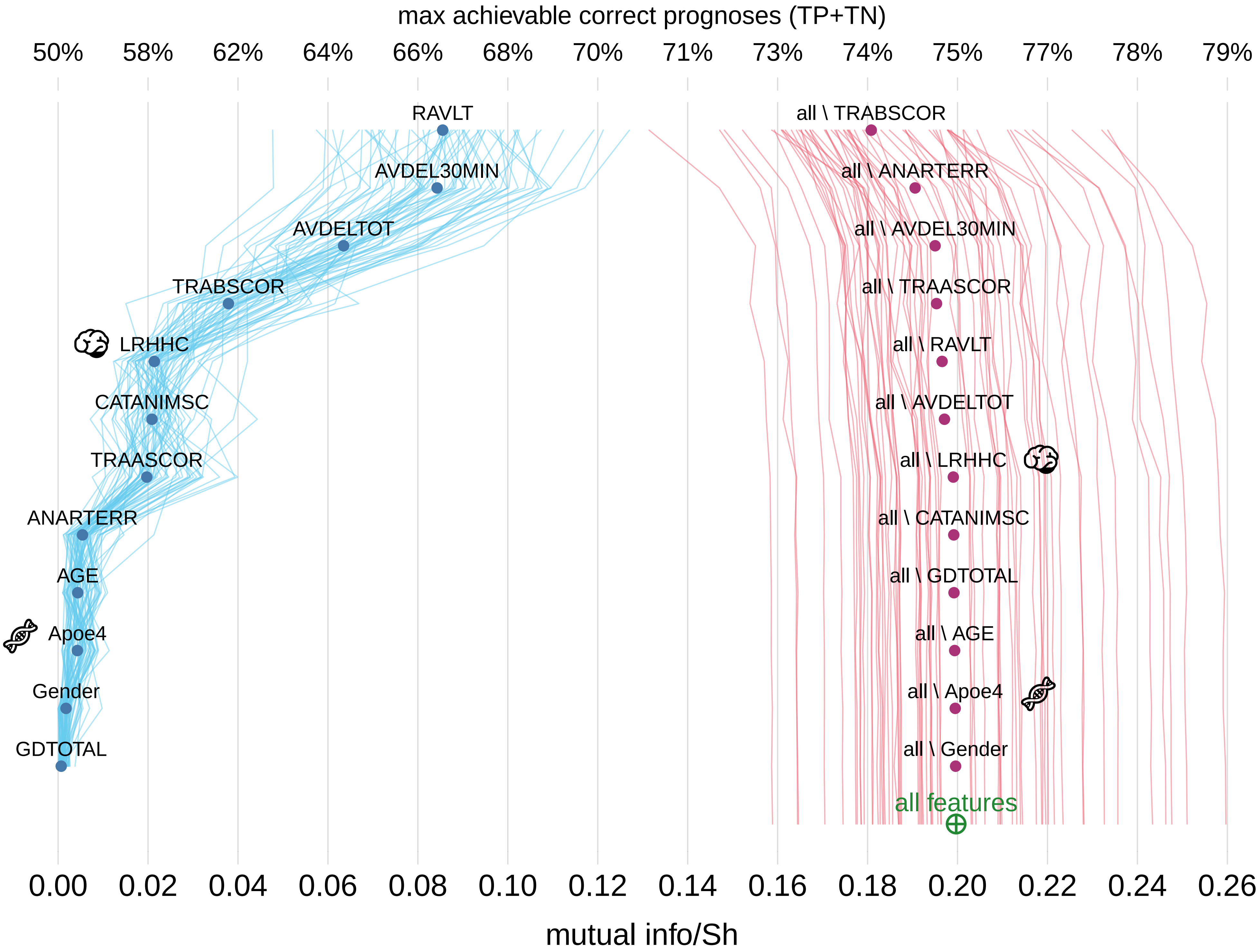
Scenario 3: discard one feature



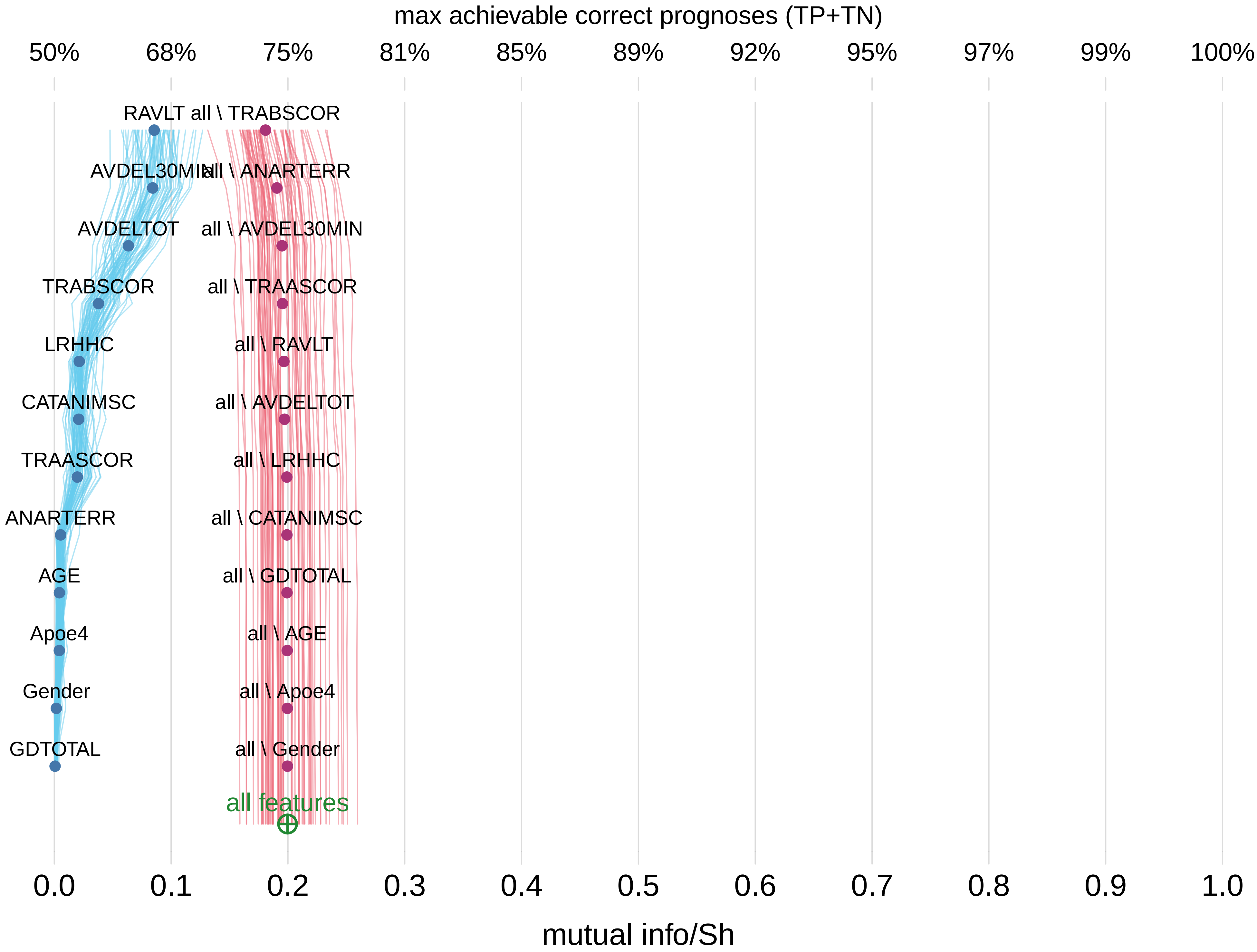



max achievable correct prognoses (TP+TN)
50%
58%
62%
64%
66%
68%
70%
71%
73%
74%
75%
77%
78%
79%
RAVLT
AVDEL30MIN
AVDELTOT
TRABSCOR
LRHHC
CATANIMSC
TRAASCOR
ANARTERR
AGE
Apoe4
Gender
GDTOTAL
all \ TRABSCOR
all \ ANARTERR
all \ AVDEL30MIN
all \ TRAASCOR
all \ RAVLT
all \ AVDELTOT
all \ LRHHC
all \ CATANIMSC
all \ GDTOTAL
all \ AGE
all \ Apoe4
all \ Gender
all features
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
0.18
0.20
0.22
0.24
0.26
mutual info/Sh

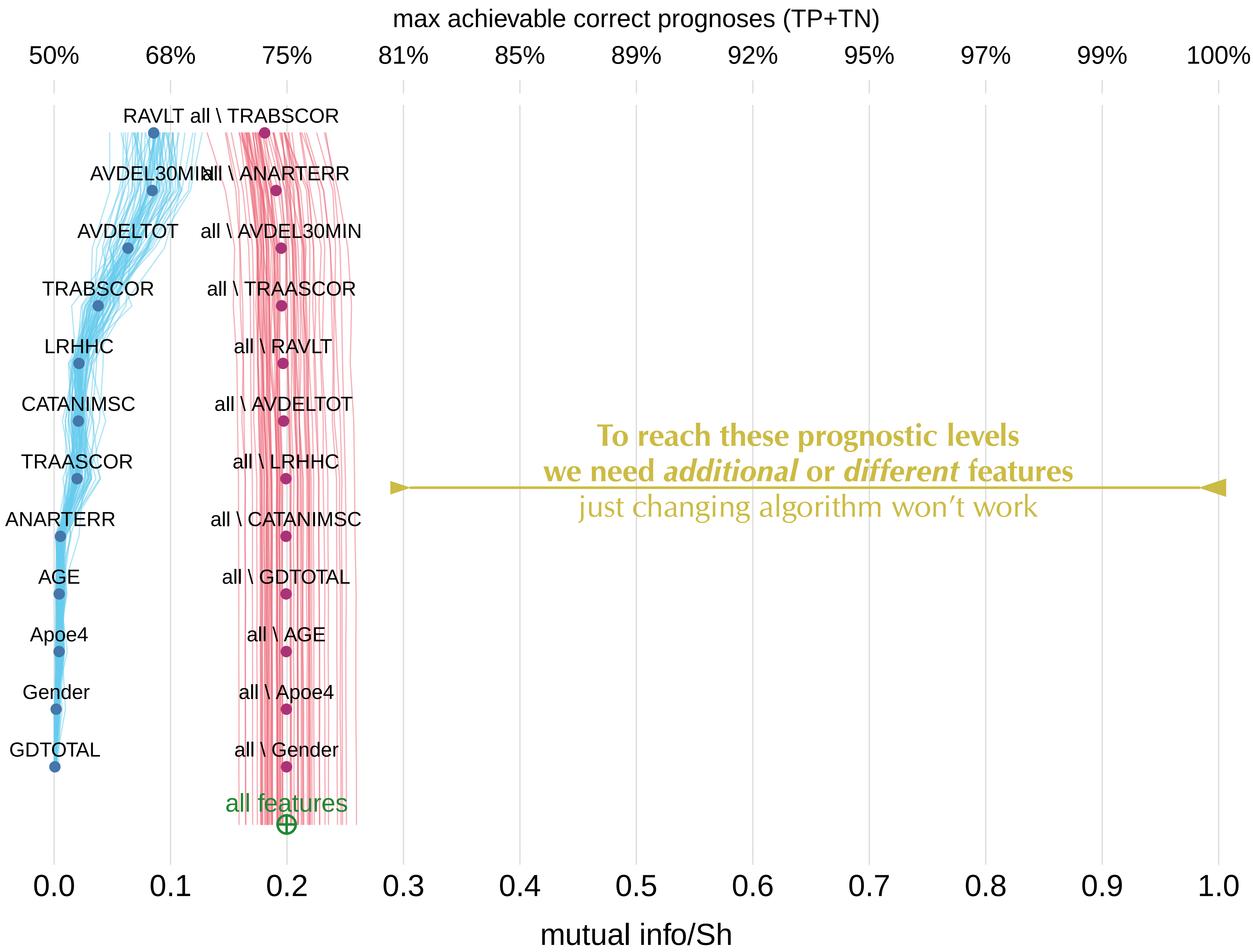




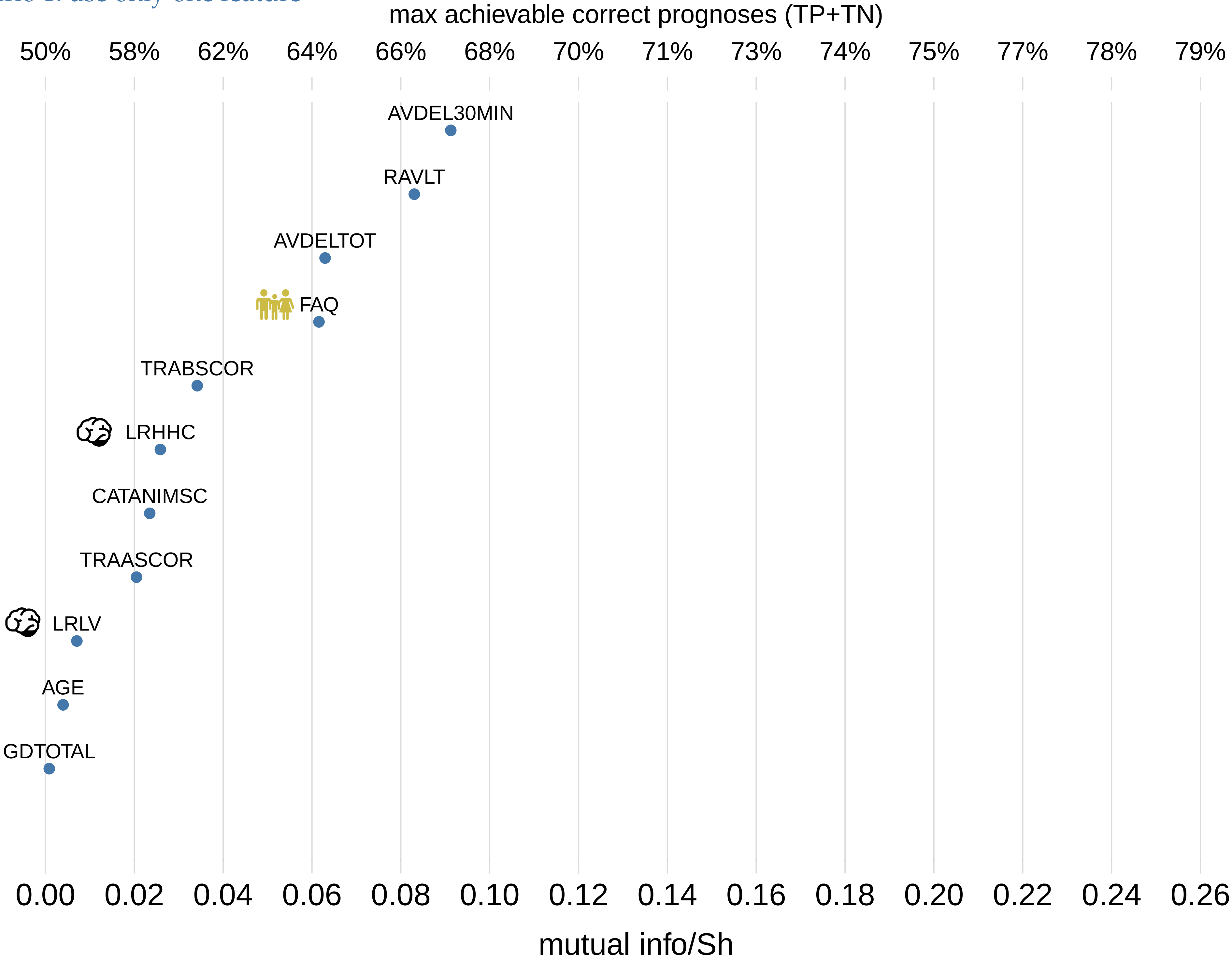
max achievable correct prognoses (TP+TN)
50%
58%
62%
64%
66%
68%
70%
71%
73%
74%
75%
77%
78%
79%
RAVLT
AVDEL30MIN
AVDELTOT
TRABSCOR
LRHHC
CATANIMSC
TRAASCOR
ANARTERR
AGE
Apoe4
Gender
GDTOTAL
all \ TRABSCOR
all \ ANARTERR
all \ AVDEL30MIN
all \ TRAASCOR
all \ RAVLT
all \ AVDELTOT
all \ LRHHC
all \ CATANIMSC
all \ GDTOTAL
all \ AGE
all \ Apoe4
all \ Gender
all features
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
0.18
0.20
0.22
0.24
0.26
mutual info/Sh



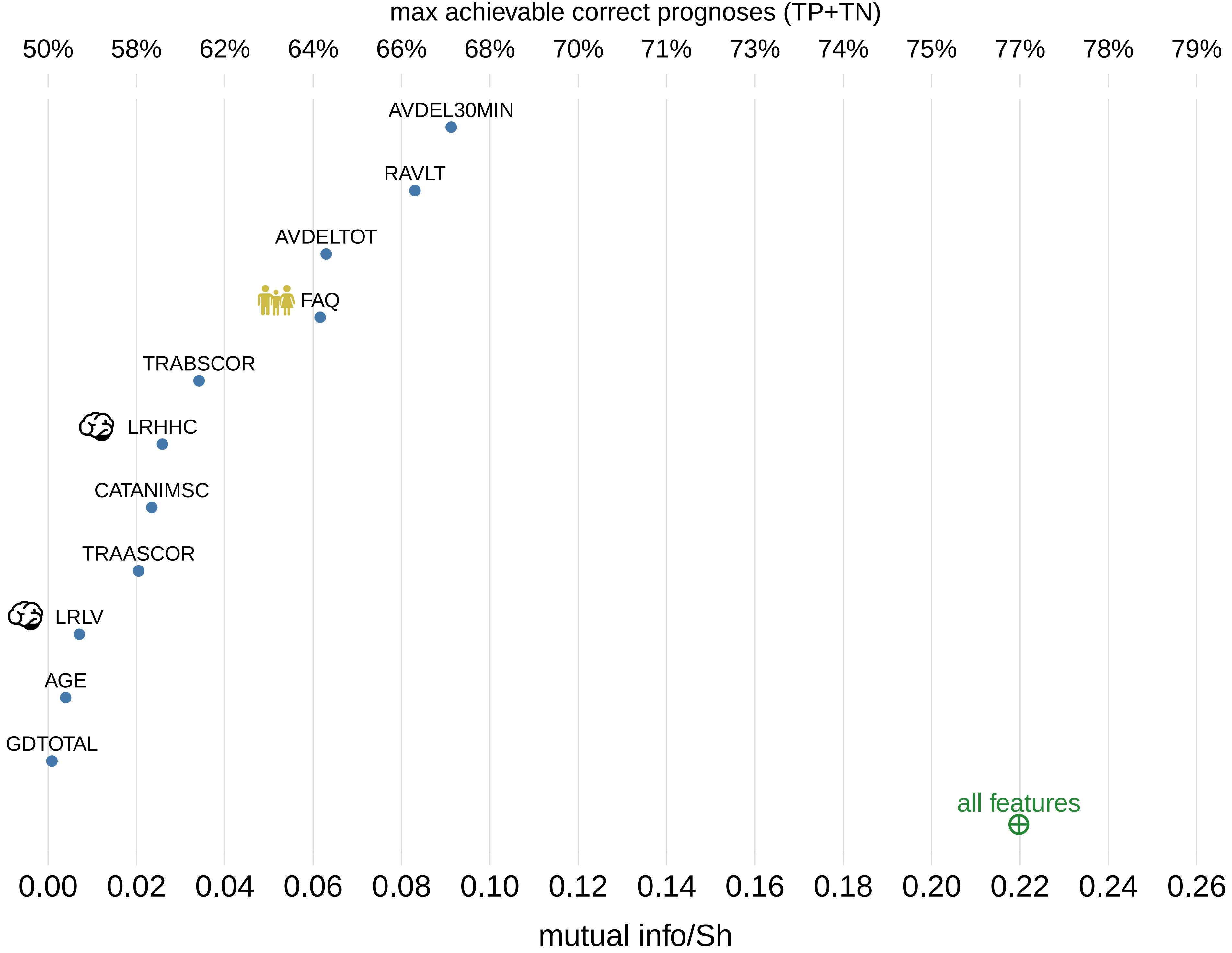
max achievable correct prognoses (TP+TN)
50%
68%
75%
81%
85%
89%
92%
95%
97%
99%
100%
RAVLT
AVDEL30MIN
AVDELTOT
TRABSCOR
LRHHC
CATANIMSC
TRAASCOR
ANARTERR
AGE
Apoe4
Gender
GDTOTAL
all \ TRABSCOR
all \ ANARTERR
all \ AVDEL30MIN
all \ TRAASCOR
all \ RAVLT
all \ AVDELTOT
all \ LRHHC
all \ CATANIMSC
all \ GDTOTAL
all \ AGE
all \ Apoe4
all \ Gender
all features
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
mutual info/Sh



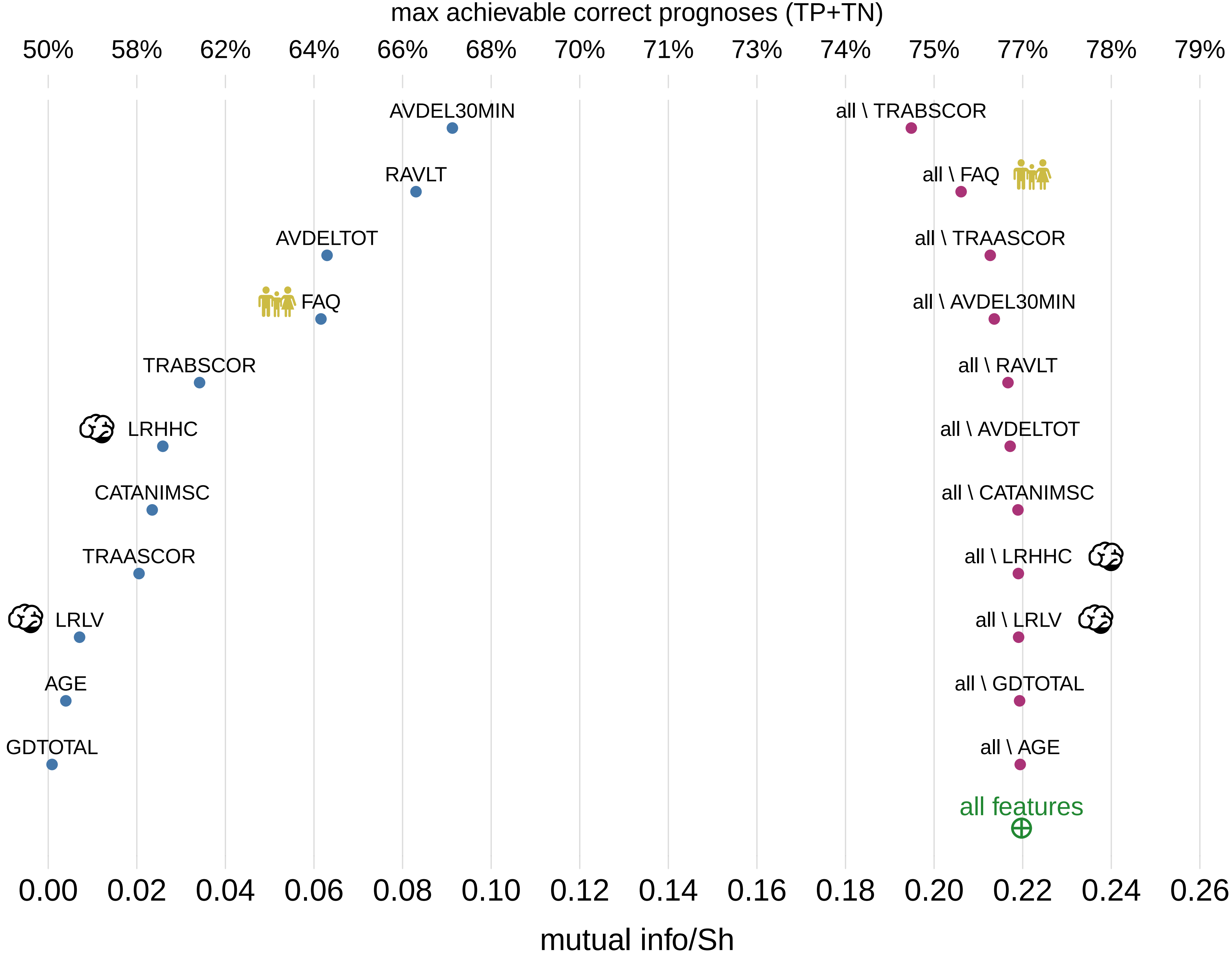
max achievable correct prognoses (TP+TN)
50%
68%
75%
81%
85%
89%
92%
95%
97%
99%
100%
RAVLT
AVDEL30MIN
AVDELTOT
TRABSCOR
LRHHC
CATANIMSC
TRAASCOR
ANARTERR
AGE
Apoe4
Gender
GDTOTAL
all \ TRABSCOR
all \ ANARTERR
all \ AVDEL30MIN
all \ TRAASCOR
all \ RAVLT
all \ AVDELTOT
all \ LRHHC
all \ CATANIMSC
all \ GDTOTAL
all \ AGE
all \ Apoe4
all \ Gender
all features
To reach these prognostic levels
we need *additional* or *different* features
just changing algorithm won't work
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
mutual info/Sh

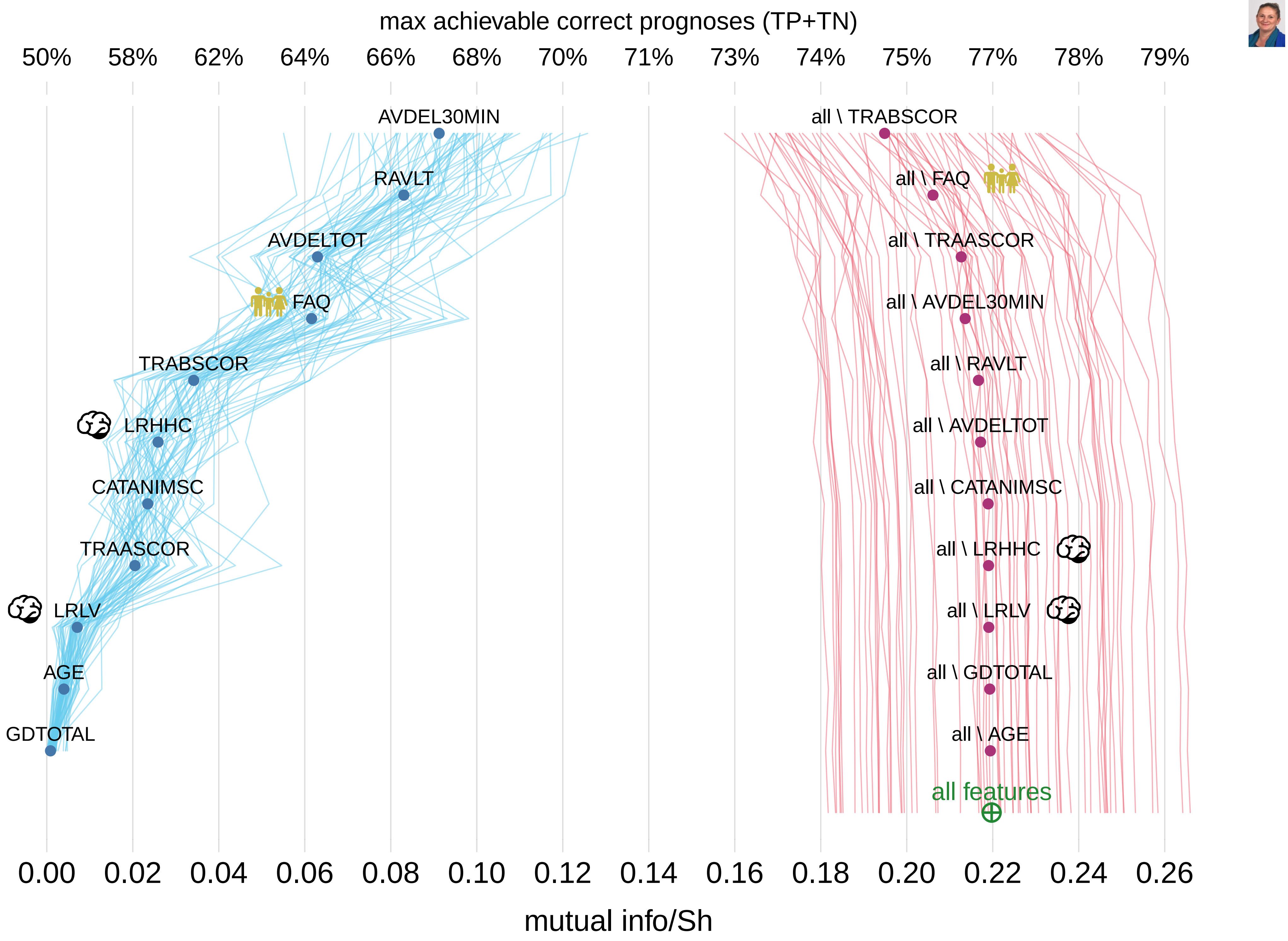

max achievable correct prognoses (TP+TN)
50%
58%
62%
64%
66%
68%
70%
71%
73%
74%
75%
77%
78%
79%
AVDEL30MIN
RAVLT
AVDELTOT
FAQ
TRABSCOR
LRHHC
CATANIMSC
TRAASCOR
LRLV
AGE
GDTOTAL
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
0.18
0.20
0.22
0.24
0.26
mutual info/Sh

Scenario 2: use all features



Scenario 3: discard one feature



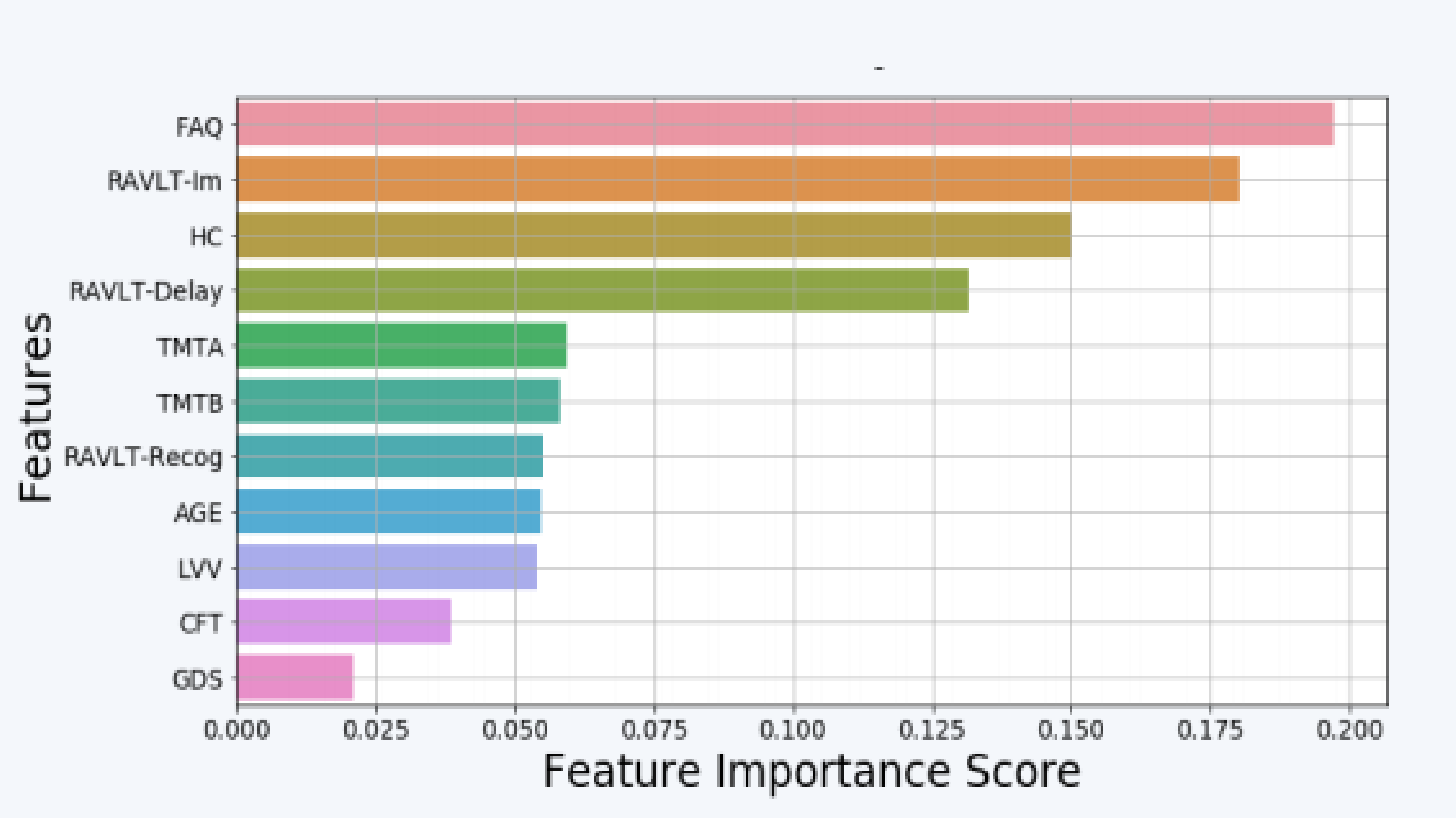

max achievable correct prognoses (TP+TN)
50%
58%
62%
64%
66%
68%
70%
71%
73%
74%
75%
77%
78%
79%
AVDEL30MIN
RAVLT
AVDELTOT
FAQ
TRABSCOR
LRHHC
CATANIMSC
TRAASCOR
LRLV
AGE
GDTOTAL
all \ TRABSCOR
all \ FAQ
all \ TRAASCOR
all \ AVDEL30MIN
all \ RAVLT
all \ AVDELTOT
all \ CATANIMSC
all \ LRHHC
all \ LRLV
all \ GDTOTAL
all \ AGE
all features
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
0.18
0.20
0.22
0.24
0.26
mutual info/Sh

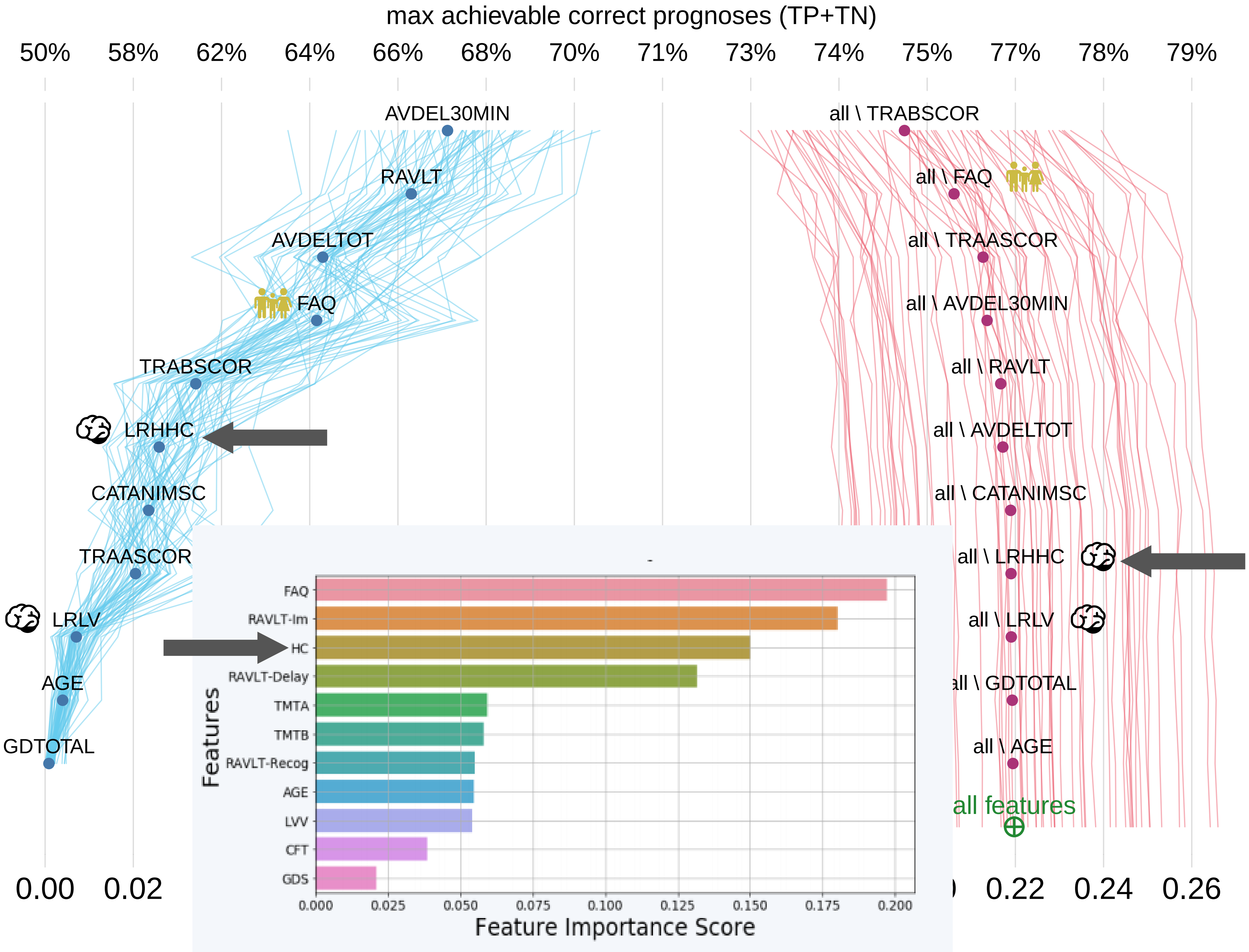
$$\mathrm{P}(Y \mid X)\,\mathrm{P}(X) \equiv \mathrm{P}(X \mid Y)\,\mathrm{P}(Y)$$

$$\mathrm{P}(Y \mid X)\,\mathrm{P}(X) \overset{!}{\equiv} \mathrm{P}(X \mid Y)\,\mathrm{P}^*(Y)$$

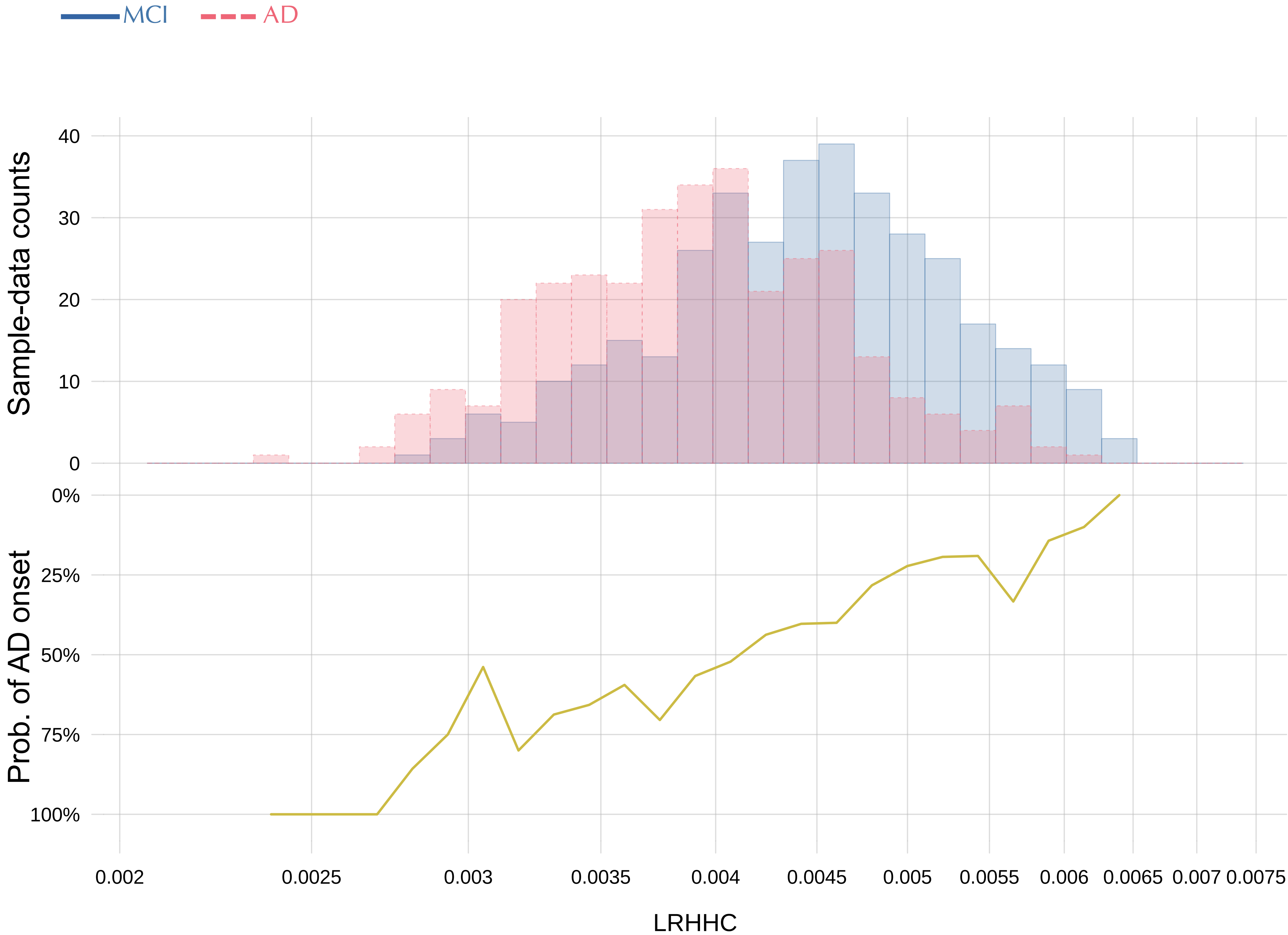
$$\overset{?}{\mathrm{P}}{}^*(Y \mid X)\, \overset{?}{\mathrm{P}}{}^*(X) \equiv \overset{?}{\mathrm{P}}{}^*(X \mid Y)\, \mathrm{P}^*(Y)$$

Alexandra’s results

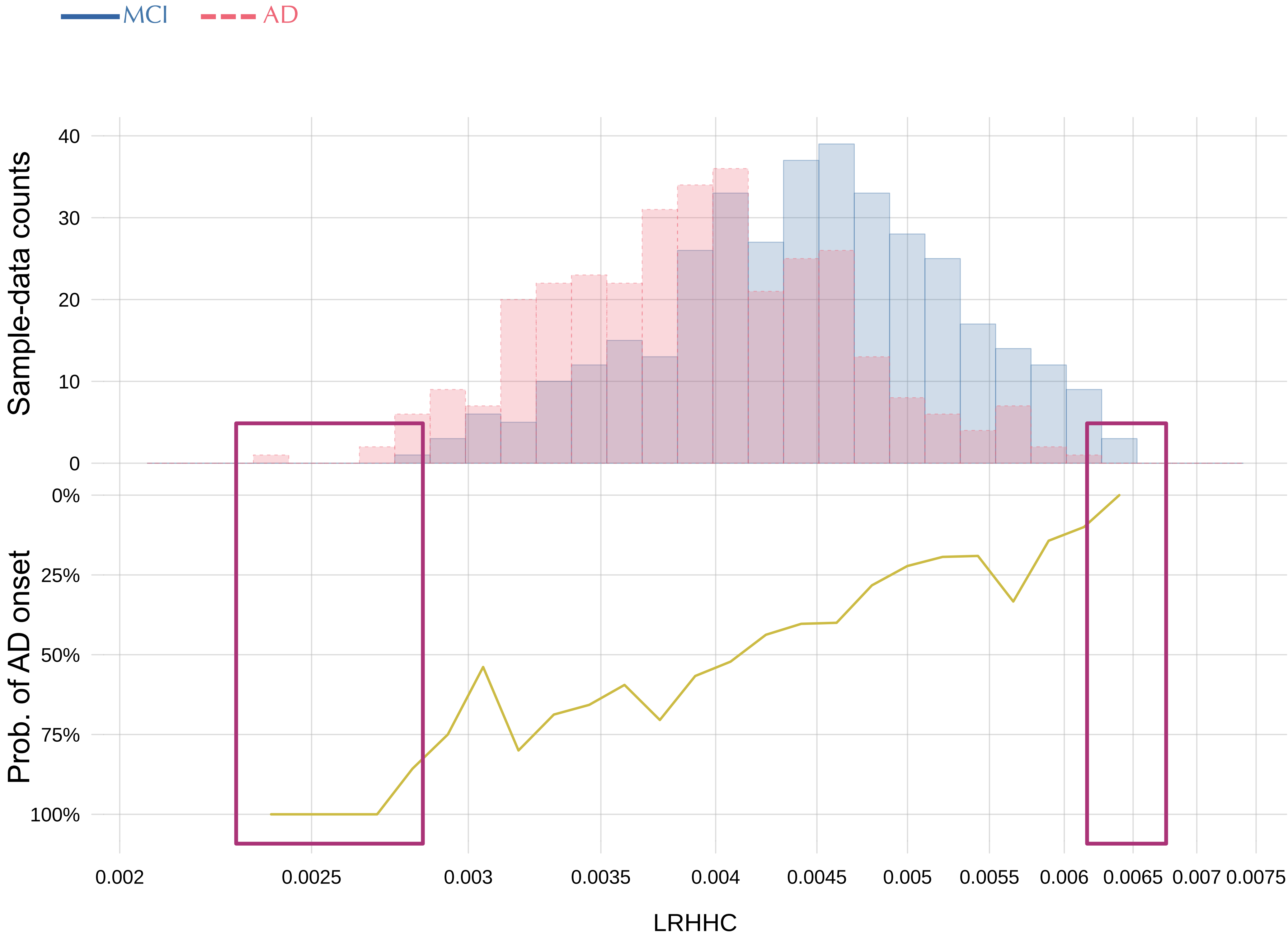


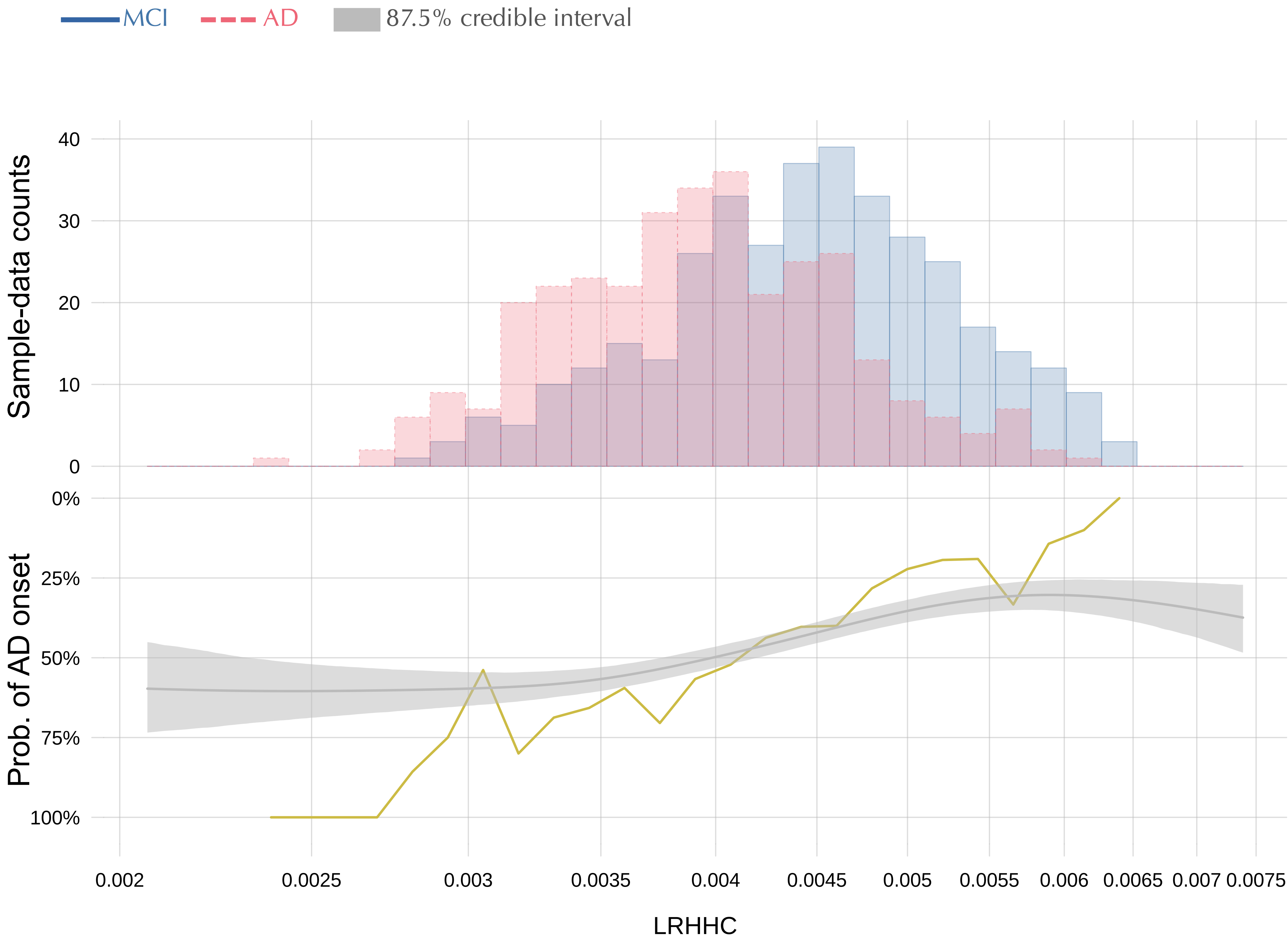


max achievable correct prognoses (TP+TN)
50% 58% 62% 64% 66% 68% 70% 71% 73% 74% 75% 77% 78% 79%
AVDEL30MIN
RAVLT
AVDELTOT
FAQ
TRABSCOR
LRHHC
CATANIMSC
TRAASCOR
LRLV
AGE
GDTOTAL
all \ TRABSCOR
all \ FAQ
all \ TRAASCOR
all \ AVDEL30MIN
all \ RAVLT
all \ AVDELTOT
all \ CATANIMSC
all \ LRHHC
all \ LRLV
all \ GDTOTAL
all \ AGE
all features
0.00 0.02 0.22 0.24 0.26
FAQ
RAVLT-Im
HC
RAVLT-Delay
TMTA
TMTB
RAVLT-Recog
AGE
LVV
CFT
GDS
Features
0.000 0.025 0.050 0.075 0.100 0.125 0.150 0.175 0.200
Feature Importance Score

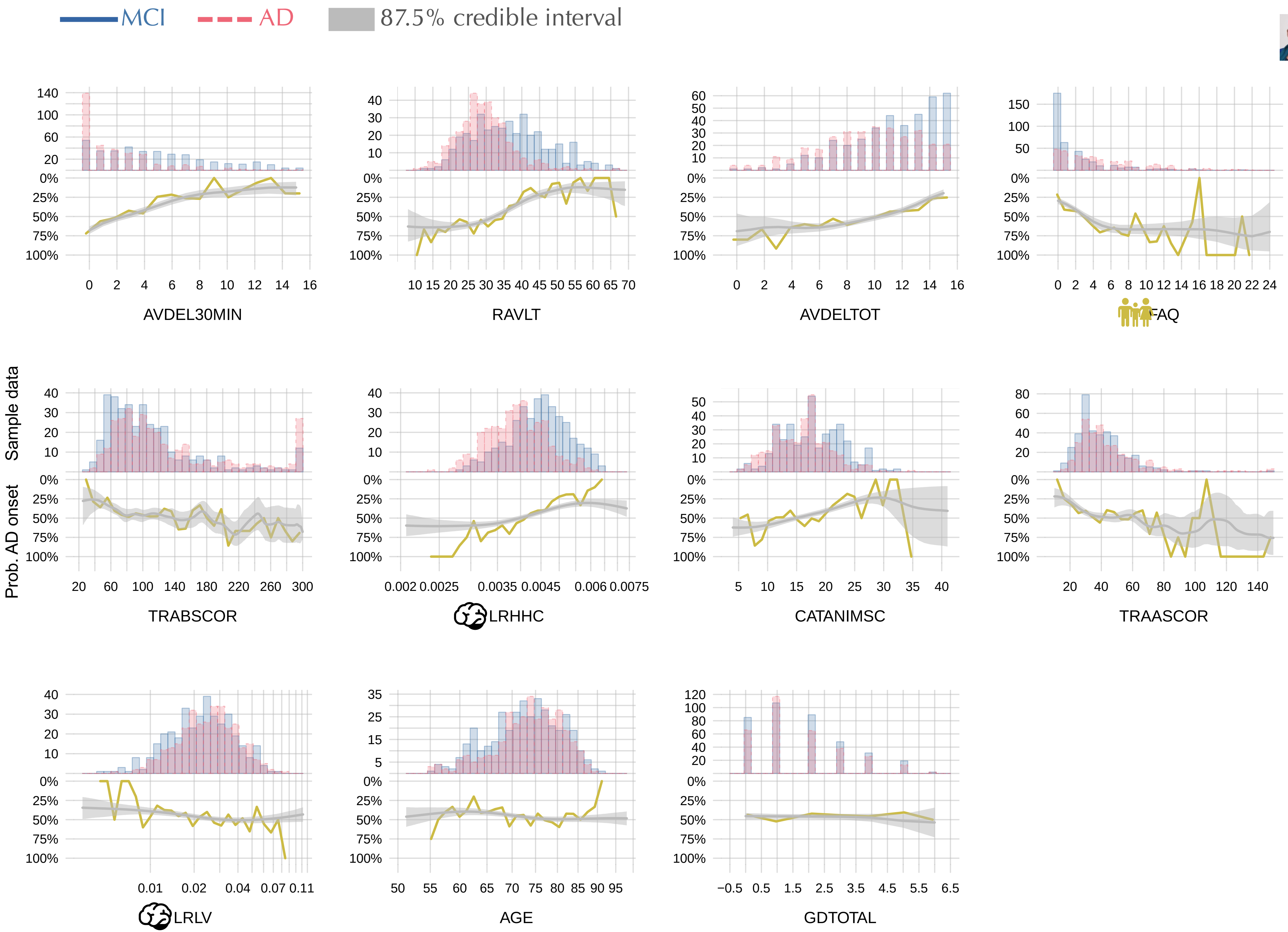


MCI
AD
Sample-data counts
40
30
20
10
0
0%
25%
50%
75%
100%
Prob. of AD onset
0.002
0.0025
0.003
0.0035
0.004
0.0045
0.005
0.0055
0.006
0.0065
0.007
0.0075
LRHHC


MCI
AD
Sample-data counts
40
30
20
10
0
0%
25%
50%
75%
100%
Prob. of AD onset
0.002
0.0025
0.003
0.0035
0.004
0.0045
0.005
0.0055
0.006
0.0065
0.007
0.0075
LRHHC


MCI
AD
87.5% credible interval
Sample-data counts
40
30
20
10
0
0%
25%
50%
75%
100%
Prob. of AD onset
0.002
0.0025
0.003
0.0035
0.004
0.0045
0.005
0.0055
0.006
0.0065
0.007
0.0075
LRHHC

MCI
AD
Sample data
Prob. AD onset
AVDEL30MIN
RAVLT
AVDELTOT
FAQ
TRABSCOR
LRHHC
CATANIMSC
TRAASCOR
LRLV
AGE
GDTOTAL


MCI
AD
87.5% credible interval
Sample data
Prob. AD onset
AVDEL30MIN
RAVLT
AVDELTOT
FAQ
TRABSCOR
LRHHC
CATANIMSC
TRAASCOR
LRLV
AGE
GDTOTAL

Thank you!