**Prototype blueprint for DR paper**

**Methods**

* **Data & Ethics**
  + Public DR Datasets to have in mind:
    - **EyePACS**
      * 38GB Data size
      * 35K Data points
      * Features – level [0-4]
    - **MESSIDOR – 2**
      * 400MB Data size
      * 1700 Data points
      * Features – diagnosis [0-4], adjudicated\_dme [0-1]
    - **APTOS 2019**
      * 8.6GB Data size
      * 3660 Data points
      * Features – diagnosis [0-4]
    - **IDRID**
      * 186MB Data size
      * 450 Data points
      * Features – diagnosis [0-4], Risk of macule edema [0-2]
  + Local Datasets:
    - Potential Collab with local hospitals (couldn’t find any data specific to here)
* **Inclusion/Exclusion Criteria**
  + Ensuring the validity of the results
  + Ensuring the adequacy of the available features
  + Removing curropt image files
* **Ethical Considerations**
  + Removing any potential personal data from patients
  + Ensuring Dataset covers all races, genders, and age groups
  + Conversation in regards to using publicly available code as a jumping point or starting from ground up
  + Example [from: Cancer survival analysis and spatial distribution during 2014–2016 in Shandong Province, China]:
    - ***Ethical approval****: All methods in our study were carried out in accordance with relevant guidelines and regulations. The entire research protocol was approved by the Ethics Committee of Preventive Medicine in the Shandong Center for Disease Control and Prevention in 2013, with approval no. 2013020. All subjects and/or their legal guardians provided verbal informed consent.*
    - ***Data availability:*** *We acquired the data through official surveillance of the Shandong Death Registration System (SDRS). All data generated and analysed during the course of this study were available from the corresponding author upon request*.
* **Preprocessing and Quality Control**
  + In regards to publicly available data, majority of them are already preprocessed and don’t need further attention
  + For potentially locally fetched datasets, work has to be done to ensure no variance between the datapoints
    - Removing/Replacing NaN values
    - Making sure there is no class imbalance ( utilizing stratifiy)
    - (?) Do we even take the patients biology data into mind or just decide based on the DR level?