Journal of the American Medical Informatics Association, 29(10), 2022, 1810–1817
https://doi.org/10.1093/jamia/ocac121
Advance Access Publication Date: 14 July 2022
Perspective

Perspective

Tasks as needs: reframing the paradigm of clinical natural language processing research for real-world decision support

Asher Lederman[1], Reeva Lederman[1], and Karin Verspoor[2]

[1]Faculty of Engineering and IT, School of Computing and Information Systems, University of Melbourne, Melbourne, Australia, and [2]STEM College, School of Computing Technologies, RMIT University, Melbourne, Australia

Corresponding Author: Karin Verspoor, BA, MSc, PhD, FAIDH, STEM College, School of Computing Technologies, RMIT University, 124 La Trobe St, Melbourne, VIC 3000, Australia; karin.verspoor@rmit.edu.au

Received 19 March 2022; Revised 6 June 2022; Editorial Decision 1 July 2022; Accepted 4 July 2022

ABSTRACT

Electronic medical records are increasingly used to store patient information in hospitals and other clinical settings. There has been a corresponding proliferation of clinical natural language processing (cNLP) systems aimed at using text data in these records to improve clinical decision-making, in comparison to manual clinician search and clinical judgment alone. However, these systems have delivered marginal practical utility and are rarely deployed into healthcare settings, leading to proposals for technical and structural improvements. In this paper, we argue that this reflects a violation of Friedman's "Fundamental Theorem of Biomedical Informatics," and that a deeper epistemological change must occur in the cNLP field, as a parallel step alongside any technical or structural improvements. We propose that researchers shift away from designing cNLP systems independent of clinical needs, in which cNLP tasks are ends in themselves—**"tasks as decisions"**—and toward systems that are directly guided by the needs of clinicians in realistic decision-making contexts—"**tasks as needs.**" A case study example illustrates the potential benefits of developing cNLP systems that are designed to more directly support clinical needs.

Key words: artificial intelligence, natural language processing, clinical decision support, clinical judgment, intersectoral collaboration

BACKGROUND

Electronic Health Records (EHR) have been rapidly adopted by hospitals and health clinics worldwide, with the intent of storing and collating data to support clinicians in making decisions at the point of care.[1] However, it can be difficult to extract knowledge from the multiple data formats in EHR, in particular from unstructured texts. Consequently, Natural Language Processing (NLP) has been deployed for automated extraction, decoding, and analysis of free-text EHR data. Since the 1960s, clinical NLP (cNLP) research has led to advances in areas such as clinical note summarization,[2,3] identifying diagnoses,[4,5] and adverse drug reactions.[6,7] However, the cNLP field has experienced an ongoing lack of deployed NLP systems in healthcare settings,[1,8–12] and this problem is arguably growing despite—or possibly because—the increasing sophistication of cNLP systems. It is important that we now work to address the factors preventing the translation of cNLP applications into real-world clinical contexts. In this paper, we survey the cNLP landscape to understand these factors and propose a way forward.

To date, several NLP researchers have explored the factors and circumstances surrounding this lack of cNLP system implementation. Researchers have focused on flaws from the perspective of deficiencies in data (eg, scalability, insufficient standardization),[13,14]

models (eg, overfitting, biases),[15–17] study designs (eg, simplifying assumptions, evaluation limitations),[16,18] and software usability.[19] All these concerns remain valid and continue to be addressed through development of suitable cNLP frameworks, standards, public datasets, and so on. However, we propose that it is the epistemological constraints within the cNLP field that most heavily detract from the intended clinical need and ultimately, deployment into clinical settings.[12,20]

Some cNLP models have been successfully deployed, where they were tasked with simple goals that directly complement clinical decision-making.[21–24] In those few cases of successful (clinically useful) NLP deployments, the models typically targeted the "low hanging fruit," such as efficient disease detection from explicit mentions, or identifying a family history of a disease.[12,22,25] However, there are more numerous examples of as-yet-unadopted NLP systems, often developed within the framework of clinical decision-support systems (CDSS)[8] that attempt to provide clinical recommendations, to use limited data to make predictions about future patient behavior or prognoses, or to audit clinical workflows.[26–31] In short, they aim to make clinical decisions automatically. This includes our own work on early prediction of diagnostic-related group classification for patients,[32] which has yet to find practical application.

These cNLP decision systems are designed under the assumption that their advanced logical or predictive power and far greater access to available clinical information, as well as the capability to synthesize this information, should facilitate more effective decisions than clinicians can make alone. It can even be said that these systems are based on an even stronger assumption, that is, that they facilitate more effective decisions than clinicians—thereby pushing the clinician outside of the decision-making. This then directly violates Friedman's "Fundamental Theorem of Biomedical Informatics,"[33] which states that "A person working in partnership with an information resource is 'better' than that same person unassisted," and from which it follows that tools that seek to be independently more effective than a person unassisted violate a core principle of informatics research.

A chief example of the failed adoption of cNLP-based systems is IBM's Watson Health platform. After the platform's announcement in 2011, IBM invested at least $5 billion into its AI healthcare initiatives and it announced over 50 partnerships with healthcare providers to develop new AI-enabled clinical tools.[20] Yet, nearly the entirety of these projects failed to lead to any useful clinical outcomes or platform deployments, often at great cost to healthcare organizations in terms of time, effort, and funds.[20] It is important to ask why this project failed, especially given the apparent large-scale access the organization had to clinicians. Contributing factors to this failure appear to be that the system was unable to provide information that was not already easily accessible to clinicians,[34] as well as lack of interoperability with EHR systems.[35] In short, the cNLP system did not satisfy the basic needs of the clinicians who would work with it; that is, they were not working in partnership.

A validating example of this problem appears in a case study of a failed AI-based clinical cognitive agent in a hospital in Germany.[36] The authors find "the cognitive agent had been given medical cocompetence with physicians, which meant that the agent wasn't just a support tool but an autonomous operator. For the physicians, this was a step too far."

We propose that the way to improve this situation is for cNLP researchers and clinicians to align in answering the following question:

How can clinical decision-making best be supported through clinical NLP?

By addressing this question, we are implicitly asking about the nature of rational clinical decision-making.[37] The question reveals a set of underlying epistemological misalignments between cNLP designers and medical clinicians. For instance, cNLP systems that aim to predict outcomes or provide recommendations may offer an unrealistic quantification of real-world uncertainty,[38] or they may aim for fixed forms of utility (ie, pre-established, stable outcomes) when these do not reflect the dynamic, real-world outcomes of the task at hand.[16] In contrast, clinicians use a variety of complex reasoning methods[39] and heuristics[40] adapted to complex, dynamic environments, and they may purposefully ignore nonsalient information when making predictions, understanding the importance of trade-offs where there are known uncertainties.[41] We argue that a deeper alignment on the objective of supporting clinical decision-making and identification of the relevant implications for cNLP systems would reduce these epistemological differences and ease the barriers to cNLP system adoption in the context of clinical decision support.

KEY DIFFERENCES BETWEEN NLP AND CLINICIAN-LED DECISION-MAKING

Over 70 years ago, Herbert Simon, a founder of the Carnegie School of business management, introduced a scientific approach to business decision-making. He developed the now well-established precept of "unbounded rationality," which states that an "ideal" decision-maker gathers, assesses, and weighs all relevant information according to some criterion, to maximize the likelihood of achieving their goal(s).[42–44] cNLP researchers such as those that developed IBM Watson appear to have similarly assumed that computers are better positioned than clinicians to achieve unbounded rationality and should (eventually) set the standard for medical decision-making quality.[20] Such claims have already been made for diagnostic imaging applications.[45]

However, individual clinician interpretation of narrowly defined language understanding tasks—rather than ideal decisions—is typically used as the "gold standard" for cNLP system development. For instance, numerous cNLP systems address information extraction of specific patient attributes, such as a patient's smoking status,[46] cardiovascular risk factors[47] or social determinants of health,[48] or focus on normalization of clinical data such as medication prescription details.[49–51] While the information targeted by these tasks clearly is relevant to clinical decision-making, the tasks themselves are nevertheless distant from higher-level clinical tasks such as estimating prognoses or selecting treatments. This then ultimately impedes effective translation of these systems into practical clinical use.

While cNLP researchers often motivate model development in terms of the potential clinical utility of the tools (ie, savings of time and cognitive savings, reduction in human error), this utility often only applies to a small, quantifiable, and simplified portion of the overall clinical problem and clinicians may not see benefits from the small gains obtained there. We are not suggesting that data extraction or standardization of clinical narratives is a practically futile endeavor, but that the research focus should aim to enhance clinical practice, rather than (only) solving narrowly defined NLP problems.[10]

Still our primary concern is not with "low hanging fruit" of diagnostic or simple information extraction tasks within cNLP, with all

of their inherent limitations. Rather, the assumption of unbounded rationality is leading NLP (and other AI researchers)[52,53] researchers too hastily toward the "high hanging fruit": NLP used directly to draw clinical conclusions and make useful recommendations.[27,54–57] In aiming high, these tools appear to be jumping right over the sweet spot of utility for clinical users.

WHERE TO FROM HERE? A NEW EPISTEMOLOGICAL MODEL FOR CNLP

We propose a more straightforward approach to address the barriers to deployment. Clinicians tend to use a "top-down" methodology to decision-making, where they fulfill the context-driven needs of relevant stakeholders using their reasoning, experience, heuristics, or protocols.[40,41] Conversely, cNLP has generally taken a "bottom-up" approach due to inherently data-driven methods, where the system's task is to optimally carry out predefined language analysis functions.[58] In the current model (seen in Figure 1A), the NLP task is seen as an end in itself, often entirely distinct from clinician needs, and divorced from the wider clinical context. We refer to this model as the "**task as decision**" model, because NLP tasks and their evaluation remain intrinsically motivated, tied directly to a narrowly defined modeling objective, and not intended to work in concert with a clinician. An alternative model (captured in Figure 1B), the "**task as need**" model, implies that NLP tasks are designed to interact with, and directly address clinical needs in making decisions. Thus, assessing system utility by whether it answers a "what" question with strong reliability only leads to *inadvertent utility*. Rather, to ensure adoptability, researchers need to shift to addressing how reliably the system answers "why" and "how" questions.[59]

We take the use case of 2 systems that address modeling of hospital readmission of patients with cardiovascular disease[31,60] as an example of a "task as decision" framing that can be reframed via the "task as need", reflecting on what NLP tools would be needed to support the clinical decision of discharging a heart-failure patient. There are numerous cNLP works that have addressed hospital readmission that could be considered for this purpose[61]; the 2 works we focus on are transparent in defining the clinical concepts they target with NLP and therefore are arguably well-suited to this reframing.

Topaz et al[31] analyzed clinical notes using a rule-based (regex-driven) NLP model. At the cohort level, they were able to successfully differentiate between re-admitted and nonreadmitted patients through an aggregated measure known as "ineffective self-management," itself derived from the presence of terms such as "difficulties with outpatient adherence," "excessive fluid intake," and "skips medicines." Navathe et al[60] target several social risk factors by identifying terms related to drug abuse, housing instability, and poor social support in a patient's clinical notes. Both systems are based on the Medical Text Extraction, Reasoning and Mapping System,[62] utilizing dictionary-based term matching and context rules for disambiguation.

In their studies, the authors start with the factors extracted from the notes of patients, and use regression to model readmission outcomes based on these factors. While the identification of relevant patient factors is a seemingly valuable use of cNLP, and arguably in line with the "Tasks as Needs" framing, there are 2 concerns. First, the systems were not designed to fully address the real-world causes of re-admission in heart failure patients, or the full range of clinical factors relevant to making discharge decisions. A cursory review of the health services literature highlights the factors affecting hospital readmissions including the patient conditions these systems primarily target, such as congestive heart failure, chest pain, anxiety and depression, but also hospital operational factors such as nurse staffing care quality, staff responsiveness, length of stay, posthospital care coordination, and medication-related events.[63–67] Many of these factors are entirely ignored by the cNLP systems, and arguably may not even be observable from a patient's data, highlighting the need to design a system that allows a clinician to bring such factors to bear. Second, and critically, the systems are designed to produce a single number (eg, probability of readmission) or to make a recommendation (eg, safe to discharge) rather than explicitly surfacing and presenting the information that will support a clinician to make an informed decision within the broader context. They therefore exemplify the "Task as Decision" framing of cNLP, where the overarching aim of the system is to leverage text processing directly to make decisions.

Clinicians on the ward are likely to ask a number of "why" and "how" questions:

> *"Why should I keep this patient in this ward bed, when there are four in the ER?"*[67]
>
> *"How can I ensure this patient's condition will be better at discharge than when she entered eight days ago?"*[63,68]

Clinicians might use heuristics to weigh up and compare the factors above (patient's physical state and attitude, length of stay, family support, likely postdischarge care) when answering these questions.[40,41] Using these heuristics, they can ignore the inherent uncertainty created by the idiosyncrasies of the individual patient and the complex, changing hospital environment.[41,69]

Therefore, we suggest that the cNLP tools should focus on those evidence-based factors, deemed relevant by clinicians, to *assist* the clinical reasoning rather than replace it. This requires both a cross-disciplinary research focus and stronger collaboration with clinicians in codesign of systems. The use of a cNLP system to detect mental health, attitudinal, or social variables deemed relevant by clinicians—and explicitly presented to them—would better support clinicians to consider the questions that they may ask at discharge.

In Figure 2, we illustrate a sample CDSS tool based on this scenario that aims to help clinicians manage a discharge decision for a chronic heart failure patient in a rehabilitation hospital. The clinician explicitly visualizes a clinical heuristic (here, a tally heuristic but other approaches may better suit specific clinical contexts) to answer "How" and "Why" questions such as those posed above. Tally heuristics have been found to be both fast and effective in complex clinical situations when compared to more complex prediction tools.[40,70] The tally heuristic categorizes factors for and against a given clinical decision and then tallies them together to help the clinician determine a course of action. In this hypothetical tool, the system determines whether the patient is performing better or worse than the average patient over a set of typical benchmarks. The cNLP system could then support this by searching the clinical narrative for each evidence point (ie, each "What" and "How many" question, such as number of social visits, patient progress, and health). The clinician can then make an informed patient care decision.

The tally heuristic used in this example is admittedly very simple. This approach does not preclude more sophisticated modeling, including statistical or predictive modeling that makes use of the same (or additional) variables that are surfaced in the CDSS system. Indeed, data visualization methods could be used to support clinician exploration of surfaced patient characteristics in the record in a

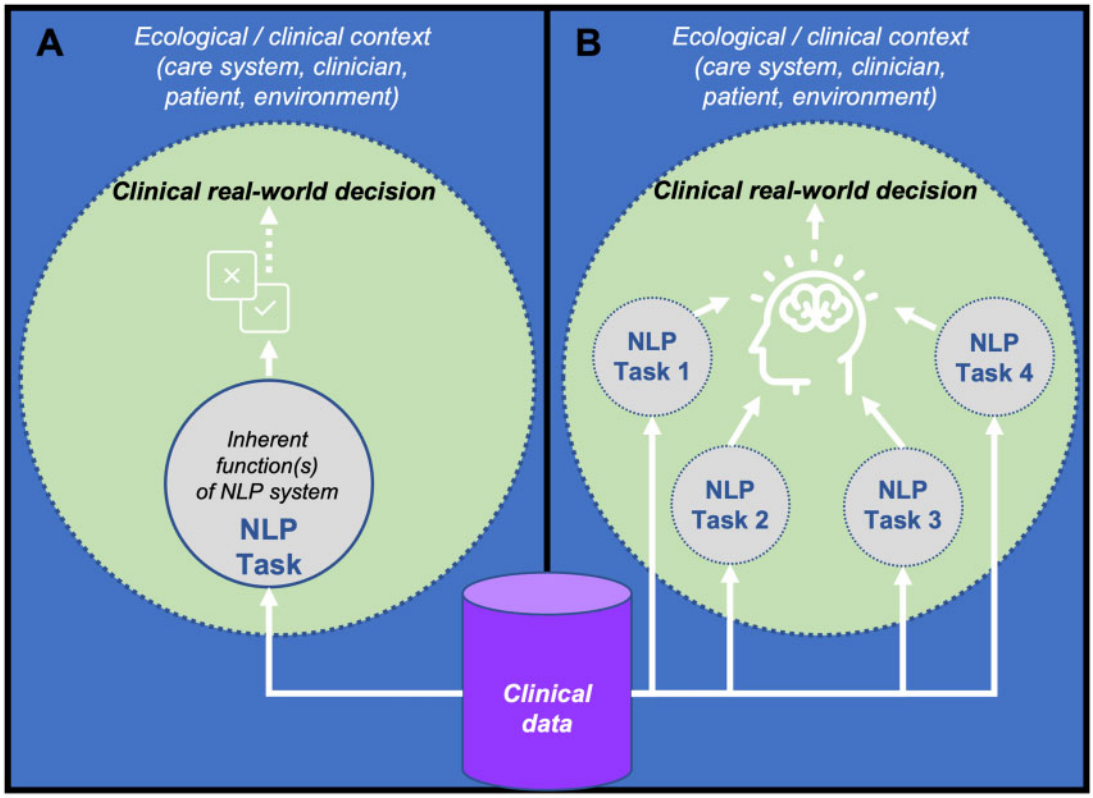

Figure 1. (**A**) Task as decision. The current model (deduced from the literature), where an NLP task (eg, classification, entity recognition) serves a discrete function that relates to a modeling objective, but is not explicitly designed to assist clinician decision-making. Dotted lines represent permeable boundaries; solid lines impermeable. (**B**) Task as need. The alternative model proposed by this paper, where one or more NLP tasks are directly designed to interact and contribute to providing evidence to support clinical decision-making. Dotted lines represent permeable boundaries.

comparative or correlative manner, or statistical weighting could be added to make the CDSS more robust.

The key points here are to identify the factors that are relevant to the clinical decision, to surface them from the notes or other patient data (using cNLP as required), and to present them concisely, providing the information that a clinician needs to make a decision. In short, we have transformed an end-to-end readmission prediction tool into a clinical tool supporting a discharge decision, and shifted the use of NLP as a feature extractor or end-to-end recommendation model in a fully automated tool to targeted evidence gathering, supporting human decisions. We emphasize that this is only a sketch of the concept; codesign with health care professionals is required to determine how such data-supported decision-making technologies can best support specific usage contexts, including identifying specific information needs and defining scenarios of use.[71]

In fact, early examples of cNLP systems exist that illustrated this approach, including the use of cNLP to populate a "Structured Narrative Database" consisting of specific fields extracted from clinical texts to support clinical audits,[72] or to enable identification of infectious patients or establish the need for inhaled anti-inflammatory agents in asthma patients[73] utilizing cNLP to encode reports.[24] The methods utilized by these systems were perhaps overly simplistic (eg, the highly literal rule [*If strings "normal," "good," or "clear" occur BEFORE term "breath sounds," score = 0*] from[73]) and they were not typically designed for active clinical use due to lack of appropriate electronic health record systems[23] or a need for retrospective surveillance or quality assurance. Nevertheless, they prioritized collaboration between medical practitioners and informaticians in defining fine-grained concepts relevant to a given clinical scenario. As methods have increased in sophistication, our sights have shifted from improving findability and organization of information to full automation of decision-making. Perhaps it is time to revisit these earlier scenarios.

DISCUSSION

Our proposal is in line with recent commentary from thought leaders in AI in Medicine. Eric Topol and colleagues have observed that "deployment of medical AI systems in routine clinical care presents an important yet largely unfulfilled opportunity."[74] They cite not only the need for systems that leverage multiple sources of medical data, including clinical texts, but also the need for human-in-the-loop setups that consider how AI can assist decision-making most effectively. While we focus here on cNLP, much of what we say is also applicable to other AI or machine learning systems. We single out cNLP for discussion here because of the clear opportunity in the context of medical texts to define (sub)tasks that enable alignment between the information that is needed for clinical decisions, and the information targeted by cNLP. In particular, the terms and concepts that constitute clinical texts, and the relationships between them, may correspond to precisely the patient details needed to assist clinical decision-making.

There are several key recognized issues that constrain the clinical utility of cNLP that merit discussion in the context of our proposal. These are elaborated below:

Patient: D Smith
Age: 64
Gender: F
Date: 28/03/2021

A

Q1. Choose Heuristic Type: Decision Tree ☐ Tally [X] Take the first good / satisfies ☐

Q1a. Choose Factors: Health-related ▼ Progress-related ▼ Non-medical ▼

Chosen: Patient Attitude/Progress, Current Overall Health, Family Visit Count (since admission)

Q1b. Choose *Tally* comparative estimates: Compare to [mean] of [10] similar inpatient presentations (condition, gender, age bracket)

Q2. Choose Data: EHR – Physician [X] Pathology [X] EHR – Allied Health [X]

B

Tally Results (calculation performed in 8.7 sec)

Heuristic Factor	Patient: D Smith			Comparison to mean of *10* previous similar inpatients
Rehab Performance (inc. mean trend of last 72 hours)	Rehabilitation Domain	**Attitude** (Normalised Clinician Sentiment - click score for source texts/sentiment trend)	**Progress** (Normalised Rehab Progress - click score for related texts/progress trend)	Patient performing slightly above average.
	Physio	+7/10 (increasing)	4/5 (stable)	
	OT	+6/10 (stable)	3/5 (stable)	
	Dietician	+3/10 (stable)	2/5 (increasing)	
Current Overall Health	Nutrition (Pathology) results (normalised)	Blood, vitals – all within reference range, but often at the lower end. Low Vit. D, B12 and Fe. 5/10		Patient has slightly below average nutrition.
	Medications (side effects/ ADRs)	Nil (10/10)		Patient has slightly better response to medications.
Family Visit Count During Hospital Stay	4	"…patient's brother visited…" (Nurse entry, 27/1/2020) "…patient's brother was sitting…" (Nurse entry, 25/1/2020) "…mother was waiting by the…" (Physician entry, 23/1/2020) "…for coffee w/ sister post-physio…" (Nurse entry, 18/1/2020)		Patient has more visits than average
Normalised Total vs Mean of *10* previous similar inpatients		**A**ttitude = 16/30. Mean Attitude = 14/30. +1 to tally **R**. Progress = 9/15. Mean Progress = 10/15. -1 to tally **N**utrition = 5/10. Mean Nutrition = 7/10. -1 to tally **M**SE/ADRs = 10/10. Mean Medications = 9/10. +1 to tally **F**amily visits = 4 per 12 days. Mean Visits = 2 per 12 days. +1 to tally. Tally = S + R + N + M + F = **+1 – 1 -1 + 1 + 1 = +1.**		Tally suggest patient is performing slightly better than average overall.
Clinical Decision Making		*Clinician reasoning:* Patient performing well. Almost reached typical progress level for this stage of rehabilitation. Poorer nutrition than similar patients. Relatively high social support. *Clinical decision:* **Propose three more days of rehabilitation to patient. Speak to dietician about B12 injection and iron sources in diet. Discuss patient's needs after discharge with siblings.**		NOTE: Similar patients stayed 5 more days (a total of 17 days)

Figure 2. An indicative CDSS relevant to a hospital discharge decision scenario. (**A**) The treating clinician flexibly selects the heuristic form (in this case, a tally), with context and task-relevant inputs and data sources, and a relevant comparative baseline. (**B**) The platform then produces and tallies the scores for all heuristic factors, leveraging NLP, helping the clinician to determine a suitable course of action.

Issue 1: under-performance of cNLP systems on complex language processing tasks

One key reason for the limited cNLP deployment is that the NLP platforms developed so far have fundamental limitations that impede their comprehension of real-world clinical data. Clinical NLP pipelines generally struggle with linguistic issues such as word or phrase ambiguity, complex semantic roles (eg, differentiating subject and object), temporality of events, or information that a clinician would intuitively flag as "missing" and fill in implicitly (eg, assumption of a positive characteristic due to a lack of documentation of the *absence* of that feature[75]). Furthermore, more basic challenges such as word misspellings or significant variations between language used across organizations impact performance.[12,20] It has also been observed that many modern NLP systems are "like a mouth without a brain"[76] or "stochastic parrots"[77]—most particularly evident in applications such as medical report generation, where reports have been optimized to "*look real* rather than to *predict right*"[78] and are biased toward normal findings.[79] cNLP systems will need to somehow incorporate a certain degree of pragmatism or "common sense" before the aforementioned issues can be fixed. Therefore, the linguistic analysis limitations and lack of pragmatic intelligence of cNLP—commonly referred to as "weak AI"[80]—limit its deployment within complex healthcare environments. Adopting an approach that focuses on more achiev-

able NLP tasks with explicit relevance to a specific clinical decision task may mitigate against this problem.

Issue 2: simplification of real-world problems

A related concern for cNLP is its simplification of real-world clinical problems. NLP model pipelines generally simplify a real-world problem into a linear "goal—task—solution" model; that is, a "task as decision" approach. As discussed above, cNLP systems are usually in line with the NLP tasks of information extraction, involving migrating unstructured data to a standardized form,[8,81] or classification of texts into categories such as for disease case detection.[82,83] While cNLP platforms (whether independent, or as components of CDSS) do not need to "understand" a task as clinicians do, they would need deeper sophistication to provide useful recommendations, predictions, or clinical workflow improvements.[20] Furthermore, simplification (at best) leads to an NLP task and outcome that comprises 1 or 2 components of the many clinical tasks involved in patient care. Thus, cNLP is typically used for a select set of simplified "what" questions, but not the more complex, and clinically useful "how" or "why" questions.[59] Additionally, cNLP models usually produce binary outputs (presence/absence, true/false) for one or more medical variables often without considering how they meaningfully associate with other medical conditions or events. Several reviews of clinical information extraction applications have found that the vast majority of cNLP models involved an attempt to automatically detect the presence or absence of a disease or injury, adverse medical or treatment events, or patient characteristics, with a small proportion also extracting numeric values from narrative text.[1,84,85] These lend themselves to "needs," better than "decisions," and may find more relevance in a decision support context.

Issue 3: explainability

What we are proposing is distinct from the current focus in the artificial intelligence community on explainability and interpretability of sophisticated *black box* statistical or machine learning-based models.[86] As Holzinger and colleagues have put it, "Explainability is at least as old as AI itself and rather a problem that has been caused by it."[87] While we wholeheartedly support efforts to make AI model decisions explainable, here, we instead suggest that not every decision is suited directly to AI. Perhaps by more carefully scoping the tasks we demand of our AI, and directly engaging powerful human intelligence capabilities, we can both arrive at more effective clinical decisions and avoid the need to immediately solve the problem of explainability.

CONCLUSION: MOVING FROM PROBLEMS TO SOLUTIONS

We have provided an indicative example that illustrates one of many possible ways to integrate NLP into existing clinical decision-making. This system could improve clinician trust by meeting clinician needs and complementing their judgment, and such systems may be a good starting point for future cNLP deployments.

Researchers need to work closely with clinicians to explicitly and flexibly incorporate and operationalize their needs into CDSS or other systems, so that cNLP tools can usefully contribute to decisions in partnership with clinicians. This allows for a shift from the "task as decision" model to a "task as need" model. Further codesign with researchers in other areas including implementation science, CDSS, user experience and clinical decision analysis would also support development of more meaningful cNLP systems.[88,89]

The "tasks as needs" model represents a paradigm shift in cNLP to focus on *supporting* clinicians rather than *emulating* clinicians, and to bring cNLP in line with Friedman's Fundamental Theorem. Under this approach, cNLP researchers will produce systems that more effectively integrate into clinical workflows and facilitate a closer working relationship between clinicians and their decision-support tools. It is in this way that cNLP researchers can realize immediate clinical gains and can increase the likelihood that clinicians will benefit from adopting cNLP systems.

FUNDING

This work was supported by the Australian National Health and Medical Research Council (NHMRC) Centre for Research Excellence in Digital Health (CREDiH), grant number APP1134919, to author KV.

AUTHOR CONTRIBUTIONS

KV and RL conceptualized the scope of the paper and planned the study. AL conducted detailed literature review, synthesized the material, defined the case study, and drafted the manuscript. All authors contributed to critically revising the manuscript. KV led the additions and revisions that resulted from the review process. All authors approved the final version to be published.

ACKNOWLEDGMENTS

The authors would like to thank Kirk Roberts and 2 anonymous reviewers who provided significant feedback on an earlier version of this manuscript that significantly shaped our thinking and helped to improve the manuscript substantially.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Wang Y, Wang L, Rastegar-Mojarad M, *et al*. Clinical information extraction applications: a literature review. *J Biomed Inform* 2018; 77: 34–49.
2. Wilcox A, Jones SS, Dorr DA, *et al*. Use and impact of a computer-generated patient summary worksheet for primary care. *AMIA Annu Symp Proc* 2005; 2005: 824–8.
3. Bashyam V, Hsu W, Watt E, Bui AA, Kangarloo H, Taira RK. Problem-centric organization and visualization of patient imaging and clinical data. *Radiographics* 2009; 29 (2): 331–43.
4. Nguyen AN, Moore J, O'Dwyer J, Philpot S. Assessing the utility of automatic cancer registry notifications data extraction from free-text pathology reports. *AMIA Annu Symp Proc* 2015: 953–62.
5. Wang M, Cyhaniuk A, Cooper DL, Iyer NN. Identification of persons with acquired hemophilia in a large electronic health record database. *Blood* 2015; 126 (23): 3271.
6. Haerian K, Varn D, Vaidya S, Ena L, Chase H, Friedman C. Detection of pharmacovigilance-related adverse events using electronic health records and automated methods. *Clin Pharmacol Ther* 2012; 92 (2): 228–34.
7. Sohn S, Kocher J-PA, Chute CG, Savova GK. Drug side effect extraction from clinical narratives of psychiatry and psychology patients. *J Am Med Inform Assoc* 2011; 18 (Supplement_1): i144–9.

8. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? *J Biomed Inform* 2009; 42 (5): 760–72.
9. Chapman W, Nadkarni P, Hirschman L, D'avolio L, Savova G, Uzuner O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *J Am Med Inform Assoc* 2011; 18 (5): 540–3.
10. Pivovarov R, Elhadad N. Automated methods for the summarization of electronic health records. *J Am Med Inform Assoc* 2015; 22 (5): 938–47.
11. Kreimeyer K, Foster M, Pandey A, *et al.* Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *J Biomed Inform* 2017; 73: 14–29.
12. Neujar Bryan T. Healthcare NLP: The Secret to Unstructured Data's Full Potential. Secondary Healthcare NLP: The Secret to Unstructured Data's Full Potential 2019. https://www.healthcatalyst.com/insights/how-healthcare-nlp-taps-unstructured-datas-potential. Accessed July 11, 2022.
13. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform* 2008; 17 (1): 128–44.
14. Stanfill MH, Williams M, Fenton SH, Jenders RA, Hersh WA. Systematic literature review of automated clinical coding and classification systems. *J Am Med Inform Assoc* 2010; 17 (6): 646–51.
15. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet* 2012; 13 (6): 395–405.
16. Friedman C, Hripcsak G. Evaluating natural language processors in the clinical domain. *Methods Inf Med* 1998; 37 (4–5): 334–44.
17. Velupillai S, Suominen H, Liakata M, *et al.* Using clinical Natural Language Processing for health outcomes research: overview and actionable suggestions for future advances. *J Biomed Inform* 2018; 88: 11–9.
18. Ferro N, Fuhr N, Grefenstette G, *et al. The Dagstuhl Perspectives Workshop on Performance Modeling and Prediction*. New York, NY: ACM SIGIR Forum: ACM; 2018: 91–101.
19. Zheng K, Vydiswaran VGV, Liu Y, *et al.* Ease of adoption of clinical natural language processing software: an evaluation of five systems. *J Biomed Inform* 2015; 58: S189–96.
20. Strickland E. IBM Watson, heal thyself: how IBM overpromised and underdelivered on AI health care. *IEEE Spectr* 2019; 56 (4): 24–31.
21. Gálvez JA, Pappas JM, Ahumada L, *et al.* The use of natural language processing on pediatric diagnostic radiology reports in the electronic health record to identify deep venous thrombosis in children. *J Thromb Thrombolysis* 2017; 44 (3): 281–90.
22. Bozkurt S, Gimenez F, Burnside ES, Gulkesen KH, Rubin DL. Using automatically extracted information from mammography reports for decision-support. *J Biomed Inform* 2016; 62: 224–31.
23. Fiszman M, Chapman WW, Aronsky D, Evans RS, Haug PJ. Automatic detection of acute bacterial pneumonia from chest X-ray reports. *J Am Med Inform Assoc* 2000; 7 (6): 593–604.
24. Knirsch CA, Jain NL, Pablos-Mendez A, Friedman C, Hripcsak G. Respiratory isolation of tuberculosis patients using clinical guidelines and an automated clinical decision support system. *Infect Control Hosp Epidemiol* 1998; 19 (2): 94–100.
25. Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton PD. Unlocking clinical data from narrative reports: a study of natural language processing. *Ann Intern Med* 1995; 122 (9): 681–8.
26. Harkema H, Chapman WW, Saul M, Dellon ES, Schoen RE, Mehrotra A. Developing a natural language processing application for measuring the quality of colonoscopy procedures. *J Am Med Inform Assoc* 2011; 18 (Supplement_1): i150–6.
27. Rumshisky A, Ghassemi M, Naumann T, *et al.* Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Transl Psychiatry* 2016; 6 (10): e921.
28. Baer B, Nguyen M, Woo E, *et al.* Can natural language processing improve the efficiency of vaccine adverse event report review? *Methods Inf Med* 2016; 55 (2): 144–50.
29. Buchan K, Filannino M, Uzuner Ö. Automatic prediction of coronary artery disease from clinical narratives. *J Biomed Inform* 2017; 72: 23–32.
30. Zhang X, Kim J, Patzer RE, Pitts SR, Patzer A, Schrager JD. Prediction of emergency department hospital admission based on natural language processing and neural networks. *Methods Inf Med* 2017; 56 (5): 377–89.
31. Topaz M, Radhakrishnan K, Blackley S, Lei V, Lai K, Zhou L. Studying associations between heart failure self-management and rehospitalizations using natural language processing. *West J Nurs Res* 2017; 39 (1): 147–65.
32. Liu J, Capurro D, Nguyen A, Verspoor K. Early prediction of diagnostic-related groups and estimation of hospital cost by processing clinical notes. *NPJ Digit Med* 2021; 4 (1): 103.
33. Friedman CP. A "Fundamental Theorem" of biomedical informatics. *J Am Med Inform Assoc* 2009; 16 (2): 169–70.
34. Hagen T, Narozniak R. IBM seeks niche for Watson for oncology. *OncologyLive* 2019; 20 (20): 76–8.
35. Herper M. MD Anderson Benches IBM Watson in Setback for Artificial Intelligence in Medicine. *Forbes*. 2017. https://www.forbes.com/sites/matthewherper/2017/02/19/md-anderson-benches-ibm-watson-in-setback-for-artificial-intelligence-in-medicine/?sh=73f02a237748. Accessed July 11, 2022.
36. Reis L, Maier C, Mattke J, Creutzenberg M, Weitzel T. Addressing user resistance would have prevented a healthcare AI project failure. *MIS Q Executive* 2020; 19 (4): 279–36.
37. Nozick R. *The Nature of Rationality*. Princeton, NJ: Princeton University Press; 1994.
38. Bradley R, Drechsler M. Types of uncertainty. *Erkenn* 2014; 79 (6): 1225–48.
39. Arocha JF, Wang D, Patel VL. Identifying reasoning strategies in medical decision making: a methodological guide. *J Biomed Inform* 2005; 38 (2): 154–71.
40. Marewski JN, Gigerenzer G. Heuristic decision making in medicine. *Dialogues Clin Neurosci* 2012; 14 (1): 77–89.
41. Kozyreva A, Hertwig R. The interpretation of uncertainty in ecological rationality. *Synthese* 2021; 198 (2): 1517–31.
42. Simon HA. Rational choice and the structure of the environment. *Psychol Rev* 1956; 63 (2): 129–38.
43. Simon HA, March JG. *Organizations*. New York: John Wiley & Sons; 1958.
44. Wilden R, Hohberger J, Devinney TM, Lumineau F. 60 years of March and Simon's organizations: an empirical examination of its impact and influence on subsequent research. *J Manage Stud* 2019; 56 (8): 1570–604.
45. Liu X, Faes L, Kale AU, *et al.* A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health* 2019; 1 (6): e271–97.
46. Caccamisi A, Jørgensen L, Dalianis H, Rosenlund M. Natural language processing and machine learning to enable automatic extraction and classification of patients' smoking status from electronic medical records. *Ups J Med Sci* 2020; 125 (4): 316–24.
47. Khalifa A, Meystre S. Adapting existing natural language processing resources for cardiovascular risk factors identification in clinical notes. *J Biomed Inform* 2015; 58: S128–32.
48. Reeves RM, Christensen L, Brown JR, *et al.* Adaptation of an NLP system to a new healthcare environment to identify social determinants of health. *J Biomed Inform* 2021; 120: 103851.
49. Sohn S, Clark C, Halgrim SR, Murphy SP, Chute CG, Liu H. MedXN: an open source medication extraction and normalization tool for clinical text. *J Am Med Inform Assoc* 2014; 21 (5): 858–65.
50. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. *J Am Med Inform Assoc* 2010; 17 (1): 19–24.
51. MacKinlay AD, Verspoor KM. Extracting structured information from free-text medication prescriptions using dependencies. In: proceedings of the ACM sixth international workshop on data and text mining in biomedical informatics. Maui, HI, USA: Association for Computing Machinery; 2012: 35–40.
52. Wiens J, Saria S, Sendak M, *et al.* Do no harm: a roadmap for responsible machine learning for health care. *Nat Med* 2019; 25 (9): 1337–40.

53. Yin J, Ngiam KY, Teo HH. Role of artificial intelligence applications in real-life clinical practice: systematic review. *J Med Internet Res* 2021; 23 (4): e25759.
54. Holzinger A, Biemann C, Pattichis, CS, Kell DB. What do we need to build explainable AI systems for the medical domain? arXiv preprint arXiv:1712.09923 2017. https://doi.org/10.48550/arXiv.1712.09923.
55. Choi GH, Yun J, Choi J, *et al*. Development of machine learning-based clinical decision support system for hepatocellular carcinoma. *Sci Rep* 2020; 10 (1): 1–10.
56. Bacchi S, Gluck S, Tan Y, *et al*. Prediction of general medical admission length of stay with natural language processing and deep learning: a pilot study. *Intern Emerg Med* 2020; 15 (6): 989–7.
57. Banerjee I, Bozkurt S, Caswell-Jin JL, Kurian AW, Rubin DL. Natural language processing approaches to detect the timeline of metastatic recurrence of breast cancer. *JCO Clin Cancer Inform* 2019; 3 (3): 1–12.
58. Agre P. The practical logic of computer work. In: Scheutz M, ed. *Computationalism*. Cambridge, MA: MIT Press; 2002: 129–42.
59. WIgnall J, Barry D. ReHumanizing hospital satisfaction data: text analysis, the lifeworld, and contesting stakeholders' beliefs in evidence. In: ethnographic praxis in industry conference proceedings. Wiley Online Library; 2018: 427–56.
60. Navathe AS, Zhong F, Lei VJ, *et al*. Hospital readmission and social risk factors identified from physician notes. *Health Serv Res* 2018; 53 (2): 1110–36.
61. Mahmoudi E, Kamdar N, Kim N, Gonzales G, Singh K, Waljee AK. Use of electronic medical records in development and validation of risk prediction models of hospital readmission: systematic review. *BMJ* 2020; 369: m958.
62. Goss FR, Plasek JM, Lau JJ, Seger DL, Chang FY, Zhou L. An evaluation of a natural language processing tool for identifying and encoding allergy information in emergency department clinical notes. *AMIA Annu Symp Proc* 2014; 2014: 580–8.
63. Taylor NJ, Lederman R, Bosua R. Medical record support for effective discharge planning. In: proceedings of the 25th European Conference on Information Systems (ECIS); June 5–10, 2017: 2825–33; Guimarães, Portugal.
64. Heggestad T. Do hospital length of stay and staffing ratio affect elderly patients' risk of readmission? A nation-wide study of Norwegian hospitals. *Health Serv Res* 2002; 37 (3): 647–65.
65. Kirby SE, Dennis SM, Jayasinghe UW, Harris MF. Patient related factors in frequent readmissions: the influence of condition, access to services and patient choice. *BMC Health Serv Res* 2010; 10 (1): 216–8.
66. Ogunneye O, Rothberg MB, Friderici J, Slawsky MT, Gadiraju VT, Stefan MS. The association between skilled nursing facility care quality and 30-day readmission rates after hospitalization for heart failure. *Am J Med Qual* 2015; 30 (3): 205–13.
67. Glette MK, Kringeland T, Røise O, Wiig S. Hospital physicians' views on discharge and readmission processes: a qualitative study from Norway. *BMJ Open* 2019; 9 (8): e031297.
68. Ommaya AK, Cipriano PF, Hoyt DB, *et al*. Care-centered clinical documentation in the digital environment: solutions to alleviate burnout. *NAM Perspectives*. Discussion Paper. Washington, DC: National Academy of Medicine; 2018.
69. Hogarth RM, Karelaia N. Heuristic and linear models of judgment: matching rules and environments. *Psychol Rev* 2007; 114 (3): 733–58.
70. Brighton H. Robust inference with simple cognitive models. In: AAAI spring symposium: between a rock and a hard place: cognitive science principles meet AI-hard problems; 2006: 17–22.
71. Tendedez H, Ferrario M-A, McNaney R, Gradinar A. Exploring human–data interaction in clinical decision-making using scenarios: co-design study. *JMIR Hum Factors* 2022; 9 (2): e32456.
72. Lyman M, Sager N, Tick L, Nhan N, Borst F, Scherrer J-R. The application of natural-language processing to healthcare quality assessment. *Med Decis Making* 1991; 11 (4_suppl): S65–8.
73. Ertle AR, Campbell EM, Hersh WR. Automated application of clinical practice guidelines for asthma management. *Proc AMIA Annu Fall Symp* 1996; 552–6.
74. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med* 2022; 28 (1): 31–8.
75. Woodward MA, Maganti N, Niziol LM, Amin S, Hou A, Singh K. Development and validation of a natural language processing algorithm to extract descriptors of microbial keratitis from the electronic health record. *Cornea* 2021; 40 (12): 1548–53.
76. Marasovic A. NLP's generalization problem, and how researchers are tackling it. *The Gradient*. Stanford, CA: Stanford Artificial Intelligence Laboratory (SAIL); 2018.
77. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? In: proceedings of the 2021 ACM conference on fairness, accountability, and transparency. Virtual Event, Canada: Association for Computing Machinery; 2021: 610–23.
78. Liu G, Hsu T-MH, McDermott M, *et al*. Clinically accurate chest x-ray report generation. In: proceedings of the 4th machine learning for healthcare conference, PMLR; 2019; Ann Arbor, MI.
79. Li CY, Liang X, Hu Z, Xing EP. Hybrid retrieval-generation reinforced agent for medical image report generation. In: p*roceedings of the 32nd international conference on neural information processing systems*. Montréal, Canada: Curran Associates Inc.; 2018: 1537–47.
80. Searle J. Minds, brains, and programs. *Behav Brain Sci* 1980; 3 (3): 417–57.
81. Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. *J Am Med Inform Assoc* 2011; 18 (2): 181–6.
82. Combi C, Pozzi G. Clinical information systems and artificial intelligence: recent research trends. *Yearb Med Inform* 2019; 28 (1): 83–94.
83. Al Mamlook RE, Chen S, Bzizi HF. Investigation of the performance of machine learning classifiers for pneumonia detection in chest X-ray images. In: 2020 IEEE international conference on electro information technology (EIT). Chicago, IL: IEEE; 2020: 98–104.
84. Wen A, Fu S, Moon S, *et al*. Desiderata for delivering NLP to accelerate healthcare AI advancement and a Mayo Clinic NLP-as-a-service implementation. *Npj Digit Med* 2019; 2 (1): 1–7.
85. Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. *JMIR Med Inform* 2020; 8 (3): e17984.
86. Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access* 2018; 6: 52138–60.
87. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip Rev Data Min Knowl Discov* 2019; 9 (4): e1312.
88. Passi S, Jackson SJ. Trust in data science: collaboration, translation, and accountability in corporate data science projects. *Proc ACM Hum-Comput Interact* 2018; 2 (CSCW): 1–28.
89. Mao Y, Wang D, Muller M, *et al*. How data scientists work together with domain experts in scientific collaborations: to find the right answer or to ask the right question? *Proc ACM Hum-Comput Interact* 2019; 3 (GROUP): 1–23.