Learning to Defer & Uncertainty

(How to Combine Different Models)

March 3, 2020

Massachusetts Institute of Technology

A De-Identification Anecdote

- In 2011, another “whack” at de-identification
- Two hypotheses
  - Multi: combine the outputs of multiple systems to come to a consensus judgment about each word in a note
  - Mega: use all features from these multiple systems in a single SVM model
- Which is better?
  - Two of my former postdocs and I had a “gentlemen’s/gentlewoman’s bet”!
- How did I bet? How did it turn out?

Multi-DeIdentifier’s Component Systems	
Stat-Deid	SVM over many features
Stanford NER	CRF over char n-grams, POS, “shape”, …
MIST Deidentifier	CRF over regex, dictionary matches
Illinois Named Entity Tagger	2-stage CRF over affixes, nearby words, caps, …
MIMIC Deidentifier	Dictionaries + rules

Mixture of Experts Models

- Define a variety of “experts” and a gating function
- Gating can be
  - “hard” if it chooses one expert
    (e.g., if different experts are good at different domains)
  - “soft” if it combines different experts’ outputs

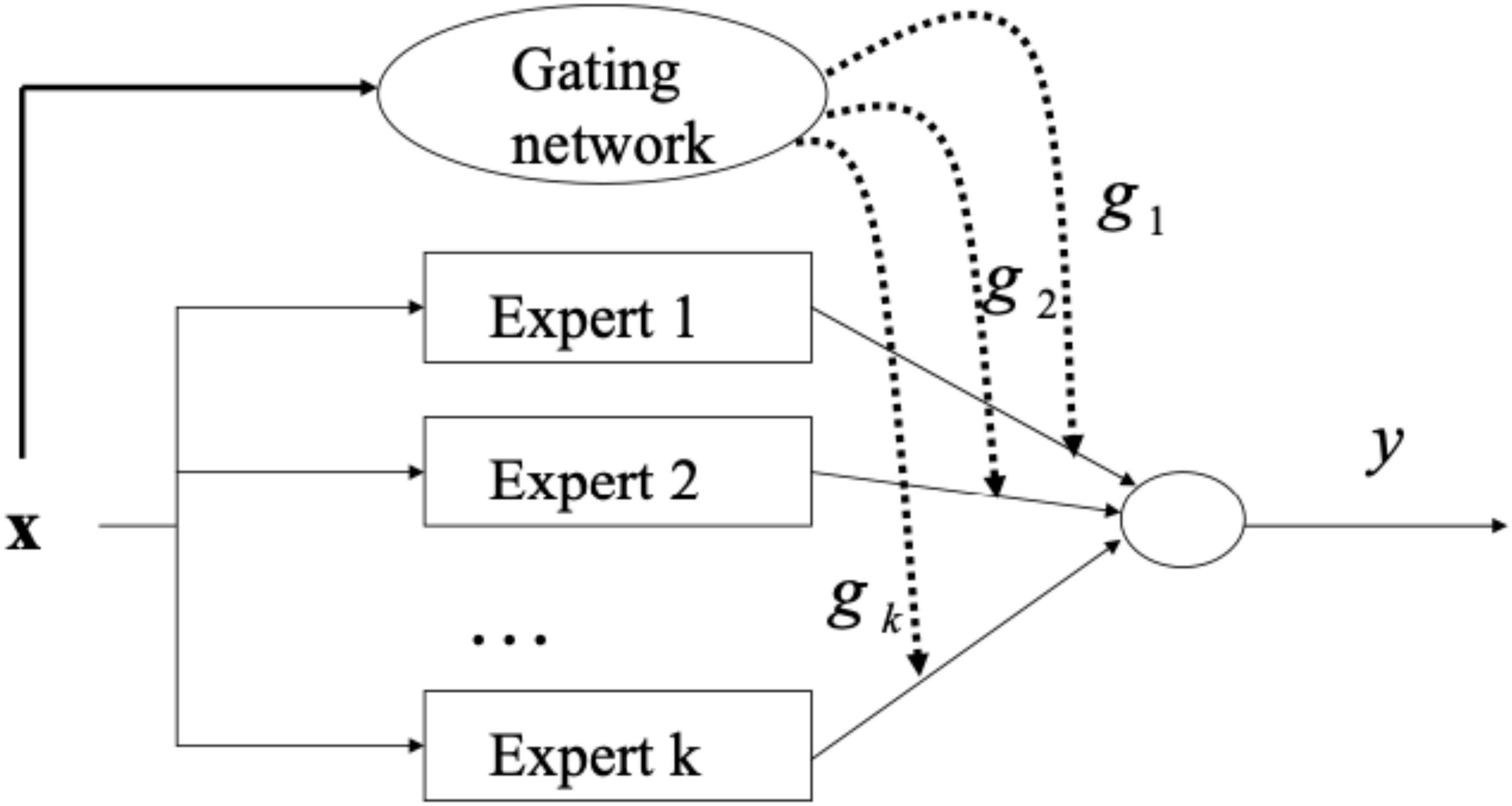
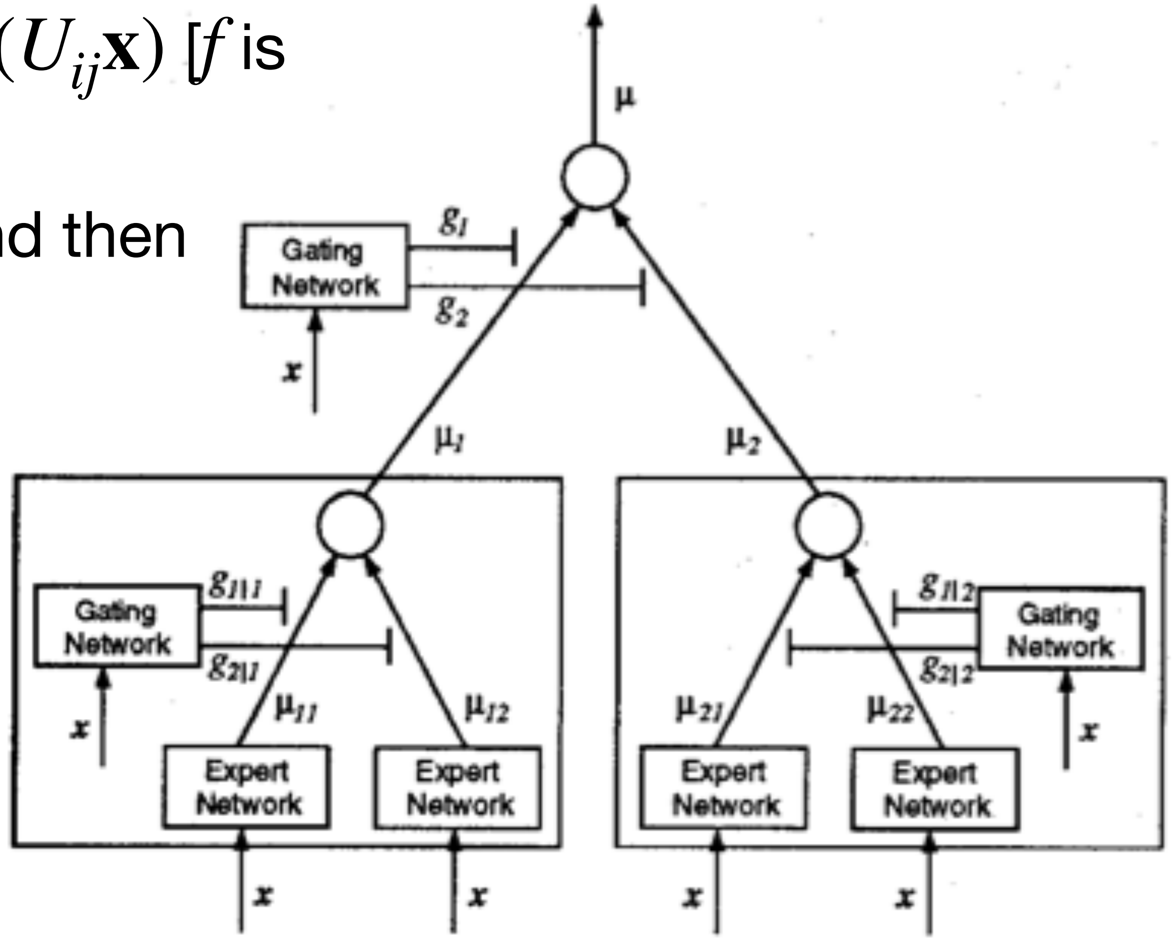


Image from Milos Hauskrecht, https://people.cs.pitt.edu/~milos/courses/cs2750-Spring04/lectures/class22.pdf

Hierarchical Mixture of Experts ("soft")

- Generalized Linear Models for experts & gating
- Expert network (i, j) produces output $\mu_{ij} = f(U_{ij}\mathbf{x})$ [f is the non-linearity]
- Gating networks compute a linear function and then softmax
  $\xi_i = \mathbf{v}_i^T\mathbf{x}$, and $g_i = e^{\xi_i} / \sum_k e^{\xi_k}$
  - Lower levels use $g_{j|i}$ and $\xi_{ij} = \mathbf{v}_{ij}^T\mathbf{x}$
  - $\mu_i = \sum_j g_{j|i}\mu_{ij}$
- Probabilistic interpretation:
  $$P(y\,|\,\mathbf{x}, \theta^0) = \sum_i g_i(\mathbf{x}, \mathbf{v}_i^0) \sum_j g_{j|i}(\mathbf{x}, \mathbf{v_{ij}}^0) P(y\,|\,\mathbf{x}, \theta_{ij}^0)$$
- Used EM to train; lots of hair, as usual!

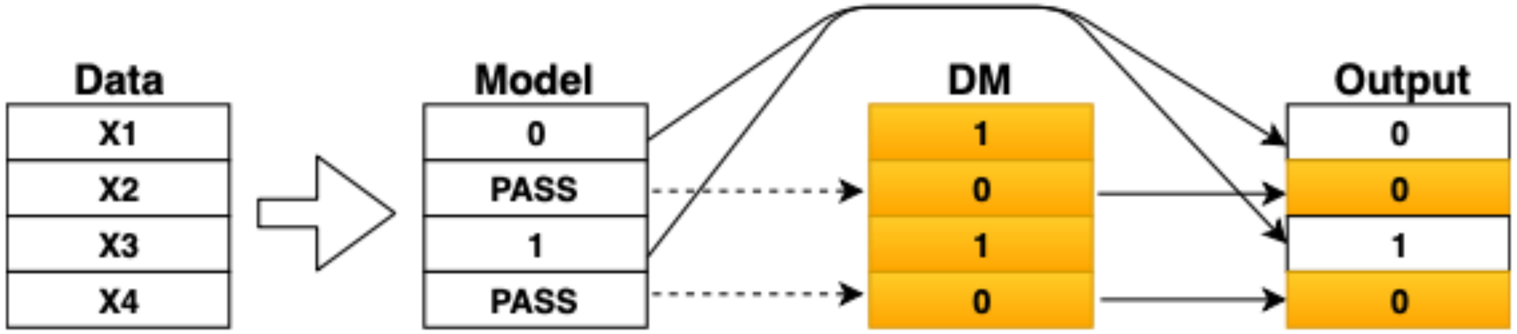


c.f., Random Forests

Jordan, M. I., & Jacobs, R. A. (1994). Hierarchical Mixtures of Experts and the EM Algorithm. Neural Computation, 6(2), 181–214. http://doi.org/10.1162/neco.1994.6.2.181

Deferring

- Assume two “systems”
  - “Model” can either decide (binary choice: 0/1) or **PASS**
  - Decision Maker (DM) may be better, but more costly (e.g., may be human)
    - May have additional inputs
    - Opaque



  - Uses of first-stage model:
    - Flag difficult cases for review
    - Cull a large pool of cases
    - Audit DM for bias
    - …

Madras, D., Pitassi, T., & Zemel, R. S. (2018). Predict Responsibly - Improving Fairness and Accuracy by Learning to Defer. NeurIPS.

Model for Deferring

- Usual set-up:
  - inputs X, output Y, additional inputs Z available only to DM
  - $X \in \mathbb{R}^n, Y \in \{0,1\}, Z \in \mathbb{R}^m$
  - $s \in \{0,1\}$ is another output of the model, 1 = PASS
-
$$P_{defer}(Y|X,Z) = \prod_i \left[P_M(Y_i = 1|X_i)^{Y_i}(1 - P_M(Y_i = 1|X_i))^{(1-Y_i)}\right]^{(1-s_i|X_i)}$$
$$[P_D(Y_i = 1|X_i, Z_i)^{Y_i}(1 - P_D(Y_i = 1|X_i, Z_i))^{(1-Y_i)}]^{(s_i|X_i)}$$
  - Model prediction: $\hat{Y}_M = f(x) = P_M(Y = 1|X) \in [0,1]$
  - DM prediction: $\hat{Y}_D = h(X,Z) = P_D(Y = 1|X,Z) \in [0,1]$; h is “black box"
  - System prediction: $\hat{Y} = (1 - s)\hat{Y}_M + s\hat{Y}_d \in [0,1]$
  - Defer decision: $s = g(X) \in 0,1$
- Learn max likelihood solution to P_{defer}, so f, g adapt to h

- Minimize negative log-likelihood

$$\mathscr{L}_{defer}(Y, \hat{Y}_M, \hat{Y}_D, s) = -\log P_{defer}(Y | X, Z)$$

- $$= \sum_i [(1 - s_i) l(Y_i, \hat{Y}_{M,i}) + s_i l(Y_i, \hat{Y}_{D,i})]$$

  where $l(Y, p) = Y \log p + (1 - Y)\log(1 - p)$, the log probability of the label wrt prediction p

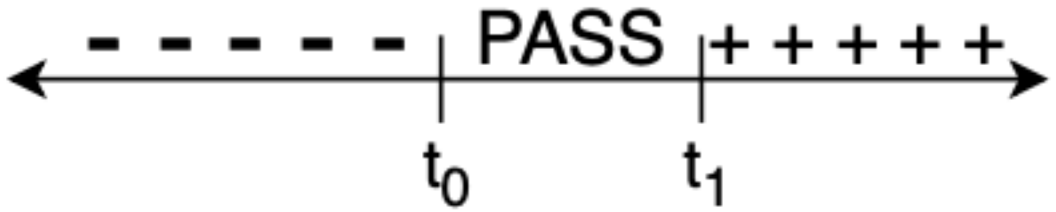
- Like a mixture-of-experts learning problem, except that we cannot learn the parameters of DM

Contrast with Learning to Reject

- Learning to reject focuses only on the accuracy of the stage 1 model:
  $$\mathscr{L}_{reject}(Y, \hat{Y}_M, s) = -\sum_i [(1 - s_i) l(Y_i, \hat{Y}_{M,i}) + s_i \gamma_{reject}]$$
  where γ_{reject} is a penalty for each rejection
- $l(Y_i, \hat{Y}_i) = \mathbf{1}[Y_i = \hat{Y}_i]$ is the classification accuracy
- $P_{reject}(Y|X) = \prod_i [\hat{Y}_{M,i}^{Y_i}(1 - \hat{Y}_{M,i})^{(1-Y_i)}]^{1-s_i} \exp(\gamma_{reject})^{s_i}$
- Rejection learning is a special case of learning to defer
  - Add a "defer" penalty to $\mathscr{L}_{defer}$ and assume DM has a constant loss; then
    $$\mathscr{L}_{defer}(Y, \hat{Y}_M, \hat{Y}_D, s) = \sum_i [(1 - s_i) l(Y_i, \hat{Y}_{M,i}) + s_i l(Y_i, \hat{Y}_{D,i}) + s_i \gamma_{defer}]$$
  - If $l(Y, \hat{Y}_D) = \alpha$, a constant, then $\gamma_{defer} = \gamma_{reject} - \alpha$

Learning a Deferral Model

- The overall model is a mixture of Y_M and Y_D
- Model the probability of deferral as π, i.e., $s \sim Ber(\pi)$
- $\hat{Y}_M, \pi$ are functions of the inputs X, parameterized by θ, which we learn
  $\mathscr{L}_{defer}(Y, \hat{Y}_M, \hat{Y}_D, \pi; \theta) = -\mathbb{E}_{s \sim Ber(\pi)} \mathscr{L}(Y, \hat{Y}_M, \hat{Y}_D, s; \theta)$
- $$= \sum_i \mathbb{E}_{s \sim Ber(\pi_i)}[(1 - s_i) l(Y_i, \hat{Y}_{M,i}; \theta) + s_i l(Y_i, \hat{Y}_{D,i})]$$
- If we can assume that $\pi = g(\hat{Y}_M)$ alone (and $\hat{Y}_M = f(X)$), then use two thresholds



Train an ANN as a binary classifier with output in [0, 1] and output according to its value compared to the thresholds.

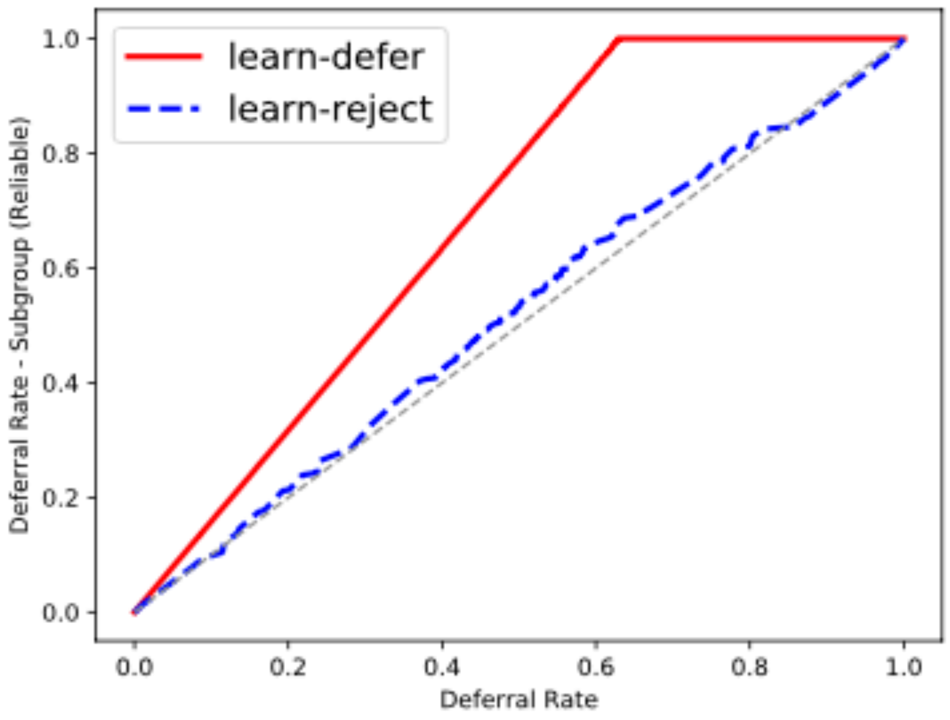
- If $\pi = g(\hat{Y}_M, X)$; useful if DM depends heterogeneously on data, differently from M
  - Train by SGD, sampling $s \sim Ber(\pi)$ during training; + lots more hair!

Fairness as Use Case vs. Inconsistency

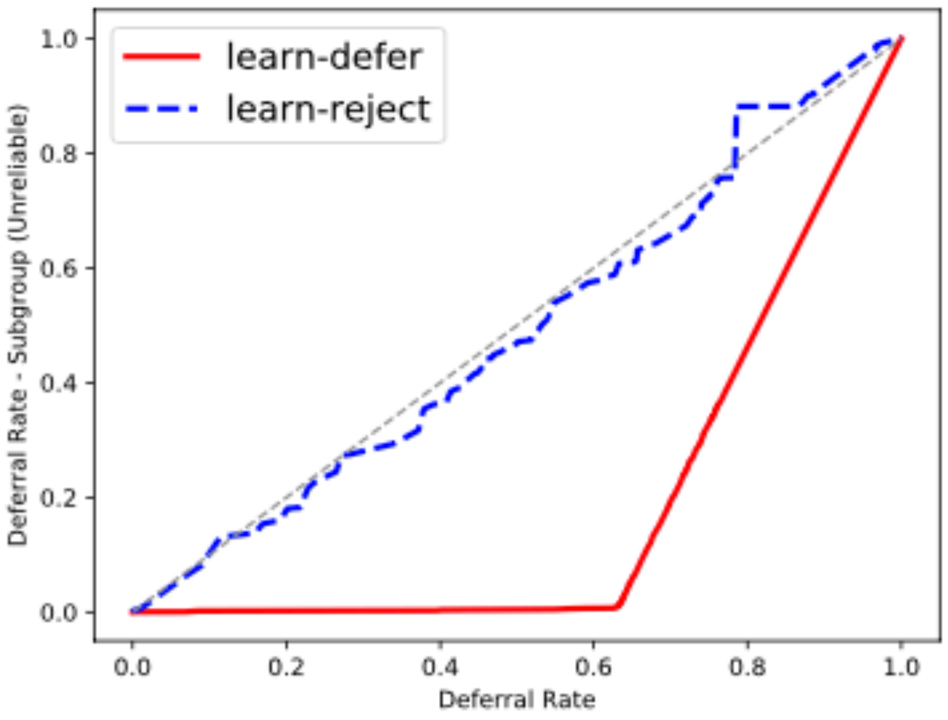
- Many of their experiments study whether one can de-bias a black-box DM such as Compas
- Today, we focus on the case where the DM does well on some subset of cases but poorly on others
  - Example: Predict the patient's Charlson Co-morbidity Index (*without discriminating by age*).
    - Probability of surviving for the next 10 years
    - De-biasing by age seems like an odd goal!
  - Extra information Z: patient's primary condition group
  - To exacerbate inconsistency, they post-hoc invert predictions $\hat{Y}_D$ on 30% of males
  - Must use $\pi = g(\hat{Y}_M, X)$ model to learn to predict where DM is reliable or not

https://www.mdcalc.com/charlson-comorbidity-index-cci

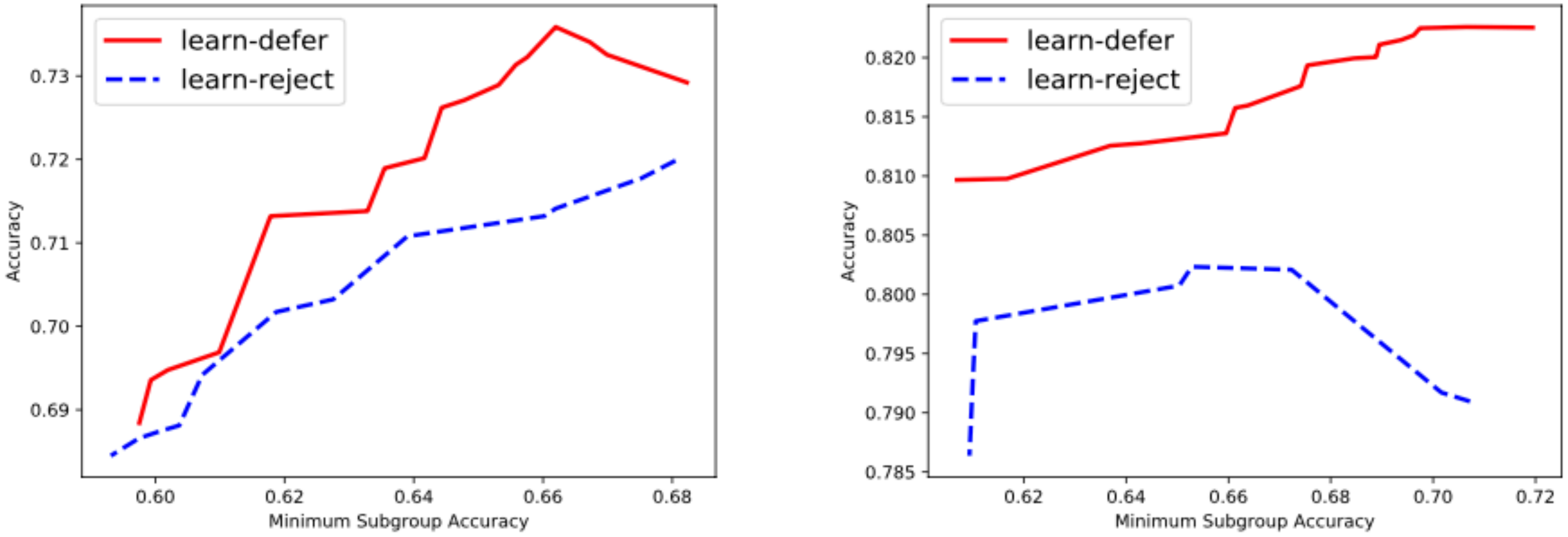
- On subset of cases where DM is reliable, system learns to defer frequently, but not on unreliable cases!



(c) Health, Reliable



(d) Health, Unreliable



(a) COMPAS dataset (b) Health dataset

Figure 5: Each point is a single run in sweep over $\gamma_{reject}/\gamma_{defer}$. X-axis is the model's lowest accuracy over 4 subgroups, defined by the cross product of binarized (sensitive attribute, unreliable attribute), which are (race, age) and (age, gender) for COMPAS and Health respectively. Y-axis is model accuracy. Only best Y-value for each X-value shown. Solid line is learning to defer; dashed line is rejection learning.

Second Opinion

- In medicine, patients facing a dangerous or expensive intervention often seek a second opinion
  - Not worthwhile for minor decisions, unless very easy/cheap
  - What are the relative expertise of the primary and secondary doctors?
  - How correlated are their opinions likely to be?
- Decision might be mapped to the “learning to defer” approach, except that we know very little about DM, since it might be almost anyone
  - Thus, difficult to train $\hat{Y}_M$
- Recall that many studies find agreement by world-class experts only about 80% of the time.
- E.g., study of medical referrals; agreement on diagnosis

final = referral	0.12
refined	0.68
different	0.21

Van Such, M., Lohr, R., Beckman, T., & Naessens, J. M. (2017). Extent of diagnostic agreement among medical referrals. Journal of Evaluation in Clinical Practice, 23(4), 870–874. http://doi.org/10.1111/jep.12747

Examples of Referral vs. Final Diagnoses

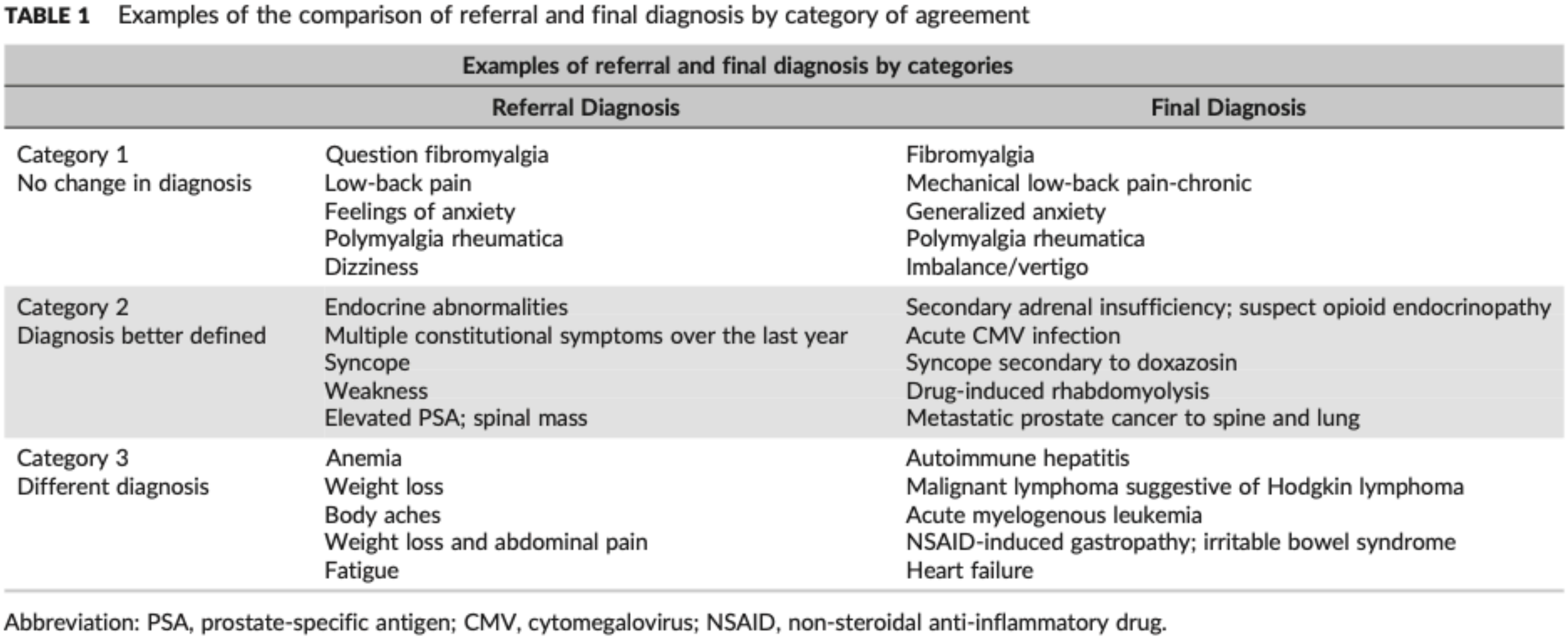


TABLE 1 Examples of the comparison of referral and final diagnosis by category of agreement

Examples of referral and final diagnosis by categories		
	Referral Diagnosis	**Final Diagnosis**
Category 1	Question fibromyalgia	Fibromyalgia
No change in diagnosis	Low-back pain	Mechanical low-back pain-chronic
	Feelings of anxiety	Generalized anxiety
	Polymyalgia rheumatica	Polymyalgia rheumatica
	Dizziness	Imbalance/vertigo
Category 2	Endocrine abnormalities	Secondary adrenal insufficiency; suspect opioid endocrinopathy
Diagnosis better defined	Multiple constitutional symptoms over the last year	Acute CMV infection
	Syncope	Syncope secondary to doxazosin
	Weakness	Drug-induced rhabdomyolysis
	Elevated PSA; spinal mass	Metastatic prostate cancer to spine and lung
Category 3	Anemia	Autoimmune hepatitis
Different diagnosis	Weight loss	Malignant lymphoma suggestive of Hodgkin lymphoma
	Body aches	Acute myelogenous leukemia
	Weight loss and abdominal pain	NSAID-induced gastropathy; irritable bowel syndrome
	Fatigue	Heart failure

Abbreviation: PSA, prostate-specific antigen; CMV, cytomegalovirus; NSAID, non-steroidal anti-inflammatory drug.

Van Such, M., Lohr, R., Beckman, T., & Naessens, J. M. (2017). Extent of diagnostic agreement among medical referrals. Journal of Evaluation in Clinical Practice, 23(4), 870–874. http://doi.org/10.1111/jep.12747

Disagreements Vary by Diagnostic Category

- "Findings from autopsies indicate that diagnostic errors contribute to approximately 10% of patient deaths and diagnostic errors account for 6% to 17% of adverse events in hospitals."
- Mayo Clinic and its referring practices (!)

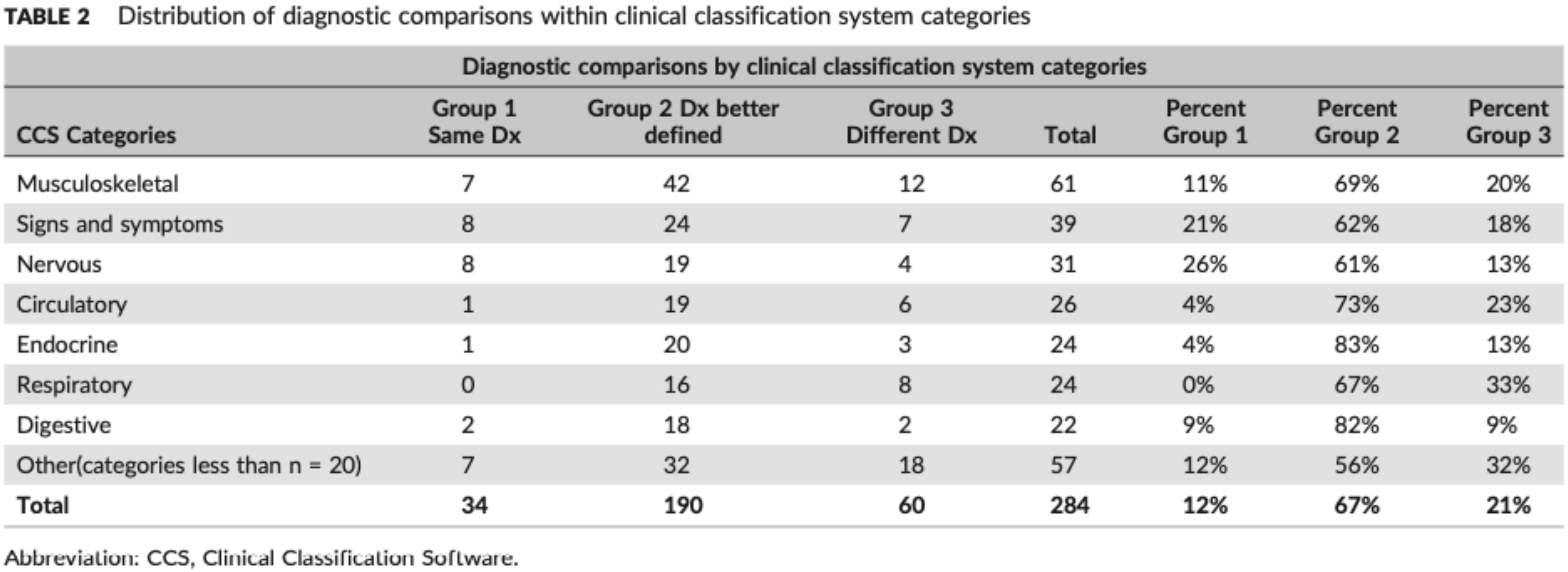


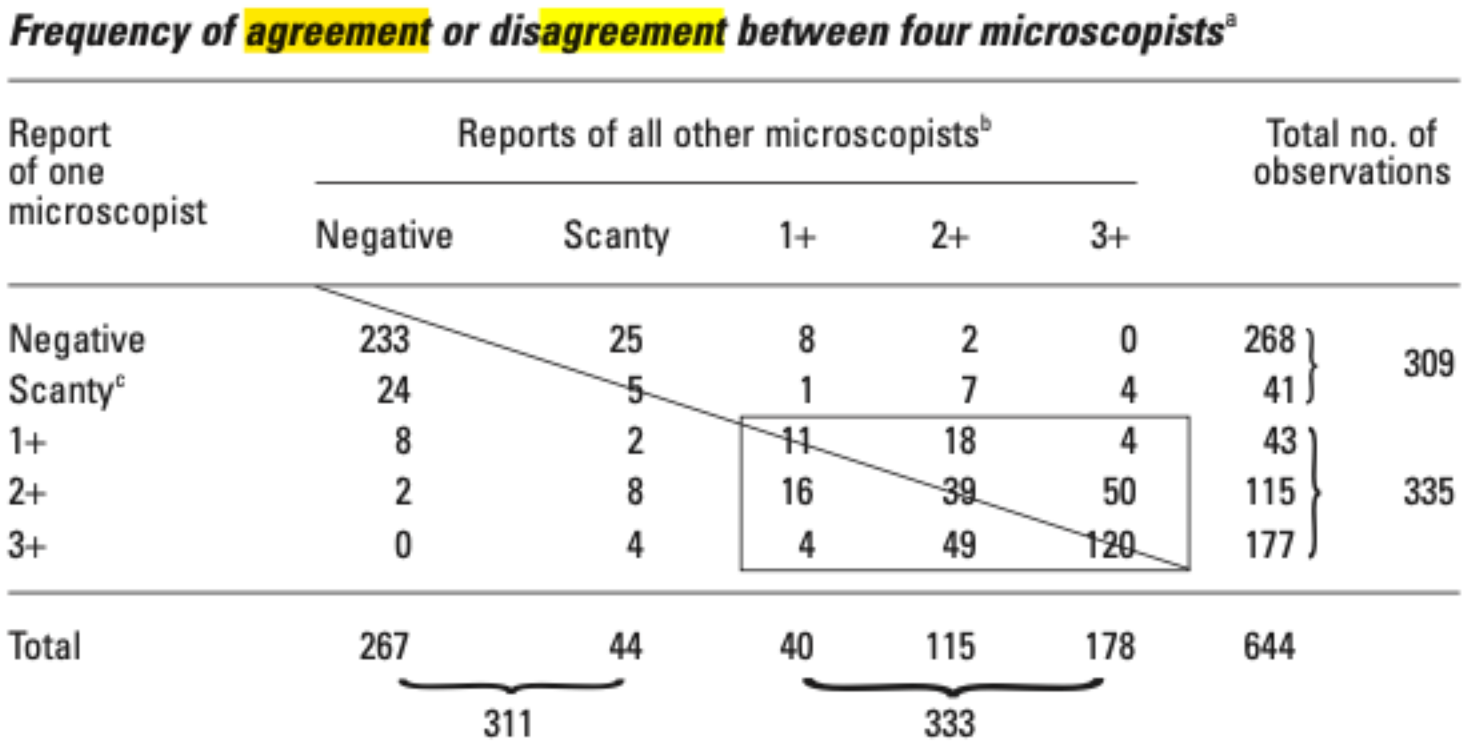
TABLE 2 Distribution of diagnostic comparisons within clinical classification system categories

	Diagnostic comparisons by clinical classification system categories						
CCS Categories	Group 1 Same Dx	Group 2 Dx better defined	Group 3 Different Dx	Total	Percent Group 1	Percent Group 2	Percent Group 3
Musculoskeletal	7	42	12	61	11%	69%	20%
Signs and symptoms	8	24	7	39	21%	62%	18%
Nervous	8	19	4	31	26%	61%	13%
Circulatory	1	19	6	26	4%	73%	23%
Endocrine	1	20	3	24	4%	83%	13%
Respiratory	0	16	8	24	0%	67%	33%
Digestive	2	18	2	22	9%	82%	9%
Other(categories less than n = 20)	7	32	18	57	12%	56%	32%
Total	**34**	**190**	**60**	**284**	**12%**	**67%**	**21%**

Abbreviation: CCS, Clinical Classification Software.

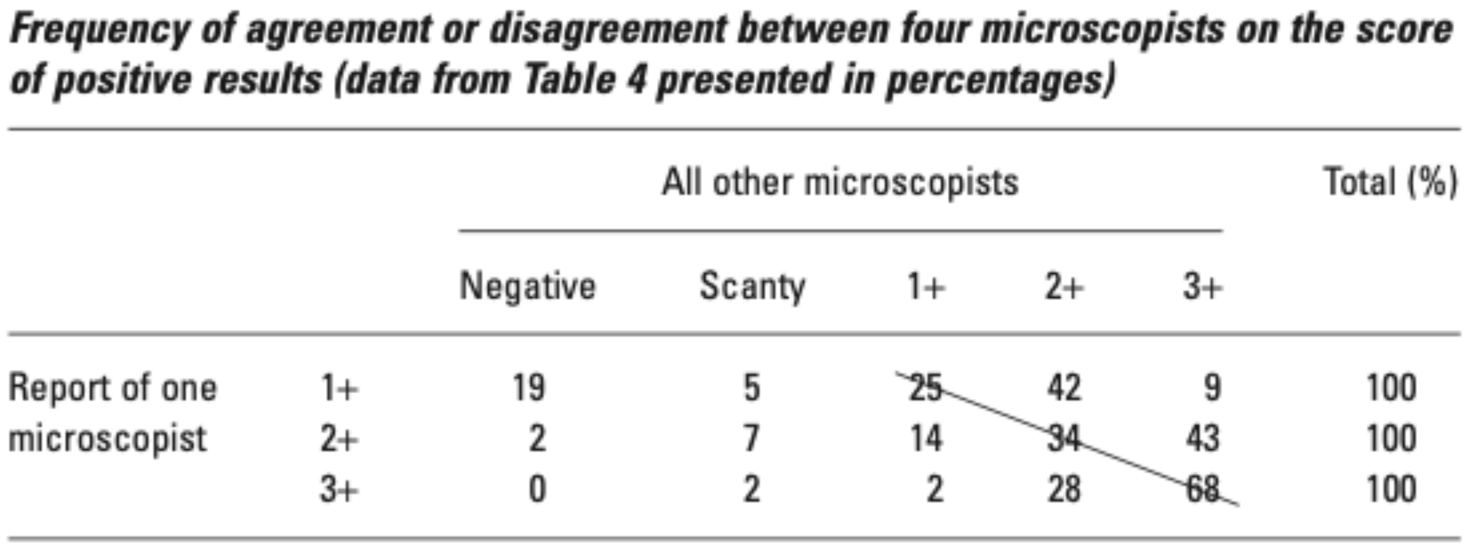
Van Such, M., Lohr, R., Beckman, T., & Naessens, J. M. (2017). Extent of diagnostic agreement among medical referrals. Journal of Evaluation in Clinical Practice, 23(4), 870–874. http://doi.org/10.1111/jep.12747

Such Disagreement are Common:
How reliable is smear microscopy (for TB)?



Frequency of agreement or disagreement between four microscopists[a]

Report of one microscopist	Reports of all other microscopists[b]					Total no. of observations	
	Negative	Scanty	1+	2+	3+		
Negative	233	25	8	2	0	268	309
Scanty[c]	24	5	1	7	4	41	
1+	8	2	11	18	4	43	335
2+	2	8	16	39	50	115	
3+	0	4	4	49	120	177	
Total	267	44	40	115	178	644	
	311		333				



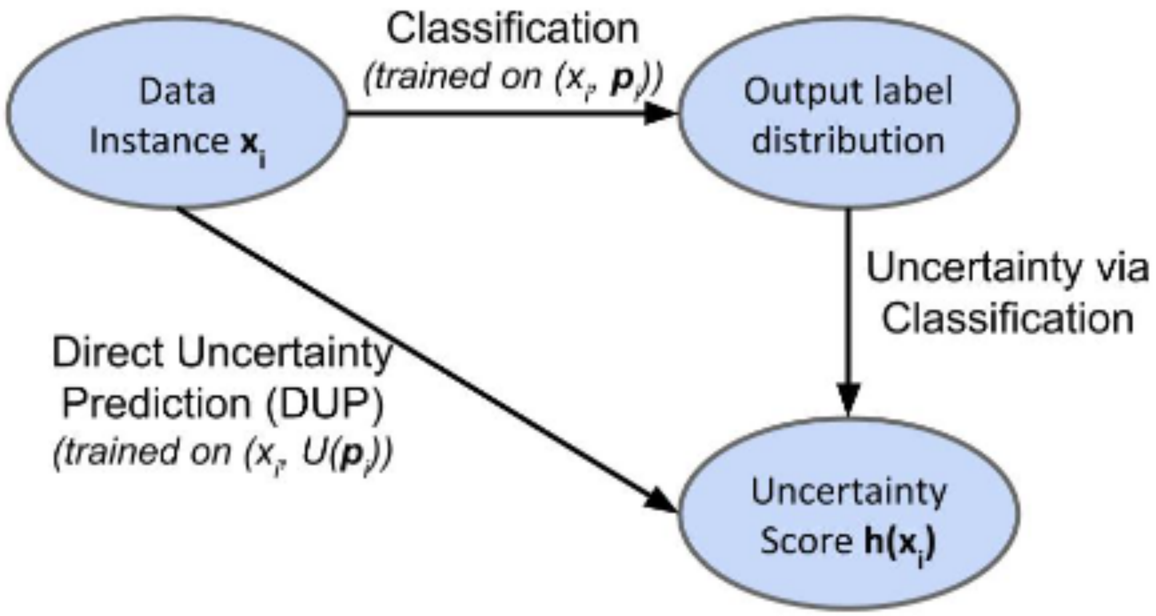
Frequency of agreement or disagreement between four microscopists on the score of positive results (data from Table 4 presented in percentages)

		All other microscopists					Total (%)
		Negative	Scanty	1+	2+	3+	
Report of one microscopist	1+	19	5	25	42	9	100
	2+	2	7	14	34	43	100
	3+	0	2	2	28	68	100

Daniel, T. M. *Toman's tuberculosis*. *Case detection, treatment and monitoring: questions and answers*. ASTMH, 2004.
https://tbrieder.org/publications/books_english/toman_2.pdf

Possible Approaches to 2nd Opinion Decision

- Uncertainty via Classification (UVC)
  - Train a classification model
  - Post-process its output distribution to estimate uncertainty
- Direct Uncertainty Prediction (DUP)
  - Train a different prediction model to estimate uncertainty directly from case inputs
- Which is better?
  - What is your intuition? Why?
  - What was mine?


Data Instance $\mathbf{x}_i$
Classification
(trained on (x_i, $\mathbf{p}_i$))
Output label distribution
Uncertainty via Classification
Direct Uncertainty Prediction (DUP)
(trained on (x_i, $U(\mathbf{p}_i)$))
Uncertainty Score $\mathbf{h}(\mathbf{x}_i)$

Raghu, M., Blumer, K., Sayres, R., Obermeyer, Z., Kleinberg, R. D., Mullainathan, S., & Kleinberg, J. M. (2019). Direct Uncertainty Prediction for Medical Second Opinions. Icml.

How to Train

- Setup
  - Cases x_i
  - multiple labels by experts, $y_i^{(1)}, y_i^{(2)}, \ldots, y_i^{(n_i)}$
  - $h(x_i)$ represents uncertainty in the prediction of experts
- UVC
  - Classifier $\hat{p}_i = f(x_i)$ gives distribution over labels
  - In absence of expert judgment, distribution of $\hat{p}_i$ (e.g., variance) can be used as estimate of uncertainty, h
- DUP
  - Assume that the $y_i^{(j)}$ are all drawn from a set of possible *grades*, $c_1, \ldots, c_k$.
  - Empirical histogram is $\hat{p}_i^{(l)} = \dfrac{\sum_j \mathbf{1}_{y_i^{(j)}=c_l}}{n_i}$
  - Target uncertainty function $U(\cdot)$ from this empirical histogram

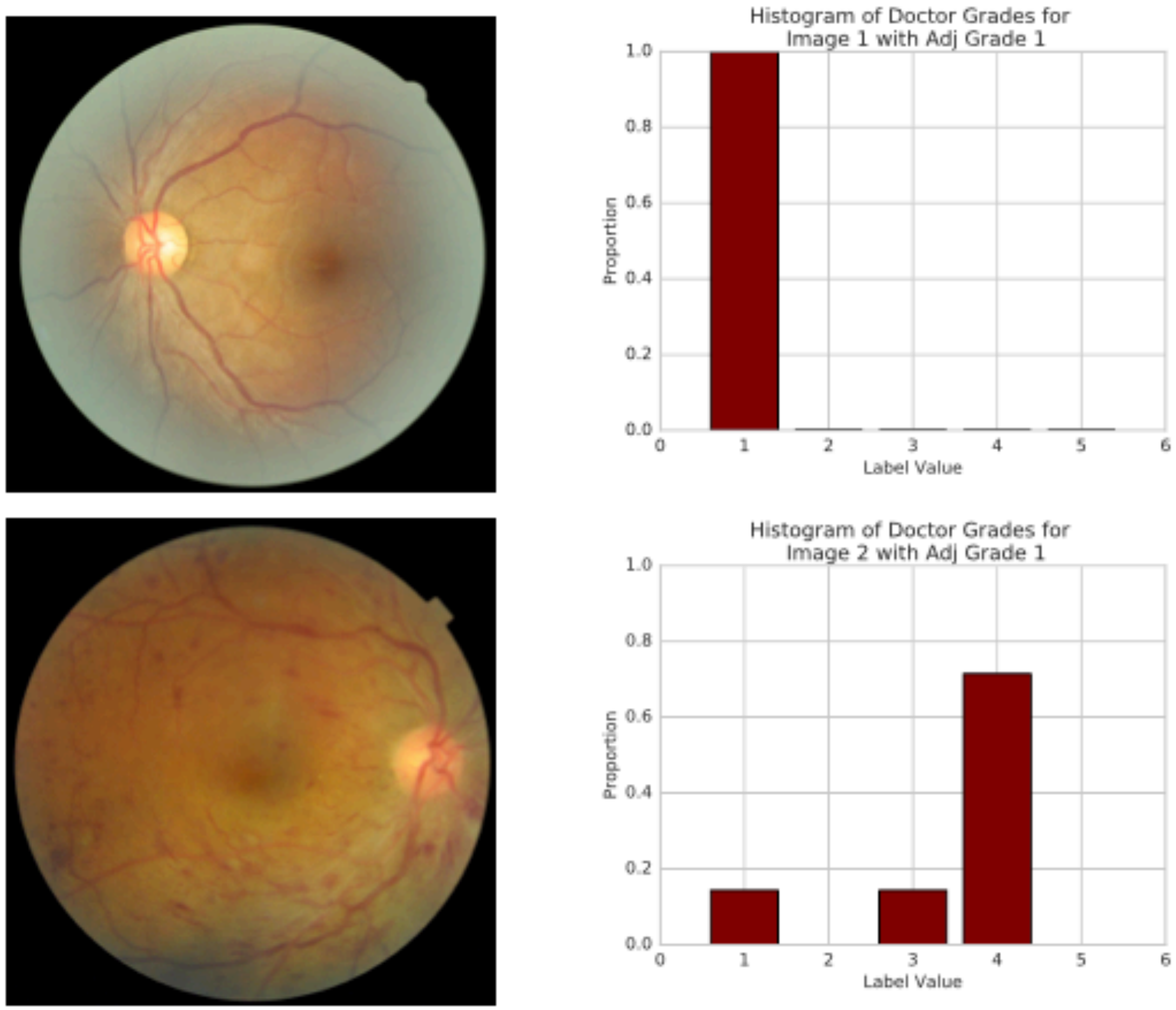

Histogram of Doctor Grades for Image 1 with Adj Grade 1
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
Label Value
Histogram of Doctor Grades for Image 2 with Adj Grade 1
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
Label Value

Figure 2. **Patient cases have features resulting in higher doctor disagreement.** The two rows give example datapoints in our dataset. The patient images x_i, x_j are in the left column, and on the right we have the empirical probability distribution (histogram) of the multiple individual doctor DR grades. For the top image, all doctors agreed that the grade should be 1, while there was a significant spread for the bottom image. When later performing an *adjudication* process (Section 5), where doctors discuss their initial diagnoses with each other and come to a consensus, both patient cases were given an *adjudicated* DR grade of 1.

Possible Versions of $U(\cdot)$

- $U_{disagree}(x_i) = U_{disagree}(\hat{p}_i) = 1 - \sum_{l=1}^{k} (\hat{p}_i^{(l)})^2$
- $U_{var}(x_i) = U_{var}(\hat{p}_i) = \sum_{l=1}^{k} c_l \cdot (\hat{p}_i^{(l)})^2 - (\sum_{l-1}^{k} c_l \cdot \hat{p}_i^{(l)})^2$
- For many versions of $U(\cdot)$ [entropy, variance, …], the paper proves that such estimators are unbiased, whereas UVC has a bias term.

- Note that, as in learning to defer, the doctors have access to more information than the model, which only sees x_i
  - Doctors also see patient and family medical history, demographics, co-morbidities, etc.
  - Let o be all data seen by doctors, and $x_i = g(o)$ where g “hides” some of the information from the model
- Assume k doctor-assigned grades, $c_1, \ldots, c_k$
- Let O be random variable for patient features, Y be the doctor labels for O
- f is a density function that assigns a probability to (patient features, doctor grade) (o, y)
- Let $Y_l = \mathbf{1}_{Y=c_l}$, and $\mathbf{Y} = [Y_1, \ldots, Y_k]$
- Then f is a density over points $f(O = o, \mathbf{Y} = \mathbf{y})$
- Marginal probability of patient features, $f_O = \int_{\mathbf{y}} f(O, \mathbf{y})$

Predict Disagreement

- Doctors have seen O
- Uncertainty of expected value of $\mathbf{Y}$ given o

$$U\left(\int_y \mathbf{y} \cdot f(\mathbf{Y} = \mathbf{y} \mid O)\right) = U(\mathbb{E}[\mathbf{Y} \mid O])$$

- For a specific patient, uncertainty is given by $U(\mathbb{E}[\mathbf{Y} \mid O = o])$
  - but model sees only $x = g(o)$
  - Assume $Y \perp g(O) \mid O$ and $g(O)$ is truly smaller than O
- $h_{dup}(x) = \mathbb{E}[U(\mathbb{E}[\mathbf{Y} \mid O]) \mid g(O) = x] = \int_o U(\mathbb{E}[\mathbf{Y} \mid O = o]) f_O(o \mid g(O) = x)$
  - Computes expectation of uncertainties of all posteriors
- $h_{uvc}(x) = U(\mathbb{E}[\mathbf{Y} \mid g(O) = x) = U\left(\int_o \mathbb{E}[\mathbf{Y} \mid O = o]\, f_O(o \mid g(O) = x)\right)$
  - Computes uncertainty of the expected posterior
- See proof in the paper's appendix.

Raghu, M., Blumer, K., Sayres, R., Obermeyer, Z., Kleinberg, R. D., Mullainathan, S., & Kleinberg, J. M. (2019). Direct Uncertainty Prediction for Medical Second Opinions. Icml.

What is the bias of Uncertainty Via Classification?

- Bias of h_{uvc} using $U_{disagree}$
  - $\mathbb{E}_{g(O)}\left[\sum_{l} Var_{O|g(O)}\left(\mathbb{E}[Y_l|O]|g(O)\right)\right]$
- Bias of h_{uvc} using U_{var}
  - $\mathbb{E}_{g(O)}\left[Var_{O|g(O)}\left(\sum_{l} l \cdot \mathbb{E}[Y_l|O]|g(O)\right)\right]$

Illustrative Simple Empirical Examples

Simple Mixture of Gaussians

$f_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, mixture probabilities q_i

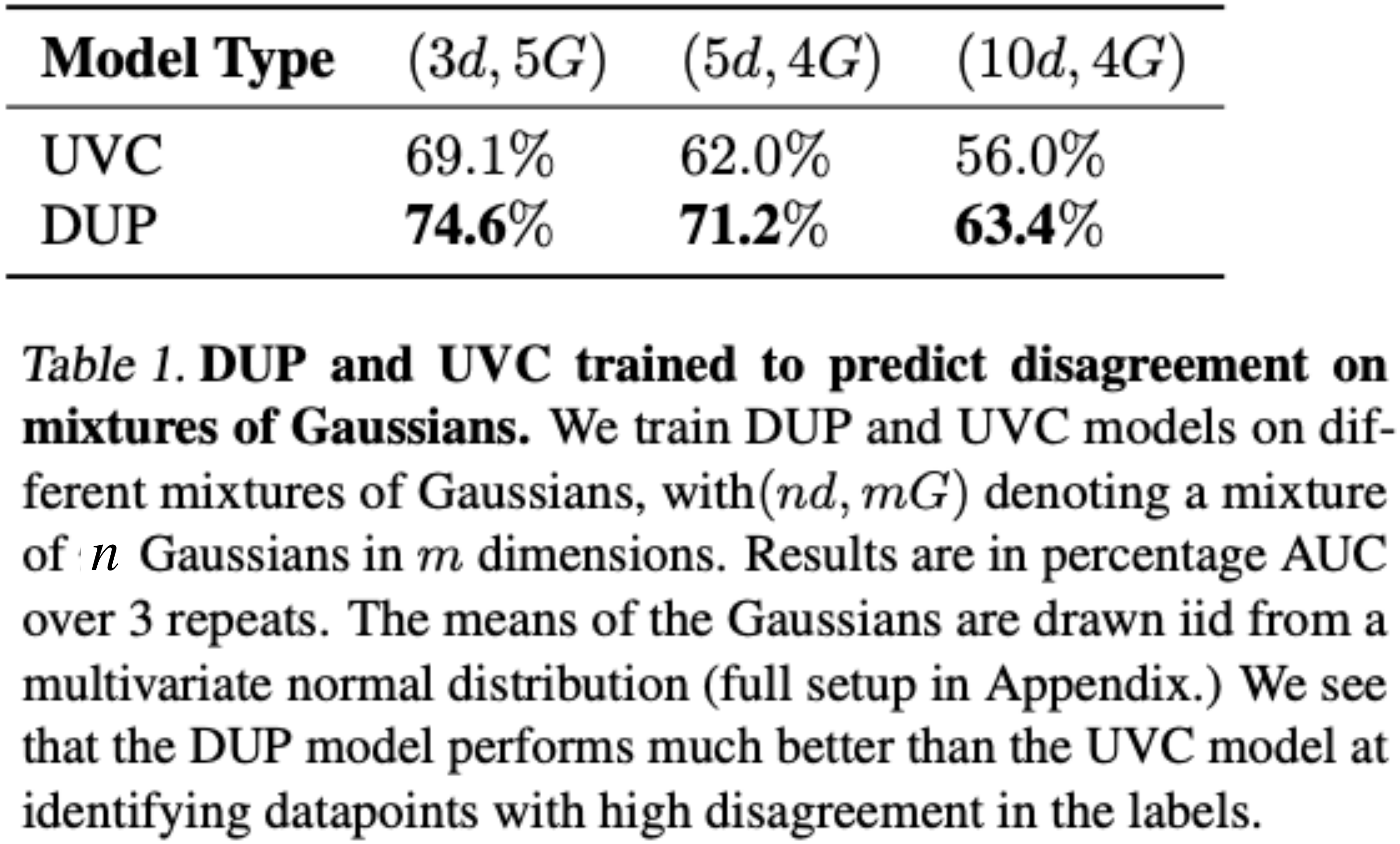
$f(o, y = i) = q_i f_i(o)$ and marginal $f_O(o) = \sum_{i=1}^{k} q_i f_i(o)$

$$f(y = l \mid o) = \frac{q_l f_l(o)}{\sum_{i=1}^{k} q_i f_i(o)}$$

Image classification

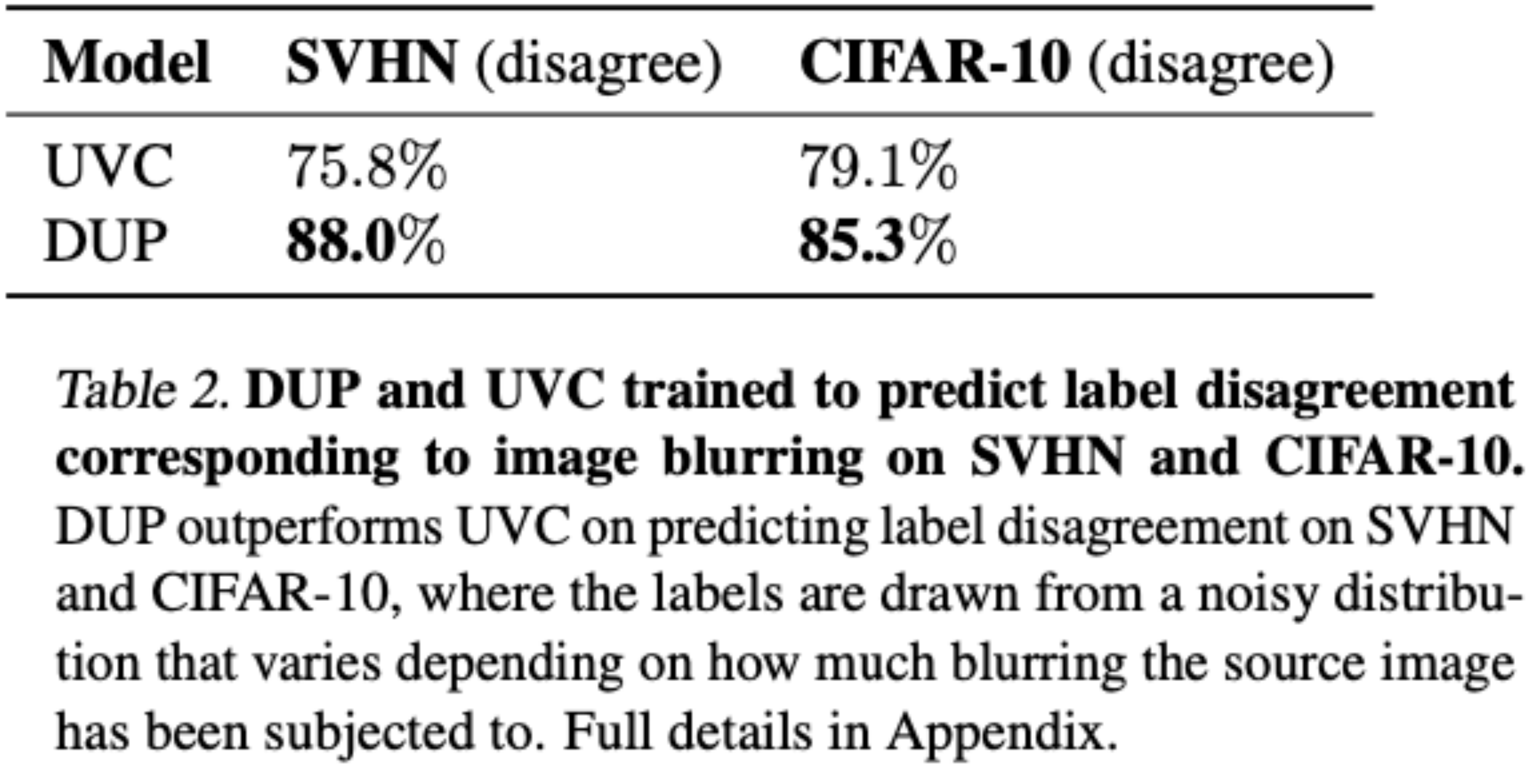
House numbers

Small images ⇒ {airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, trucks}



Model Type	$(3d, 5G)$	$(5d, 4G)$	$(10d, 4G)$
UVC	69.1%	62.0%	56.0%
DUP	**74.6%**	**71.2%**	**63.4%**

Table 1. **DUP and UVC trained to predict disagreement on mixtures of Gaussians.** We train DUP and UVC models on different mixtures of Gaussians, with(nd, mG) denoting a mixture of n Gaussians in m dimensions. Results are in percentage AUC over 3 repeats. The means of the Gaussians are drawn iid from a multivariate normal distribution (full setup in Appendix.) We see that the DUP model performs much better than the UVC model at identifying datapoints with high disagreement in the labels.



Model	**SVHN** (disagree)	**CIFAR-10** (disagree)
UVC	75.8%	79.1%
DUP	**88.0%**	**85.3%**

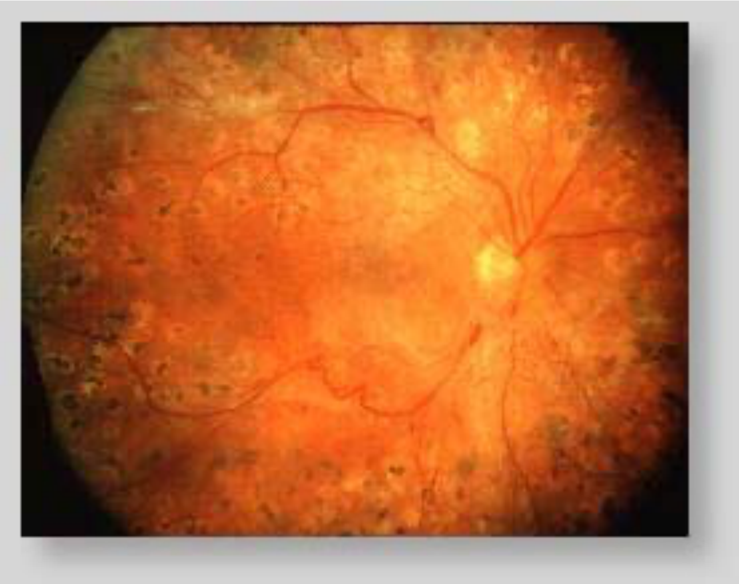
Table 2. **DUP and UVC trained to predict label disagreement corresponding to image blurring on SVHN and CIFAR-10.** DUP outperforms UVC on predicting label disagreement on SVHN and CIFAR-10, where the labels are drawn from a noisy distribution that varies depending on how much blurring the source image has been subjected to. Full details in Appendix.

CIFAR-10: 60K 32x32 color images, 10 labels, balanced

SVHN: Google Street View house number images, 600K, 32x32

Doctor Disagreement on Diabetic Retinopathy Images

- 587x587 retinal fundus images (back of the eye)
  - Can diagnose Diabetic Retinopathy (DR)
    - Leading cause of blindness
    - Treatable if caught early
  - DR graded on scale: {1=no, 2=mild, 3=moderate, 4=severe, 5=proliferative}
    - ≥ 3 is *referable*
  - Unclear how many images they used. Gulshan *et al.* has 2 datasets of
  - ~10K images from ~5K patients and
  - ~2K images from ~900 patients,
    - 7.8%, 14.6% referable in the two datasets

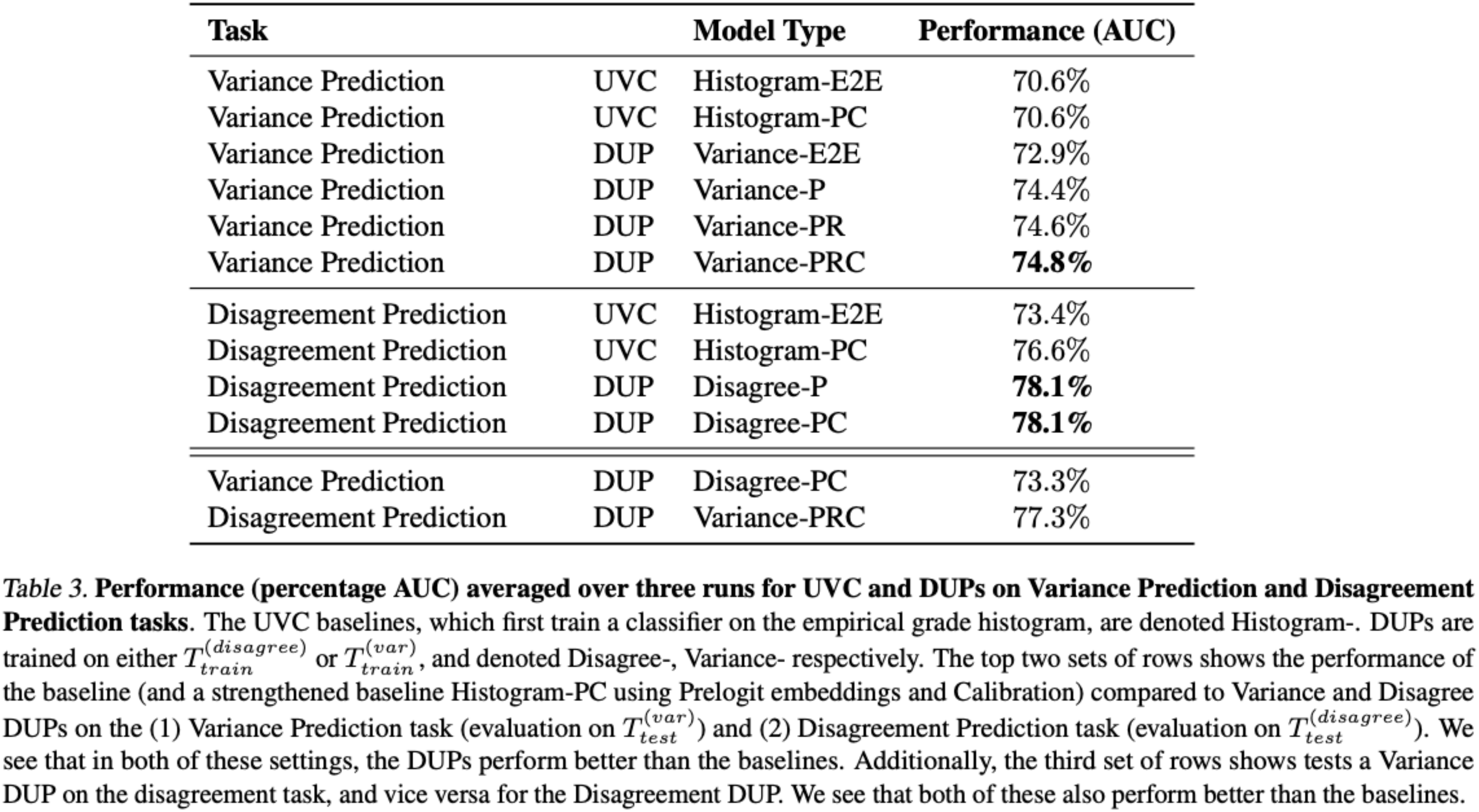


DR photo
https://www.opsweb.org/page/fundusphotography

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., et al. (2016). Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA : the Journal of the American Medical Association, 1–9. http://doi.org/10.1001/jama.2016.17216

DR Experiment

- Train: Test in 80:20 ratio, by patient id
  - (avoid correlations from multiple images of same patient)
  - focus on cases with more than one doctor-assigned label
- Interpreting DR grades:
  - Categorical: E.g., grade 2 always means micro aneurysms, grade 5 may refer to lesions or laser scars. $U_{disagree}$ is the appropriate measure
  - Continuous: Patients tend to progress sequentially through the grades. U_{var} is the appropriate measure
- Train $U_{disagree}$ and U_{var} on their Train data [using pre-trained ImageNet models]
- Consider Test and Train data sets specialized to $(x_i, U_{disagree}(\hat{\mathbf{p}}_i))$ and $(x_i, U_{var}(\hat{\mathbf{p}}_i))$
- Binarize uncertainty measures, at 0.3 and 2/9, respectively
- Train a classifier, h_c, on the Train data pairs $(x_i, \hat{\mathbf{p}}_i)$
- UVC models trained on $U \circ h_c(x_i)$
- DUP models trained directly on pairs $(x_i, U^B_{disagree}(\hat{\mathbf{p}}_i))$ and $(x_i, U^B_{var}(\hat{\mathbf{p}}_i))$, where U^B is the binarized version of U



Task		Model Type	Performance (AUC)
Variance Prediction	UVC	Histogram-E2E	70.6%
Variance Prediction	UVC	Histogram-PC	70.6%
Variance Prediction	DUP	Variance-E2E	72.9%
Variance Prediction	DUP	Variance-P	74.4%
Variance Prediction	DUP	Variance-PR	74.6%
Variance Prediction	DUP	Variance-PRC	**74.8%**
Disagreement Prediction	UVC	Histogram-E2E	73.4%
Disagreement Prediction	UVC	Histogram-PC	76.6%
Disagreement Prediction	DUP	Disagree-P	**78.1%**
Disagreement Prediction	DUP	Disagree-PC	**78.1%**
Variance Prediction	DUP	Disagree-PC	73.3%
Disagreement Prediction	DUP	Variance-PRC	77.3%

Table 3. **Performance (percentage AUC) averaged over three runs for UVC and DUPs on Variance Prediction and Disagreement Prediction tasks**. The UVC baselines, which first train a classifier on the empirical grade histogram, are denoted Histogram-. DUPs are trained on either $T_{train}^{(disagree)}$ or $T_{train}^{(var)}$, and denoted Disagree-, Variance- respectively. The top two sets of rows shows the performance of the baseline (and a strengthened baseline Histogram-PC using Prelogit embeddings and Calibration) compared to Variance and Disagree DUPs on the (1) Variance Prediction task (evaluation on $T_{test}^{(var)}$) and (2) Disagreement Prediction task (evaluation on $T_{test}^{(disagree)}$). We see that in both of these settings, the DUPs perform better than the baselines. Additionally, the third set of rows shows tests a Variance DUP on the disagreement task, and vice versa for the Disagreement DUP. We see that both of these also perform better than the baselines.

Adjudicated Evaluation

- Do high uncertainty scores correspond to cases where average doctor grade differs from adjudicated grade?

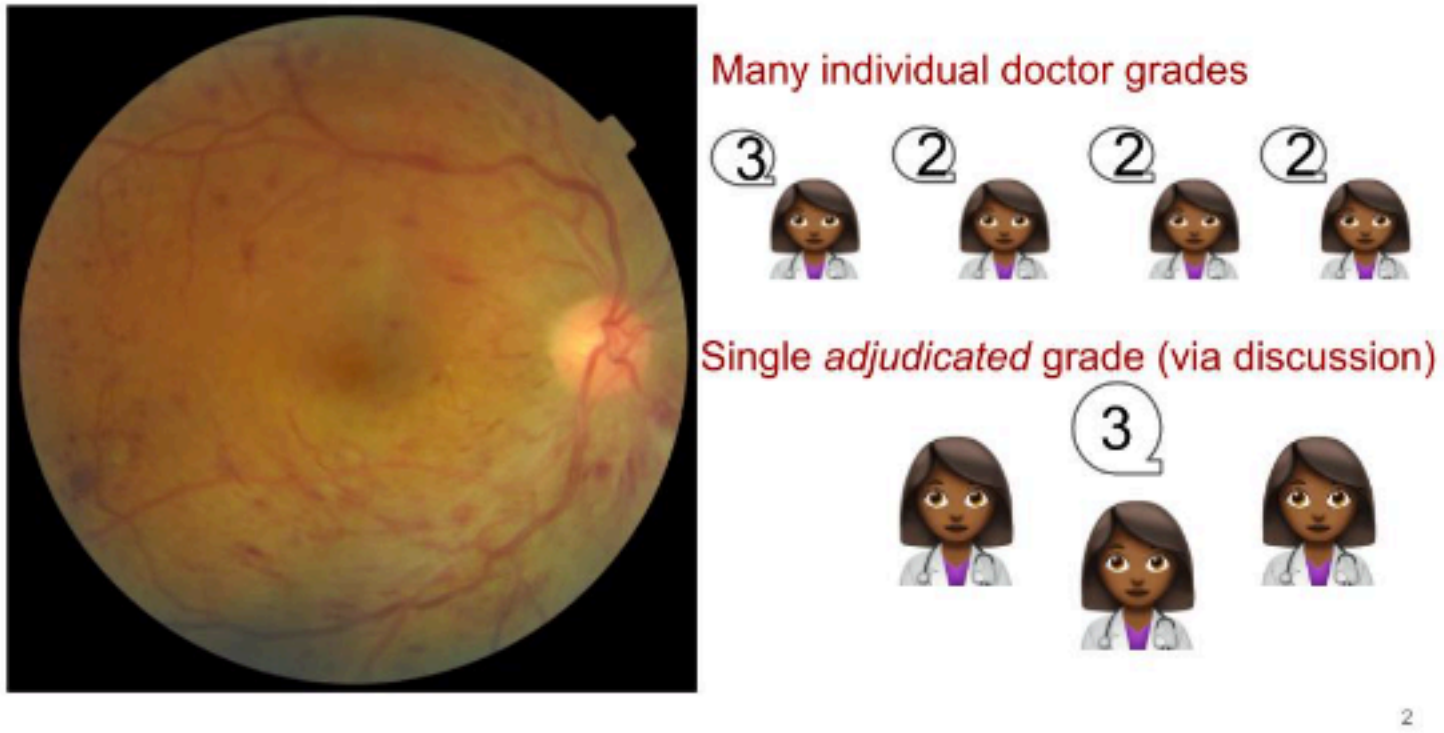
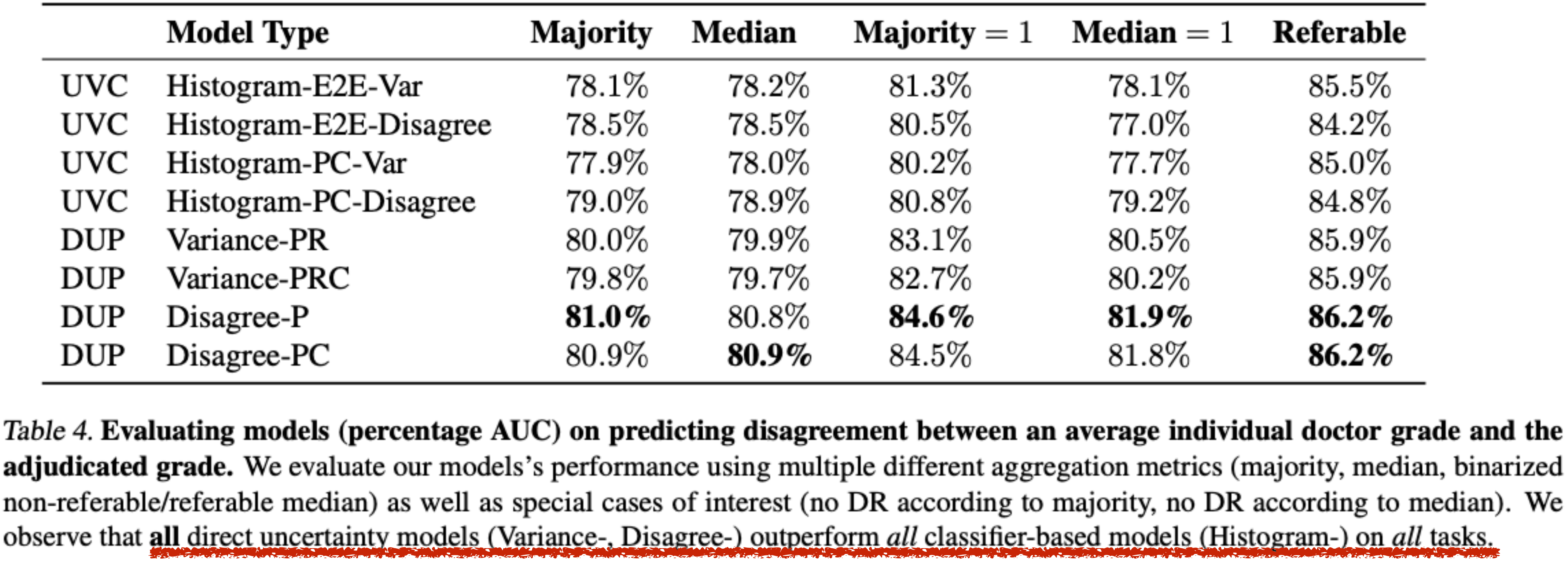

Many individual doctor grades
3
2
2
2
Single *adjudicated* grade (via discussion)
3
2

Figure 3. **Labels for the adjudicated dataset** A**.** The small, gold standard adjudicated dataset A has a very different label structure to the main dataset T. Each image has many individual doctor grades (typically more than 10 grades). These doctors also tend to be specialists, with higher rates of agreement. Additionally, each image has a single *adjudicated* grade, where three doctors first grade the image individually, and then come together to discuss the diagnosis and finally give a single, *consensus diagnosis*.

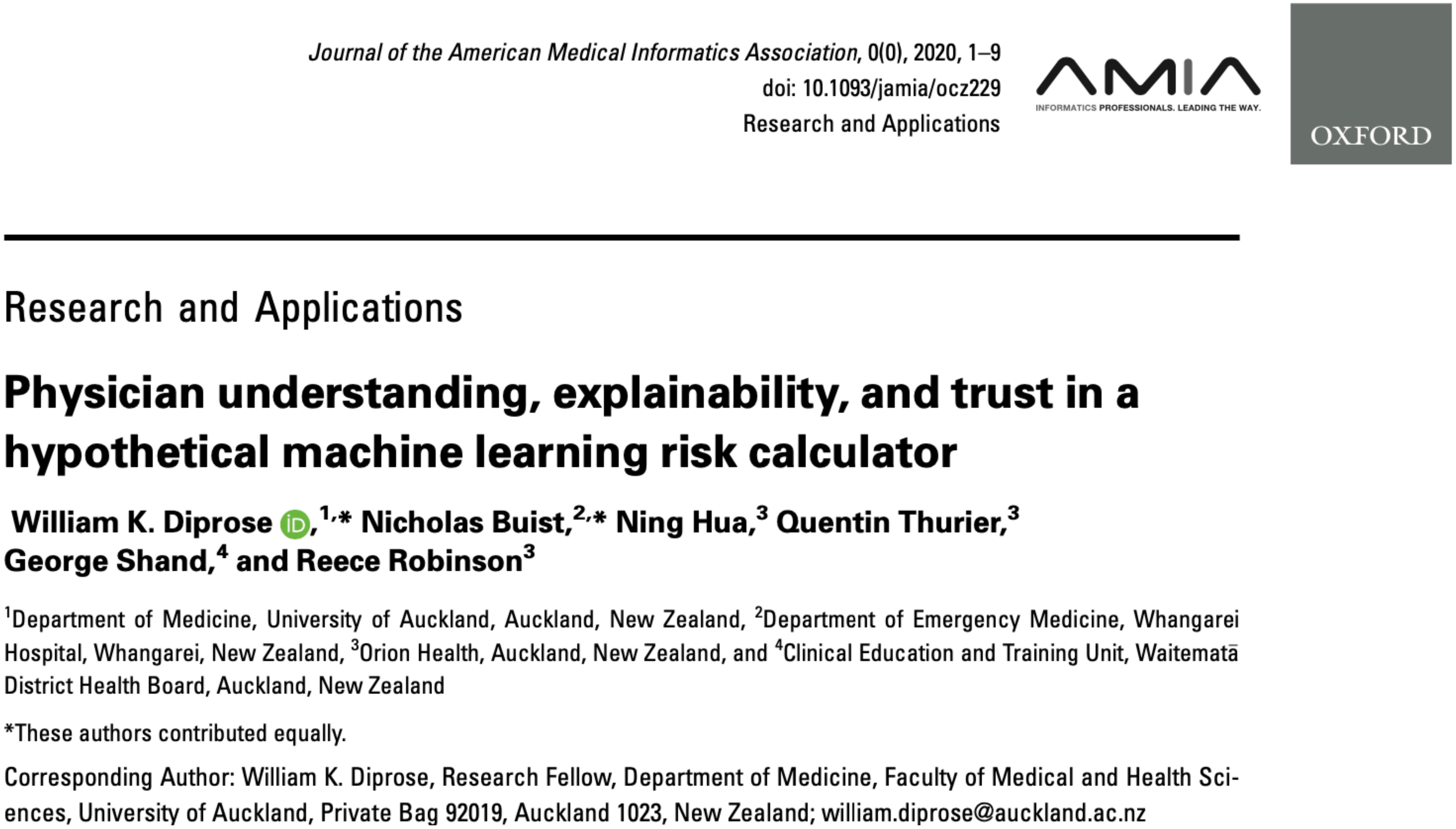


	Model Type	Majority	Median	Majority $= 1$	Median $= 1$	Referable
UVC	Histogram-E2E-Var	78.1%	78.2%	81.3%	78.1%	85.5%
UVC	Histogram-E2E-Disagree	78.5%	78.5%	80.5%	77.0%	84.2%
UVC	Histogram-PC-Var	77.9%	78.0%	80.2%	77.7%	85.0%
UVC	Histogram-PC-Disagree	79.0%	78.9%	80.8%	79.2%	84.8%
DUP	Variance-PR	80.0%	79.9%	83.1%	80.5%	85.9%
DUP	Variance-PRC	79.8%	79.7%	82.7%	80.2%	85.9%
DUP	Disagree-P	**81.0%**	80.8%	**84.6%**	**81.9%**	**86.2%**
DUP	Disagree-PC	80.9%	**80.9%**	84.5%	81.8%	**86.2%**

Table 4. **Evaluating models (percentage AUC) on predicting disagreement between an average individual doctor grade and the adjudicated grade.** We evaluate our models's performance using multiple different aggregation metrics (majority, median, binarized non-referable/referable median) as well as special cases of interest (no DR according to majority, no DR according to median). We observe that **all** direct uncertainty models (Variance-, Disagree-) outperform *all* classifier-based models (Histogram-) on *all* tasks.

Larger Theme:
Decision Support between Machine and Human

- Explanation
- Trust
- Optimization over entire system



Journal of the American Medical Informatics Association, 0(0), 2020, 1–9
doi: 10.1093/jamia/ocz229
Research and Applications

Research and Applications

Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator

William K. Diprose,[1,*] Nicholas Buist,[2,*] Ning Hua,[3] Quentin Thurier,[3] George Shand,[4] and Reece Robinson[3]

[1]Department of Medicine, University of Auckland, Auckland, New Zealand, [2]Department of Emergency Medicine, Whangarei Hospital, Whangarei, New Zealand, [3]Orion Health, Auckland, New Zealand, and [4]Clinical Education and Training Unit, Waitematā District Health Board, Auckland, New Zealand

*These authors contributed equally.

Corresponding Author: William K. Diprose, Research Fellow, Department of Medicine, Faculty of Medical and Health Sciences, University of Auckland, Private Bag 92019, Auckland 1023, New Zealand; william.diprose@auckland.ac.nz

Empirical Evaluation of Understanding and Trust

- GDPR requires patients to be able to receive “meaningful information about the logic involved” in an automated tool.
- Hypothetical ML risk calculator for pulmonary embolism
    - “You are a GP [general practitioner] who has reviewed a 50-year-old woman presenting with shortness of breath. After a history, examination, laboratory tests, ECG [electrocardiogram], and a chest x-ray, you are comfortable you have excluded the most concerning diagnoses. However, you are still considering pulmonary embolism. The practice has installed a piece of software that uses artificial intelligence to assist with ruling out pulmonary embolism. It can stratify patients as either; (1) low risk of pulmonary embolism: reassurance and discharge recommended; or (2) not low risk of pulmonary embolism: computed tomography pulmonary angio- gram recommended. The software automatically analyses the electronic record, including your documented history, examination, and laboratory tests, and provides its recommendation.”
    - Control: “Your patient has a low risk (<1% chance) of pulmonary embolism. They should be reassured and followed up in the community as you deem appropriate. This recommendation is based on a cohort of 10 000 patients who were investigated for pulmonary embolism, of whom 1000 had a similar risk profile. The software has been externally validated.”

Four Possible Explanations (Control + one of these)

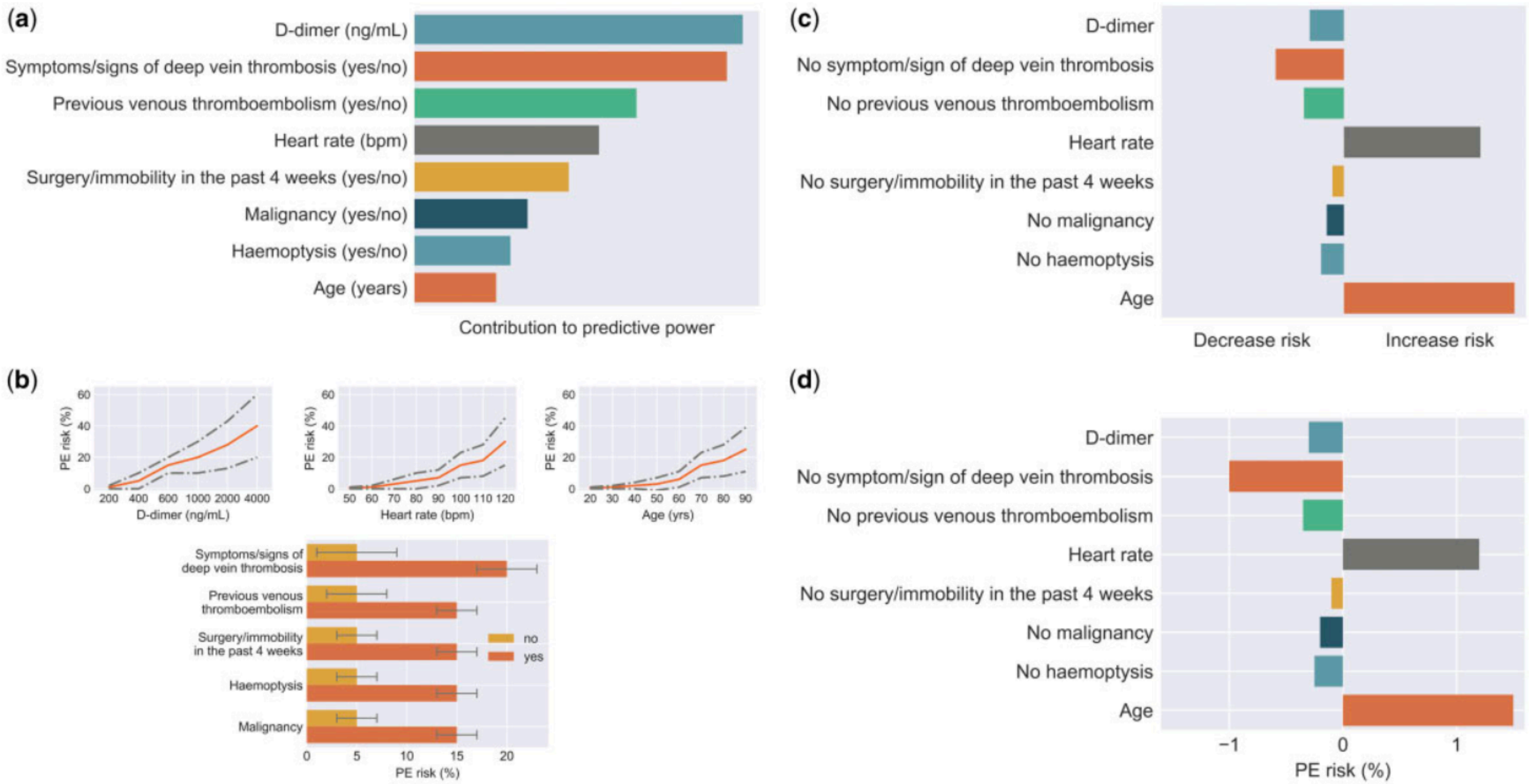
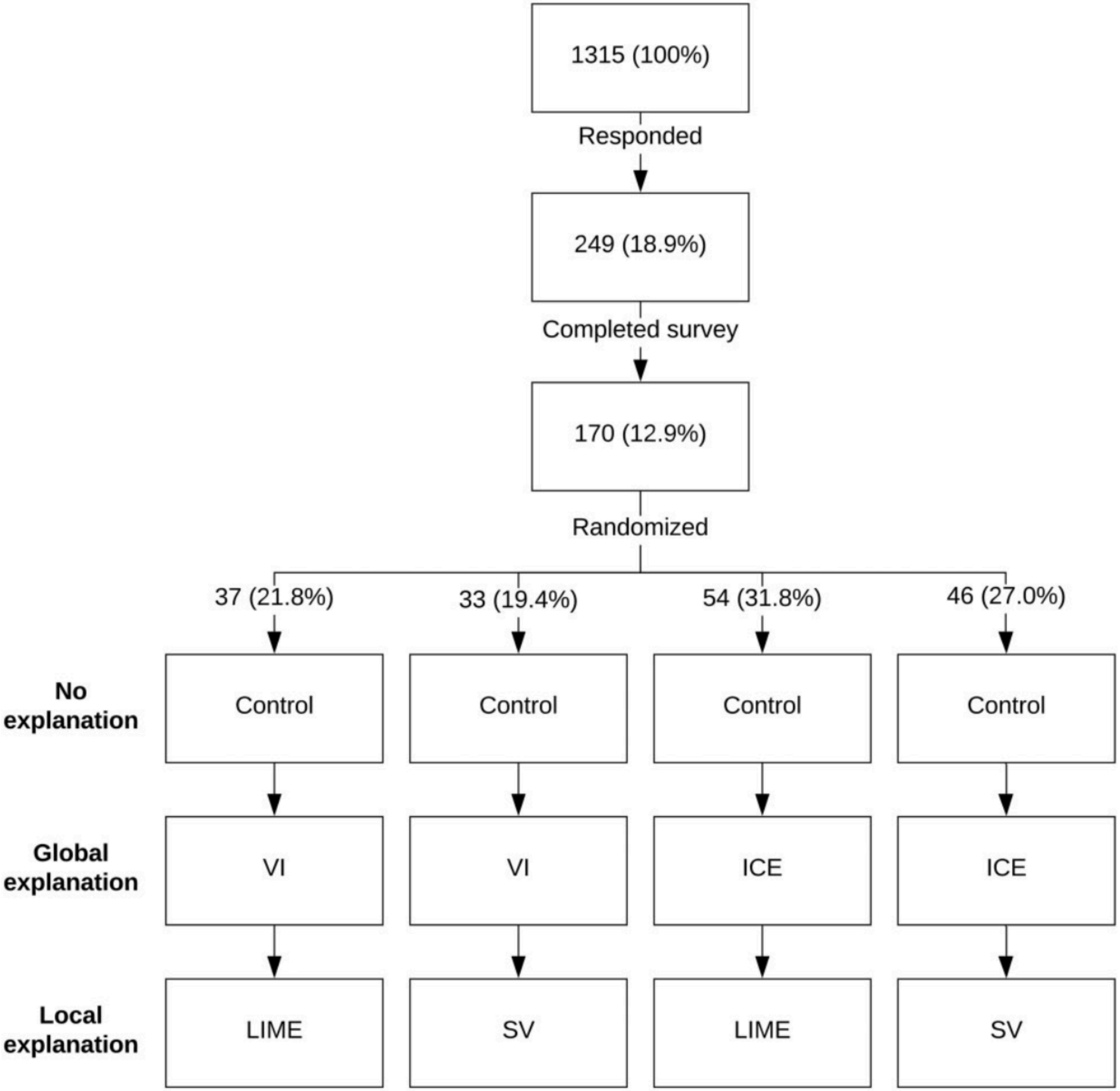


Figure 1. Graphical visualizations shown to participants with the following complementary explanation. (A) For variable importance (VI), the visualization shows the relative importance of each clinical factor used by the model. This is general information about the software's logic. (B) For individual conditional expectation plots (ICE), the visualizations show the average and standard deviation of the software's predictions for different values of the most influential clinical factors used by the model. This is general information about the software's logic. (C) For local interpretable model-agnostic explanations (LIME), the visualization shows the positive or negative relative impact of the most influential clinical factors used by the software to estimate your patient's risk of pulmonary embolism (PE). This profile is specific to your patient. (D) For Shapely values (SVs), the visualization shows the positive or negative contribution of each clinical factor to the risk estimated by the software. The sum of the bars is equal to your patient's PE risk. This profile is specific to your patient.

Survey

- Weak experiment. Only 249/1315 subjects responded, and only 170 completed survey.



Results

- Physicians who reported “Yes” to the question “Would you be able to explain the software’s decision to the patient” were more likely to respond “Yes” to the question “Would you follow the software recommendation?” … No particular ML explainability method had a greater influence on intended physician behavior
- 87.8% of physicians preferred an ML output which contained a model-agnostic explanation, compared with 12.2% who preferred the control output (no model-agnostic explanation)
- Local explanations (LIME [32.1%] and SVs [29.9%]) were preferred over global explanations (VI [18.2%] and ICE [19.7%])

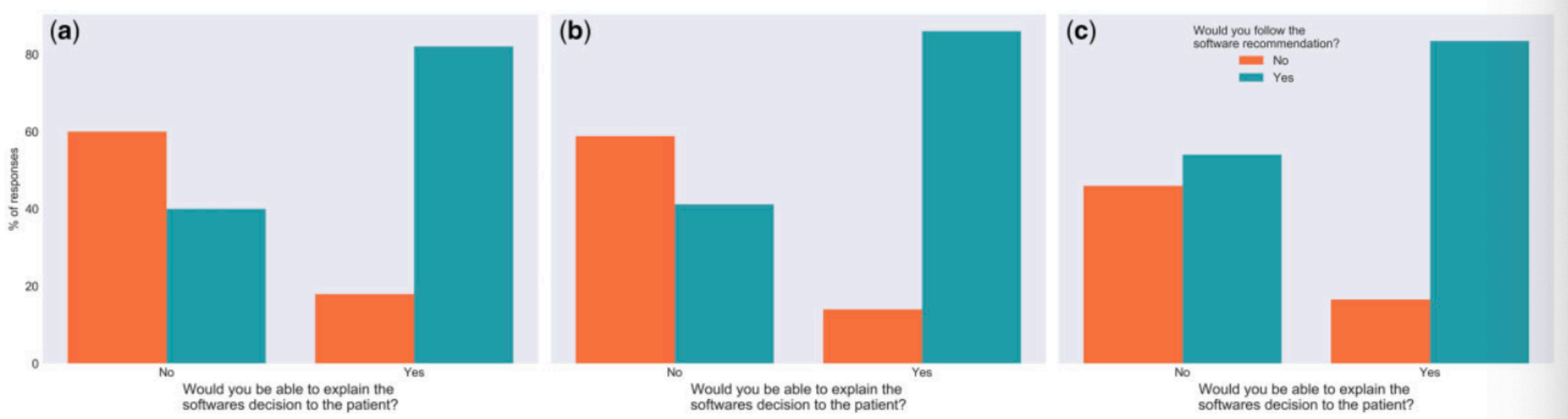


Figure 5. Relationship between explainability and trust. (A) Control condition. (B) Global explanation. (C) Local explanation.