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?
Classification Models

Text Preprocessing

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.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Key Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active surveillance</td>
<td>Likelihood, outcome, discussion, benefit, approach, surveillance</td>
</tr>
<tr>
<td>Surgery</td>
<td>Robotic, erectile, small_incision, lymph_node, pad, sew, recover</td>
</tr>
<tr>
<td>Radiation</td>
<td>External_beam, implant, therapy, rectal, needle, radiate, brachytherapy</td>
</tr>
<tr>
<td>Family history</td>
<td>Brother, family, bump, continence, difficulty, blood_pressure, speak, free</td>
</tr>
</tbody>
</table>
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.