How does work environment relate to diagnostic quality? A prospective, mixed methods study in primary care

Maram Khazen, Erin E Sullivan, Sophia Arabadjis, Jason Ramos, Maria Mirica, Andrew Olson, Mark Linzer, Gordon D Schiff

ABSTRACT

Objectives The quest to measure and improve diagnosis has proven challenging; new approaches are needed to better understand and measure key elements of the diagnostic process in clinical encounters. The aim of this study was to develop a tool assessing key elements of the diagnostic assessment process and apply it to a series of diagnostic encounters examining clinical notes and encounters’ recorded transcripts. Additionally, we aimed to correlate and contextualise these findings with measures of encounter time and physician burnout.

Design We audio-recorded encounters, reviewed their transcripts and associated them with their clinical notes and findings were correlated with concurrent Mini Z Worklife measures and physician burnout.

Setting Three primary urgent-care settings.

Participants We conducted in-depth evaluations of 28 clinical encounters delivered by seven physicians.

Results Comparing encounter transcripts with clinical notes, in 24 of 28 (86%) there was high note/transcript concordance for the diagnostic elements on our tool. Reliably included elements were red flags (92% of notes/encounters), aetiologies (88%), likelihood/uncertainties (71%) and follow-up contingencies (71%), whereas psychosocial/contextual information (35%) and mentioning common pitfalls (7%) were often missing. In 22% of encounters, follow-up contingencies were in the note, but absent from the recorded encounter. There was a trend for higher burnout scores being associated with physicians less likely to address key diagnosis items, such as psychosocial history/context.

Conclusions A new tool shows promise as a means of assessing key elements of diagnostic quality in clinical encounters. Work conditions and physician reactions appear to correlate with diagnostic behaviours. Future research should continue to assess relationships between time pressure and diagnostic quality.

INTRODUCTION

Improving diagnosis is a top priority for medical quality and patient safety. Based on data from patient surveys, malpractice claims and polling of safety professionals, reinforced by a report from the National Academy of Medicine (NAM), diagnostic errors and the imperative for diagnosis quality improvement have been recognised as important needs. While there have been multiple reports on epidemiology and magnitude of diagnostic errors, there has been a paucity of in-depth qualitative studies examining diagnostic processes to clarify what actually occurs during clinical encounters, as well as clinicians’ decision-making as represented in their clinical notes. Further, most qualitative studies examining clinician behaviour are paper scenario-based exercises rather than examining the more complex realm of real-world diagnostic encounters.

A central challenge in improving diagnosis has been the issue of measurement: how to reliably, efficiently and effectively measure diagnosis quality. Unlike other areas of quality and safety, creating a standardised method for measuring diagnostic errors has proven challenging. In this regard, NAM and the National Quality Forum cautioned that the field of diagnostic metrics is still ‘emerging’ and ‘needing further exploration and development’, before any implementation. Measuring simpler, narrower aspects of the diagnostic process, such as what per cent of ordered tests are performed and/or followed-up when abnormal, is easier to conceptualise and measure than crafting metrics evaluating clinicians’ diagnostic reasoning or verbal or written communication, including evaluating quality of clinical
notes in the electronic health record. In this study, we aimed to examine these more difficult-to-measure elements of diagnostic process by designing and applying a tool to evaluate physicians’ diagnostic assessments during urgent care encounters as well as in their clinical notes.

Because diagnosis often occurs in time-constrained, stressful and chaotic settings, we sought to contextualise our findings with a more complete understanding of conditions where clinicians performed their diagnostic work. In addition to applying our tool to assess elements of the diagnostic process, we collected data on time physicians took for encounters as well as perceptions of time adequacy and work conditions by querying them regarding the time they felt was required for specific patient encounters. We concurrently measured these physicians’ level of stress and burnout using previously validated standardised tools.

Thus, our study aimed to:
1. Iteratively develop, refine and apply a tool to assess key elements of diagnostic assessment in actual clinical encounters,
2. Apply the tool to evaluate a series of diagnostic encounters by examining audiotaped encounter transcripts and the associated clinical notes and
3. Correlate and contextualise findings with measures of encounter time and clinician perceptions of work conditions and burnout.

METHODS
Study design
The MD-SOS study (Medical Diagnosis: Safety or Stress) was designed to better understand diagnosis quality and its relationship to clinician stress and burnout. We iteratively developed and refined a tool, based on a previously piloted instrument, to examine selected elements of clinical notes and that visit’s audiotaped encounter. We conceptualised and built the tool based on key diagnostic elements suggested in the literature and prior work of the investigator team around diagnosis quality (Table 1).

These included elements such as explicitly addressing chief concern(s), listing a differential diagnosis, commenting on probabilities (degree of certainty and associated uncertainties), mention of presence (or absence) of diagnostic red flags, consideration of ‘don’t miss’ diagnoses, mention of relevant pitfalls (eg, test limitations or atypical presentations) to consider, diagnostic follow-up testing, plans and time frames, mention of specific parameters for patients to monitor, with expected time frames, and avoidance of pejorative language. The tool, Inclusion of Diagnostic Elements Assessment (IDEA) was designed to evaluate diagnostic assessment and documentation and consisted of five categories with 15 items for assessing diagnostic quality: (1) completeness and quality of assessment, (2) diagnostic situational awareness with potential red flags and don’t miss diagnoses, (3) adequacy of diagnostic follow-up plan,

<p>| Table 1 | Inclusion of diagnostic elements assessment tool items of the encounter and note |</p>
<table>
<thead>
<tr>
<th>Assessment method</th>
<th>Clinical notes</th>
<th>Transcripts of audiotaped encounters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scoring</td>
<td>5-point scale</td>
<td>Present, absent</td>
</tr>
<tr>
<td></td>
<td>0—absent,</td>
<td>Present, absent</td>
</tr>
<tr>
<td></td>
<td>1—minimal,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2—less than</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3—neutral,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4—good,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5—excellent,</td>
<td></td>
</tr>
<tr>
<td>Items assessed</td>
<td>20 items</td>
<td>15 items</td>
</tr>
<tr>
<td>Comparing between transcripts and corresponding notes</td>
<td>1, 2—absent.</td>
<td>3, 4, 5—present.</td>
</tr>
<tr>
<td>Diagnostic assessment items</td>
<td>5 items</td>
<td>6 items</td>
</tr>
<tr>
<td>Addresses chief issue</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Includes differential diagnosis</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Mentions aetiologies</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Delving into aetiologies</td>
<td>–</td>
<td>+</td>
</tr>
<tr>
<td>Contextual/ psychosocial info mentioned</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Comment on likelihood, uncertainties</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Diagnostic follow-up plan items mentioned/noted</td>
<td>4 items</td>
<td>5 items</td>
</tr>
<tr>
<td>Diagnostic test</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Explanation of test/rationale</td>
<td>–</td>
<td>+</td>
</tr>
<tr>
<td>Contingencies: what to watch for</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Expected time frames: (eg, for sx improvement)</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Rational</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Situational awareness and ‘safety net’ items</td>
<td>3 items</td>
<td>3 items</td>
</tr>
<tr>
<td>Red flags</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Do not miss diagnosis</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Pitfalls</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Global assessment of documentation quality</td>
<td>5 items</td>
<td>1 item</td>
</tr>
</tbody>
</table>

Continued
(4) overall note quality (organisation, succinctness) and (5) global subjective rating of documentation diagnosis quality. Data collection took place between November 2019 and March 2020. (The study was terminated early due to institutional COVID-19 restrictions imposed on in-person visits and research).

We examined the tool’s construct validity, including valid content, response process, internal structure, pairwise correlation and relations to other variables and consequences.37 Our analysis suggested that we reached good content validity given that the items were based on our previous work and the extensive experience of experts in the field. The items of the tool were rated according to a standard 5-point Likert scale, where the scoring process was carefully documented, with all raters keeping track of their comments and questions. The internal consistency/reliability was satisfactory with a kappa for inter-rater reliability (kappa=0.8), and consequences aspects of validity (clinical utility) showing favourable ability to categorise charts for diagnostic quality. We could not assess relationship to other variables (external validity) since our tool is a novel instrument. We also examined pairwise correlations across the 12 items that were used to assess both the clinical notes and transcribed encounters. The correlation ranged from −0.18 (items: differential diagnosis and chief issue) to 0.69 (items: psychosocial context and differential diagnosis). The mean absolute correlation across all items was 0.26. Given the low pairwise correlation (r<0.8), we assume low collinearity between items.

**Study setting, participants and recruitment**

Study participants were primary care physicians and their patients, recruited from Brigham and Women’s Hospital at three different urgent care settings. Two sites were affiliated with primary care clinics, and one was a walk-in urgent care centre affiliated with the emergency department.

Through a convenience sample, physicians were approached by the principal investigator (PI) (GS) via an email explaining the study and inviting them to participate. Of the 10 physicians invited, 7 agreed to participate. They completed a written consent form which explained the study and permitted them to raise and resolve any questions or concerns prior to patient recruitment. After obtaining consent, two researchers (MK and JR) attended sessions in the physician’s clinic where they approached patients in the waiting room prior to encounters to request permission to record their encounter. The researchers explained to patients that the study was about quality improvement, and they reviewed with them protections to ensure confidentiality. Demographics about patients were not collected, to fully anonymise the audiotaped encounters. To mitigate any bias, the researchers mentioned their specific role in the study without providing additional information about personal views. Participating patients and physicians were provided US$25 and US$75 honoraria, respectively.

**Procedures**

For consenting patients, we audio-recorded clinical encounters and collected accompanying visit notes from the electronic health record, or EHR (Epic, Verona, Wisconsin, USA). A research assistant started the recorder and timed the visit length but was not in the room during the encounter. At the end of each session, digital speech files from the encounters were collected and stored in secure password-protected file areas. Digital audio recordings were transcribed using a specially designed interface with Amazon Web Services that redacts HIPAA (Health Insurance Portability and Accountability Act) - protected health information such as patient and physician names. A trained research assistant (JR) manually edited each transcript for readability and any speech recognition errors. Only encounters with English-speaking patients were included.

To explore connections between work conditions, stress and burnout and diagnostic quality (element inclusion), physicians completed a Worklife and Wellness Mini Z survey form shortly after the session where their visits were recorded. At the end of each encounter, researchers asked physicians about the perceived time needed for the encounter.

**Table 1** Continued

<table>
<thead>
<tr>
<th>Assessment method</th>
<th>Clinical notes</th>
<th>Transcripts of audiotaped encounters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Succinctness</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>Clinician readability</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>Pt readability</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>Pejorative Comments</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Excessive Templating/Copy-paste</td>
<td>+</td>
<td>–</td>
</tr>
</tbody>
</table>

Global assessment of diagnostic assessment in note

3 items –

Quality of diagnosis

+ –

Needed tests ordered

+ –

Avoids over testing

+ –

*Excluded from analysis because no instances of pejorative language were found in any note or transcript.

pt, patient; Sx, symptom.
Data analysis

The study included two groups of coders. Clinical notes were analysed by the first group comprising two coders (a trained male research assistant (JR) and a general internist PI (GS) who reviewed the written notes. Encounter transcripts were independently coded by two PhD-level social science female researchers (MK and EES) using a modified version of the IDEA tool along with additional qualitative coding using NVivo V.12 software.

Clinical notes

We adapted the IDEA chart review tool from our previous study (Assessment of The Assessment) and pilot-tested the coding agreement and clarified operational definitions by having five team members each analyse three notes and then resolve discrepancies. The two coders then coded five additional notes applying the tool’s 15 items. Disagreements were resolved to ensure consistent coding. The research assistant coded the remaining notes, conferring with the PI (GS) to resolve uncertainties, difficult judgements or questions. Issues arising in the tool refinement and coding of notes were further discussed by the entire team (three general internists and three PhD qualitative researchers). Ten charts were independently re-reviewed by one study physician to assess inter-rater reliability (kappa=0.8).

Encounter transcripts

Transcripts were entered into NVivo and deductively and inductively analysed. The researchers met after independently coding five transcripts to compare findings and reconcile discrepancies. The PhD-level researchers met bi-weekly with the larger research team to revise the coding scheme, ensure consistent coding application and reach consensus on disagreements. The clinical dialogue transcripts were qualitatively analysed by two PhD researchers to (1) identify diagnostic utterances in the conversation, noting when (using the transcript time stamps) and in what general phase of the encounter they occurred (eg, initial history taking, physical examination, wrap up assessment and plan) and their content, and (2) evaluate presence or absence of elements from the IDEA tool. Given the difficulty of more precisely rating the conversations using the tool (designed for EHR note review) with its 5-point Likert scale, for the verbal conversations, we instead dichotomised applicable IDEA items as present versus absent. Hence, if an item was present anywhere in the conversation, it was marked as present. Two researchers independently coded each transcript, then met to reconcile coding decisions. The researchers reached 100% agreement after each review and reconciliation meeting. IDEA items which were applicable only to the written note (eg, documentation completeness, succinctness, readability) were analysed separately from transcribed conversations.

Comparing clinical notes and encounter transcripts

After analysis was completed, two coders met to cross reference coded ratings between notes and transcripts to compare similarities and discrepancies.

Comparison with encounter time and Mini Z and Worklife responses

To evaluate relationships between note and encounter quality, time taken, perceived adequacy of time allotted and physician burnout, we analysed this data for each physician and encounter and compared results to determine extent to which (1) diagnostic process elements were present in verbal encounter transcripts and the accompanying written notes (concordance), (2) actual time taken and perceived time needed related to each other, (3) strength of relationships between encounter diagnostic process element inclusion (using a summary measure incorporating both verbal and written note quality) and physician burnout measured by the well validated Mini Z Worklife survey instrument (with the single item burnout measure correlating most strongly, in prior studies, with emotional exhaustion). We rated physicians as highly stressed if they answered agree or strongly agree to ‘I feel a great deal of stress because of my job’ and burned out if they answered affirmatively to one or more of the following items: ‘I am definitely burning out and have one or more symptoms of burnout’, ‘The symptoms of burnout that I’m experiencing won’t go away, I think about work frustrations a lot’ or ‘I feel completely burned out, I am at the point where I may need to seek help’.

Because diagnostic reasoning is linked to individual physicians, data were aggregated at the physician level which limited the power and scope of any statistical analysis. Hence all results presented here are descriptive and trend oriented.

Diagnostic elements concordance

The IDEA tool was applied to transcripts and notes to provide insight regarding which items were frequently communicated by physicians only verbally during the encounter versus only in the clinical note, versus present in both. If items were present in more than two-thirds of notes and encounters, we considered them ‘reliably included’, whereas items that were present in less than two-thirds of encounters were referred to as ‘often missed’. Items that were almost never present were deemed ‘widely missed’.

Patient and public involvement

None.

RESULTS

This study was originally designed to include 100 encounters by audiotaping 5 encounters of 20 physicians, in order to detect a correlation coefficient of 0.3 at the significance level of 0.05. Due to COVID-19 constraints,
we collected data on 28 patient encounters among seven physicians before March 2020, after which the data collection was halted. A total of 42 patients were approached and 28 (67%) agreed to participate. For each physician, between one and five patients were consented. Most physicians practiced in general medicine (five), whereas two practiced in emergency medicine (two). Table 2 shows diagnostic process elements captured in notes and transcripts across each physician as compared with the average across the seven physicians. In figure 1, physicians reporting high stress and/or burnout are represented in red outlines and tended to demonstrate different documentation profiles. For example, physician 2 compared with physician 3 verbally communicated and documented in their notes fewer psychosocial and contextual elements, mentions of diagnostic uncertainty comments on time frame for expected improvement.

Elements reliably included or rarely missed
Addressing the patient’s chief issue was found to be present in all but one encounter and reliably across physicians (75%–100% of visits). ‘Red flags’ were frequently included in both the note and encounter across physicians. Another prominently featured item was exploration of aetiology, as well as how clinical issues/symptoms influenced patients’ daily routines. Importantly, addressing likely diagnoses along with degree of certainty or uncertainty was noted in both the transcript and notes in 75% of encounters. Mention of contingencies (what to watch for if symptoms worsen) was present in 70% of encounters and corresponding notes. However, in 22% of encounters, such contingencies were present in the note but absent from the verbal encounter conversation (although in 6 of 28 notes it was simply a generic sentence ‘to return if worse’; this was often not communicated verbally to the patient during the visit).

Elements sometimes missing
Four items were frequently absent from encounter conversations as well as the notes. The ‘time frames’ of expected duration of symptoms and desired time frame within which the patient should follow-up was present only 50% of the time across physicians and visits. Two additional less-frequently noted items were ‘don’t miss diagnoses’ and ‘differential diagnosis’ (means of 58%). The item ‘discussing diagnostic testing’ and commenting on tests performed at this visit or on tests that the provider was ordering as part of the plan was noted in transcripts and notes on average 60% of the time across physicians and encounters.

Often missed items
Two items were absent in most of the cases. ‘Psychosocial/contextual information’ which we defined broadly as any mention of these issue (including financial context, mention of mental health, relationships and emotional context or simply how the symptom/illness impacted day-to-day functioning) was present in only 29% of visits. Less stressed physicians addressed psychosocial context more frequently (almost 67% of encounters) than burned out/

---

Table 2  Diagnostic process elements present in both the clinical note and corresponding transcribed audiotaped encounter, by physician

<table>
<thead>
<tr>
<th>Physician ID</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
<th>P6</th>
<th>P7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of patient visits</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Self-reported burnout status (1, 0)</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Diagnostic assessment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Addresses chief issue (element present in visits/total N. of visits)</td>
<td>2/3</td>
<td>4/4</td>
<td>5/5</td>
<td>5/5</td>
<td>1/1</td>
<td>5/5</td>
<td>5/5</td>
</tr>
<tr>
<td>Includes differential diagnosis related to chief issue</td>
<td>1/3</td>
<td>2/4</td>
<td>4/5</td>
<td>4/5</td>
<td>5/5</td>
<td>3/5</td>
<td></td>
</tr>
<tr>
<td>Addresses notes contextual and/or psychosocial information</td>
<td>0</td>
<td>0</td>
<td>3/5</td>
<td>4/5</td>
<td>0</td>
<td>0</td>
<td>3/5</td>
</tr>
<tr>
<td>Discusses possible aetiologies of the issues</td>
<td>3/3</td>
<td>3/4</td>
<td>5/5</td>
<td>5/5</td>
<td>1/1</td>
<td>4/5</td>
<td></td>
</tr>
<tr>
<td>Addresses degree of certainty/uncertainty</td>
<td>1/3</td>
<td>2/4</td>
<td>4/5</td>
<td>5/5</td>
<td>1/1</td>
<td>0</td>
<td>5/5</td>
</tr>
<tr>
<td>Diagnostic follow-up plan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mentions diagnostic tests (laboratory, imaging)</td>
<td>1/3</td>
<td>3/4</td>
<td>2/5</td>
<td>4/5</td>
<td>1/1</td>
<td>2/5</td>
<td>3/5</td>
</tr>
<tr>
<td>Contingencies discussed</td>
<td>0</td>
<td>2/4</td>
<td>5/5</td>
<td>4/5</td>
<td>1/1</td>
<td>3/5</td>
<td>5/5</td>
</tr>
<tr>
<td>Time frames discussed</td>
<td>1/3</td>
<td>3/4</td>
<td>5/5</td>
<td>3/5</td>
<td>0</td>
<td>0</td>
<td>5/5</td>
</tr>
<tr>
<td>Includes rational</td>
<td>0</td>
<td>3/4</td>
<td>3/5</td>
<td>5/5</td>
<td>1/1</td>
<td>3/5</td>
<td>5/5</td>
</tr>
<tr>
<td>Situational awareness/safety nets</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Red flags</td>
<td>3/3</td>
<td>3/4</td>
<td>5/5</td>
<td>5/5</td>
<td>1/1</td>
<td>4/5</td>
<td>5/5</td>
</tr>
<tr>
<td>‘Don’t miss diagnoses’ (worst case scenarios) considered/noted</td>
<td>0</td>
<td>2/4</td>
<td>3/5</td>
<td>4/5</td>
<td>1/1</td>
<td>3/5</td>
<td>3/5</td>
</tr>
<tr>
<td>Pitfalls considered, noted, commented</td>
<td>0</td>
<td>1/4</td>
<td>1/5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Global assessment of quality</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avoid legal liability pejorative</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
highly stressed physicians (for whom psychosocial information was not recorded in any encounters). Diagnostic pitfalls (eg, test limitations, atypical presentation) were noted in only 6% of encounters.

**Work conditions**

In analysing work conditions, including time pressure, we included five physicians, as two physicians practiced in urgent care embedded in emergency departments and did not have prespecified visit lengths. Figure 2 shows the relationship between actual time taken per visit (face-to-face time with the physician) and perceived time needed for the visit. Of 18 visits, 11 were measured as taking between 15 and 25 min, and 7 were longer (longest visit 38 min). Most of the longer-lasting...
visits matched the perception of additional time needed made by each physician, but this was not uniform. Burnout status (reporting high stress or burnout) did not appear to be related in a systematic way to either perceived or actual time taken for the visit.

Additional qualitative review of encounter transcripts: physicians’ diagnostic engagement during the clinical visit

Transcripts were analysed inductively to better understand physicians’ diagnostic engagement with patients especially ways they inquired about their clinical problem. This analysis focused only on the transcripts, analysing the conversation and extent to which physicians solicited and communicated diagnostic information with the patients. We examined three aspects: (1) depth of delving into aetiologies (scored present, if physicians asked more than one question about an aetiology; or absent if no inquiries were made); (2) acknowledging psychosocial elements (scored present if physicians acknowledged patients’ psychosocial expressions or absent if not acknowledged); (3) whether red flags symptoms were queried during the encounter (dichotomised as present if 4 or 5, or absent if less than 3). As shown in table 3, non-burned out/low stress physicians (white columns) were inclined to inquire and acknowledge what patients shared more than burned out/highly stressed physicians (grey columns).

DISCUSSION

Defining diagnosis quality is challenging for multiple reasons in including the highly subjective and variable nature of patients symptoms, clinical interactions and disease manifestations. However, there are key elements of good diagnostic process that have been recommended by experts that can serve as the basis for designing a tool for measuring conformance with these recommended qualities, at least in physicians note. Further, we reassuringly found that most (86%) of the recorded clinical encounters were highly concordant with
their corresponding notes in terms of inclusions of key diagnostic elements as assessed by the tool we developed. This is reassuring for both health services and quality research that often must solely rely on chart review.17 48

Our findings from both note and transcript review suggest that clinicians could do a better job in paying attention to several recommended items that were omitted in half or more of the encounters. These include attention to psychosocial history and context, consideration of ‘don’t miss diagnoses’ and documenting and communicating follow-up contingencies.31–34 In addition, despite their oft-lauded importance, there is room to improve consideration of ‘don’t miss’ diagnoses and development of a differential diagnosis.

This study is novel in examining diagnostic elements present in a clinical note while triangulating data from recorded encounters and measures of work conditions and clinician reactions (eg, length of encounter, burnout and stress). Although limited by our small sample size, the findings show a trend toward higher note quality and what is communicated to the patient during the encounter when physicians have and take the time that they feel is needed. When perceived time required more closely matched actual time taken (though not necessarily allotted), IDEA scores were higher. These findings also indicate a trend between low stress/non-burned out physicians, and documenting diagnostic elements regarding diagnostic tests, delving into aetiologies, and most importantly physicians addressing psychosocial elements, suggesting there may be measurable differences in diagnosis quality between non-burned out/low stress and burned out/highly stressed clinicians, particularly around communicating key elements with their patients, verbally and in their notes. It is noteworthy that this study is the first to measure burnout in real time. While stress is a factor that clearly varies day-to-day and from patient-to-patient, burnout is considered a longer-term stress reaction. Thus, the impact of acute stresses on burnout is uncertain and has not been studied.

These worklife findings should serve as a signal for concern, adding to other warning signs related to stresses in primary care.49–53 Diagnosis is foundational for high-quality care, and linking lower diagnosis quality with burnout is an important hypothesis generated by this work. The interaction between burnout and diagnosis documentation quality—as well as the vexing interaction between documentation itself and burnout. This two-way interaction, between feeling burned out and better diagnosis documentation on one hand, and documentation burdens themselves contributing to burnout on the other, calls for more transformative thinking about the EHR and workflow redesign.54 55

LIMITATIONS
The study has several limitations. First, it was an exploratory qualitative study based on a small convenience sample of physicians which limited us from performing a multilevel (clustered) statistical analysis. However, we thought it was nonetheless valuable to share our timely findings even at this stage for the benefit of other researchers and also to set the stage for future work. Second, physicians and patients were aware of being recorded during the encounter which might have altered clinician’s conversations and documentation behaviours, (eg, more thorough conversations or documentation). However, physicians were not aware of the specific quality elements being examined. Anecdotally, several physicians commented that having the recorder in the room was not noticeable and at times they forgot the conversation was being recorded. The IDEA tool, while iteratively developed by a team of physicians, health services researchers and communication PhD’s, has not undergone extensive psychometric testing. It focuses only on selected diagnostic process elements, rather than actual diagnoses made by physicians or patients’ clinical outcomes. In focusing on key diagnostic elements, we did not examine other important nuances in the encounter, such as non-verbal cues or power dynamics in patient-physicians’ relations which could have influenced trust as well as the sharing and communication of diagnostic information.56–58

CONCLUSIONS
A new tool to assess diagnostic process elements permitted us to conceptualise and measure aspects of diagnostic assessment in clinical encounters of acute problems, measure concordance between encounter transcripts and written documentation and correlate these findings with clinician burnout and stress. Our findings show reassuring concordance between notes and transcribed encounters and highlight areas that may be fruitful for future work in improving diagnosis, including better documenting psychosocial information, explaining proposed diagnostic tests and consideration of potential diagnostic pitfalls. Finally, the findings suggest time pressures and clinician burnout may correlate with poorer diagnostic performance, with a need for further research and support for primary care clinicians and practices. Given the paucity of metrics to assess diagnosis quality, the tool offers a template to model good diagnostic assessments, and a metric for offices and clinics to use and adapt for quality improvement and feedback.

Author affiliations
1Harvard Medical School, Center for Primary Care, Boston, Massachusetts, USA
2The Max Stern Yezreel Valley College, Emek Yezreel, Northern, Israel
3Suffolk University Sawyer Business School, Boston, Massachusetts, USA
4Harvard Medical School Department of Global Health and Social Medicine, Boston, Massachusetts, USA
5University of California Santa Barbara, Santa Barbara, California, USA
6Emory University School of Medicine, Atlanta, Georgia, USA
7Brigham and Women’s Hospital, Boston, Massachusetts, USA
8University of Minnesota Medical School Twin Cities, Minneapolis, Minnesota, USA
9Hennepin Healthcare System Inc, Minneapolis, Minnesota, USA

Contributors MK was a major contributor in gathering, analysing and interpreting the data and writing the manuscript. EES analysed and interpreted the data and

was a major contributor in writing the manuscript. SA analysed and interpreted the data and contributed to writing the manuscript. JR was a contributor to obtaining and analysing the data. MM was the project manager and helped in obtaining and interpreting the data. ML was a consultant and contributed to interpreting the data and writing the manuscript. GS was the principal investigator, contributed to analysing and interpreting the data and writing the manuscript, and is the guarantor. AO was a consultant and contributed to analysing and interpreting the data and writing the manuscript. All authors read and approved the final manuscript.

Funding This work was supported by CRICO (Harvard Risk Management Foundation)—grant number 072. The funders had no role in planning the design of the study, the data collection, management, analysis and interpretation of data. They had no part in the writing of the manuscript and no influence on the decision to choose a journal for publication.

Competing interests ML declares support through his place of employment (Hennepin Healthcare) by the American Medical Association (AMA), American College of Physicians (ACP), the Optum Office for Provider Advancement (OPA), Essentia Health Systems, Gillette Children’s Hospital, the California Area Health Education Centers (AHEC), the Institute for Healthcare Improvement (IHI) and the American Board of Internal Medicine Foundation (ABIMF) for burnout prevention research, projects and training. He is also supported for scholarly work by the NIH and the US Federal Agency for Healthcare Quality and Research. The other authors declare that there are no competing interests.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Consent obtained directly from patient(s).

Ethics approval This study involves human participants and was approved by the Institutional Review Board at Mass General Brigham (IRB approval number: 2018P000466). Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. The data sets during and/or analysed during the current study available from the corresponding author on reasonable request.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

ORCID iD Maram Khazen http://orcid.org/0000-0003-3165-3496

REFERENCES

BMC Prim Care 2022;23:153.


41 Miles MB, Huberman AM, Saldaña J. *Qualitative data analysis: A methods sourcebook*. 2014: 3.

42 Rohland BM, Kruse GR, Rohrer JE. Validation of a single-item measure of burnout against the maslach burnout inventory among physicians. *Stress and Health* 2004;20:75–9.


