Situation

- Searching real-time clinical data for patterns or trends for testing clinical hypothesis is currently a manual process, requiring extensive manual intervention and prone to bias

- Early (semi or fully) automated detection of important patterns or trends in study data may help clinical teams across a broad variety of TAs to adapt study conduct and/or study oversight,

Solution Benefits

- Enable timely interventions during trial conduct - refinements to protocol document, investigator guideline updates

- Support to review large and diverse volumes of data and efficiently identify data patterns at scale

Our Approach

- A deep-learning, based hypothesis generation and prioritization module, which will leverage a range of unsupervised, deep-learning methods to identify hypothesis of potential interest, followed by synthesis and human validation
Data inputs

- Data Ingestion (Historical Trial Data + Protocol + External Data Sources)

Configuration and Data Engineering

Configuration will include specification of input/output variables, join specifications etc.

Hypothesis Generation and Prioritization

- Synthesis/Summarization (for Prioritized Hypothesis)

Synthesis and Feedback

- Human in the loop / expert review

Feedback to Hypothesis Generation and Prioritization Module

Validation by Clinical SMEs

Insights for Future Studies

Smart Clinical Signal Detector – Solution outline for proposed proof of concept (1/2)
Smart Clinical Signal Detector – Solution outline for proposed proof of concept (2/2)

Data Ingestion

Primary Data Inputs
- Treatment blinded clinical trial data – SDTM, ePRO, eCOA
- Protocol document
- Known safety concerns

Reference Data Inputs
- Historic column mappings and domain ontologies

Configuration & Data Engineering

Data Configuration (Human expert inputs)
- Primary, secondary end-points identification
- Column mappings using human expert inputs/ontologies

Data Engineering
- Generation of multiple databases, as needed, combining identifiers, patient demographics, eCOA responses, safety and efficacy indicators, at a patient-time or patient level as needed

Generation of Candidate Insights
- Leveraging established and exploratory algorithms, such as
  - Deep learning based phenotyping methods
  - Generalized mixed effects models
  - Evolutionary Search Algorithms
  - Recursive tree-based partitioning
  - Exploratory algorithms including Deep Reinforcement Learning, Deep Learning, Frequent pattern mining etc.¹ ²

Prioritization of candidate insights
- Lift, Statistical significance tests
- Sub-group effect size estimation and comparison vs. global threshold
- Reward functions defined on the basis of interestingness, diversity and coherence

Synthesis and Feedback

Synthesis/Summarization

Key Trends and Patterns such as subgroups with higher/lower efficacy rates

Human in the loop / expert review

Validation and Review

- Expert based qualification of signals to share back feedback

Case Study – an ensemble-based approach to identify historical trends or patterns in various studies and detecting safety and efficacy signals

**Situation**

- Client has near real-time clinical data like CRF, PRO etc. and currently rely on manual data review process which is time-consuming and laborious; wanted an AI/Machine Learning based technology platform with the ability to surface important and clinically meaningful trends or patterns during study conduct, such as
  - Cluster patients into treatment/placebo cohorts based on treatment response
  - Early prediction of treatment response
  - Identify safety signals

**Outcomes**

- Was able to identify patients with high/low response early in the trial (utilizing data available in the first snapshot), along with drivers; subsequent data improved the accuracy
- Uncovered clinically meaningful trends or patterns during study conduct, such as (but not limited to) -
  - High response in patients with a high dose of NSAIDs
  - High Probability of TEAE such as Arthralgia, Osteoarthritis, Back Pain, associated with ongoing medications such as ASA, Supplements, Chondroitin/Glucosamine, Levothyroxine, etc.
  - Correlation between liver adverse event, and treatment, possibly due to non-linear combination of concomitant medication and the treatment in question

**Key Learnings**

- Focus on explainability key as clinical SMEs are key consumers – use of effective visualizations key
- Ensemble of methods needed to capture diverse signals of interest across TAs
Approach #1 - Computational phenotyping for identifying insights in complex, multi-dimensional temporal data

**Vector Representation**

Numerical representations of individual patient/visit level data are automatically generated using embedding techniques to help render a digital portrait enabling unsupervised subtyping via graph modularity or other clustering methods.

**Pattern Detection**

Clusters in the representations are identified, which determine why certain groups of patients lie in a tight vector space distinct from others and further what signals are of importance within each of these clusters (potential trends/patterns).

**Post-profiling & Validation**

Identifying the optimal number of subtypes using unsupervised methods. Clusters in the representations are identified, which determine why certain groups of patients lie in a tight vector space distinct from others and further what signals are of importance within each of these clusters (potential trends/patterns).

Descriptive summaries to aid in hypothesis prioritization:

- Cohort characteristics to understand patterns and trends from blinded trial data
- CNNs to model temporal aspects coupled with autoencoders to enable unsupervised architectures (ConvAE)
- Tanh-LSTM based encoder-decoder architectures
- Stacked denoising autoencoders (DeepPatient) architecture
- Graph-based methods which leverage hidden structures in the data
### Approach # 2 - Proposed exploratory method for an Automated Insight Engine, utilizing optimized frequent pattern mining

<table>
<thead>
<tr>
<th>Master Database</th>
<th>Relationship Extractor</th>
<th>Knowledge Extractor</th>
<th>Insight Generator</th>
<th>Signals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Master dataset generated by combining identifiers, patient demographics, eCOA responses, safety and efficacy indicators, at a patient-time level</td>
<td>Enumerate possible data aggregations and corresponding operations, applicable for aggregated subgroups</td>
<td>Compute possible relationships from the data</td>
<td>Generate insights based on aggregated data and ranks them based on <strong>Domain Knowledge</strong> and <strong>Statistical Analysis</strong></td>
<td>Top-K insights for Clinical SME review</td>
</tr>
</tbody>
</table>

#### Configuration
- Configurable taxonomy for aggregations and operators
- Configurable statistical tests and domain related scoring functions

This proposed exploratory approach aims to extract top insights from any transactional dataset, using optimized frequent pattern mining methods – the approach requires configuration of data taxonomy, necessary operations and scoring functions to automatically traverse through potential insights and highlight top statistically significant insights.