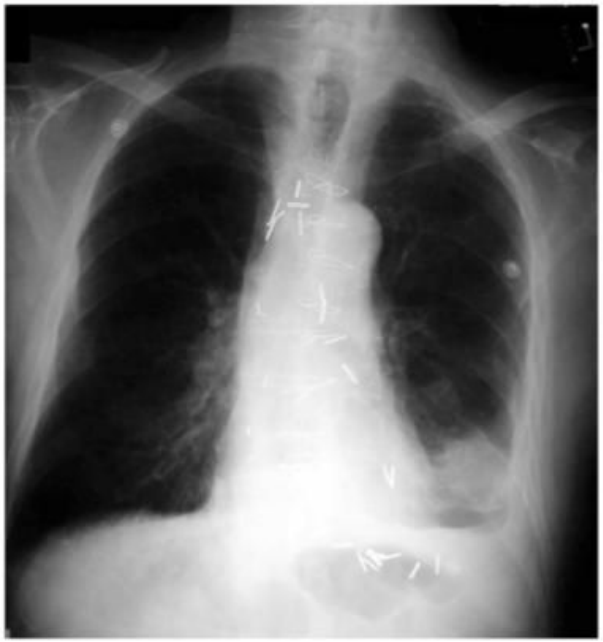
Direct Uncertainty Prediction for Medical Second Opinions

Maithra Raghu, Katy Blumer, Rory Sayres, Ziad Obermeyer, Sendhil Mullainathan, Jon Kleinberg

Poster #246

Human Expert Disagreements

Human Expert Disagreements



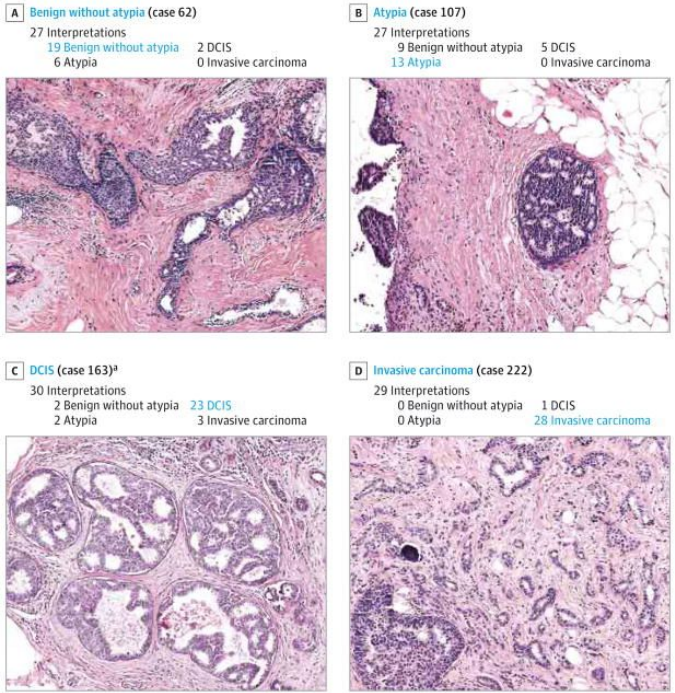
No follow up exam

Follow up exam

Doctor Disagreements

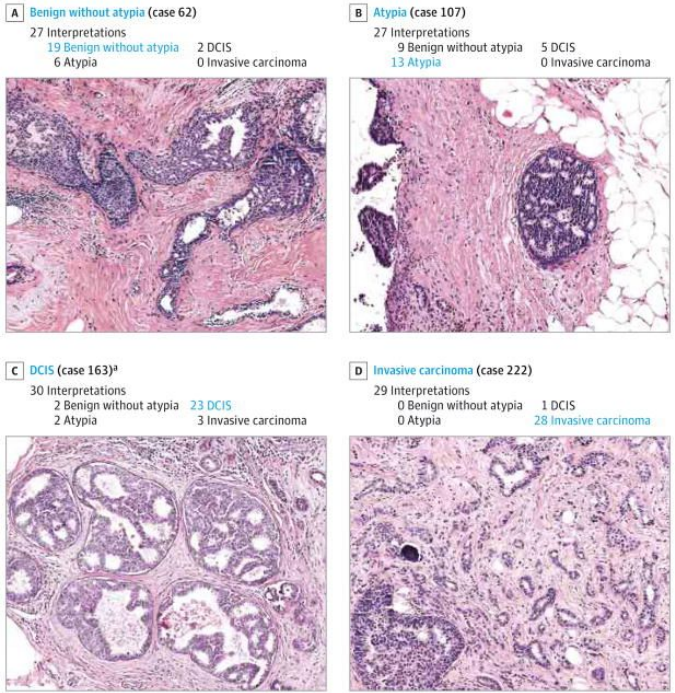
Doctor Disagreements

Diagnostic Concordance Amongst Pathologists Interpreting Breast Biopsy Specimens, UW School of Medicine, JAMA, 2015



Doctor Disagreements

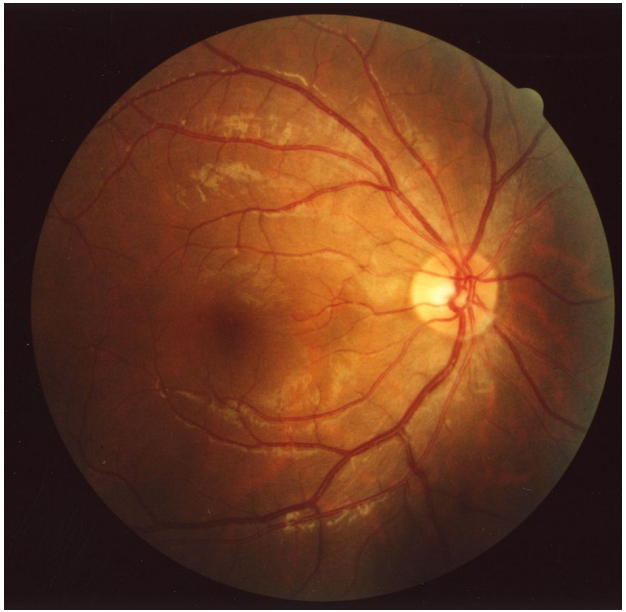
Diagnostic Concordance Amongst Pathologists Interpreting Breast Biopsy Specimens, UW School of Medicine, JAMA, 2015



- Agreement between individual pathologist grade and a panel consensus score on ~240 breast biopsies, 6900 individual case diagnoses
- **25% disagreement** between pathologists and consensus

Doctor Disagreements

Ophthalmology: Diagnosis from Fundus Photographs

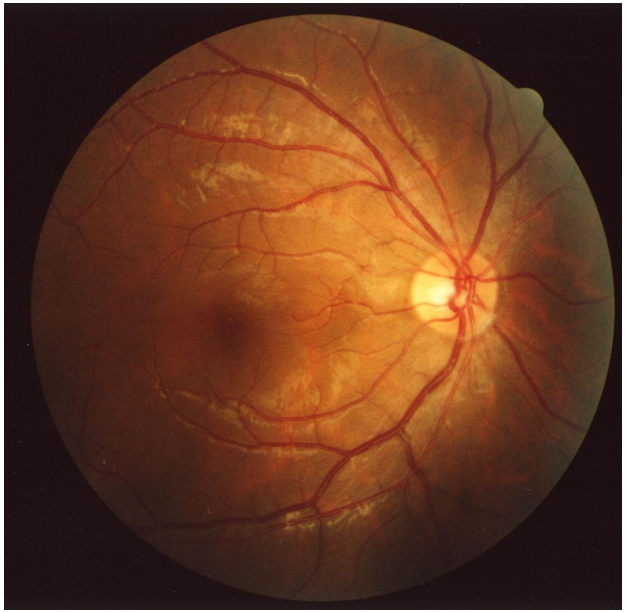


Grade 2: Mild Diabetic Retinopathy

Grade 3: Moderate Diabetic Retinopathy

The Source of Disagreements

Ophthalmology: Diagnosis from Fundus Photographs

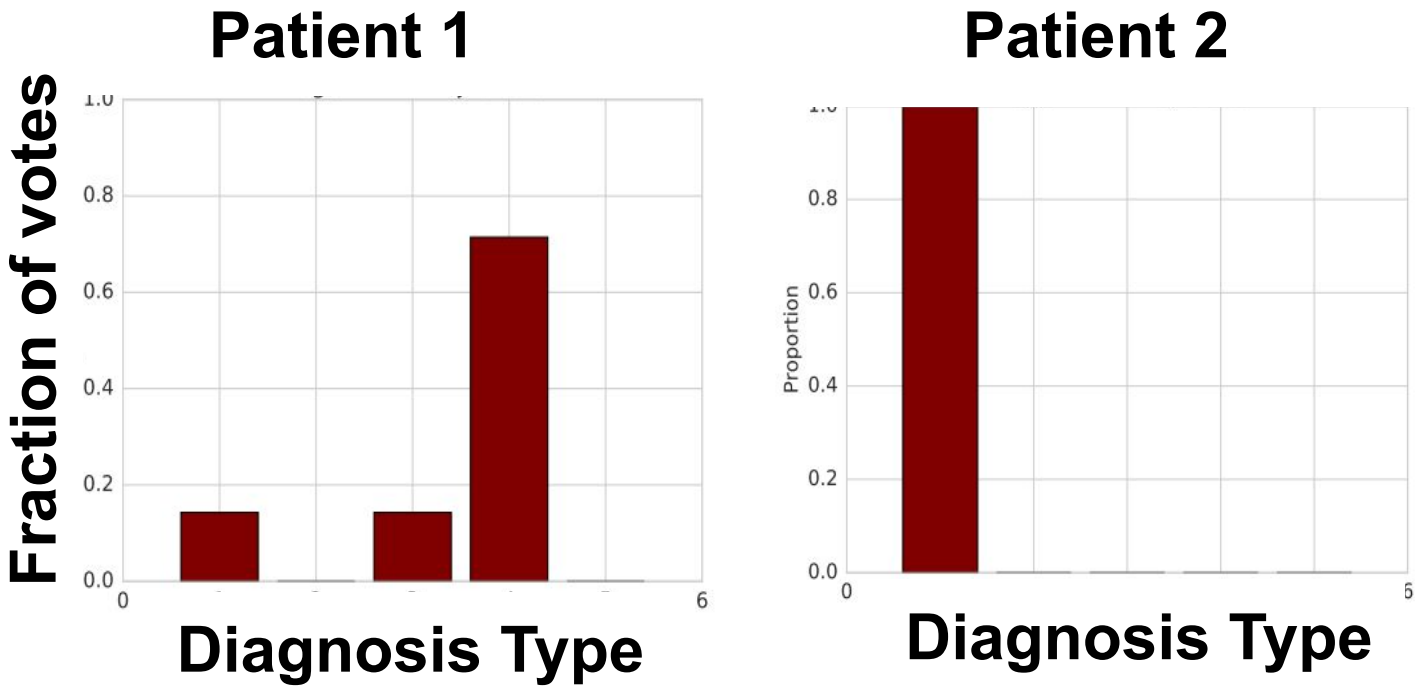


Grade 2: Mild Diabetic Retinopathy

Grade 3: Moderate Diabetic Retinopathy

Random Mistakes?

The Source of Disagreements



ML for Doctor Disagreement Prediction

Given input (image) x, predict the amount of disagreement. Flag patients for *medical second opinions.*

ML for Doctor Disagreement Prediction

Given input (image) x, predict the amount of disagreement. Flag patients for *medical second opinions.*

Training data: x_i, with multiple labels $y^{(i)}_1,...,y^{(i)}_k$ (different doctors) I.e. (x_i, $\mathbf{p}_i$), $\mathbf{p}_i$ grade distribution, target $U(\mathbf{p}_i)$ (e.g. U() = entropy)

ML for Doctor Disagreement Prediction

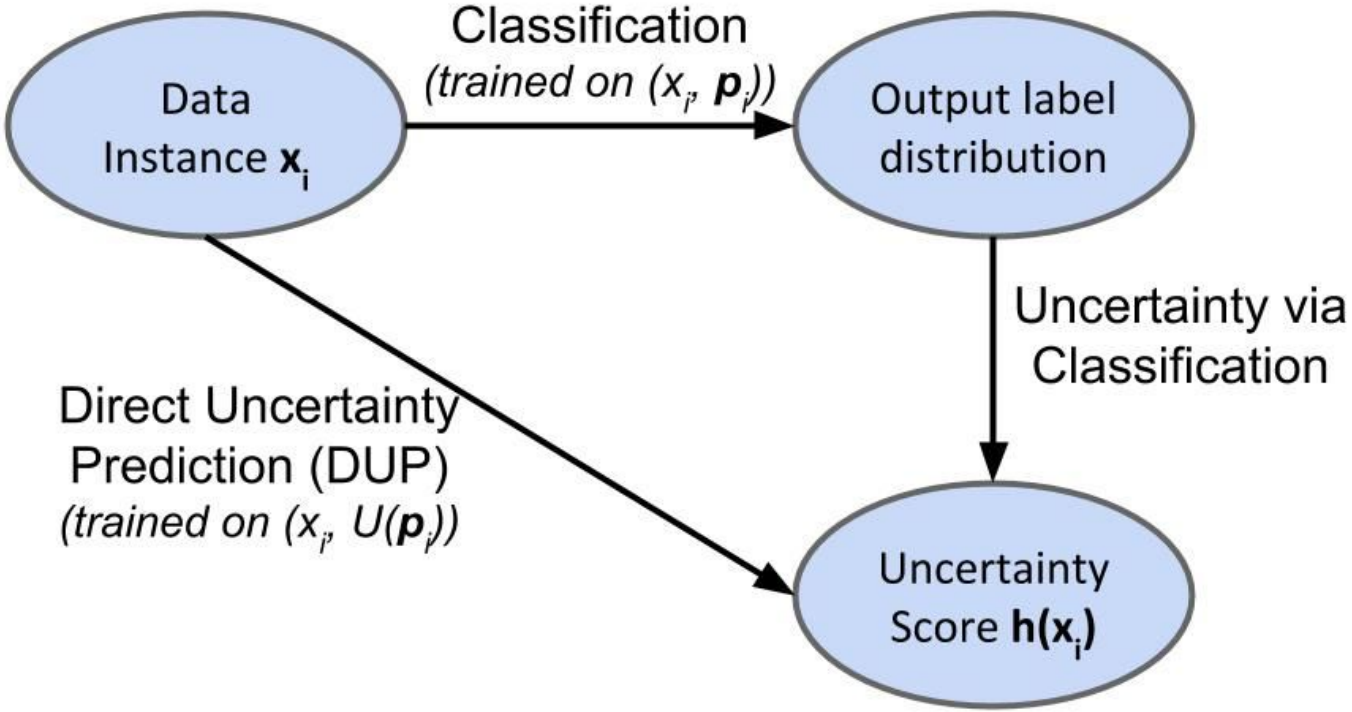
Given input (image) x, predict the amount of disagreement. Flag patients for *medical second opinions.*

Training data: x_i, with multiple labels $y^{(i)}_1,...,y^{(i)}_k$ (different doctors) I.e. (x_i, $\mathbf{p}_i$), $\mathbf{p}_i$ grade distribution, target $U(\mathbf{p}_i)$ (e.g. U() = entropy)

1) **Uncertainty Via Classification (UVC)**: (i) train *classifier* on empirical distribution of labels (x_i, $\mathbf{p}_i$) (ii) postprocess with U()

2) **Direct Uncertainty Prediction (DUP):** directly predict scalar uncertainty score (x_i, $U(\mathbf{p}_i)$)

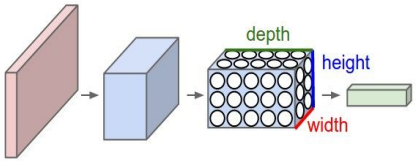
Direct Uncertainty Prediction



Direct Uncertainty Prediction

Hidden information:

61 (age) F (gender) medical history



Direct Uncertainty Prediction

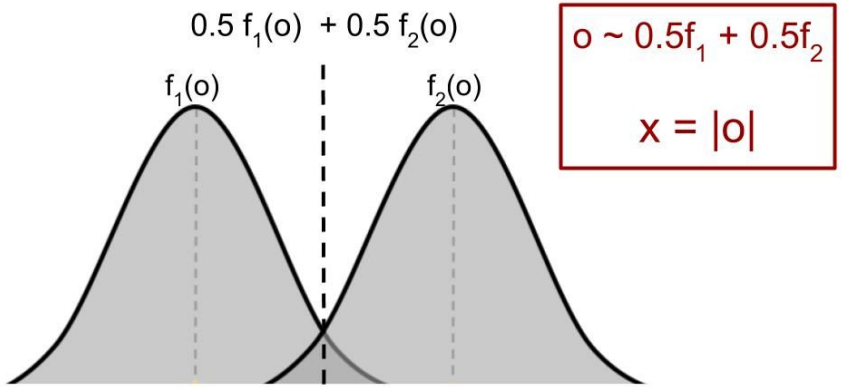
$$h_{dup}(x) = \int U(\mathbb{E}[\mathbf{Y}|O = o]) f_O(o|g(O) = x)$$

$$h_{uvc}(x) = U\left(\int \mathbb{E}[\mathbf{Y}|O = o] f_O(o|g(O) = x)\right)$$

Theorem: DUP gives an unbiased estimate of true uncertainty

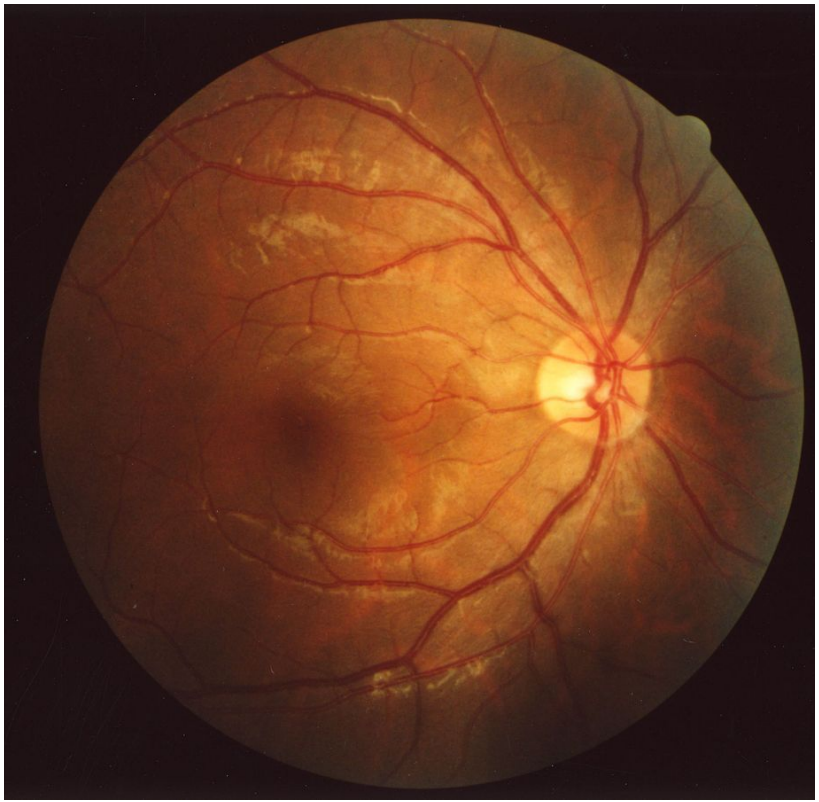
Empirical Results: Synthetic Examples

Mixture of Gaussians



SVHN and CIFAR-10: Image Blurring Application

Large Scale Medical Application



Diabetic Retinopathy (DR)
5 class scale:

1 None
2 Mild

Referable

3 Moderate
4 Severe
5 Proliferative

Large Scale Medical Application

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%

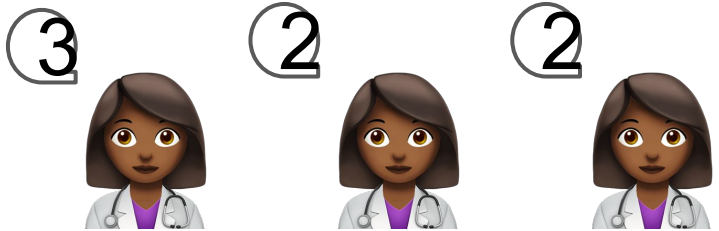
Poster #246: Direct Uncertainty Prediction for Medical Second Opinions

Large Scale Medical Application

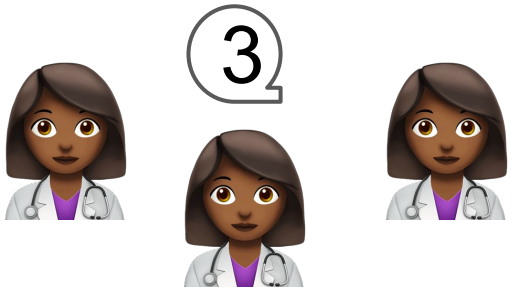
Poster #246

Small Gold Standard Evaluation Set

Individual Grades by Specialists



Single, Consensus, Adjudicated Grade



	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%**