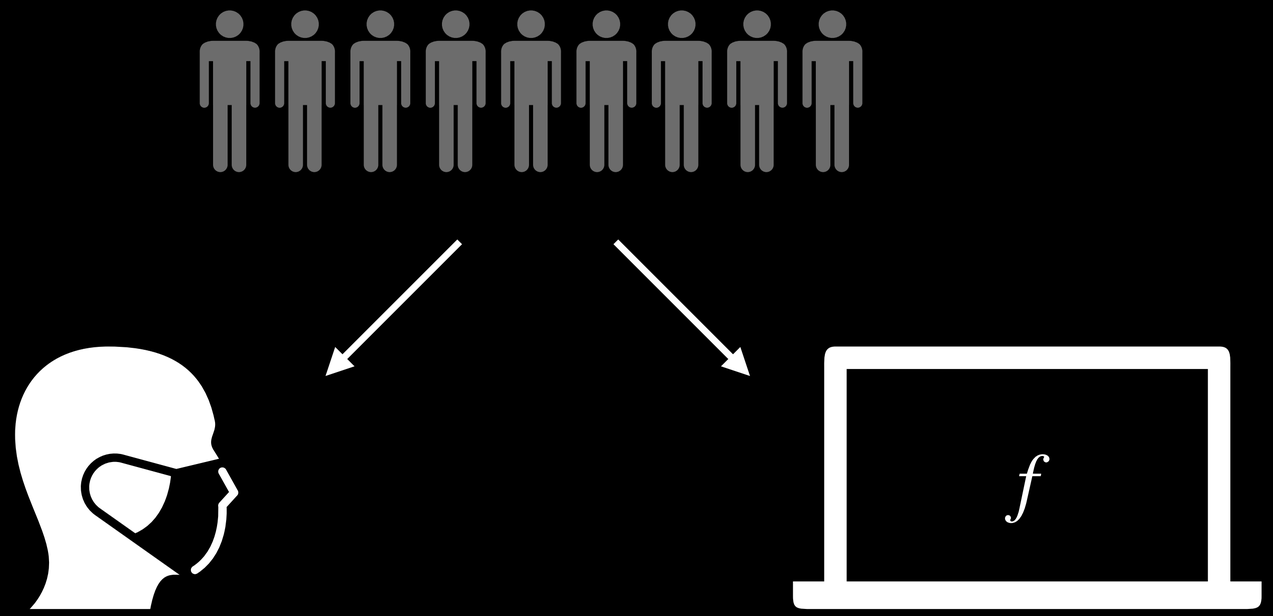
Updating Clinical Risk Stratification Models Using Rank-Based Compatibility

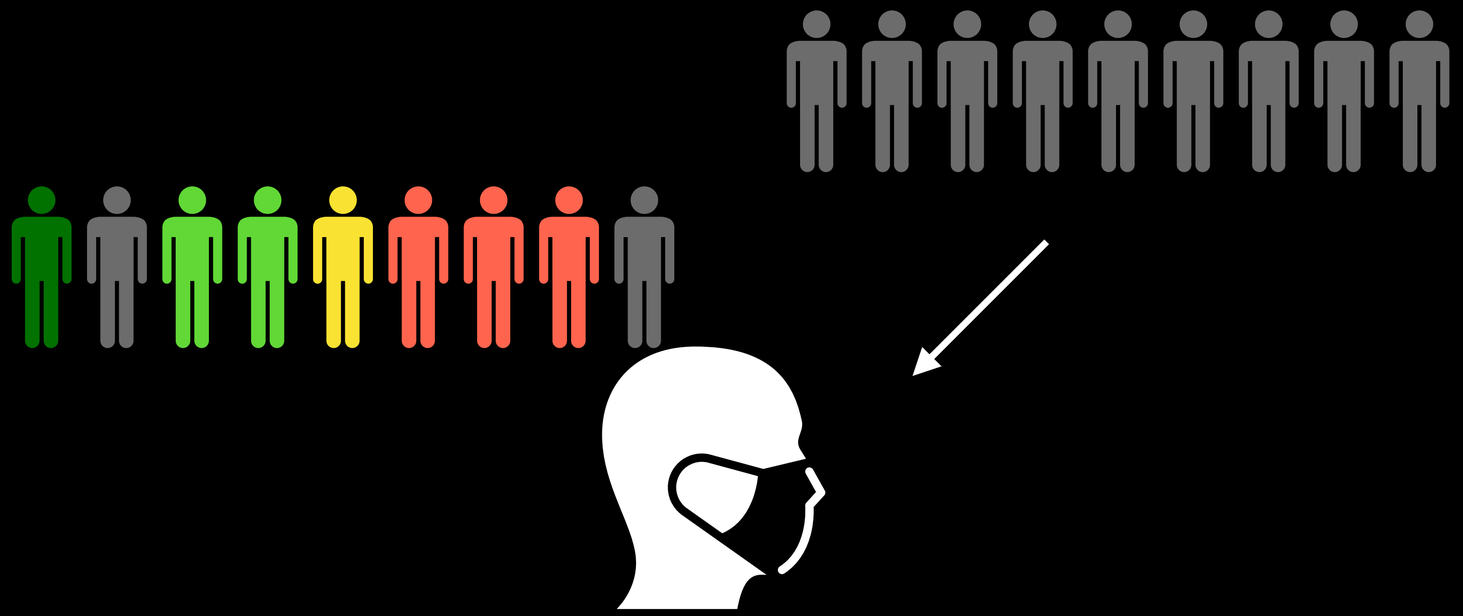
Evaluating & Optimizing Clinician-Model Team Performance

MLHC 2023
Erkin Ötleş, Brian T. Denton, Jenna Wiens
August 2023

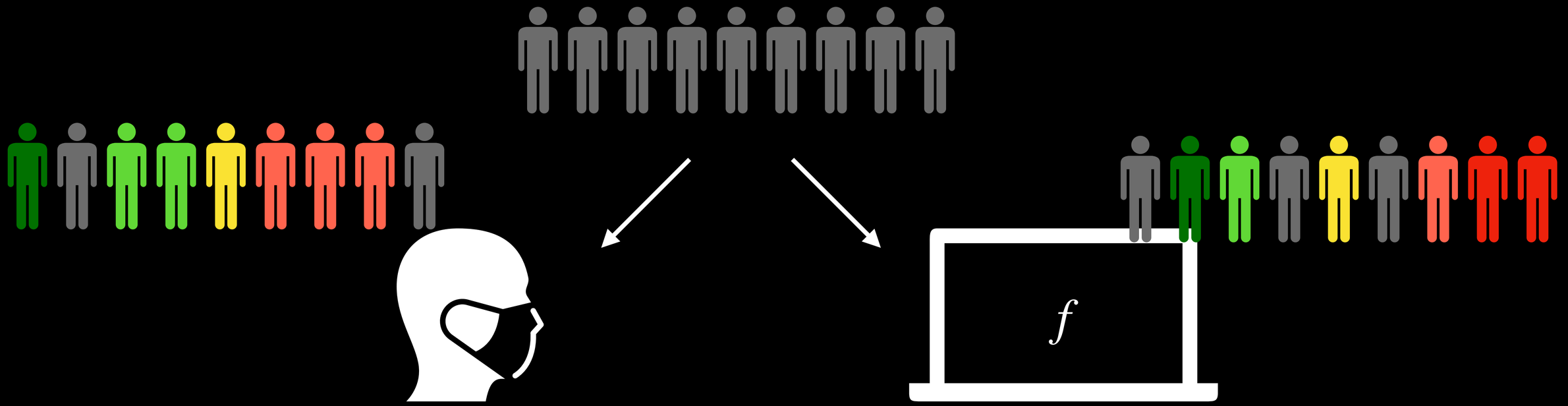
Physicians and models function as a team in healthcare settings.



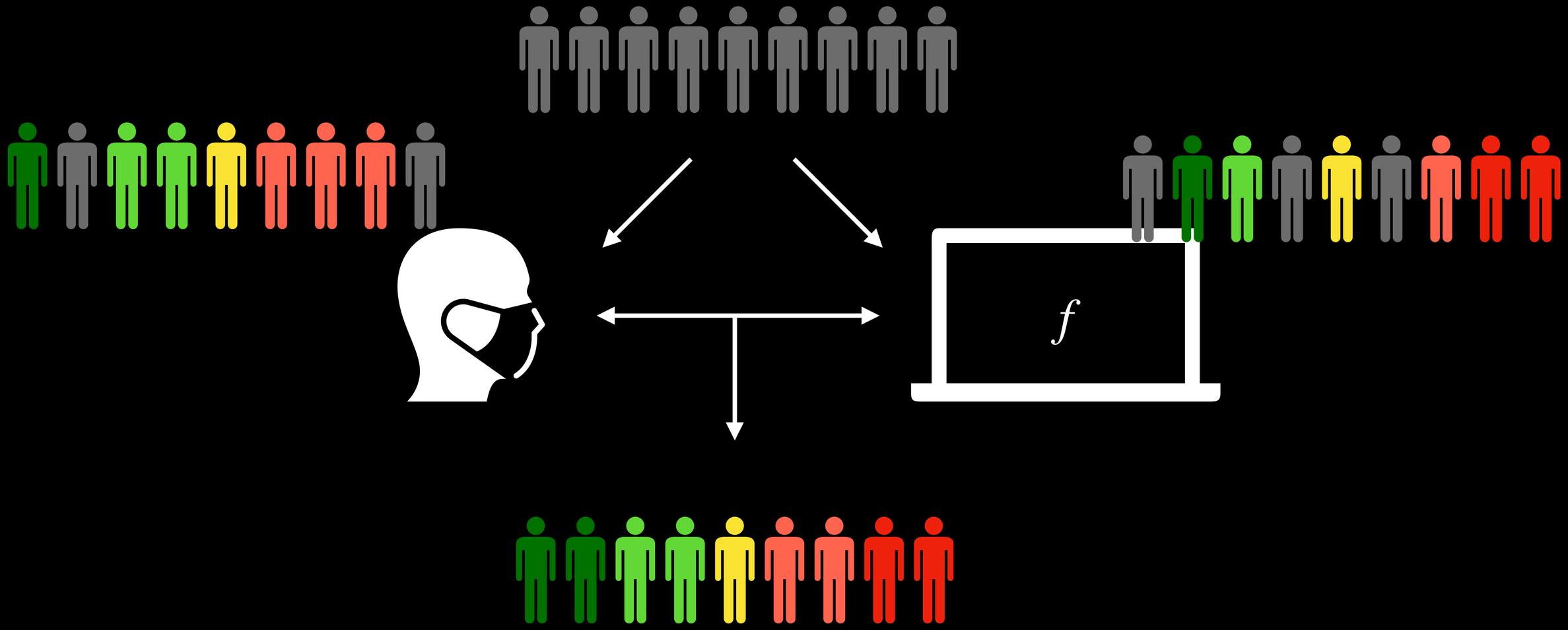
Physicians and models function as a team in healthcare settings.



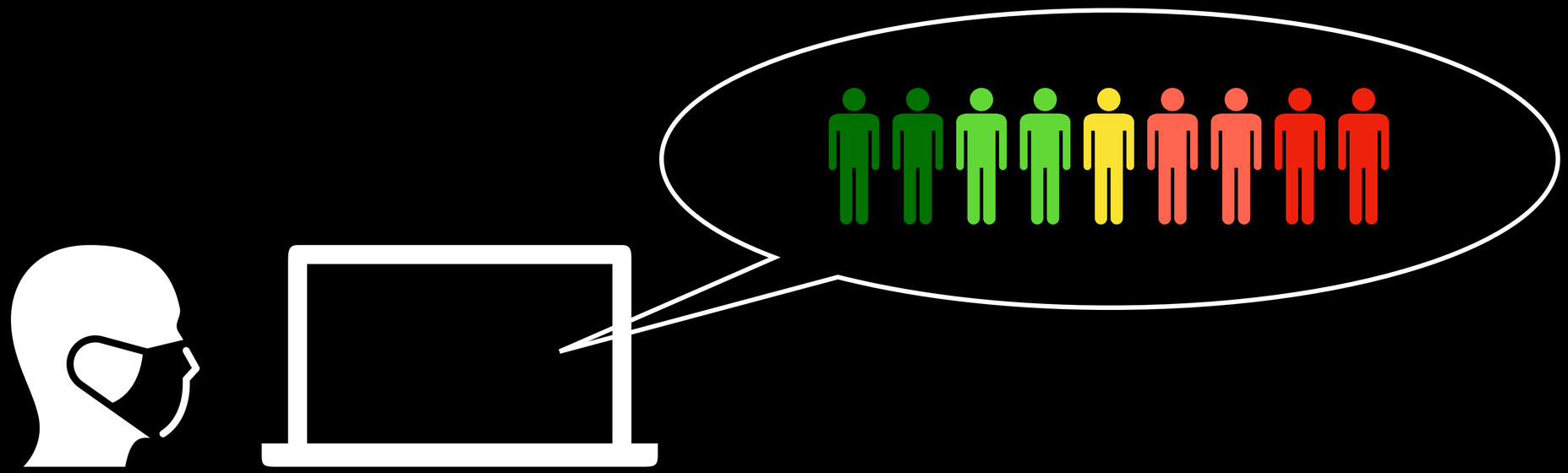
Physicians and models function as a team in healthcare settings.



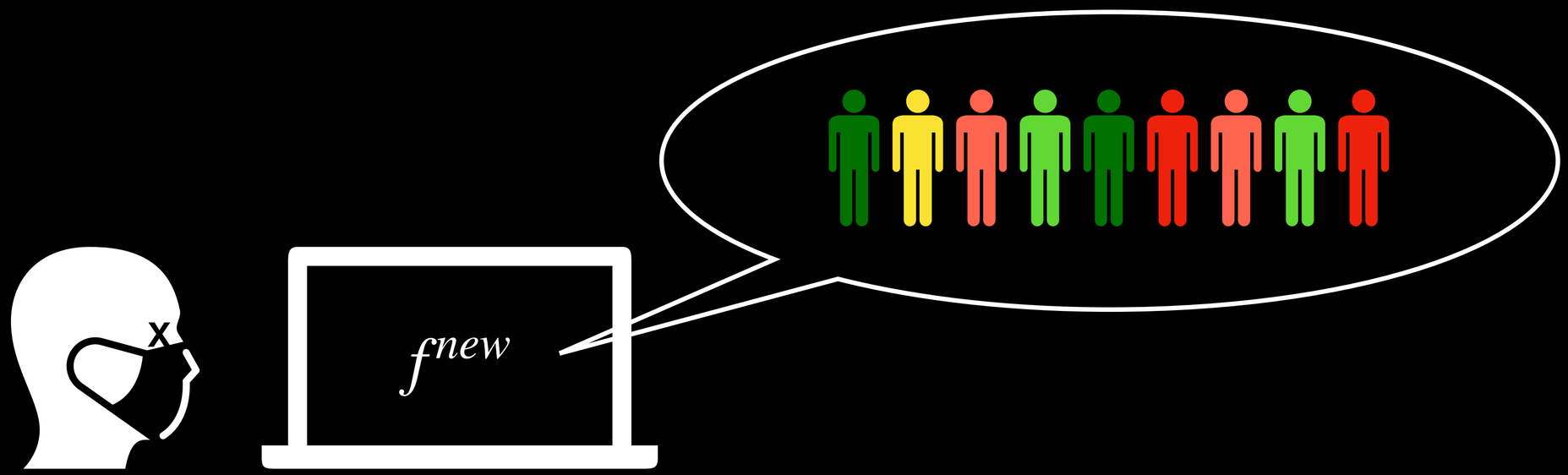
Physicians and models function as a team in healthcare settings.



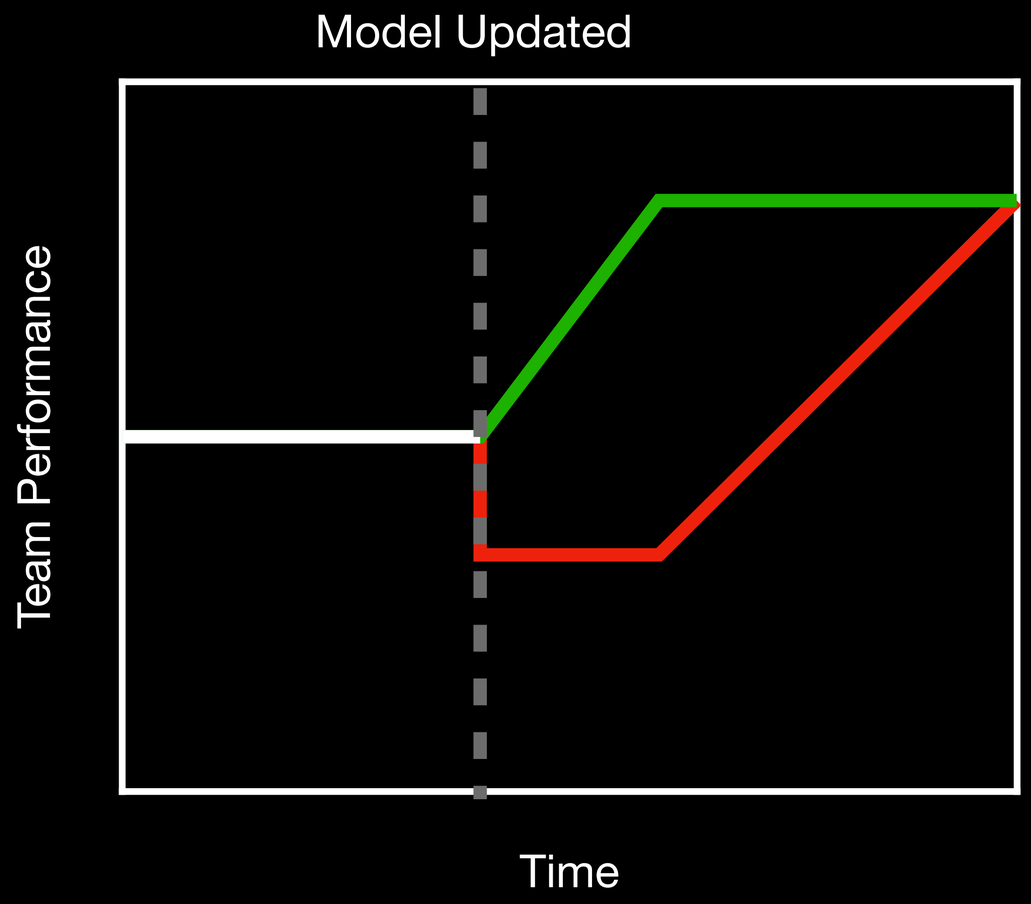
Updates can mess with user expectations.



Updates can mess with user expectations.



Team performance may suffer if models don’t meet user expectations.



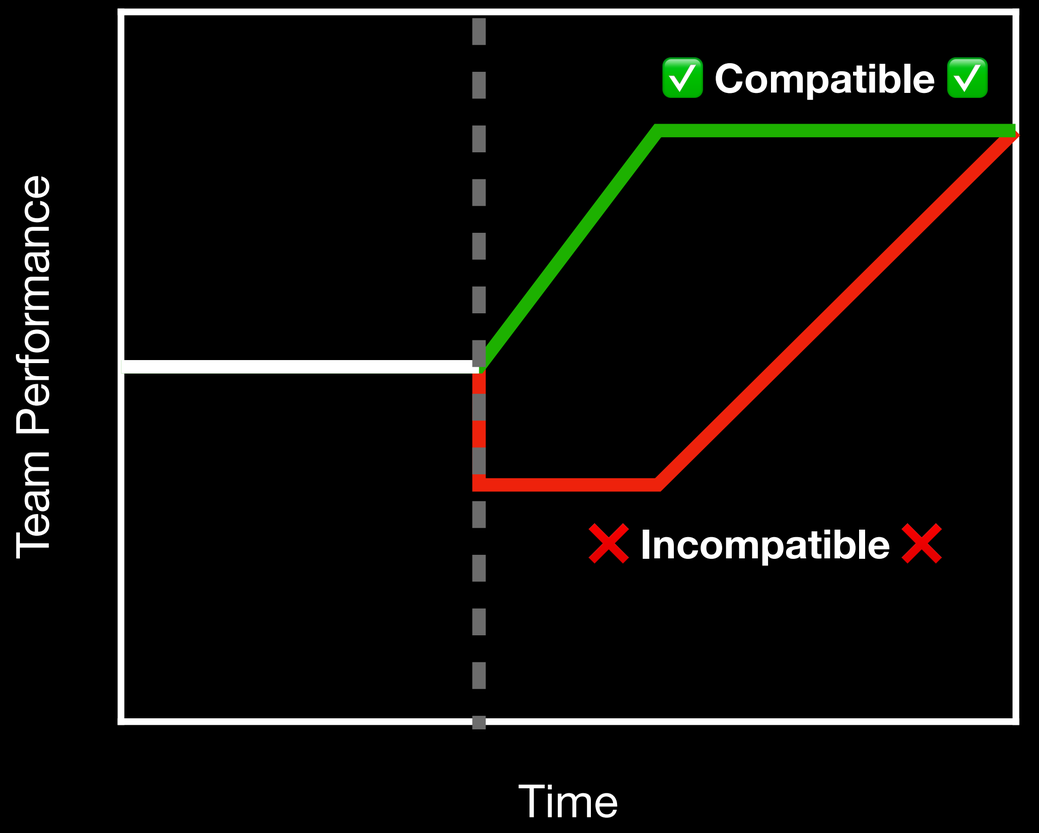
Adapted from Bansal 2019

Ideally updated models meet the expectations of users

Compatibility: the amount an updated model continues the correct behavior of an original model

Way to measure user expectations

Goal: updated models should have high compatibility



Bansal 2019

Problems with existing compatibility measures.

Existing measure depends on equality comparison

Problematic for use in risk stratification model & healthcare settings

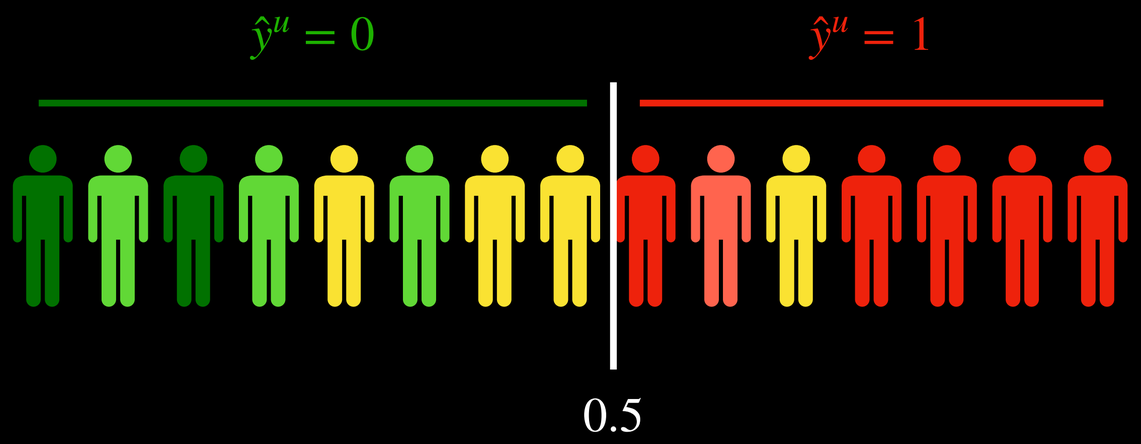
Depends on setting a single decision threshold

Problems with existing compatibility measures.

Existing measure depends on equality comparison

Problematic for use in risk stratification model & healthcare settings

Depends on setting a single decision threshold

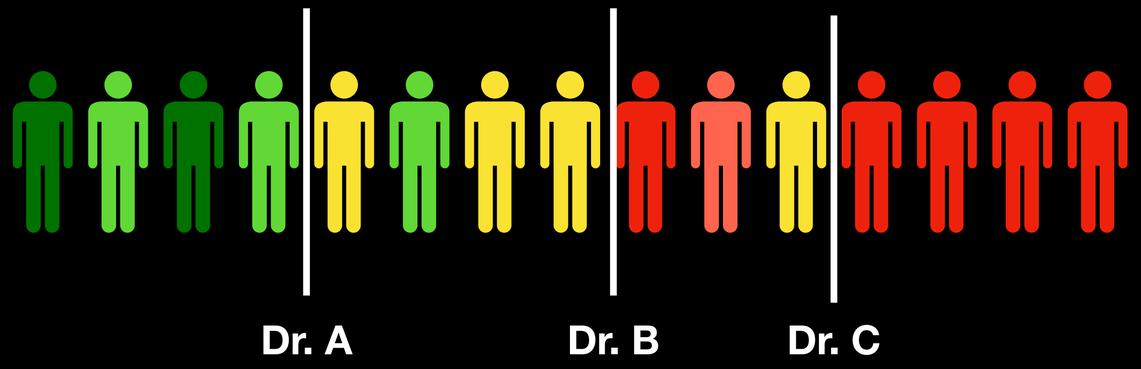


Problems with existing compatibility measures.

Existing measure depends on equality comparison

Problematic for use in risk stratification model & healthcare settings

Depends on setting a single decision threshold

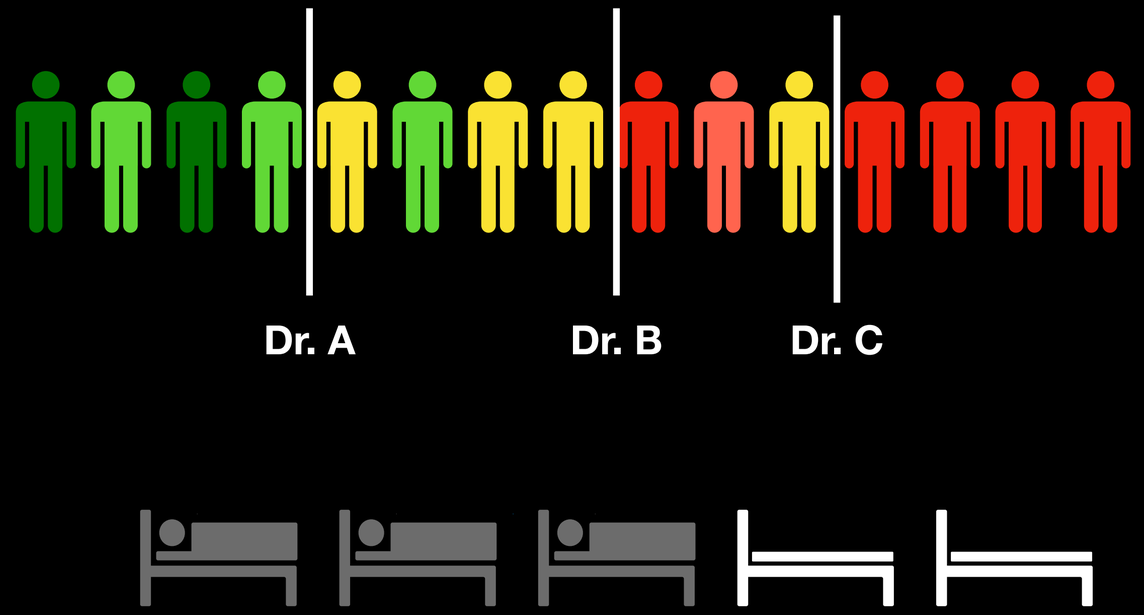


Problems with existing compatibility measures.

Existing measure depends on equality comparison

Problematic for use in risk stratification model & healthcare settings

Depends on setting a single decision threshold



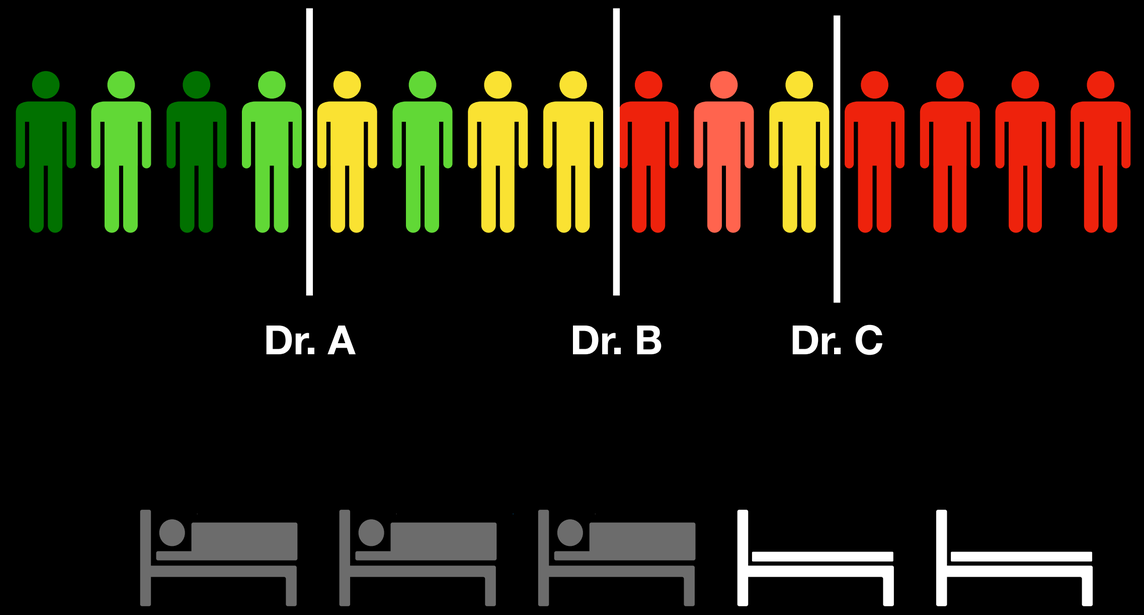
Problems with existing compatibility measures.

Existing measure depends on equality comparison

Problematic for use in risk stratification model & healthcare settings

- Depends on setting a single decision threshold

- No direct relationship with AUROC



Our contributions

Define a new rank-based compatibility measure (C^R)

Characterize C^R and its relationship with AUROC

Custom loss function to engineer model updates with improved C^R

Rank-based compatibility C^R

Agreement of risk estimate rankings produced by original & updated models given original ranked correctly:

$$C^R(f^o, f^u) = \frac{\sum_{i \in I^0} \sum_{j \in I^1} \mathbf{1}(\hat{p}_i^o < \hat{p}_j^o) \cdot \mathbf{1}(\hat{p}_i^u < \hat{p}_j^u)}{\sum_{i \in I^0} \sum_{j \in I^1} \mathbf{1}(\hat{p}_i^o < \hat{p}_j^o)}$$

Rank-based compatibility C^R

Agreement of risk estimate rankings produced by original & updated models given original ranked correctly:

$$C^R(f^o, f^u) = \frac{\sum_{i \in I^0} \sum_{j \in I^1} \mathbf{1}(\hat{p}_i^o < \hat{p}_j^o) \cdot \mathbf{1}(\hat{p}_i^u < \hat{p}_j^u)}{\sum_{i \in I^0} \sum_{j \in I^1} \mathbf{1}(\hat{p}_i^o < \hat{p}_j^o)}$$

original model

updated model

Rank-based compatibility C^R

Agreement of risk estimate rankings produced by original & updated models given original ranked correctly:

$$C^R(f^o, f^u) = \frac{\sum_{i \in I^0} \sum_{j \in I^1} \mathbf{1}(\hat{p}_i^o < \hat{p}_j^o) \cdot \mathbf{1}(\hat{p}_i^u < \hat{p}_j^u)}{\sum_{i \in I^0} \sum_{j \in I^1} \mathbf{1}(\hat{p}_i^o < \hat{p}_j^o)}$$

original model ranks correctly

Rank-based compatibility C^R

Agreement of risk estimate rankings produced by original & updated models given original ranked correctly:

$$C^R(f^o, f^u) = \frac{\sum_{i \in I^0} \sum_{j \in I^1} \mathbf{1}(\hat{p}_i^o < \hat{p}_j^o) \cdot \mathbf{1}(\hat{p}_i^u < \hat{p}_j^u)}{\sum_{i \in I^0} \sum_{j \in I^1} \mathbf{1}(\hat{p}_i^o < \hat{p}_j^o)}$$

updated model ranks correctly

original model ranks correctly

Rank-based compatibility C^R

Agreement of risk estimate rankings produced by original & updated models given original ranked correctly:

$$C^R(f^o, f^u) = \frac{\sum_{i \in I^0} \sum_{j \in I^1} \mathbf{1}(\hat{p}_i^o < \hat{p}_j^o) \cdot \mathbf{1}(\hat{p}_i^u < \hat{p}_j^u)}{\sum_{i \in I^0} \sum_{j \in I^1} \mathbf{1}(\hat{p}_i^o < \hat{p}_j^o)}$$

both models rank correctly

original model ranks correctly

Rank-based compatibility C^R

Agreement of risk estimate rankings produced by original & updated models given original ranked correctly:

$$C^R(f^o, f^u) = \frac{\sum_{i \in I^0} \sum_{j \in I^1} \mathbf{1}(\hat{p}_i^o < \hat{p}_j^o) \cdot \mathbf{1}(\hat{p}_i^u < \hat{p}_j^u)}{\sum_{i \in I^0} \sum_{j \in I^1} \mathbf{1}(\hat{p}_i^o < \hat{p}_j^o)}$$

1 → perfect compatibility, 0 → perfect incompatibility

We introduce rank-based incompatibility loss.

Rank-based incompatibility loss:

$$\mathscr{L}^R(f^o, f^u) = 1 - C^R(f^o, f^u)$$

Minimization of $\mathscr{L}^R$ will lead to higher levels of C^R.

Differentiable approximation $\widetilde{\mathscr{L}^R}$ for SGD.

C^R is a new compatibility measure inspired by AUROC

Not threshold dependent: ↑ clinical utility

Has a direct relationship with AUROC which we can balance against compatibility

In the paper you'll find…

Empirical results characterize C^R

Using $\widetilde{\mathscr{L}^R} \rightarrow \uparrow C^R$ & ↑AUROC

eotles@umich.edu
@eotles