BEK platform

BEKhealth AI outperforms Google, Amazon, and other leading medical AI

BEKhealth's patient-matching large language models (LLMs) demonstrate higher accuracy in matching patients to clinical research than other leading medical models

Highly accurate AI models can significantly improve both the speed and accuracy of the patient recruitment process for clinical trials. AI's ability to quickly sift through vast amounts of data to identify trends and extract valuable information can help to more efficiently find potential candidates who match study criteria. Traditional methods of patient recruitment are often slow, labor intensive and error prone; relying on highly manual processes to review patient records for eligibility. AI can automate and refine many of these processes, analyzing electronic health records (EHRs), genetic information, and other relevant data to find suitable candidates more quickly and accurately. This promises not only to speed up the recruitment process but also to improve the quality of matches. The ability to do these things quickly and cost-effectively is becoming increasingly vital as researchers seek to expand clinical trial access to more diverse and global populations of patients.

Key Benefits of Highly Accurate AI in Patient Recruitment

Personalized Patient Engagement

AI-driven tools can personalize the way information about clinical trials is presented to potential participants. For instance, AI can tailor communications based on the patient's medical history, demographic information, and even preferred communication channels. This personalized approach can increase patient engagement and aid in retention.

Real-time Monitoring and Faster, Informed Adjustments

AI models can monitor the progress of recruitment in real time, providing insights into which strategies are working well and which are not. This allows trial coordinators to make timely adjustments to their recruitment strategies, like changing advertising channels, revising inclusion criteria, or addressing other barriers to recruitment. This real-time feedback loop can significantly optimize the recruitment process.

Predictive Analytics for Recruitment Success

AI models can forecast the success rate of patient recruitment in various demographics and geographies. By analyzing past trials and recruitment patterns, AI can provide insights into where and how to target recruitment efforts.

Comprehensive Consideration

AI models efficiently and comprehensively review every available candidate. This can be particularly useful in identifying underrepresented groups in clinical trials, ensuring a more diverse and representative participant pool. Diversifying trial participation is crucial for understanding how treatments work across different populations.

Building Accurate AI Models – Validating Existing Technology

BEKhealth sought to evaluate the capabilities of leading AI solutions for natural language processing (NLP) to determine their suitability for use in clinical trial recruitment. These included Amazon Web Service's Comprehend Medical (AWS), Google Healthcare Natural Language API (Google), John Snow Lab's Spark NLP (Spark), and the medspaCy (medspaCy) NLP models.

To begin, BEKhealth deployed each of the solutions to extract information of interest from volumes of labeled unstructured and deidentified patient documents. These documents, combined with digital records from EMR databases and patient charts, would allow the team to gauge the performance of the existing NLP tools and provide a foundation for the kinds of AI models the company sought to build.

Evaluation of Existing NLP Infrastructures – Four Key Areas

There is a great deal of unstructured data that needs to be reviewed when it comes to vetting potential clinical trial participants. For an AI to be useful, it must be able to accurately decipher

inconsistent terminology, incorrect coding, handwritten notes, context dependent acronyms, unstandardized units of measure, as well as a myriad of other issues. It is estimated that 70-80% of patient medical history details are contained in the unstructured sections of the electronic medical record systems[1]. Clinical providers generate nearly 140 terabytes of data every day, most of it unstructured[2], and as much as half of that data may be duplicated due to human error.[3]

The clinical team quantitatively evaluated the predictions made by the AWS, Google, Spark, and medspaCy NLP solutions, measuring their ability to accurately identify and codify medical information of interest. The AI predictions were then benchmarked against BEKhealth's human team's confirmed records of diagnoses, medications, lab tests, biometric measures, clinical observations, and procedures. This data, corrected and validated by BEKhealth's team of experts, established a new Gold Standard for accuracy that would serve as a performance target as the team sought to build its own clinical trial recruitment AI solution.

The team at BEKhealth investigated how accurately each of the four solutions identified medical procedures (including surgeries), lab tests, biometric measurements, medications, and diagnoses present in patient histories. Performance was summarized using standard performance metrics, each chosen for its ability to provide insights into different aspects of the solution's effectiveness:

- **Accuracy:** The overall correctness of the model, i.e., how often the AI's predictions match the true outcomes.
- **Threat Score:** The model's ability to correctly predict the cases of interest, balancing the importance of hits against false alarms.
- **Recall:** How many of the actual positive cases were correctly identified by the model.
- **Precision:** How many of the instances identified as positive by the model were actually positive.
- **Specificity:** The true negative rate, i.e., the proportion of actual negative cases that the model correctly identified.
- **F1-Score:** A harmonic mean of precision and recall, offering a balance between them.

They also tracked precision for the high-confidence subset of AI predictions. The metrics for each solution were normalized as a percentage on a 0 to 100 scale.

1 Negro-Calduch E, Azzopardi-Muscat N, Krishnamurthy RS, Novillo-Ortiz D. Technological progress in electronic health record system optimization: Systematic review of systematic literature reviews. Int J Med Inform. 2021 Aug;152:104507. doi: 10.1016/j.ijmedinf.2021.104507. Epub 2021 May 21. PMID: 34049051; PMCID: PMC8223493.
2 Brian Eastwood. (2023, May 22). How to navigate structured and unstructured data as a healthcare organization. Technology Solutions That Drive Healthcare. https://healthtechmagazine.net/article/2023/05/structured-vs-unstructured-data-in-healthcare-perfcon
3 Jauhar, S. (2023, June 19). Bloated patient records are filled with false information, thanks to copy-paste. STAT. https://www.statnews.com/2023/06/20/medical-records-errors-copy-paste/

Medical Procedures/Surgeries

Solution performance for this category was measured based on the ability to accurately identify medical procedures and surgeries from unstructured medical histories and codify those events using standardized medical vocabularies such as the Healthcare Common Procedure Coding System (HCPCS). Performance was poor across all existing solutions; AWS, Google, and Spark failed to extract any of the labeled procedures from the processed records, while the MedspaCy model had an accuracy score of 0.5%.

Labs and Biometrics

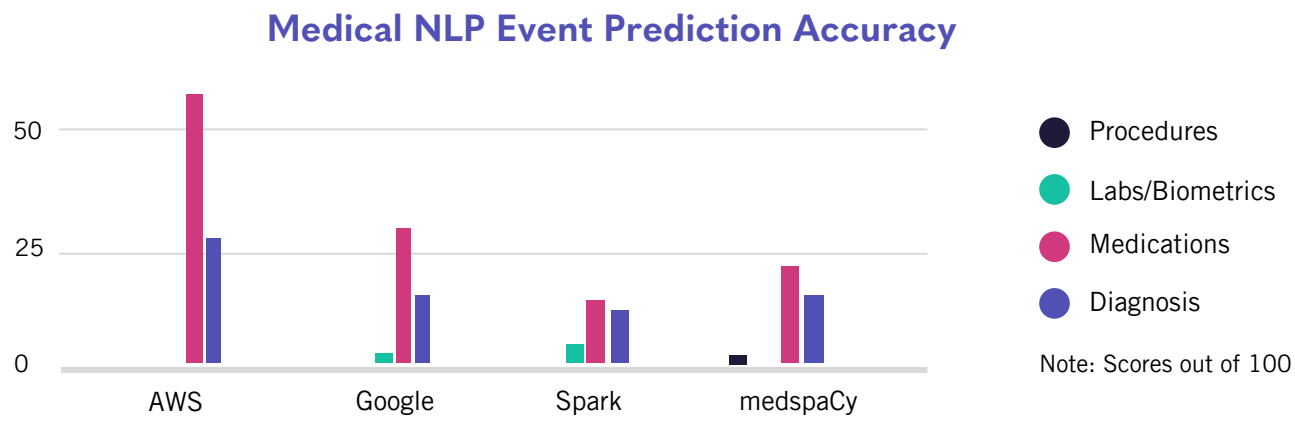
Solution performance for this category was measured based on the ability to accurately identify biometrics, clinical observations, and lab test measurements and codify those events using the Logical Observation Identifiers, Names, and Codes (LOINC) vocabulary. Performance was poor on this task as well; AWS failed to identify any of the labeled tests or biometric measures, while the other three solutions all had accuracies under 5%.

Medications

Solution performance for this category was measured based on the ability to accurately identify medication exposures and codify those events using the RxNorm and NDC vocabularies. The Medications category saw a significant increase in performance across the board. This appears to be largely attributable to the fact that the classification of mediations is more standardized than that of procedures and labs. AWS achieved an accuracy of over 50%, while the other solutions fell in the 15%-30% range.

Diagnosis/Diseases/Disorders/Conditions/Adverse Events/Comorbidities

Solution performance for this category was measured based on the ability to accurately identify medical conditions and codify those events using the International Classification of Diseases (ICD-10) and the Systematized Nomenclature of Medicine-Clinical Terminology (SNOMED) vocabularies. Again, a high degree of standardization due to the robust diagnosis classification system led to better performance. AWS led the way overall, falling just above 25% accuracy. Accuracies for each tested solution across the four categories can be visualized in the chart below.



Limitations and Room for Improvement

The analytical and clinical validity scores for the four established NLP AI solutions fell far short of acceptable thresholds, establishing the need for a more specialized approach. The AI offerings being evaluated, while robust and versatile for a broad range of applications, clearly struggled to provide useful output when applied to the challenges unique to clinical research teams and clinical trial recruitment. The reasons for these limitations can be categorized into several key areas:

Generalization vs Specialization:

Most NLP solutions are designed to be performant across a wide range of data domains. This generalization, while beneficial for broad use cases, leads to a lack of nuance needed to be useful in highly specific domains like clinical research. Clinical trials demand an AI system that understands complex and frequently evolving medical terminology, patient histories, and highly-specific, time-bound criteria for trial eligibility, all of which are inadequately addressed by generalist AI solutions.

Interpreting Unstructured Medical Data:

Clinical trial recruitment relies heavily on accurately interpreting unstructured medical data such as patient notes, which are often filled with domain-specific jargon, abbreviations, and varying formats. Clinical terminology is a high-dimensional space, covering tens of millions of medical terms, synonyms, and lexemes, that also spans many orders of magnitude difference in incidence and prevalence rates. Furthermore, medical notes often describe the absence of something, using various forms of negation to do so. It is difficult to detect these negations, leading to the positive extraction of events that were actually described as being absent. Existing generalist AI is not optimized for the deep understanding and contextual analysis required to accurately parse and interpret these types of statements.

Regulatory Compliance and Data Privacy:

Clinical trials are subject to strict regulatory compliance and data privacy concerns. AI tools that require the transmission of data via API may not fully align with the specific compliance requirements of global regulatory bodies (like HIPAA in the U.S. or GDPR in Europe) or offer the level of data privacy needed for handling sensitive patient data.

Customization and Flexibility:

Clinical research teams often require AI solutions that can be highly customized to their specific protocols, study designs, and patient populations. AI solutions like AWS, Google, Spark, and spaCy offer limited customization options compared to a bespoke AI system built specifically for clinical trials, which can be tailored to specific research needs and continuously refined.

Accuracy and Reliability in Clinical Contexts:

The stakes in clinical trials are incredibly high, and the cost of inaccuracies can be significant in terms of patient safety and data quality. Existing AI solutions do not offer the level of accuracy and reliability needed for clinical contexts, especially when it comes to understanding complex patient eligibility criteria and making nuanced recommendations or judgments.

Building Accurate AI Models – Developing an AI Solution That Works for Trial Recruitment

Seeking to approach human levels of accuracy, BEKhealth developed its own AI pipeline. This began with a large-scale data labeling effort by their team of clinical experts, wherein sanitized electronic health records were manually annotated to capture every relevant piece of information. Using this high quality data, their Data Science team rigorously trained and fine-tuned an array of specialized models on millions of data entities and built out their AI pipeline.

The BEKhealth AI leverages an orchestra of fine-tuned large language models (LLMs), deep transformer-based neural networks, siamese networks, and classical models to achieve greater than state-of-the-art performance on clinical tasks. The models are continuously retrained as more data becomes available and has been improving for over three years with no plateau in sight given the vastness and complexity of the medical corpus. BEKhealth employs a human-in-the-loop feedback mechanism (Deep-Learning with Clinical Expert Feedback) to enable clinical experts to hone the models' outputs, ensuring a high level of reliability and relevance to clinical contexts.

The team built its tool with a self-evaluation feature; the AI provides a confidence level metric for each of its predictions. Only predictions with at least 80% confidence are accepted, as the team found that predictions with 80% or higher confidence scores aligned with over 90% of its human medical staff assessments.

Comparing the BEKhealth AI Solution with the Other Models

This commitment to constant evaluation of prediction performance routinely returns accuracy rates in the high 70%-to-80% range. Benchmarking its solution versus AWS, Google, Spark, and medspaCy, BEKhealth found that its solution consistently outperformed competing models when applied specifically to clinical trial data.

Procedure Data

The BEKhealth solution is significantly better than the other tools at accurately deriving procedure events and matching them to the HCPCS vocabulary. As a reminder, the AWS, Spark, and Google tools failed to provide any results in this data category, and medspaCy achieved an accuracy of just 0.5%. BEKhealth scored 37.5% in overall accuracy, and 47.9% when the confidence threshold was set to 0.8.

Lab Data

The BEKhealth solution also proves to be much more reliable at accurately identifying biometric, lab and observation data, which tends to be largely unstructured. BEKhealth's model outperforms the other solutions by wide margins, achieving 61.5% overall accuracy and 87.5% when the confidence threshold is set to 0.8.

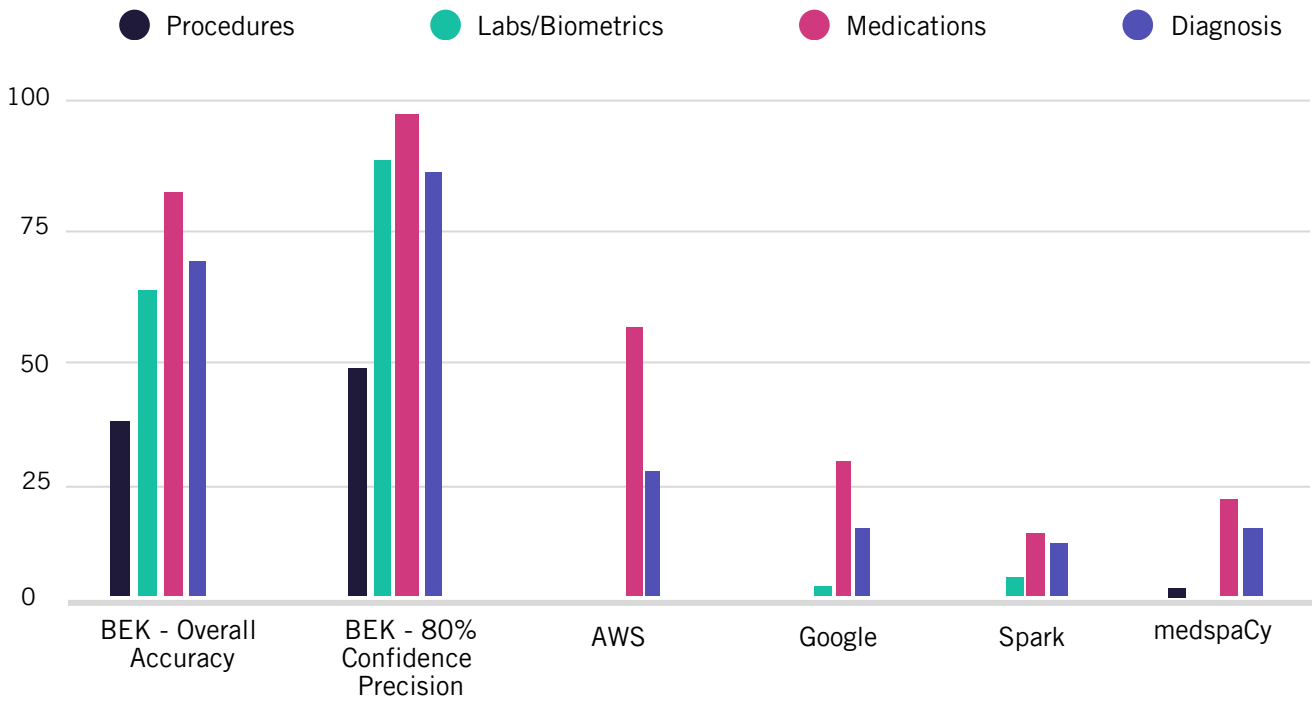
Medications

With medication data offering more standardization and, thus, less variability, all AI solutions generated much higher results than in previous categories. AWS, notably, showed a marked performance improvement in this category relative to procedure and lab data. Still, the BEKhealth solution remained the top performer, achieving 79.6% overall medication prediction accuracy and 95.7% when the confidence threshold is set to 0.8.

Diagnoses

BEKhealth's solution also led the way in all diagnostic performance subcategories, outperforming its nearest competitor by 40 percentage points with an overall accuracy of 66.8% and 86.1% when the confidence threshold is set to 0.8.

Comparison of Medical NLP Event Prediction Accuracy



Note: Scores out of 100

Conclusion

As clinical researchers seek to understand how AI-powered tools can help drive efficiency and maximize the value of study data in their trial recruitment and enrollment efforts, they must understand the capabilities and limitations of the various AI models available. Generalist AI tools, even those directed toward healthcare data challenges, are largely still unable to reliably and accurately parse through the volumes of unique and often unstructured data common in clinical research. BEKhealth, through its rigorous and stringent approach to continuous expert validation, has successfully developed a best-in-class AI solution that significantly outperforms other medical NLP solutions. This achievement is attributed to an intentional, mindful approach to AI development along with its expert human team of experienced clinicians and end-user research staff evaluating each clinical trial candidate identified by the BEK platform. This ensures that BEKhealth's AI solution maintains the highest level of accuracy. As time goes on, the model will only become more accurate and useful as the BEKhealth team remains dedicated to their meticulous training approach.

For more information on BEKhealth's highly accurate AI models and how you can use them to facilitate more effective clinical trial recruitment programs, visit **bekhealth.com** or contact us at **info@bekhealth.com**