ChirokHealth
Where every story matters

EVALUATING THE ROLE OF AI IN VALUE BASED CARE PREMIUM ADJUSTMENT

White Paper, October 2024

Kate Casaday MPH, MBA
Chief Strategy Officer & Chief Financial Officer
Chirok Health LLC

EXECUTIVE SUMMARY

This study aims to understand the ways in which augmenting AI with clinical reviewers can unlock faster adoption, increase efficiency, and improve compliance in the HCC coding process. Probing the relative benefit of AI as a standalone solution, versus augmenting with a service-based product is especially relevant. The healthcare product market offers many solutions that deploy either services or technology, but few solutions pair technology with optimized, customized clinical coding and documentation review services. This study intends to deliver an evidence-backed rationale for augmenting technology with high-skill labor to support risk adjustment processes.

Clinically trained reviewers audited AI generated suspected hierarchical condition categories (HCCs) against coding guidelines using electronic health record (EHR) data. The audit assessed the accuracy, productivity, interpretive power, and learning capability of using AI as a stand-alone solution and demonstrated the benefit of combining AI with a clinical review team. Key findings include:

1. Accuracy: 46 percent of HCCs suggested by AI required modification or removal by clinical reviewers to either align with coding guidelines or maximize the performance of the process.
2. Productivity: 0.73 HCCs per patient were suggested by AI alone, compared to 1.06 HCCs per patient when used in conjunction with clinical reviewers.
3. Interpretive Power: 23 percent of all AI errors mapped to misinterpretation of data in progress notes.
4. Learning Capacity: Incorporating clinical reviewer findings into model calibration activities improved AI accuracy by 17 percent.

BACKGROUND

Medical coding, documentation, and billing have long been a focus for computer assistance and are a prime target for AI. However, healthcare organizations participating in either fee-for-service (FFS) or value-based care (VBC) reimbursement models[1] historically have two options to meet administrative requirements:

1. Clinicians are responsible for writing clinical notes and assigning Current Procedural Technology (CPT) and International Classification of Disease (ICD) codes to bill a patient's insurance.
2. A team of certified coders and billers is staffed to review clinicians' notes, translate the information into the relevant CPT and ICD codes, and bill the patient's insurance.

Complying with best practice documentation, coding, and billing requirements is a major source of clinician burn-out[2]. Navigating Electronic Health Record (EHR) and Electronic Practice Management (EPM) systems requires 44 percent of primary care providers' (PCP) total work time[3]. This mandate decreases the time available for patient interactions and contributes significantly to clinician frustration and dissatisfaction[4].

Hiring a team to audit documentation and perform coding and billing functions may reduce clinician burnout, however, it has a significant impact on practice finances. It is estimated that 14.5 percent of PCP annual revenue can be attributed to the costs associated with billing insurance companies[5]. Primary care practices also face financial pressure from third-party payors in the form of price controls, utilization management, and total cost of care oversight[6]. Narrowing margins discourages investment in staffing for support services like billing and coding. Selling a practice to a larger health system or merging under a roll-up of small providers remain the most common way to improve financial performance and access economies of scale[7].

Assuming organizations have sufficient capital to invest in AI, adopting this technology can both reduce overhead and improve documentation and billing quality. Legacy coding and clinical documentation improvement (CDI) teams can re-purpose their qualifications and tribal knowledge to support automation and model accuracy. However, in-house coding teams often lack the experience or desire to support AI model performance for fear of losing their jobs as the technology improves.

Relevance to Value-Based Care

VBC continues to place heavy emphasis on providers' proficiency at documenting and coding HCCs. The tactics used to maximize the reporting of HCCs have received increasing amounts of negative media attention[8,9], and regulatory updates add stress to Medicare Advantage margins in particular[11]. The compression will drive increased investment in risk adjustment where workable solutions present themselves, with additional attention paid to solutions' ability to adapt to future regulatory changes.

Use of Tech in Clinician Billing Workflow

To compliantly document HCCs, clinicians must change ingrained note taking habits and adhere to the rules of risk adjustment. Supporting clinician behavior change compliantly, requires consistent and timely feedback. The highest impact place in the workflow to provide feedback exists in the narrow window after the chart note is signed, and prior to the charge being released to the clearing house. At this point, reviewing the documentation against the charges in the EPM can flag compliance risk, and provide opportunity to clarify conflicting documentation before the claim is submitted to insurance.

While clinical reviewers can access this window of intervention from the front end of the EPM system, technology solutions must integrate with both EHR and EPM systems to install the necessary logic. This level of integration provides technology companies with significant access to a provider's billing systems and demands extensive support from information technology (IT) staff at a provider organization to implement and maintain.

METHOD

The review involved 3,703 patient records randomly selected from a Medicare population at a multi-provider primary care clinic. All patient records existed in a single EHR, and the same data was made available to AI and the clinical reviewers for analysis. All AI errors were communicated back to the AI product development team for model improvement.

For three HCCs: HCC 85 Congestive Heart Failure, HCC 55 Drug/Alcohol Dependence, and HCC 27 End-stage Liver Disease, clinical reviewers pinpointed the EHR report type used to reject the AI suggestions in the event they disagreed (Figure 1).

The distribution of errors by report type was tested against two null hypotheses using a Chi-Squared Goodness of Fit Test. The hypothesized AI error distributions were as follows:

1. AI more seriously considers information from report types that both produce discrete data and are highly relevant to a particular HCC (e.g. echocardiograms for coronary heart failure). Clinical reviewers report infrequently that these types of reports contained data that overturned an AI suggestion (Figure 2).
2. AI equally weights all EHR data independent of report type when suggesting HCCs. Clinical reviewers overturn AI suggestions at the same rate, independent of report type (Figure 3).

FINDINGS

The study's findings are summarized below:

Suspect Identification Performance

On average, AI identified 0.73 conditions per patient. The clinical reviewers agreed with 54 percent of the suggestions. The clinical reviewers identified an average of 1.06 suspected, or potentially uncoded, conditions per patient when paired with AI.

Rapid Model Learning

Clinical reviewers identified specific instances of AI failing to ingest relevant clinical data and report the shortfall to the AI product development team for quick resolution. After performing an update, AI performance improved by 17 percent from 54 to a 65 percent agreement rate.

Distribution of AI Errors by EHR Report Type

The observed distribution of evidence types used to overturn AI predictions across three HCC categories (HCC 85, HCC 55, and HCC 27) demonstrated significant deviations from the expected distributions: HCC 85 ($\chi^2 = 7317.26$, $p < 0.001$), HCC 55 ($\chi^2 = 371.90$, $p < 0.001$), and HCC 27 ($\chi^2 = 1070.13$, $p < 0.001$). These findings indicate strong divergence between the observed result and both expected evidence distributions.

DISCUSSION

The results suggest AI is neither equally likely to error across all EHR report types, nor least error prone when interpreting report types with discrete data highly relevant to the HCC at hand. Instead, the errors appeared concentrated in a limited set of report types.

Interpreting progress notes drove 23 percent of errors made by AI across the three HCCs reviewed. This suggests a shortfall in parsing free text to contextualize, rule-out, or refine suggestions initially made based off discrete lab or imaging data.

For HCC 85 Congestive Heart Failure, 66 percent of AI suggestions were overturned by information in an echocardiogram, followed by data in the progress notes (22 percent). This finding was counterintuitive. The ejection fraction reading taken during echocardiograms provide a clear measure of heart health, but AI performance interpreting demonstrated neither accuracy nor precision.

For HCC 55 Drug/Alcohol Dependence, 46 percent of AI suggestions overturned by information in the medications list, followed by the progress notes and social history (39 and 14 percent, respectively). This performance was also significantly worse than expected, give there are four commonly prescribed drugs to treat drug and alcohol use disorder. Recent prescription of any of these drugs is a strong signal that drug or alcohol dependence is currently being treated, but AI was not able to use the medication list to accurately and consistently suggest HCC 55.

For HCC 27 End-stage Liver Disease, AI errors demonstrated greater dispersion across report types than the other two HCCs tested, but lab reports drove the highest proportion of errors (23 percent). Blood tests are used to assess liver function and can differentiate between acute and chronic liver disorders, but this data was not used by AI to make accurate suggestions.

AI Errors Categorized by Type

Overall, differences in suggested HCCs between the clinical reviewers and AI fell into three categories:

1. AI was unable to ingest certain sections of the medical record, often involving hand-written notes or information stored in an atypical location. This is referred to as “error of omission”.
2. AI struggled to parse complex medical terminology involving various forms of abbreviation and shorthand. This is referred to as “error of interpretation”.
3. AI failed to correctly sequence a series of clinical events into a narrative leading to suggestions without relevant context or HCCs miscoded to reflect a transient historical state. This is referred to as “error of conceptualization”.

Errors of omission are not explicitly included in the data. At the start of the analysis, clinical reviewers identified a section of the progress note where AI repeatedly missed HCCs with a clear body of historical evidence. The failure to ingest was reported to the AI product development team, who updated the model to include previously excluded reports. Once the updates were made to align with the clinical reviewers’ findings, this error of omission ceased.

Errors of interpretation were observed most frequently in echocardiograms and the medication list. These errors occur when clear clinical evidence from a lab report, image, or prescription is not properly applied to a clinical guideline to suggest an HCC.

Errors of contextualization appeared most prominently in analyzing progress notes. Errors of contextualization appear where AI failed to reconcile the free text of a chart note with other data like labs or imaging to form a narrative of historically sequenced clinical events. This caused AI to make suggestions that ignored content in recent progress notes that indicated a rule out or resolution of a condition.

CONCLUSION & RECOMMENDATIONS

The rate at which AI suggests HCCs, and the distribution of errors suggests that technology alone is not sufficient to providing a reliable solution to risk adjustment.

There are three ways organizations can increase the usefulness of AI in this area; ensuring training data used to calibrate the models is the most applicable to the deployment environment, improving the feedback to clinicians within the documentation workflow, and augmenting AI with the support of clinical reviewers at scale.

Understand Model Training Methods

There are two data training set curation models with significant impact on model predictions:

1. Calibrating a model using the historical data of a single organization. This method excels at recognizing and applying organization specific patterns, but relatively small sample sets may increase confidence intervals around predictions.
2. Calibrating a model using aggregated data. This method allows for sufficient volume to support predictions with small confidence intervals. Aggregated data can also adjust for specific institutional biases but may fail to replicate the nuances of real-world clinical documentation.

Understand and Counter Limitations of AI Within the Billing Workflow

There are three major considerations for maximizing clinician coding and billing effectiveness and minimizing the effects of burn-out and administrative overwhelm:

1. Provide clinicians real-time, succinct feedback on documentation and coding entries to support rapid learning of VBC and FFS reimbursement requirements. The American clinical workforce is highly proficient at learning but must be afforded the opportunity to learn from their own clerical errors in a way that does not contribute to further notification and decision fatigue. Clinical reviewers can intervene after the note is closed but before the charge is sent to the clearing house.
2. Provide clinicians relevant, brief, and highly curated information with a discernable audit trail at the point of care. Coding suggestions provided without reference to the rationale for inclusion erodes trust in a sensitive workflow. Clinicians must be able to follow along with the logic applied.
3. Orient suggestions around clinical care. Financial return is a significant decision driver for the American healthcare system. However, support in rendering a patient diagnosis must also serve as a clinically useful tool that provides for quick and efficient consumption of the information required for medical decision making.

Augment Technology with Expertise

Deploying clinical chart reviewers alongside AI can accelerate model accuracy and operational progress. Augmentation has three key areas of impact:

1. Clinical chart reviewers can quickly identify errors of omission. AI may fail to ingest or misidentify data sources when rendering a decision. Clinical chart reviewers work inside the EMR and are positioned to identify these errors and report the issue to AI product development teams for resolution.
2. Clinical reviewers can identify specific errors of interpretation and suggest model improvements. A subsequent audit on 3,096 patients saw the agreement rate jump to 65 percent.
3. Clinical reviewers can work with clinicians directly to correct documentation errors at the highest impact place in the workflow, prior to charge being sent to the clearinghouse. Providing hands-on support for the correction process eases the burden of clerical rework. Real-time notifications support the learning process more effectively than performance reporting that summarizes errors after the time to act has passed.

ABOUT THE AUTHOR

Kate A. Casaday is the Chief Strategy Officer and Chief Financial Officer at Chirok Health LLC. With extensive experience in healthcare financing, strategy, and operations, she has held leadership roles across the healthcare spectrum, including managing product development, risk adjustment, and value-based care initiatives. Kate holds an MBA from The Wharton School at the University of Pennsylvania and a master's in public health from Columbia University. Her expertise in financial analysis, acquisitions, and strategic decision-making has been pivotal in driving growth and operational efficiency in healthcare organizations.

REFERENCES

1. de Silva Etges APB, Liu HH, Jones P, Polanczyk CA. Value-based Reimbursement as a Mechanism to Achieve Social and Financial Impact in the Healthcare System. J Health Econ Outcomes Res. 2023 Oct 31;10(2):100-103. doi: 10.36469/001c.89151. PMID: 37928822; PMCID: PMC10621730.Ozeran L, Schreiber R. Reduce Burnout by Eliminating Billing Documentation Rules to Let Clinicians be Clinicians: A Clarion Call to Informaticists. Appl Clin Inform. 2021 Jan;12(1):73-75. doi: 10.1055/s-0041-1722872. Epub 2021 Feb 3. PMID: 33535252; PMCID: PMC7857966.
2. National Academies of Sciences, Engineering, and Medicine; National Academy of Medicine; Committee on Systems Approaches to Improve Patient Care by Supporting Clinician Well-Being. Taking Action Against Clinician Burnout: A Systems Approach to Professional Well-Being. Washington (DC): National Academies Press (US); 2019 Oct 23. 7, Health Information Technology.Available from: https://www.ncbi.nlm.nih.gov/books/NBK552608/
3. Roman LC, Ancker JS, Johnson SB, Senathirajah Y. Navigation in the electronic health record: A review of the safety and usability literature. Journal of Biomedical Informatics. 2017;67:69–79. [PubMed] [Reference list]
4. Tseng P, Kaplan RS, Richman BD, Shah MA, Schulman KA. Administrative Costs Associated With Physician Billing and Insurance-Related Activities at an Academic Health Care System. JAMA. 2018 Feb 20;319(7):691-697. doi: 10.1001/jama.2017.19148. PMID: 29466590; PMCID: PMC5839285.
5. Zismer DK, Schwartz GS, Zismer ED. Finding the Financial Margin Expansion Leverage in the Medical Practice. Healthcare Administration Leadership & Management Journal. (2023);1(1)17/21.https://doi.org/10.55834/halmj.6064036896
6. Modern Healthcare. (2023). Health care M&A insights: Q4 2023 and outlook for 2024 deal activity. Modern Healthcare. Retrieved from https://www.modernhealthcare.com/mergers-acquisitions/health-care-ma-insights-q4-2023-outlook-2024-deal-activity
7. Pear, R. (2022, October 8). Medicare Advantage's cost to taxpayers rivals the Pentagon's, critics say. The New York Times. Retrieved from https://www.nytimes.com/2022/10/08/upshot/medicare-advantage-fraud-allegations.html
8. Mathews, A. W. (2023). Medicare's diagnosis-based payments are proving costly. The Wall Street Journal. Retrieved from https://www.wsj.com/health/healthcare/medicare-health-insurance-diagnosis-payments-b4d99a5d?mod=article_inline
9. Centers for Medicare & Medicaid Services. (n.d.). Medicare risk adjustment data validation (RADV) program. Centers for Medicare & Medicaid Services. Retrieved from https://www.cms.gov/data-research/monitoring-programs/medicare-risk-adjustment-data-validation-program#:~:text=The%20Medicare%20Advantage%20Risk%20Adjustment,in%20the%20enrollee's%20medical%20record
10. AAPC. (2023). Get ready for CMS-HCC v28. AAPC. Retrieved from https://www.aapc.com/blog/88300-get-ready-for-cms-hcc-v28/
11. Khullar, D., Casalino, L. P., & Bond, A. M. (2024). Vertical integration and the transformation of American medicine. The New England Journal of Medicine, 390(11), 965-967. https://doi.org/10.1056/NEJMp2313406

EXHIBITS

Figure 1. Actual Distribution of Overturn Evidence by Document Type

Report Type	HCC 85	HCC 55	HCC 27
Echocardiogram	225	0	0
Medication List	0	133	0
Lab Reports	0	0	103
Progress Note	74	111	18
CT Abdomen	0	0	40
Social History	0	40	0
US Abdomen	0	0	31
CT Chest	13	0	18
Transthoracic Echocardiogram (TTE)	17	0	0
Comprehensive Metabolic Panel (CMP)	0	0	11
Cardiac Stress Test	7	0	0
Total Disagreements	336	284	221

Figure 2. Expected Distribution of Overturn Evidence Based on Weighted Report Type

Report Type	HCC 85	HCC 55	HCC 27
Echocardiogram	6.7	5.7	4.4
Medication List	40.3	34.1	26.5
Lab Reports	16.8	14.2	11.1
Progress Note	151.2	127.8	99.5
CT Abdomen	16.8	14.2	11.1
Social History	50.4	42.6	33.2
US Abdomen	16.8	14.2	11.1
CT Chest	6.7	5.7	4.4
Transthoracic Echocardiogram (TTE)	6.7	5.7	4.4
Comprehensive Metabolic Panel (CMP)	16.8	14.2	11.1
Cardiac Stress Test	7	6	4

Total Disagreements	336	284	221

Figure 3. Expected Distribution of Overturn Evidence Based on Equal Probability of Error Across Report Types

Report Type	HCC 85	HCC 55	HCC 27
Echocardiogram	30.5	25.8	20.1
Medication List	30.5	25.8	20.1
Lab Reports	30.5	25.8	20.1
Progress Note	30.5	25.8	20.1
CT Abdomen	30.5	25.8	20.1
Social History	30.5	25.8	20.1
US Abdomen	30.5	25.8	20.1
CT Chest	30.5	25.8	20.1
Transthoracic Echocardiogram (TTE)	30.5	25.8	20.1
Comprehensive Metabolic Panel (CMP)	30.5	25.8	20.1
Cardiac Stress Test	30.5	25.8	20.1
Total Disagreements	336	284	221