# AI-Powered Health

## Balancing Algorithmic Accuracy with Human-Centricity in Healthcare

**Sridhar Turaga**
Head of Data & Analytics Practices, CitiusTech

**Yogesh Parte, PhD**
Sr. Practice Data Scientist, CitiusTech

**Shitang Patel**
Payer Consulting Leader, CitiusTech

**Yunguo Yu, MD, PhD**
Sr. Practice Data Scientist, CitiusTech

## Insights

- Acceptance of AI has crossed a threshold across healthcare. Today, almost every organization, be it process-based, function-based, or product-based, is looking to natively adopt AI for better results.

- However, a common challenge faced by AI sponsors and business executives is determining the success criteria of AI initiatives.

- Intuitively, everyone focuses on accuracy and error reduction – as if higher accuracy is all that matters. But quality, success, and value in healthcare are nuanced and multi-dimensional.

- Relying on statistical accuracy alone doesn't account for the multiple (often conflicting) objectives we deal with in this industry.

- A multi-dimensional framework that CitiusTech teams have used over the years has been a valuable tool in helping clients frame the right measure of success and drive impact.

## Abstract

Healthcare leaders are increasingly experiencing the transformative power of Analytics & Artificial Intelligence (AI) across the care continuum and healthcare operations. So much so that AI is becoming core to almost every digital transformation initiative, adding intelligence to every product and will soon transform every process in healthcare. For example, customer experience teams of payer organizations are implementing intelligent solutions, such as provider search, chatbots, and plan recommendations, to enhance member portals. Clinical teams are increasingly utilizing AI-based diagnostics to deliver better patient care. Product and application engineering managers are modernizing workflows by incorporating AI to work smartly. Examples range from automated claims adjudication and task prioritization to smarter targeting of members for care gap closure. Population health management teams are relying on AI and machine learning (ML) algorithms to generate patient-level risk and impact-ability scores. Additionally, MedTech and medical device product managers are commonly adding AI algorithms for improved diagnostics.

AI and analytics are now ubiquitous in all healthcare processes and products.

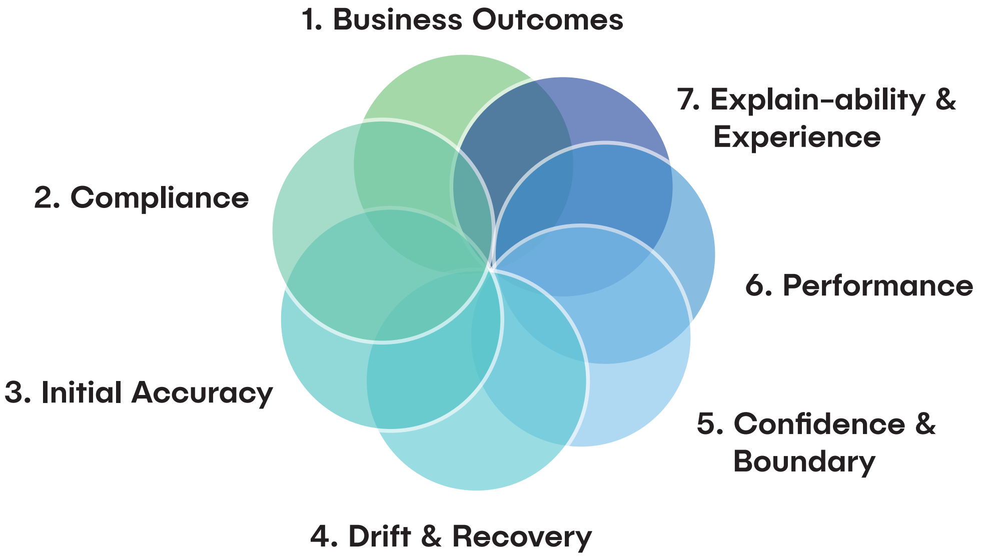
However, AI sponsors and users face a common challenge in determining the success criteria for their projects/ initiatives. Merely relying on statistical accuracy as the sole metric is insufficient, as success and outcomes in healthcare are complex and multi-dimensional. The narrow pursuit of accuracy will prove to be too simplistic where decisions often need to be made balancing other equally important aspects such as patient safety, avoiding biases, and delivering sustainable value. Unlike other forms of automation (through IT), AI-driven solutions are not evolving and changing constantly, as they learn from new data/ decisions. One needs a more continuous monitoring mindset, than a once-and-done approach.

While analytics and data science teams are majorly involved in defining the measure of success, healthcare organizations must develop a language beyond accuracy and adopt a framework tailored to navigate healthcare's unique challenges.

**The Healthcare-Measure of Success Framework (framework-1)** has been used in many client engagements at CitiusTech. It has not only resulted in more sustainable ROI (Return on Investments) but also acted as an enabler for improved collaboration across stakeholders with varied capabilities, language, and priorities. The framework tries to walk a tightrope between being simple enough that it is easy enough to be used but not too simplistic that it disrespects the complexity of the situations that innovators face.

## Healthcare-Measure of Success Framework (H-MSF)

The H-MSF helps healthcare innovators and collaborators define measures of success for their AI/Analytics solutions they develop across seven crucial dimensions. Balancing these dimensions requires careful consideration for each one and necessitates trade-offs between them.



*Framework-1: A multi-dimensional framework for measuring success in AI solutions*

**1. Business Outcomes**

Does the solution help achieve broader organizational goals, across stakeholders and as a whole?

**2. Compliance**

Will the solution keep people safe?

**3. Initial Accuracy**

How accurate are the results at the point and time of use of the solution? How accurate and reliable are the results from a clinical/functional lens? What is the statistical/mathematical basis for using the accuracy measure?

**4. Drift & Recovery**

How does the accuracy of the solution vary over time? Does it drift with new information and data? Does it auto-recover to its expected and defined accuracy levels?

**5. Confidence & Boundary**

How 'confident' are we about the predictions? What is the error range/'tolerance'? Can we measure it and specify the 'confidence interval' upfront? Is there a well-defined boundary or expected scenarios beyond which the solution will say 'I don't know' instead of making a prediction?

**6. Performance**

Does the AI solution deliver its recommendations and predictions without deteriorating the user experience or acceptable performance of the broader system/process it is embedded into?

**7. Explain-ability & Experience**

Does the end user/consumer feel satisfied with the result? What kind of experience do they get from the solution? Do they understand the recommendations and can make decisions using them? How explainable is the final solution and its recommendations?

Now, how can we apply this framework when multiple measures are taken under each dimension? The following use cases show how this framework is applied in real-life situations. (Exact details have been masked and abstracted for client confidentiality).

## Applied Scenario #1: Claims Renegotiation

A key challenge in the healthcare claim negotiation process is optimizing negotiators' time and prioritizing claims for maximum financial impact. To address this, an ML tool is developed that enables claims negotiators to efficiently identify the most impactful claims, resulting in increased customer savings and enhanced negotiator throughput. The tool uses years of quality data to fuel its predictive ML models, delivering immediate results and improving productivity while considering claim financials, clinical aspects, negotiation history with providers, etc. It then automatically assigns a priority score to each claim, streamlining the claim queue.

### Business Outcomes:

- Improve healthcare negotiation efficiency by prioritizing the most impactful claims
- Reduce claim turn-around-time by reducing clutter (deprioritize low likelihood of renegotiation claims)
- Improve over bottom line by putting the negotiators' time to best use on successful and higher potential saving claims

Hence, the outcome of the use case reduces human labor and increases negotiation efficiency, vastly shortens the turn-around time of claim processing, significantly increases the negotiation success rate, and improves the financial outcomes.

### Compliance:

- Adherence to the best practice of responsible AI/ML
- HIPAA and relevant compliance for healthcare
- Adherence to Healthcare IT security and process standards

### Initial Accuracy:

- Accuracy
- Precision and Recall
- The mean standard error for an estimated saving

Since this is a worklist prioritization, the overall accuracy and recall are foundational to the success of the use case.

### Drift & Recovery:

- Concept drifting due to policy changes, for instance, COVID claims
- Data drifting due to the new payer data that are different from the data used for modeling

The machine learning models can be re-trained if the data-drifting is discovered. For example, COVID claims models needed to be regularly retrained to retain high accuracy due to federal policy changes on COVID.

### Confidence & Boundary:

The process is in place that establishes the confidence ranges, which define the model quality – at a cohort level of member and claim type.

### Performance:

- Saving per negotiator
- The average number of claims processed by a negotiator

- Cycle time (Incoming to case closure in days)
- Real-time API response rates, as any deterioration will result in degradation of user experience and hurt usage

### Explain-ability & Experience:

- The claim characteristics that led to estimation change in renegotiation and projected savings

## Applied Scenario #2: Coding from Clinical Notes

Extraction and classification of clinical entities, their relationship from clinical notes, reports and mapping them to desired ontologies such as ICD, Current Procedural Terminology (CPT), Snomed, LoINC, RxNorm codes, etc., is an important task in healthcare information extraction. Natural Language Processing (NLP) techniques can be used to automatically extract and classify these codes from unstructured clinical notes or reports.

The process involves first identifying relevant information from the clinical notes using techniques such as part-of-speech tagging, named entity recognition, dependency parsing, syntactic parsing and relationship extraction, and negation detection, to name a few. The identified information is then used to match the corresponding codes in a database using supervised or unsupervised techniques and labeled data based on the context and semantic meaning of the clinical notes. This automatic extraction of codes from clinical notes can help to improve accuracy and efficiency in healthcare billing and coding, mapping clinical concepts in EHR workflows, indexing healthcare data, transforming unstructured data to structured data, de-identifying data, building healthcare knowledge graph, ultimately leading to better patient care and outcomes.

### Business Outcomes:

- Improved coding efficiency and speed with accuracy
- Better patient care and outcomes
- Reduced revenue cycle management
- Higher human productivity

### Compliance:

- Meet HIPPA standards
- HIRUST certified system

### Initial Accuracy:

- Precision
- Recall
- F1-score
- Measured at an entity extraction level such as diagnosis or procedure code

### Drift & Recovery:

- Presence across clinical notes (such as discharge summary, prescription notes) to be handled with equal accuracy
- Changes/updates in ontologies, and codebooks to be accounted for and adapted to
- Be able to handle styles and guidelines followed by various institutions (technically this falls under generalizability, than drift)
- Clinical ontologies evolve and so do the codes annually. Some codes get suppressed and some new codes get introduced. The extraction pipeline must account for these changes to maintain the accuracy

### Confidence & Boundary:

- The extraction was limited to 14 clinical entities such as disease, procedure, tests, medication, age, gender, member ID, etc., and 4 different ontologies such as ICD, CPT, SNOMED, RxNORM/one needs to define for their use case/application clearly
- Each of the extracted entities had an associated confidence measure

### Performance:

- Number of clinical notes processed per minute
- Typically, this is not used in real-time scenarios

### Explain-ability & Experience:

- LIME and Shapley based explain-ability measures were utilized to showcase context and dependency of words/sentences that influence the class of extracted entities
- Semantic matching score and explainer API in Elasticsearch was utilized to explain mapping of concepts to codes

## Applied Scenario #3: STARS Optimization

Here is another use case where AI was leveraged for a more comprehensive insight into care gap management.

CMS Star ratings are used to assess the performance of health plans in serving Medicare Advantage beneficiaries. The ratings are based on five broad categories – outcomes, intermediate outcomes, patient experience, access, and process which are updated annually.

- Achieving higher star ratings can be challenging due to varying methodologies and complexities. This involves continuous monitoring and improvement.
- Our solution prioritizes members and providers to reach the next star threshold. Prioritized listing is generated based on a local plan's capabilities to address specific measures and considers historical performance to generate prioritized listings.
- The solution allows for 'what if' analysis at the measure level, facilitating planning and investments in specific measure categories. For example, it enables strategies like improving access to care through retail clinics or sending in-home screening kits to eligible members.
- The solution provides a single source of truth for enterprises, offering readily available insights. It thus promotes consistent decision-making across all stakeholder-facing teams, enhancing both stakeholder and employee experiences.

### Business Outcomes:

- Prioritize members to close care gaps/ improve rating

  Members' engage-ability (likelihood member will be compliant if approached) scores help allocate resources and get higher ratings (ROI-driven insight).
- Prioritize providers to close care gaps/ improve rating

  Provider impact-ability score and underlying member/measures inform value-based contracts for higher network performance.
- Develop initiatives (3 year strategy)

  What-if analysis, based on 'ease of

closure' by measure, helps develop measure-specific strategies and initiatives (e.g., investing in COL screening kits, aligning medication adherence with clinical programs).

## Compliance:

- CMS Stars is a regulatory requirement

  Stars is a 'zero sum' game – high-performing plans receive bonuses and low-performing plans are terminated. ML-based outputs help health plans understand current and projected ratings for continuous monitoring and strategy adjustments, all within the parameters of CMS Star regulations.

- Bias toward compliance

  Stars solution identifies ROI-driven insights for attaining higher ratings, however, the primary objective for health plans is also improving health outcomes across their entire populations (e.g., close care gaps, improve access to care).

## Initial Accuracy:

- Star Projections, Member Engage-ability and Provider Impact-ability scores

  The combination of these outputs allows ROI-driven decision-making to improve health outcomes, increase Star ratings, and optimize network performance. All the above outputs are the result of 7 distinct algorithms, all intricately designed (output of one model is input for another) to provide users with insights and help with necessary decisions (strategy, operations, resource allocation, etc.).

- Focus Measures

  Calibration and recommendation models produce a set of 'focus measures,' or prioritized Star measures to attack for a given MA contract (impact overall Star rating).

## Drift & Recovery:

- Star projections improve year over year

  New data is introduced every year based on the prior year's performance. As such, models are trained annually and weekly/monthly depending on the data refresh cycles; methodology is reviewed annually for maintaining/improving accuracy.

- Member and Provider score should improve monthly

  Member utilization and provider performance data are added each week/month. As such, both engage-ability and impact-ability models are retrained frequently, but engagement data is required to validate accuracy. Engagement data includes what actions were taken post outreach (outcome) and can capture additional variables for determining channel effectiveness and other engage-ability characteristics.

## Confidence & Boundary:

- Important to revisit methodology to maintain/improve accuracy

  Health plans need to measure and monitor network performance, and member actions (post outreach to prioritized members) to validate solution predictions. Non-prioritized measures/members/providers could have equal or more impact on network performance and member health outcomes.

## Performance:

- Insights for improving Star ratings vs. improving health outcomes

- Time to process and generate recommendations
- Speed of providing real-time what-if capabilities

A prioritized list of members can potentially ignore or take minimal action for all members with open care gaps. Health plans need to improve health outcomes for all MA members they serve.

### Explain-ability & Experience:

- Readily available insights

  Users have the necessary insights for decision-making (prioritizing measures, informing value-based contracts) at their fingertips. Users don't need to wait a month to get the reports required to design initiatives and adjust strategies.

- Single source of truth for enterprise

  Users across different departments (customer service, network management, quality/HEDIS) can make decisions based on the same underlying data and probabilities.

- Continuous improvement

  New data are introduced annually, presenting an opportunity to re-train models and update algorithms/ methodology. If health plans can document engagement data (what happened post outreach or execution of value-based contracts), models and predictions can be improved and 'stay relevant' for business decision-making.

## Applied Scenario #4: Smart Search

Hospitals are known for their multidisciplinary collaborative effort to address the needs of the patient. As hospital staff grows, it becomes increasingly challenging to find the right set of care team experts and maintain a repository of self-referred expertise. While web-based tools exist to facilitate these connections, many physicians find them to be inefective, outdated, prone to bias owing to self-reported expertise and often do not account for availablity of care team member(s). This is where the Find an Expert (FaE) search and recommendation solution plugs the known gaps, and connects care team members with the appropriate expert for a given patient case, quickly and easily. It allows for seamless connections between clinical team members, whether they are in the same building, across campuses, or in different states, allowing for optimal care by leveraging collective expertise and elevating patient care quality.

User can search for an expert care team member by inputting a natural language-based query, a phrase, a sentence, or even a couple of sentences describing the scenario, diagnosis, or procedure for which he/she is seeking the expert.

### Business Outcomes:

- Improved Click Through Rates (CTR)
- Increased User Engagement (Amount of time a user spends on the application)
- Effect on easing physician appointment workload, availability
- Adoption and Conversion to expert call, chat, curbside consultations, appointment scheduled
- User Behavior and Engagement

The above outcomes are aligned with the central goal of seamless connections between clinical colleagues, whether they are in the same building, across campuses, or in different states, allowing for optimal care by leveraging collective expertise and elevating patient care quality.

### Compliance:

- Adherence to hospital IT system security best practices
- HIPAA and relevant compliances for Healthcare IT systems
- Best practices for adversarial search and information retrieval systems

### Initial Accuracy:

- Technical metrics: such as ndcg@5, ndcg@10, precision@5, precision@10
- These are established minimum standard values for search and retrieval systems out of BIER benchmark suits
- Other measures such as BLEU, ROGUE, BERT-Score, Normalized Discounted Cumulative Gain (NDCG) and Mean Reciprocal Rank (MRR) help measure search & retrieval accuracy

### Drift & Recovery:

- Changes in ndcg@5, ndcg@10, precision@5, precision@10 using CTR and conversion data
- Recovery is affected using a number of relevancy tuning techniques such as boosting, ML model update, reinforcement learning, matrix factorization

This metric measures how a user perceives the recommended results. Though the user may key in an identical query, its intent, and context can evolve over time. For example: a query to find an expert for lymphedema can change its intent from finding a plastic surgeon to finding a psychologist and physical therapy expert during remission. What people search remains same, but what people search for, its intent evolves over time and therefore drifts in the system.

### Confidence & Boundary:

- The confidence is measured by technical KPI and explicit feedback from physicians and care teams
- Micro-feedback (thumbs up/down) for results proved to be very useful and tracking quality of results

### Performance

- Response time for results
- Speed of auto-complete

This speed is crucial as there is a ‘consumer’ expectation for any search solution since we are all tuned to see auto-complete and search results in micro-seconds.

### Explain-ability & Experience:

- Use of “Explain-ability API” to explain why a particular physician expert is suggested
  - It highlights the regions, texts, and summary of the data/similar keywords that led to the relevancy score. Depending on verbosity level provides deep auditing workflows to know exactly why particular physicians were ranked higher than other physicians. Provides even mathematical calculation details and decomposition of relevancy score, readily available in the debug/ Insight mode
  - Traceability and transparency of the search results back to research papers published or prior history goes a long way in justifying the results and rankings

## Successful Design and Implementation of AI in Healthcare Needs a New Language

The use of this framework has led to greater success rates for many of our AI engagements and projects. It has also been an invaluable tool in helping create a common language for collaboration across various stakeholders.

Each organization and its stakeholders must engage in discussions and determine the importance, goals, and trade-offs associated with the measures of success across these seven dimensions. The true value is achieved only through debate and discussion, as there are no right or universal answers.

The absence of such a framework and the failure to openly communicate goals under multiple dimensions upfront often leads to conflict and confusion within the organization when new AI-based solutions are implemented. Unrealistic expectations, disappointments, and potential dangers are common pitfalls of this lack of clarity.

The simplicity of the framework has also been a reason for it’s adoption. We are not advocating that this framework applies in every situation and for every organization. But, it hopefully demonstrated that a similar framework can go a long way in the success of your AI initiatives.