# Team 17: Complex Decisions, Real Numbers: Medical Decision-Making

Undergraduates: Samantha Garland, Grant Kim, and Preethi Seshadri
Project Leads: Dr. Ilona Fridman, Dr. Marc Ryser, and Dr. Peter Ubel
Project Manager: Dr. Jorge Silva

## Introduction:

When men are diagnosed with intermediate-stage prostate cancer, they can choose between treating their cancer or going into active surveillance. Active surveillance allows men to put off surgery and monitor the cancer's progression until if or when their cancer advances enough to impact their health.

**Active Surveillance:**

▷ **Pros:** no side effects from treatment
▷ **Cons:** cancer remains in body

**Treatment:**

▷ **Pros:** cancer is removed from the body
▷ **Cons:** side effects such as incontinence or erectile dysfunction

## Objectives:

We want to find what factors influence a patient's decision to choose active surveillance over treatment by using clinical appointments transcripts. Specifically,

▷ Does adding text (represented by topic distributions) as features to our baseline classification model improve the model's ability to predict final choice?
▷ Do these topic distributions perform as well as the manually-coded advice?
▷ Which topics are most predictive of patient choice?

## Data:

**Baseline variables**

Variables like age, cancer grade, initial treatment preference, and Control/DVD from before the clinical appointment

**Physician's advice**

Manually-coded by external coders based on the conversation transcripts

**Topic distributions**

From topic modeling on the clinical appointment transcripts

# Text Preprocessing



Sample sentence:
"Is getting surgery best for me?"

↓

Remove all punctuation, convert to lowercase:
"is getting surgery best for me"

↓

Tokenize and remove stop words:
["getting", "surgery", "best"]

↓

Lemmatize words:
["get", "surgery", "best"]

↓

Remove any remaining stop words:
["surgery", "best"]

# Topic Modeling

▷ We used Latent Dirichlet Allocation (**LDA**), a generative model that assumes **all documents are a mixture of topics**, with each topic then **represented as a distribution over words**.

▷ Our topic model uses twelve topics. Of these, we find **four identifiable topics** using the top words (on the right). The example below is an active surveillance topic. Other topics were surgery, radiation, and medical history.
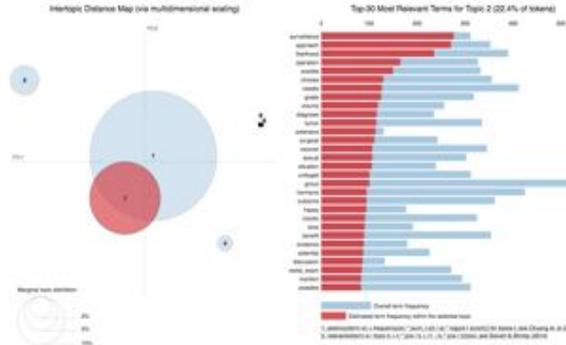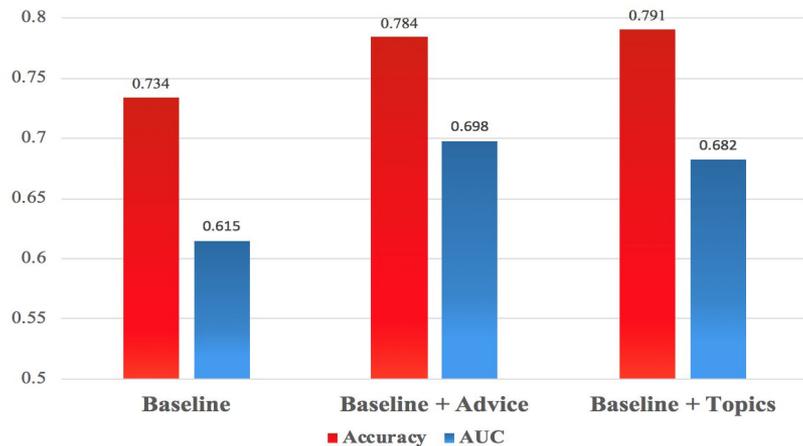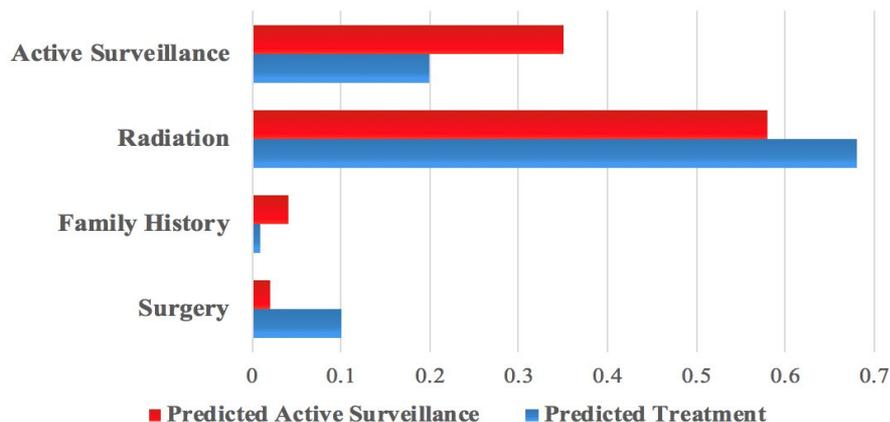


# Classification Models

▷ **Model 1 (Baseline):** Age + Cancer Grade + Study Condition + Initial Patient Preference

▷ **Model 2:** Baseline Variables + Physician's Recommendations for Active Surveillance, Surgery, and Radiation

▷ **Model 3:** Baseline Variables + Topic Distributions for Topics 1 to 12

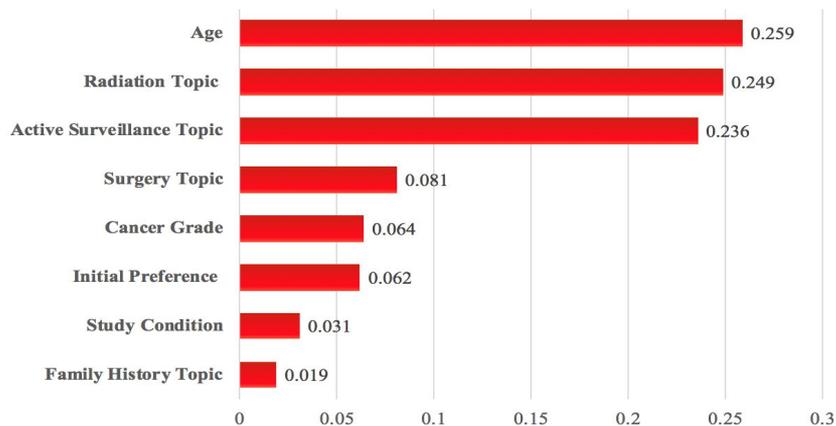| Topic | Key Words |
|---|---|
| Active surveillance | Likelihood, outcome, discussion, benefit, approach, surveillance |
| Surgery | Robotic, erectile, small_incision, lymph_node, pad, sew, recover |
| Radiation | External_beam, implant, therapy, rectal, needle, radiate, brachytherapy |
| Family history | Brother, family, bump, continence, difficulty, blood_pressure, speak, free |

## Model Performances

Baseline — Accuracy 0.734, AUC 0.615
Baseline + Advice — Accuracy 0.784, AUC 0.698
Baseline + Topics — Accuracy 0.791, AUC 0.682

## Average Topic Distributions by Predicted Choice

Legend: ■ Predicted Active Surveillance ■ Predicted Treatment

## Feature Importances

- Age — 0.259
- Radiation Topic — 0.249
- Active Surveillance Topic — 0.236
- Surgery Topic — 0.081
- Cancer Grade — 0.064
- Initial Preference — 0.062
- Study Condition — 0.031
- Family History Topic — 0.019

## Conclusions:

▷ Topic distributions substantially improved the baseline model.

▷ Baseline + topics performed comparably to the baseline + advice, meaning that topic distributions capture similar predictive performance to manually-coded advice.

▷ We can quantitatively identify how much physicians and patients talk about each topic.

▷ We gain insight into which topics in patient transcripts are most influential in patient decisions, which can aid future researchers in unpacking the medical decision-making process.