saama

Product Advisory Board
Data Management & Submission Session Notes

Product Roadmap Notes:

- Prasanna: The "big whitespace" before clinical conduct which has not been captured
  - Q: "what is saama's plan for building tools in that space, or is Saama looking for partners"
  - A (Nekzad): Study Planning Insights covers feasibility, site selection, patient identification, study diversity, etc.
- Jeremy Cronin (Dicerna) asked about centralized data management for integrating clinical and commercial data (RWE, bio-bank data, master metadata set)
  - Nekzad: Commercial insights coming to the fore (post approval) with use cases: physician targeting, KOL identification, contracting, sales messaging, claims datasets, GPO contracting, doctor site contracting, etc.
  - Jeremy also asked questions about unstructured data source ingestion
- Sam Tomioka - asked nekzad to elaborate on clinical neural networks
  - Dr. DaLIA network
- Prasana: when does diversity in clinical trials come in?
  - Nekzad: built into study planning insights, part of a Q1 2021 launch

Data Management & Submissions Breakout Room Notes

- Joe Fitzgerald (Regeneron) asked about the enhanced data quality engine rules.

  Srini explained there will be a four-level library, including global
  DQ rules can be auto-selected, used from global library or updated
  Global library and ML-generated rules don't exist yet.
- Suman from IQVIA asked if the data sources were EDC. Answer was yes.
- Is there an intent to get adapters to get connected device information and eCOA (?) Capability is there. For now we aren't actively building it out, but we will after we finish specific engagements.
- Site and subject mastering are being explored. We have a lite version already. How much should we invest in this?

- Prasanna: Are roadmap items specifically targeted for certain apps from an AI perspective?
    - The whole SDQ extension is focused on Machine Learning, as per DQ rules. Medical monitoring is a strong ML use case as well. All the core features of SDQ will leverage Machine Learning.
- Suman: I don't see automapping as part of the patient pipeline.
    - Srini: It's implicit. Our biggest differentiator is automapping using global libraries in the patient pipeline and to some extent in the operations pipeline.
- Prasanna: When you talk about ML/NLP, you need training data. Who besides Pfizer is providing this? Multiple sponsors.
    - For SDQ we are starting with Gilead. For clinical data hub we're leveraging IQVIA data. Astex is another. Each use case has individual sponsors.
    - Tim Riley clarified: Saama is not using IQVIA data to train the algorithms
    - Srini: Models trained today won't be deployed elsewhere because they are trained on sponsors' data.
    - Prasanna: When I hear model, I think training data (ours) and algorithm (yours)
    - Malai: Data never goes out; our algorithm comes in and we fine tune it based on specific sponsor data.

- SDQ today we have standard in standard production. Logical way to extend is to focus on third-party data sets. What are your recommendations?
    - Question: Where does the dM team spend time in manual review?
    - Question: Where should we invest our time and effort?
    - Ashley from Pfizer: biggest one is to **incorporate more external data models;** patterns between external and EDC. Amount of external data is increasing exponentially. Big scope applying SDQ there, looking for patterns not just rules-based.
    - Suman: Accelerate cleansing between labs? And EDC.
- Smart Coder. Not in the roadmap. We've had multiple sessions sponsors aren't happy with the accuracy of the coding tools. Drug related coding is 30-40% accuracy. Workflow doesn't work. eCoding.
- How much value would this give you?

    - Prasanna: **There's a cost-saving opportunity.** EDC vendors typically have rules based coding features, but there's always a manual curator to overcome semantic problems. Accuracy can't be lower than 90%. **Key strategy is to introduce ML-based coding. Stepping stone for future data cleansing using similar technologies. Short-term operational goal; long-term strategic goal.**
    - Sam Tomioka: **There's a need for coding outside EDC.** Safety coding in pharmacovigilance. Manual coding doesn't work with massive amounts of data.
- It's challenging for us to deal with conmeds.
- License a model using different sponsors' data.
- If you're willing to share your model, we'd be willing to share our data as well.
    - Malai: We actually were able to get to a decent level of accuracy on conmeds with one sponsor. Good UI, coupled with human-in-the-loop. We would like to take you up on your offer.
- We know that medical monitoring is a different ballgame from data management, but we think SDQ can be leveraged to identify safety signals. Trying to find a needle in the haystack. Is it even a good idea to do this?
    - Prasanna: **My opinion is that's a quantum leap. Missing steps in the middle.** A lot of sponsor use cases need to be solved and stabilized. We're in the early stages of how this tech can transform data management. Data discrepancies are looked at from different lenses. Clinicians are looking beyond accuracy. I feel that clinicians are an important stakeholder, and so many other internal teams that data management can span across. We want to explore sponsor-related use cases before taking this quantum leap.

Snapshot Management:

- Sam Tomioka: **Need both: date/time and events.**
- Olek Czepla: **Snapshot management is important for oncology.** Once the number of events are reached, we assign the date. All assessments on or before that date will be included. Another example would be Adverse Events. If an AE occurred before the date, we need the datapoint for analysis. If it happens after, it won't matter as much.

Fine Grained Access: Now you can define your own rule where a specific user can access specific rows or columns. Should we do the scrambling?

- Joseph Fitzgerald: It could be valuable. Now it's done in a more programmatic way. How do you get it to the point where it's really scrambled? As long as those kinds of rules can be available, it would be a value-added. **It looks appealing but there's always a downstream concern about data protection.** It's a bit of a threshold that would need to be crossed with stakeholders.
- Srini: put it as a backlog item. Biostatistics teams want to look at all the data.

MDM: Do we need it? Build or buy? Is it required in patient pipeline?

- Srini example: Actual enrollment data can come from CTMS, EDC or IVRS. Which is the golden standard of truth? The same type of thing doesn't happen in patient pipeline.
- Prasanna: I have not seen a use case for MDM. EDC is our single source of truth. I don't see a need of creating a golden record from different systems. I'm trying to understand your strategy.
- Basant: Lab data comes from lab and from EDC. If there's a difference, what do you do?
    - Prasanna: Always trust the lab if the data differs in the EDC.
- Tim: **Very light MDM solution** could solve problems around site numbers. Minor things like country codes create manual activity. Anything bigger might be overkill.

ADaM:

- Prasanna: **Flexibility is good. Why just Python?** Why not R? Multiple programming languages is good. But what is the compelling reason to rip out Sas? Why would we change to a different language?
- Sam Tomioka agrees. Sas is necessary; it has things Python doesn't. If R is used for statistical analysis, it might be good for ADaM. Python isn't used for statistical analysis.

SDQ: incorporate more external data.

- Some questions about the model training: offer the algorithm in return for data
- FGA in the patient pipeline had concerns about data protection.