Supplementary Digital Content

Chan A, Chien I, Moseley E, Salman S, Kaminer Bourland S, Lamas D, Walling A, Tulsky J, and Lindvall C. Deep Learning Algorithms to Identify Documentation of Serious Illness Conversations During Intensive Care Unit Admissions. 2018.

eAppendix. Abstraction Guidelines

eTable 1. Performance of Each Clinician Coder Against the Gold Standard

eTable 2. Algorithm Performance at the Token Level

eFigure 1. Algorithm Performance for Detection of Note Level Documentation of Care Preferences (NQF #1626 content) as a Function of the Number of Notes Used in the Development Set

This supplementary material has been provided by the authors to give readers additional information about their work.

**eAppendix**. Supplemental Information and Abstraction Guidelines

**Abstraction Guidelines**

In order to ensure consistent annotation, a set of abstraction guidelines was developed for the annotators. Each annotator identified text from clinical notes that fit specified domain of serious illness conversation. The portion of text that was deemed relevant to the domain was labeled in its entirety, with no restrictions on length of a single annotation. Although the selected text was only used once for one indicator, the same text could be used for multiple domains if deemed appropriate by the annotator.

Serious illness conversation domains:

1. Goals of care conversations
2. Code Status Limitations
3. Communication with Family/Caregiver/Health Care Proxy
4. Full Code Status
5. Ambiguous
6. None

The sections below describe in detail the criteria that were used to identify and label text that matched each pre-specified domain. The examples are exact excerpts from the clinical notes.

**Labeling of Goals of Care Conversations**

For this indicator, statements that met *both* of the following criteria were included:

1. Mention of a conversation with either a family member and/or the patient.
2. Mention of a specific care preference pertaining to hospital care.

Family members included immediate family members (e.g. spouse, partner, child, sibling) as well as health care proxies and formal care givers (e.g. distant family member caring for patient).

Instances were considered as care preferences if they included decisions relating to the diagnostic and management strategies of the patient. Examples include, but are not limited to: decisions regarding code status preference, elective procedures, chemotherapy, inotropes, noninvasive ventilation, hemodialysis, or requests for palliative care.

Examples of Exclusions:

There were many instances of patients not wanting to use an oxygen mask. An example note excerpt is:

*“would refuse to wear mask”*

Such instances were not considered a patient care preference. There was general consensus among the healthcare personnel on our team that while many patients may find such parts of their care uncomfortable, these refusals are often temporary.

Furthermore, care preferences that were not within the scope of the inpatient medical team were not included as an indicator. For example, if a patient had initially refused to come to the hospital, this would not be labeled under this indicator. An example of this would be:

*“patient refusing to come to hospital.”*

**Labeling Code Status Limitations**

For this indicator, any statements mentioning restrictions in the medical team’s ability to engage in resuscitative efforts were included. This includes CPR, intubation and defibrillation/cardioversion. Comfort care and comfort measures only (CMO) were also included.

Examples include:

*“Code: DNI/R”*

*“Family was present and decided to make patient CMO”*

The second example would also be labeled under the “Communication with Family/Caregiver/Health Care Proxy” and “Goals of care” domains.

**Labeling Communication with Family/Caregiver/Health Care Proxy**

For this category, we included statements that met *both* the following criteria:

1. Any type of communication that has or will happen with a family member or other designated decision-maker.
2. The communication involves a discussion of the patient’s healthcare. Examples of this include, but are not limited to, goals of care, code status, as well as updates on the condition of the patient.

Instances that showed a clear intent to arrange communication were included in this category. For example, the following statements would be marked under this domain.

Examples Include:

*“will meet with family today”*

*“Comments: communication with son by team frequently over past day”*

While the reasons for establishing contact were considered, the content of the communication itself was not used as a prerequisite for labeling. Therefore, instances that indicated communication with the family, without detailing what was discussed, were included in this indicator.

Examples of Exclusions:

When communication was initiated for purposes other than discussing patient care, we disqualified them as indicators. Instances of communication with the family that were not included in this criteria include:

1. Obtaining the patient’s medical history from a family member.
2. Any communication with the family after the death of the patient.
3. Using a family member as an interpreter.
4. Documenting a family member’s contact information.

For example, the following instances would not be included in this indicator:

*“confirmed med list and functional status with wife”*

*“Communication: Patient and daughter”*

Many of the clinical notes contained a “communication” section that was used to list contact information for family members or other decision-makers. Information on how to contact family members does not meet our criteria, thus these instances were excluded.

**Labeling Full Code Status**

For this domain, any documented evidence that there were no restrictions on the medical team’s ability to resuscitate the patient, including CPR, intubation and defibrillation/cardioversion.

Examples Include:

*“Code status: Full code”*

*“Code: Full (discussed with daughter)”*

As described earlier, there was no restriction in place for the number of indicators a statement may be labeled as. The second example could also be labeled under the “Communication with Family/Caregiver/Health Care Proxy” and “Patient and Family Care Preferences” indicators. This is because it meets the criteria for both indicators- there is mention of a family member, intent to discuss patient care, and a preference related to hospital-care has been explicitly stated.

**Ambiguous Statements**

Statements that were either incomprehensible, or appeared to contradict previous documentation, were included in these criteria.

One common example of contradictory text was when the same patient note contained documentation of  “Full Code” in one section while another section included “DNR”.

**Labeling Notes with None of the Domains**

If none of the text within a patient note met any of the criteria for any of the pre-specified domain, annotators would mark a box labeled “none” before moving on to the next patient note. This category enabled tracking of how many clinical notes had no relevant documentation on any of the designated advanced care planning indicators.

**Individual annotator abstraction’s agreement with Gold Standard**

There was variability between each of four annotators’ abstraction and the final gold standard. The gold standard was generated by two clinician coders and a final validation by a third clinician without time constraint on every note with real-time consultations with US board certified hospice and palliative medicine attending physician-expert reviewers (CL and AW). We show the agreement between each annotator and the Gold Standard in eTable 1 below.

**eTable 1**. Performance Metrics of Each Annotator Against the Gold Standard

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Clinician coder | Domain | Kappa | Positive Predictive Value | Sensitivity | Specificity | F1-score |
|  |  Full code  | 0.84 | 1.00 | 0.89 | 0.99 | 0.94 |
|  |  Code status limitations  | 0.82 | 0.83 | 0.94 | 0.91 | 0.88 |
| Clinician 1 |  Goals of care   | 0.55 | 0.94 | 0.49 | 0.99 | 0.64 |
|  |  Communication with family  | 0.74 | 0.92 | 0.77 | 0.95 | 0.84 |
|  |  Care preferences  | 0.70 | 0.88 | 0.77 | 0.92 | 0.82 |
|  |  Full code  | 0.79 | 0.99 | 0.87 | 0.97 | 0.92 |
|  |  Code status limitations  | 0.81 | 0.80 | 0.95 | 0.91 | 0.86 |
| Clinician 2 |  Goals of care   | 0.69 | 0.90 | 0.67 | 0.97 | 0.77 |
|  |  Communication with family  | 0.55 | 0.99 | 0.51 | 1.00 | 0.67 |
|  |  Care preferences  | 0.73 | 0.82 | 0.87 | 0.87 | 0.84 |
|  |  Full code  | 0.71 | 1.00 | 0.76 | 1.00 | 0.87 |
|  |  Code status limitations  | 0.73 | 0.75 | 0.96 | 0.81 | 0.84 |
| Clinician 3 |  Goals of care   | 0.80 | 0.92 | 0.83 | 0.95 | 0.87 |
|  |  Communication with family  | 0.59 | 0.93 | 0.59 | 0.97 | 0.72 |
|  |  Care preferences  | 0.71 | 0.81 | 0.92 | 0.78 | 0.86 |
|  |  Full code  | 0.92 | 0.98 | 0.96 | 0.97 | 0.97 |
|  |  Code status limitations  | 0.97 | 0.97 | 1.00 | 0.98 | 0.98 |
| Clinician 4 |  Goals of care   | 0.81 | 0.96 | 0.80 | 0.98 | 0.87 |
|  |  Communication with family  | 0.71 | 0.91 | 0.75 | 0.95 | 0.82 |
|  |  Care preferences  | 0.87 | 0.96 | 0.90 | 0.97 | 0.93 |

**Assessing algorithm performance at the token level**

We processed 282,788 tokens in the validation set to assess the algorithm’s performance at the token level. The criteria used to determine matches require an exact match by token. The criteria are much more stringent than note level performance assessment. For example, if the phrase “Code: DNI/R” was labelled as Code Status Limitation by the Gold Standard in a note but only “DNI/R” was labeled by our algorithm as Code Status Limitation, the note level prediction would still score a perfect true positive prediction that the note has documentation of Code Status Limitation whereas the token level prediction would only have a partial match as the tokens within the “Code:” part of “Code: DNI/R” was not labelled by the algorithm. The following results reflects this more stringent criteria.

It is worth noting that the very high accuracy reflects the fact that most tokens within a given are “negatives” (not a token that represents any of our domains of interest) and are correctly labelled as such.

**eTable 2**. Algorithm Performance at the Token Level

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Domain | F-1 Score | Accuracy | Sensitivity | Specificity | Positive Predictive Value |
| Patient care preferences a | 0.76 | 99.6% | 75.8% | 99.8% | 75.2% |
| Goals of care | 0.70 | 99.6% | 70.0% | 99.8% | 69.9% |
| Code status limitations | 0.76 | 99.8% | 72.7% | 99.9% | 80.5% |
| Communication with Family | 0.68 | 99.7% | 62.0% | 99.9% | 76.4% |
| Full code status | 0.91 | 99.9% | 88.3% | 99.8% | 93.6% |

a Goals of care or code status limitations (NQF measure #1626 content)

**eFigure 1**. Algorithm Performance for Detection of Note Level Documentation of Patient Care Preferences as a Function of the Number of Notes Used in the Training Set