(1) Research the generic combinations of ingredients commonly found in oral contraceptives

(2) Collect data → look for sources already consolidated (PubMed, Emabase, LIVIVO) and then consider scraping the web. (WebMD or Drugs.com)

  (a) Ingredient profile

  (b) User ratings

  (c) User comments

  (d) Duration of Use

  (e) Monophasic, Biphasic, Triphasic

  (f) User age

(3) Issue Handling During Data Collection

  (a) *Multi-phasic Products.* What hormone concentrations should be entered in the table for multi-phasic products? This will have an effect on any regressions performed on the data → handled by using the dominant concentration (the concentration the persists the longest in the cycle)

  (b) *Generic vs. Brand-name.* Even the "generic" brands have names and it is difficult to distinguish between what's name brand and what's not, with the exception of pills produced by Teva (Junel, Apri, etc). Additionally, sometimes the data available for certain brands (i.e. amount of reviews) is limited. → handled by listing multiple brands of the same drug (active ingredients) and sampling reviews from the two brands with the most reviews. In the future, data may be collected for other brands if time allows. Another solution: include the product multiple times with its varying concentrations

(4) Tokenize each comment and determine which side effects are most prevalent in each combination birth control

(5) **Topic Modelling:** Determine which combinations produce the greatest volume of each given side effect

(6) In-depth analysis

  (a) Control for duration of use

  (b) **Regression Analysis:**

    (i) Examine the concentration of each ingredient in the pills that produce the most side effects

(ii) Perhaps use a logistic regression to classify the likelihood that a certain birth control will produce a side effect or not