Comments slides for Tuesday, October 27: Healthcare, part 1

Matthew J. Salganik

COS 597E/SOC 555 Limits to prediction
Fall 2020, Princeton University

Observations/comments/questions/provocations based on Dawes, Faust and Meehl (clinical versus actuarial judgement):

- Clinic judgement seems bad . . .

Observations/comments/questions/provocations based on Dawes, Faust and Meehl (clinical versus actuarial judgement):

- Clinic judgement seems bad . . .
- How could clinical judgement be better than actuarial judgement if they both have access to the same data?

Observations/comments/questions/provocations based on Dawes, Faust and Meehl (clinical versus actuarial judgement):

- Clinic judgement seems bad . . .
- How could clinical judgement be better than actuarial judgement if they both have access to the same data?
- The power of labels: “Actuarial” vs “statistical” vs “algorithmic”

Observations/comments/questions/provocations based on Dawes, Faust and Meehl (clinical versus actuarial judgement):

- Clinic judgement seems bad . . .
- How could clinical judgement be better than actuarial judgement if they both have access to the same data?
- The power of labels: "Actuarial" vs "statistical" vs "algorithmic"
- "broken leg" problem: when an expert can override even a good actuarial model? Are "broken legs" a source of limits to prediction?

Observations/comments/questions/provocations based on Dawes, Faust and Meehl (clinical versus actuarial judgement):

- Clinic judgement seems bad . . .
- How could clinical judgement be better than actuarial judgement if they both have access to the same data?
- The power of labels: "Actuarial" vs "statistical" vs "algorithmic"
- "broken leg" problem: when an expert can override even a good actuarial model? Are "broken legs" a source of limits to prediction?
- They seem aware of the limits of actuarial prediction in some cases, but this is not explored in details. When do we care about actuarial vs clinical and when we do we care about absolute level of accuracy of actuarial?

Observations/comments/questions/provocations based on Dawes, Faust and Meehl (clinical versus actuarial judgement):

- Clinic judgement seems bad . . .
- How could clinical judgement be better than actuarial judgement if they both have access to the same data?
- The power of labels: "Actuarial" vs "statistical" vs "algorithmic"
- "broken leg" problem: when an expert can override even a good actuarial model? Are "broken legs" a source of limits to prediction?
- They seem aware of the limits of actuarial prediction in some cases, but this is not explored in details. When do we care about actuarial vs clinical and when we do we care about absolute level of accuracy of actuarial?
- Nothing here about bias in-bias out or differential accuracy by subgroups. How does that compare for clinical vs actuarial methods?

Observations/comments/questions/provocations based on Dawes, Faust and Meehl (clinical versus actuarial judgement):

- Clinic judgement seems bad . . .
- How could clinical judgement be better than actuarial judgement if they both have access to the same data?
- The power of labels: "Actuarial" vs "statistical" vs "algorithmic"
- "broken leg" problem: when an expert can override even a good actuarial model? Are "broken legs" a source of limits to prediction?
- They seem aware of the limits of actuarial prediction in some cases, but this is not explored in details. When do we care about actuarial vs clinical and when we do we care about absolute level of accuracy of actuarial?
- Nothing here about bias in-bias out or differential accuracy by subgroups. How does that compare for clinical vs actuarial methods?
- What is better evidence of a limit failure of actuarial (with bounded information) or clinical (with unlimited information)?

Observations/comments/questions/provocations based on Caruana et al. (pneumonia and hospital readmission):

- If treatment intensity is a confounder, why not include it in the model?

Observations/comments/questions/provocations based on Caruana et al. (pneumonia and hospital readmission):

- If treatment intensity is a confounder, why not include it in the model?
- What does it mean to "repair" a model?

Observations/comments/questions/provocations based on Caruana et al. (pneumonia and hospital readmission):

- If treatment intensity is a confounder, why not include it in the model?
- What does it mean to "repair" a model?
- If these models were deployed would they become less accurate? Does that make them less useful? If not, is accuracy a good measure?

Observations/comments/questions/provocations based on Caruana et al. (pneumonia and hospital readmission):

- If treatment intensity is a confounder, why not include it in the model?
- What does it mean to "repair" a model?
- If these models were deployed would they become less accurate? Does that make them less useful? If not, is accuracy a good measure?
- They claim term importance follows a power law distribution: a few terms are very important and a lot of terms are a little important. Has anyone seen systematic evidence of this?

Observations/comments/questions/provocations based on Caruana et al. (pneumonia and hospital readmission):

- If treatment intensity is a confounder, why not include it in the model?
- What does it mean to "repair" a model?
- If these models were deployed would they become less accurate? Does that make them less useful? If not, is accuracy a good measure?
- They claim term importance follows a power law distribution: a few terms are very important and a lot of terms are a little important. Has anyone seen systematic evidence of this?
- Missing data seemed to be handled very roughly

Observations/comments/questions/provocations based on Caruana et al. (pneumonia and hospital readmission):

- If treatment intensity is a confounder, why not include it in the model?
- What does it mean to "repair" a model?
- If these models were deployed would they become less accurate? Does that make them less useful? If not, is accuracy a good measure?
- They claim term importance follows a power law distribution: a few terms are very important and a lot of terms are a little important. Has anyone seen systematic evidence of this?
- Missing data seemed to be handled very roughly
- I wonder how both of these claims have held up: 1) GA^{2}Ms give state-of-the-art predictive performance and 2) GA^{2}Ms are interpretable

Observations/comments/questions/provocations based on Gulshan et al (diabetic reinopathy):

- A breakdown of “predicting the present” vs “predicting the future”

Observations/comments/questions/provocations based on Gulshan et al (diabetic reinopathy):

- A breakdown of "predicting the present" vs "predicting the future"
- What is the value of "proof of concept" / "hype maximizing" studies? This is a very specially selected problem (Wong and Bressler).

Observations/comments/questions/provocations based on Gulshan et al (diabetic reinopathy):

- A breakdown of "predicting the present" vs "predicting the future"
- What is the value of "proof of concept" / "hype maximizing" studies? This is a very specially selected problem (Wong and Bressler).
- Is this clinical or actuarial? It seems like building a model of a clinician.

Observations/comments/questions/provocations based on Gulshan et al (diabetic reinopathy):

- A breakdown of "predicting the present" vs "predicting the future"
- What is the value of "proof of concept" / "hype maximizing" studies? This is a very specially selected problem (Wong and Bressler).
- Is this clinical or actuarial? It seems like building a model of a clinician.
- How much extra confidence do you have in the findings given that the training data comes from two settings (US and India) and some of the test data comes from a third setting (France)? How can this extra confidence be quantified?

Observations/comments/questions/provocations based on Gulshan et al (diabetic reinopathy):

- A breakdown of "predicting the present" vs "predicting the future"
- What is the value of "proof of concept" / "hype maximizing" studies? This is a very specially selected problem (Wong and Bressler).
- Is this clinical or actuarial? It seems like building a model of a clinician.
- How much extra confidence do you have in the findings given that the training data comes from two settings (US and India) and some of the test data comes from a third setting (France)? How can this extra confidence be quantified?
- What should we conclude that the ophthalmologists don't agree? Can we use the same averaging procedure in other settings (e.g., civil war)?

Observations/comments/questions/provocations based on Gulshan et al (diabetic reinopathy):

- A breakdown of "predicting the present" vs "predicting the future"
- What is the value of "proof of concept" / "hype maximizing" studies? This is a very specially selected problem (Wong and Bressler).
- Is this clinical or actuarial? It seems like building a model of a clinician.
- How much extra confidence do you have in the findings given that the training data comes from two settings (US and India) and some of the test data comes from a third setting (France)? How can this extra confidence be quantified?
- What should we conclude that the ophthalmologists don't agree? Can we use the same averaging procedure in other settings (e.g., civil war)?
- Reporting of data and results seems better than typical ML paper

Observations/comments/questions/provocations based on Gulshan et al (diabetic reinopathy):

- A breakdown of "predicting the present" vs "predicting the future"
- What is the value of "proof of concept" / "hype maximizing" studies? This is a very specially selected problem (Wong and Bressler).
- Is this clinical or actuarial? It seems like building a model of a clinician.
- How much extra confidence do you have in the findings given that the training data comes from two settings (US and India) and some of the test data comes from a third setting (France)? How can this extra confidence be quantified?
- What should we conclude that the ophthalmologists don't agree? Can we use the same averaging procedure in other settings (e.g., civil war)?
- Reporting of data and results seems better than typical ML paper
- Nice that sensitivity/specificity trade-off can be tuned for the setting

Observations/comments/questions/provocations based on Gulshan et al (diabetic reinopathy):

- A breakdown of "predicting the present" vs "predicting the future"
- What is the value of "proof of concept" / "hype maximizing" studies? This is a very specially selected problem (Wong and Bressler).
- Is this clinical or actuarial? It seems like building a model of a clinician.
- How much extra confidence do you have in the findings given that the training data comes from two settings (US and India) and some of the test data comes from a third setting (France)? How can this extra confidence be quantified?
- What should we conclude that the ophthalmologists don't agree? Can we use the same averaging procedure in other settings (e.g., civil war)?
- Reporting of data and results seems better than typical ML paper
- Nice that sensitivity/specificity trade-off can be tuned for the setting
- Surprising use of ImageNet

Observations/comments/questions/provocations based on Gulshan et al (diabetic reinopathy):

- A breakdown of "predicting the present" vs "predicting the future"
- What is the value of "proof of concept" / "hype maximizing" studies? This is a very specially selected problem (Wong and Bressler).
- Is this clinical or actuarial? It seems like building a model of a clinician.
- How much extra confidence do you have in the findings given that the training data comes from two settings (US and India) and some of the test data comes from a third setting (France)? How can this extra confidence be quantified?
- What should we conclude that the ophthalmologists don't agree? Can we use the same averaging procedure in other settings (e.g., civil war)?
- Reporting of data and results seems better than typical ML paper
- Nice that sensitivity/specificity trade-off can be tuned for the setting
- Surprising use of ImageNet
- Are we more willing to accept uninterpretable models when the input is images?

Comments slides for Tuesday, October 27: Healthcare, part 1

Matthew J. Salganik

COS 597E/SOC 555 Limits to prediction
Fall 2020, Princeton University