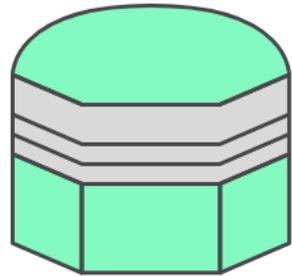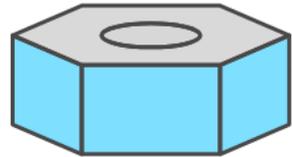# Key Steps for a Human Assisted Machine Learning Approach

## 1. Import 'Unlabelled' big qual data into Excel

2. Data preparation

a) Data cleaning - check responses, correct mistakes, remove empty responses & duplicates.

b) Create a coding key to explain medical terms/slang for non-medical team members.

c) Quantitative free-text analysis can be performed using AI methods via the Pulsar Platform. Conduct segmentation to identify key actors such as gender or stage in career. Conduct keyword analysis such as frequency analysis to identify the common topics of concern.
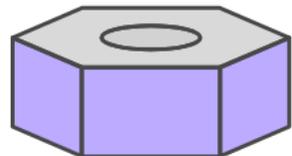
3. Data analysis – Sentiment & thematic analysis (Caplena)

a) Import cleaned data into the software Caplena to conduct the sentiment & thematic analysis - sentiment analysis focuses on identifying & categorising the emotions expressed in text, whilst thematic analysis identifies & explores recurring themes within a dataset.
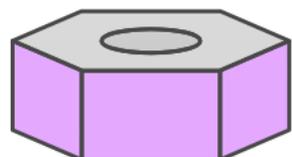
4. Validity checks of the sentiment & thematic analysis

a) Develop a sentiment analysis framework to provide clear guidance on what should be coded under each sentiment for the types of themes in the data.

b) At least two researchers to train the sentiment analysis tool by manually coding the data using the framework. Each response receives a percentage of positive, negative, neutral & ambiguous (total of 100% for each response).

c) Assess agreement between machine-generated sentiments & human annotators. Discussion & resolve until agreement between two researchers has reached over 82%.

d) Themes generated by Caplena should also be reviewed by at least two researchers. Train the model by reviewing at least a 10% subset of the responses. If necessary, make changes to the Caplena generated coding of themes, names of themes & subthemes to allow the model to learn from the changes.
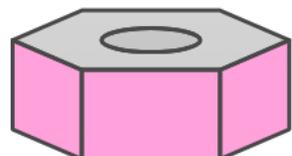
5. Topic & discourse analysis (InfraNodus)

a) Conduct a sub-analysis of the main themes identified by Caplena in text network analysis software InfraNodus. Import data into InfraNodus by theme from Caplena.

b) Highlight keyword clusters of interest & the most influential topics as well as linked betweenness centrality (this shows how often a node connects other nodes in a network, highlighting its role in enabling interactions) & topic diversity.
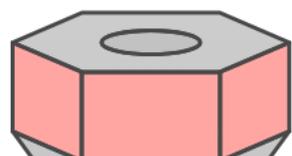
6. Validity checks of the topic & discourse analysis

a) At least two researchers to review topics identified by InfraNodus. Researchers use InfraNodus outputs for final revision of themes & subthemes from Caplena.

b) Output is the coded data & list of themes & subthemes identified through Caplena.

7. Large language model (LLM)

a) Export the machine-assisted coded data from Caplena into a GPT-powered LLM to train upon with the aim of developing a predictive model.