Supplement B. Generating symptom likelihoods from LDA models

Latent Dirichlet allocation (LDA) models do not estimate the probability that a symptom would be present in a patient in a given subtype. Instead, assuming a patient has a symptom, they estimate the probabilities that the patient has any particular symptom. To convert this into a proxy for symptom likelihoods, we consider the likelihood that a patient would have a particular symptom if they had a fixed number of symptoms. We assume that

1. For a patient in a given subtype, the number of symptoms n is equal to the average number of symptoms in the patient group that corresponds to the given subtype;
2. Symptoms are generated independently;
3. A patient has a particular symptom if it is generated at least once.

These assumptions make it possible to model symptom likelihoods by calculating binomial probabilities; that is, given the relative probability P_r given by an LDA model for a symptom, the symptom likelihood P_s is

$$P_s = 1 - (1 - P_r)^n$$

where n is the number of symptoms described above.