A scoping review of AI, speech and natural language processing methods for assessment of clinician-patient communication

Pierre Albert[a], Brian McKinstry[b], Saturnino Luz[b]

[a] *National Institute for Public Health and the Environment, Bilthoven, 3721 MA, The Netherlands*

[b] *Usher Institute, Edinburgh Medical School, The University of Edinburgh, 5-7 Bioquarter, Edinburgh, EH16 4UX, UK*

Abstract

Introduction. There is growing research interest in applying Artificial Intelligence (AI) methods to medicine and healthcare. Analysis of communication in healthcare has become a target for AI research, particularly in the field of analysis of medical consultations, an area that so far has been dominated by manual rating using measures. This opens new perspectives for automation and large scale appraisal of clinicians' communication skills. In this scoping review we summarised existing methods and systems for the assessment of patient doctor communication in consultations.

Methods. We searched EMBASE, MEDLINE/PubMed, the Cochrane Central Register of Controlled Trials, and the ACM digital library for papers describing methods or systems that employ artificial intelligence or speech and natural language processing (NLP) techniques with a view to automating the assessment of patient-clinician communication, in full or in part. The search covered three main concepts: dyadic communication, clinician-patient interaction, and systematic assessment.

Results. We found that while much work has been done which employs AI and machine learning methods in the analysis of patient-clinician communication in medical encounters, this evolving research field is uneven and presents significant challenges to researchers, developers and prospective users. Most of the studies reviewed focused on linguistic analysis of transcribed consultations. Research on non-verbal aspects of these encounters are fewer,

and often hindered by lack of methodological standardisation. This is true especially of studies that investigate the effects of acoustic (paralinguistic) features of speech in communication but also affects studies of visual aspects of interaction (gestures, facial expressions, gaze, etc). We also found that most studies employed small data sets, often consisting of interactions with simulated patients (actors).

Conclusions. While our results point to promising opportunities for the use of AI, more work is needed for collecting larger, standardised, and more easily available data sets, as well as on better documentation and sharing of methods, protocols and code to improve reproducibility of research in this area.

Keywords: Patient-clinician communication, Medical consultations, Clinical Encounters, Artificial Intelligence, Machine Learning, Communication Analysis

1. Introduction

Clinician-patient communication has been the focus of considerable research efforts by the health community. The assessment of communication skills in medical consultations and teamwork among clinicians has been studied for more than sixty years, and numerous models of the medical consultation have been proposed. Better understanding of the patient's motivations and expectations (attitude to illness, psychological aspects), and new insights on the sequence of the consultation itself have led to changes in practice, such as taking social history or safety netting during medical consultations, alongside the formalisation of phases and tasks of the consultation. This has led to the creation of guides and assessment tools for learning and training purposes, such as the Calgary Cambridge Guide to the Medical Interview [21].

Changes and discoveries in models naturally led to their integration in the training of health professionals. However, the assessment of doctors' communication skills is a complex and time consuming process, performed by human experts. While innovations in automated processing have allowed an initial explorations of this domain, the development of automatic assessment of communication skills in clinical settings remains a challenging task.

Clinician-patient communication is a synchronous, usually dyadic communication: a dialogue between two participants interacting dynamically

with each other, or tryadic — a clinician interacting with a patient and another person, such as the patient's carer or a relative.

Sociological studies of the consultation have investigated many general traits of social behaviours of clinicians and patients. This includes the role of the patient, for instance, the definition and discussion of a patient's "sick role" (normative expectations around illness) [36, 51], the relationship between clinician and patient [9], the influence of the general organisation of the healthcare system [50, 9], social aspects of health and disease [9], and social factors determining the health of individuals, groups, and large populations [9].

This general picture has been refined by actual observations of interaction patterns contrasted with patient expectations. Such patterns have been the subject of investigation over many years. Davis [11], for instance, analysed recordings of medical consultations combined with interviews and questionnaires to identify patterns of communication explaining non-compliance (tension between the patient and the clinician, lack of rapport, seeking information without giving feedback). Regarding patients' expectations, McKinstry [29] (in a cross-sectional survey) and Elwyn et al. [14] found that patients varied in their desire for involvement in decision making, stressing the need for doctors to determine the level of involvement desired by a patient.

Communication in a dialogue can be divided into different modalities, verbal (speech), paralinguistic (tone, use of silence) and non-verbal (gestures, smiles, showing concern), and between *content*, i.e. the semantics of the interaction, and *content-free* aspects, the form of the interaction. The distinction between verbal and non-verbal aspects of an utterance refers to the distinction between the semantic content, and its paralinguistic content [16].

This review concerns technology that aims to extract meaning from patient-clinician interactions using existing tools, new methods, and their combination in a processing pipeline able to provide data that can be turned into metrics and feedback. While some work exists on the automatic assessment of parts of the communication, this domain is in its infancy and more is still needed for its practical applications, e.g. to teach and train communication skills. Nonetheless recent studies have demonstrated the capacity of current systems produce meaningful results, such as prediction of student's success based on communication and domain skills [6], identification and assessment of suicidal risk using verbal and nonverbal cues during interviews with adolescents [52], characterisation of semantic similarity of the patient's

and physician's language [53], etc.

At the acoustic level, speech processing focuses on content-free patterns that may be helpful in structuring the communication, such as *prosody* and *segmentation*. Prosody and assessment of voice quality have been used in clinical training using staged scenarios [55]. Spoken dialogue can be segmented by monitoring *turn taking patterns* or using *vocalisation patterns*. Vocalisation patterns [25, 24] are Markov diagrams encoding transition probabilities between vocalisation events of both participants, providing patterns of interaction. In medical applications, they have been used in the context of mental health to characterise power dynamics during dementia diagnosis disclosure conversations [45].

Semantic processing is the content-rich approach to speech processing. For the analysis of consultations, it first requires the transcription or automatic speech recognition (ASR), and its understanding using different semantic processing. A typical semantic pipeline includes diarisation (determining who spoke when), ASR, syntactic analysys and semantic interpretation. Variations to this typical architecture presented in this review include the use of machine learning (ML) methods for detection of dialogue acts, analysis gestures, facial expressions and other non-verbal signals which affect communication [23].

2. Methods

The reporting of this scoping review follows the recommendations set by the preferred reporting items for systematic reviews and meta-analyses (PRISMA) statement [33].

2.1. Eligibility criteria

To be included all studies had to have three main characteristics. First, the studies had to analyse the interaction or communication between clinicians and patients (dyadic) in primary care settings. Second, the analyses had to be done by using automated methodologies including but not limited to machine learning and deep learning methodologies. Third, the studies had to report performance measure for these analyses.

2.1.1. Dyadic clinician-patient interactions

Dyadic refers to the interaction between a clinician and one patient. Studies including a third person (e.g. carer or relatives) were not deemed eligible.

These conversations should occur only in primary care settings and the term clinician refers to any professional who provides care to patients in these settings such as general practitioners, and nurses.

The patient-clinician interaction had to occur in real time (synchronous) through spoken natural language, either face-to-face or remotely using video-conferencing technology. The interactions had to be spontaneous. We included in this category semi-structured interviews and studies that enrolled simulated patients.

2.1.2. Automated analysis

Automated analysis include machine learning and deep learning methodologies that automatically extract features from dyadic communications. Other type of automated analysis are those that describe precise algorithms or specific instructions which needed to be followed to analyse the interaction.

2.1.3. Performance measures

We distinguished three types of measures to assess performance of automated analyses of clinician-patient interactions: 1) intrinsic evaluations, such as F1-score, recall, precision, sensitivity, area under the receiver operating characteristic curve (AUC), 2) medical communication evaluation meeting certain criteria from medical frameworks, and 3) correlation with human assessment, such as the patient's assessment, as reflected in questionnaires or structured interviews that yield a numerical score (e.g. [48, 13, 15].

2.2. Information sources

A systematic search was performed on Cochrane Central Register of Controlled Trials, Cochrane Database of Systematic Reviews, Embase and Medline through Pubmed from their inception to January, 2021. Because the importance of the Association for Computing Machinery (ACM) Digital Library as one of the world's most comprehensive bibliographical databases in the field of computing we searched this digital library from inception to January 2021. We also included grey literature such as dissertation and theses as our information sources.

The search strategy included terms related to the following terms: 1) dyadic communication, 2) clinician-patient interaction, and 3)automated analysis. We also performed snowballing of included articles. We searched through the references of these articles and assess them against our eligibility criteria.

2.3. Article selection

All authors participated in abstract screening. Full-text screening was performed by one author (PA), with random samples assigned to SL and BM for confirmation of selection. A pilot for this last step was performed to homogenise the eligibility criteria of included studies. Borderline papers were identified regarding the interpretation of automation and patient-clinician consultations, and a stricter application of the criteria was advised. Following this, the definitions were clarified and every full text paper was reviewed a second time. Twenty-two studies were rejected and one additional study was included.

2.4. Data collection

We extracted the following overall information: 1) general characteristics such as publication date, first author, and location of the study, 2) baseline characteristics such as sample size, inclusion criteria, age, sex, ethnicity, socioeconomic information. We also extracted the following methodology-related data: 1) datasets' characteristics: language, availability, annotated data, 2) the purpose of the interaction: palliative, risk-benefits of treatment options, etc. 3) the name of the framework or theory used to analyse the interaction, 4) type of input material: transcripts of video-recordings or audio-recordings, semi-structured interview transcripts, video-recordings, audio-recordings, manual annotation of non-linguistic features, etc, 5) features of interest: discourse acts.

In terms of results, we extracted the following information:

1. performance metrics: kappa, accuracy, sensitivity, specificity correlation scores, F-scores, AUC, and error scores;
2. dataset characteristics: language of the dataset, availability, type of collected information (transcription of video- or audio-recordings), input material for preprocessing or cleaning (transcripts, audio, video), preprocessing techniques (e.g. stop-word removal), type of system input information (video, text, or audio), and extracted features;
3. analysis characteristics: theory or framework behind the analysis, machine learning or statistical method (supervised, unsupervised, semi-supervised), type of analysis (statistical, machine learning, deep learning).

2.5. Critical appraisal of included studies

The search was centred around the three main concepts of the review: dyadic communication, clinician-patient interaction, and systematic assessment.

The concepts grouped under *systematic assessment* of the communication by this review are diverse. Specific terms used in the language processing community are not always used by the medical community, and broader terms needed to be included. In addition, systematic assessment differs from automated assessment, encompassing studies that may have not used computational methods to extract features of the communication. Additional relevant terms were identified during a preliminary search on a subset of studies and reference lists. The final list (figure 1, item 1) includes terms from both medical and speech processing fields.

No previous review on a topic similar to this review was found during initial searches. The scoping and search strategy for this review was developed from scratch to identify studies using systematic approaches or automated processing to support the assessment of patient-clinician communication. Search updates were done as the search protocol was refined.

2.6. Sources searched

Searches were not restricted by location, date of study, or language. Grey literature (dissertations and theses) was also included for screening.

2.6.1. Medical libraries

Systematic searches were performed in the main electronic databases, using the search strategy presented in figure 1.

Dates and issues of the medical databases searched for the review are:

Cochrane Cochrane Central Register of Controlled Trials Issue 1 of 12, January 2021 and Cochrane Database of Systematic Reviews Issue 1 of 12, January 2021. 44 reviews and 41 controlled trials were found.

Embase Embase 1980 to 2021 Week 03

MEDLINE/PubMed Accessed 2021-01-25

ACM Full-Text Collection was searched to retrieve studies with a strong focus set on the language processing aspect that may not have been reported in medical journals. The search was performed on full texts

1. "machine learning" or "natural language processing" or "speech processing" or "artificial intelligence" or "video analysis" or "visual analytics" or "text analytics" or "text analysis" or "speech analysis"
2. communication or consultation or interview
3. clinician or doctor or nurse or gp or "general practitioner" or "general practice" or physician or "primary care" or "family practice" or "family practitioner"
4. 1 and 2 and 3

Figure 1: Search terms.

until 2019-08-30, then updated with searches on titles and abstracts until 2021-01-25.

Reference lists from eligible studies identified using the developed search strategy were searched manually for additional studies. Within-paper references were searched using Google Scholar[1] or DuckDuckGo[2] when not referenced to find further relevant studies. Search updates were conducted until January 2021, as mentioned in the respective online libraries search protocols.

2.7. Inclusion criteria

- Primary research study;
- The study is on clinician and/or patient communication;
- The studied interaction is based on synchronous interactive communication using spoken natural language (face-to-face or remotely), spontaneous (including semi-structured interviews), staged or not (e.g. a simulated patients acted a predefined scenario);

[1]Google Scholar is a specialised search engine for published scientific literature. It is a valuable resource to lookup specific references, e.g. cited articles in a publication.

[2]DuckDuckGo is a generic web search engine with a strong focus on keeping user's privacy. It is an alternative to the more popular web search engine by Google.

- Direct signals processing or their interpretation (e.g. speech or transcripts). Secondary interpretation, such as studies that extract patterns from manual annotations were included;
- Automated analysis is used. Therefore statistical analyses only based on manual annotations were discarded. Automation includes manual analysis in which a precise algorithmic methodology was described and used (following objective instructions, e.g. if ... then ... else ...);
- Study must report evaluation measures. The measures can be classified into three types of evaluation: technical evaluation (e.g. standard NLP metrics), medical communication evaluation (e.g. using medical frameworks), and correlation with assessment (e.g. patient's assessment).

Secondary sources were screened for the identification of additional material. Some studies in foreign language were included (French and German).

2.7.1. Exclusion criteria

- Studies based on asynchronous communication: clinical narratives, medical notes (discharge summaries, nursing notes), speech notes using ASR;
- Automatic analysis of medical expert systems (diagnosis systems) and electronic health records without an interactive component;
- Studies using patient interviews or focus group discussions by researchers that were conducted after the interaction with clinicians for qualitative studies;
- Studies without a strong focus on communication between a clinician and a patient (e.g. team communication in presence of a patient);
- Studies reporting manual annotation and observation of the results whithout automation;
- Opinion and prospective papers;
- Studies with no full text available, or full text not in English, French or German.

2.8. Screening Procedure

A search of the main medical databases was conducted using the search strategy described in Figure 1. Results were automatically merged and duplicates removed using a specific tool[3], then screened for relevance using the title, keywords, and abstracts. Relevance was established where studies discussed analysis of communication in a primary care setting or in a clinical setting similar to primary care (e.g. consultation with a surgeon). Full texts of identified studies were retrieved, and eligibility was screened against review inclusion and exclusion criteria outlined above.

2.8.1. Updates

The search was updated four times — every 6 months — from July 2018 until January 2021. Retrieved results were merged and filtered, and previously screened references were discarded using the aforementioned automated tool. Potentially relevant studies uncovered during article screening (retrieved using Google Scholar) were also screened for eligibility.

2.8.2. Results from the search

Due to the heterogeneity of the systems, aspects of communication, and interventions, a meta-analysis was not attempted. A detailed visualisation of the result of the search and screening procedure is provided in figure 2, formatted in accordance with the PRISMA flow diagram for screening [32].

We can group studies by themes according to the type of communication investigated. The first and largest group of studies explored *verbal communication*: the semantic content of the interaction. In this group, the first theme is the structures of the discourse, either task-specific [3] (VR-CoDES) or general [49] (behavioural codes), [52] (conversation dynamic), [4] (characterisation of utterances, sequential information), [47] (questions and answers), [27] (sequences of discourse elements). Related themes were task-based categories (interaction elements) [4] and the general structure of the dialogue (Speech acts) [56], [28]. The second theme focuses on topics - what was discussed - [57, 58, 35, 8, 10, 6]. The third theme relates to words: embeddings (use and context of a word) [39, 46, 10], types (e.g. part of speech) [28], frequency [47], and polarities (positive and negative words, e.g. related to gain and loss) [34, 17]. A final theme was the expression of affect (sentiment analysis) [47].

[3] `https://edin.ac/3OIvxVW`

The other main group relates to the **non-verbal** components of the interaction: the part of the communication conveyed by other channels than the speech. Most use the visual modality: the face of participants [40, 39], gestures and movements [27, 18, 39], gaze [38, 39], and posture [7]. The other studies observed activities performed during the consultation: clinician's activities [20] and computer / screen interactions [38].

The last group of investigation relates to the **paralinguistic** components: the part of the communication conveyed by the speech but not its content. Acoustic features (verbal dominance) [52] [46], pauses [26], and silence [12, 30, 26].

The type of interactions are shown in Table A.8. The *type of interaction* relates to the active participants in the interaction. Constrained analysis is specified when applying (e.g. only dyadic interactions are analysed). The value can be dyadic (2 persons) or triadic (3 persons). The *medical interaction* describes the context in which the clinician patient communication occurred, e.g. GP consultation, outpatient visits, etc.

Information of the eligible articles is summarised in two tables. First, studies following the "participants, interventions, comparisons and outcomes" (PICOS) framework [42] are shown in tables 1 and 2. Additionally, we included three columns: a brief description of the aim of the study, an outline of the methodology, and a summary of results.

A summary of tasks related to clinician-patient communication assessment performed in each study is provided in table A.5. The *frameworks* column contains the medical and/or annotation that were used or referenced, the *type of material* is the type of data on which the study was conducted (e.g. audio, video). The *task performed* lists the processing applied to the data, either manual and automated (e.g. emotion recognition). The *performance* variable summarises the main quantified results, and the *dataset* variable describes the collected data. The information is then developed using six tables to extract detailed information relevant to this review (available in appendix, see section Appendix A).

The population of each study is described in two tables: table A.6 for the patients, and table A.7 for the clinicians. Both tables include the same demographic information in addition to the population included in the study: age, sex, ethnicity, location and socio-economical information. Patient-specific information relates to the personal, socioeconomic attributes, and medical condition of the cohort. Clinician-specific information regards speciality and experience. The analysis conducted in each study is then detailed in table

A.9, which contains the following columns:

- *Preprocessing* list the procedures undertaken on raw data (text, audio, video) as preparation steps for subsequent extraction of features analysis. Text processing usually include transcription, in which case the method is reported (by professionals or by researchers). Since no instance of the use of ASR was found, all reported transcriptions were manually produced. It must be stressed however that instances of uses of ASR to help generate transcripts were found: Alloatti et al. [1] for instance used manually corrected ASR output on 30 physiotherapy sessions. Other text preprocessing methods include cleaning of transcripts and removal of unwanted events, such as stop-words or disfluency. preprocessing of audio and video can include segmentation, extraction of parts (beginning, end), signal processing (background noise removal, normalisation, colour balance etc.). Finally, any manual processing is also listed.

- *Feature extraction* reports on automated processing. The generation of the features was documented either from raw data (acoustic or video analysis) or manually generated data (through text analysis: tokenisation, part-of-speech tagging, etc.)

- *Task and method* reports on the task that was performed (e.g. classification of a sentence) and on the methodology that was used. This includes supervised learning or unsupervised learning, a detail of any analysis used used: machine learning algorithms, clustering, feature set reduction, classification, etc. An accompanying table of abbreviations is provided in section A.3.

- *Evaluation* reports how the results were assessed in the study. This is broken down in four items, if present in the study — B: baseline, PM: performance metric, CV: cross-validation technique used, T: test set held out and its size.

- *Results* are numerical results of the reported performance metrics.

An assessment of the research potential and applications of each of the study is presented in table A.10. It is structured around the following columns:

- *Research implications* regroups three general characteristics. Novelty (yes/no): whether the study implemented a new method or applied an existing one — no is assigned where the study uses an existing tool or method. Replicability (low/partial/full): whether the reported procedure is described in sufficient details and data is available — low is assigned where both data is n4ot available and method description is incomplete; partial where either is the case, and full where both data and detailed methods are available. Generalisability (low/medium/high): whether the analysis is specific to the task — low is assigned where the method can only be applied to similar settings; medium where the analysis can be applied to other settings (i.e. type of medical encounters) with adaptations (e.g. changing a dictionary of terms); high where the analysis can be applied directly to other settings.

- *Risk of bias* Real life (RL, yes/partial/no): whether the interaction featured real interactions (e.g. between patients and doctors) or simulated interactions (e.g. training sessions with an actor). Feature balance (FB, yes/no): whether reported individual features were balanced across classes. Suitable metrics (SM, yes/no): whether metrics other than overall accuracy are reported when data are class-imbalanced. Contextualised results (CR, yes/no): whether a baseline is provided to put the results into perspective. Overfitting (yes/no): whether cross-validation and/or hold-out set were used. Sample size (S): three ranges are reported: ≤ 50, ≤ 100, and ≥ 100).

- *Strengths/Limitations* five characteristics are reported with yes/partial/no assessment, each yes indicating a strength, each no indicating a limitation. Spontaneous speech: whether speech was naturally generated or prompted in response to open-answer questions. Conversational speech: whether the study is based dialogue. Automation: whether the automation (other than the machine learning tasks) was complete (excluding preprocessing) or only some aspects of the procedure used in the study. Transcription-free: whether the method required transcription of the dialogue. Content-independence: whether the method is content-based or not.

Finally the dataset table (table A.11) summarises details of the datasets used in the reviewed studies. It contains the following columns:

- *Data set/Subset size* Quantification of the number of documents details by groups of participants, including number of minutes recorded and number of words when available.

- *Data type* Data recorded and used in the study. Two types of data are reported. Data streams (audio, video) and derived data (e.g. transcripts — with information about the transcription when available). Other type of data (patients' information, questionnaires, etc.). The type of interaction during the dialogue is characterised as either structured, semi-structured, or conversational.

- *Data annotation* Type of annotations with details about the annotation set.

- *Data balance* reports whether the dataset is balanced in terms of age (a), gender (g), and socio-professional class (s). Yes reports balance for both between and within class balance when applicable.

- *Data availability* whether the dataset has been published or made available.

- *Language* is the spoken language used during the interactions. It can differ from the main language of the country where the collection took place.

3. Results and discussion

Before analysing their content, a look at the distribution of the dates of publications of the included articles (see figure 3) provides a sense of how recently the field has emerged. All studies were published after 2005, and more than half of the articles were published after 2018.

A total of 27 studies are included in the final selection. While they cover a wide range of aspects of clinician-patient communication, with only a limited number of studies having been dedicated to each aspect. A wide range of medical speciality are featured: General Practice, dentistry, radiography, language pathology, psychometry, oncology, urology, palliative care, psychotherapy, home medical care. In five occurrences, the interacting clinicians were medical students. A single study used an actor to perform the role of the

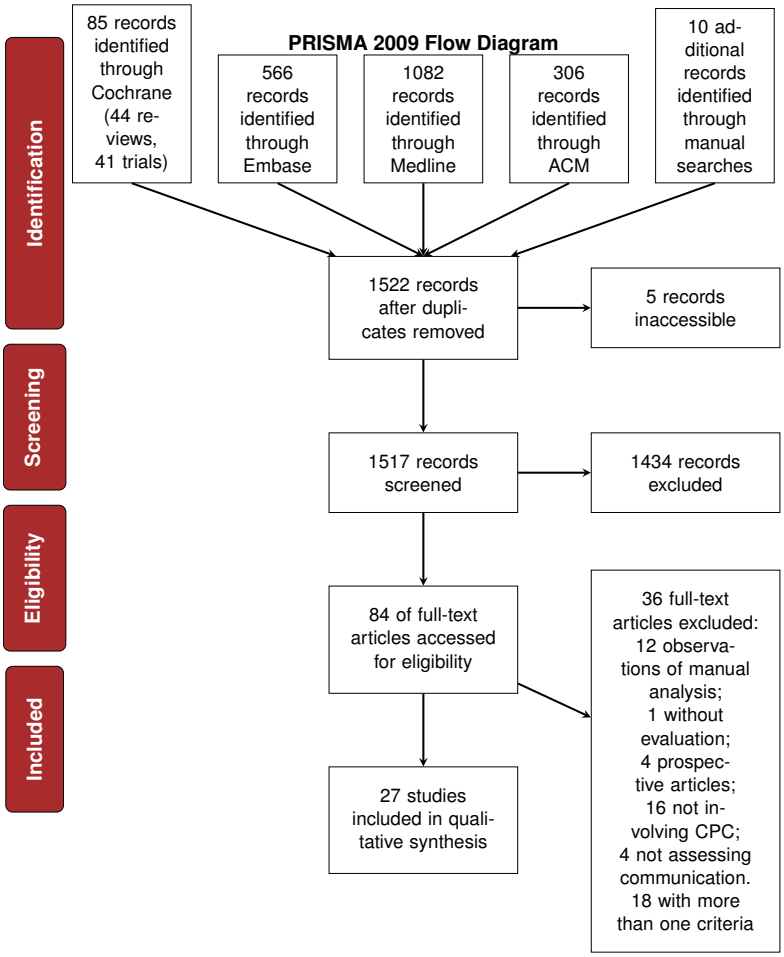


Figure 2: Detailed result screening procedure.

doctor [18], in order to control the behaviour in preset scenarios (engaged or disengaged).

The retrieved studies feature several types of clinician-patient interactions. Twenty-two studies were conducted on real interactions and five were simulated, including one with a virtual avatar. The dialogues during medical interactions can be grouped in three different type. Twenty studies are based on conversational interactions, i.e. free form interactions during which the participants exchange freely without constraints over the content. Five studies used semi-structured interviews, i.e. an open discussion with a set of themes or questions to direct the interaction or elicit answers. Finally two

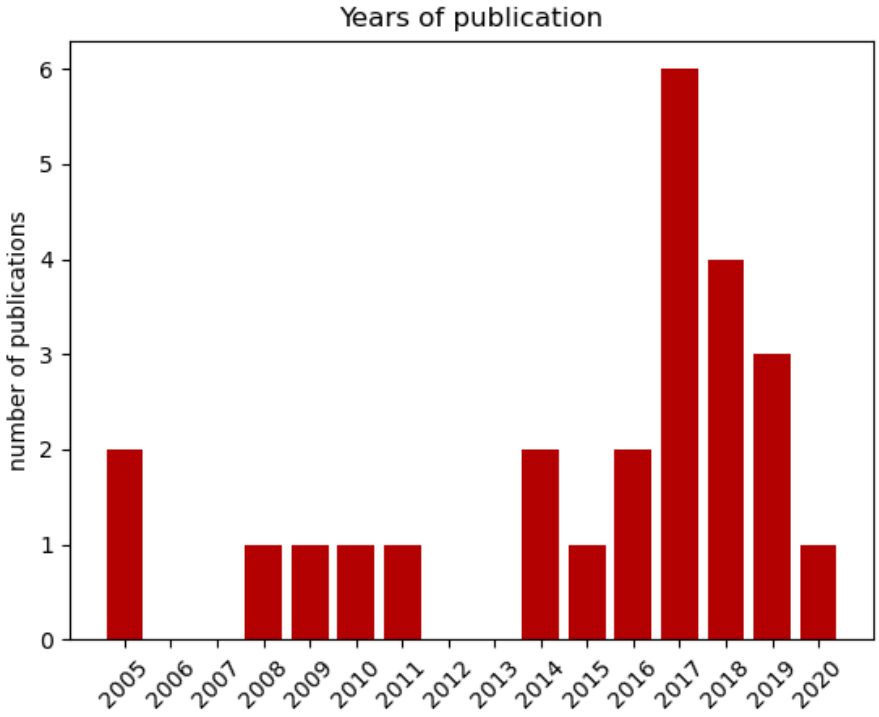


Figure 3: Distribution of the years of publication of included studies.

studies used structured interactions, i.e. a planned, constrained discussion during which the same set of predefined questions is asked to each participant.

Regarding settings, 17 studies investigated medical consultations, either during GP consultations or routine patient visits (e.g. dental care), of which 13 were dyadic consultations (clinician and patient) and three were triadic interactions (a patient's helper or a second clinician). One study features mainly dyadic interactions with a small fraction of triads. Overall, 7 studies report triadic interactions. A majority of the triads concerns an additional caregiver (e.g. parent). Only one feature an active second clinician although a few report non-interactive clinicians (passive, observing) or interacting before or after the studied interaction.

Five studies used clinical interviews (intake interviews, diagnoses or assessment of a particular condition). Two featured motivational interviewing

(one on substance use, one on adherence dialogues). Two investigated disclosure interactions and the breaking bad news. Finally, one used a instruction session (on how to use a specific drug).

Of the 17 studies investigating medical consultations, seven used constrained topics and one investigated only a specific phase of the interaction. The cultural context of the studies (see table A.6) was fairly restricted. More than half of the studies (fifteen) were conducted in the USA, and all but five were conducted in western countries (USA, UK, Scotland, Australia, France, Germany). Of the others, none were conducted in developing countries: four were conducted in Asian countries (Japan, Singapore (PRC), Hong Kong) and one in Israel. The socio-cultural diversity was also quite lacking. Reported age and sex were generally balanced (featuring patients of all age, from children to elderly people). The distribution of ethnicity seemed balanced when reported, but the information is missing in more than half of the studies. While some patients of lower income or lower education were included in some studies, with two studies specifically on low-income cohorts, the information is also often lacking.

Most of the studies investigated cohorts of patients with cancer (seven studies) or patients for general consultations (five). Five studies were conducted with patients suffering from psychological issues such as suicidal thoughts or patients with Dementia, which could potentially influence their speech.

Regarding clinicians (see table A.7), most studies do not report information beyond sex distribution, and even this is missing for eighteen of the studies. Out of the studies reporting those, sex distribution was equal, which can simply signal that studies which paid attention to this metric paid attention to the sex distribution while recruiting the cohort. This is further illustrated regarding the ethnicity, where out of the four studies reporting it, only one was featuring a cohort of white only clinicians. Interactions featured a wide range of clinicians: nurses, oncologists, GPs, etc. Five studies were conducted with students clinicians, and two with resident doctors.

3.1. Investigated aspects

Most studies, 20 out of 27, investigated the semantics of the interaction. Ten studies used the global semantic space (such as topics in Carnell et al. [6] or participants' semantic space in Vrana et al. [53]), i.e. the spoken content of the participants as a whole to characterise the communication.

Nine studies investigated topics or closely related concepts in conversations [57, 58, 35, 8, 10], either investigating consequences of differences in their presence or frequency, or evaluating internal structures, e.g. tracking reuse by participants. One additional study addressed the use and presence of more specific task-based categories: Blomqvist et al. [4] investigated interaction elements, characterising syntactic roles of utterances (statement/information, question, request).

Word-based studies are another type of unstructured characterisation, based on the quantification of used words: [39, 46, 10], the words used and their context (word embeddings), their type (part of speech) [28], or their frequency [47]. Investigations of emotions using verbal features were undertaken with two objectives: the classification of positive or negative speech using word polarities [34, 17], and the detection of sentiments from text [47].

While 15 studies investigated unstructured content (e.g. occurrences of topics), five studies investigated the discourse structure of the interaction, tracking the use, presence and absence of predefined sets of structuring elements: either task-specific structure in Birkett et al. [3]) based on VRCoDES [59], a system for coding the patient's expressions of emotional distress, or using a more general linguistic approach (using behavioural codes in Tanana et al. [49], or a set of conversation dynamic features in Venek et al. [52]).

Other studies interpreted the interaction in a more global way by investigating the structure of the interaction, identifying links between its elements and their sequences: Sen et al. [47] tracked the questions and answers between participants, and Mase et al. [27] extracted patterns of interaction from sequences of discourse elements. Blomqvist et al. [4] combined the characterisation of utterances (syntax and type) with sequential information: source (spoke, focus) what was the aim of the utterance, response.

Some studies used concepts stemming directly from theoretical linguistics, such as speech acts [56, 28], although the precise definition of what constitutes a speech act varies across studies. Although both Mayfield et al. [28] and Wallace et al. [56] defined the speech act as a social act embodied in an utterance, and both restricted the possible acts to the categories listed by the Generalized Medical Interaction System (GMIAS) [22], Mayfield et al. [28] aggregated multiple categories into two acts: information-giving and information-requesting.

Further paralinguistic analysis uses acoustic features for the characterisation of speech, generally for its classification, e.g. between healthy and unhealthy patients, [52], but also for investigating non-verbal aspects of the

interaction such as the types of pauses [26]. Another paralinguistic aspect of the interaction is the characterisation of the sequences of spoken interaction, or speakers turns: silences [12, 30] and verbal dominance [46] (calculated indirectly by quantifying the words of each participants).

Manukyan et al. [26] and Durieux et al. [12] are based on the same parent cohort study. Their experiments were conducted by the same team and complement each other: speech and silence detection, characterisation of the silences. Manukyan et al. [26] extracted and aggregated of acoustic features for the identification of conversational pauses. The random forest classifier achieved slightly lower accuracy than manual annotators (94.4% vs 99.1% over a ground truth defined as the consensus of three human coders) but it was much faster than the human coders (two orders of magnitude, requiring minutes instead of hours). Durieux et al. [12] used similar acoustic features with statistical aggregators to classify types of connectional silences (emotional, compassionate, invitational). While the automated identification misidentified 41.3% of the clips, its use to semi-automate the annotation task for human annotators was significantly more efficient, manual annotation requiring 61% more time.

In the evaluation of non-verbal element of the interaction, studies prominently investigated the visual modality: studies have used face [40], gestures and movements [27, 18], gaze [38], posture [7], and a combination of them (head movements, posture, gaze, eyebrow, hand gesture, smile) in [39]. Another element of the communication investigated was the ongoing activity of the participant while the interaction was taking place. This included the clinician's activities in Kocaballi et al. [20], and computer / screen interactions in Pearce et al. [38], both of which are known to affect consultations.

3.2. Theoretical background

The theoretical background for the evaluation of the communication was diverse. Eleven studies used ad-hoc coding systems, either designed and tailored for the study, derived from previous works by the same authors (e.g. Pearce et al. [38]), or inspired by concepts defined by existing framework but heavily modified (Two studies, [35]: modified Multi-Dimensional Interaction Analysis, [26]: ad-hoc set of acoustic features including mel-frequency cepstral coefficients (MFCC). The frameworks used in the studies can be separated into four (+ one) types.

The papers of the first type comprise assessment criteria of a medical authority (e.g. the Australian Open Disclosure Standard in [57]) and nor-

malised assessment tools such as patients' feedback tools, such as scales used to quantify anxiety (20-item State-Trait Anxiety Inventory), depression (15-item Geriatric Depression Scale), and satisfaction with the appointment (Dementia Care Satisfaction Questionnaire) used by Sakai and Carpenter [46]. Two other studies used these scales to assess interaction quality: [58] (Dental Patient Feedback on Consultation skills), and [47].

The second type of framework are medical scales, used to evaluate the medical condition of patients such as the NSA16 in Chakraborty et al. [7]. Four different medical scales were used in the reviewed studies. The list is provided in table A.4.

The third type is frameworks for aspect-specific elements of the communication. The largest subset concerns semantic analysis of the interaction and linguistic or word based dictionaries, e.g. MetaMap for medical terms. Watson et al. [57] used Discursis, a visualisation tool for the analysis of term reuse. Sakai and Carpenter [46], Fridman et al. [17], Carnell et al. [6] and Venek et al. [52] used the LIWC, a word-based framework to quantify the frequency of terms and word categories (e.g. to quantify the use of possessives pronouns). Watson et al. [57] used a generic conversation and dialogue analysis tools, CAT, providing higher level structuring of the dialogue in terms of interpretability, discourse management, interpersonal control and emotional expression.

While a number of studies used acoustic and prosodic features, all studies have used their own set of features [26, 31], usually selected from a combination of sets used in other studies making it very difficult to compare their findings. It must be noted however that part of the feature selected in Manukyan et al. [26] is MFCC, a common set of acoustic features. The study of other non-verbal and paralinguistic aspects of the communication can be similarly depicted, i.e. extraction and study of ad-hoc sets of features, however one study [40] used the Emotion Facial Action Coding System (EmFACS) to code expressions of affects (happiness, social smiles, sadness, fear, anger, disgust and contempt) as well as social smiles and combinations of different affects.

The fourth type of framework used are the medical frameworks designed to study patient-clinician communication: VRCoDES [3], GMIAS [28, 56], the Comprehensive Analysis of the Structure of Encounters System CASES [28], and the Motivational Interviewing Skill Code (MISC) [49].

The Roter interaction analysis (RIAS) framework [43] is also referenced by Carnell et al. [6], although only its distinction between biomedical utter-

ances and psychosocial utterances is used. Finally, six studies (e.g. [58]) did not use medical or conversational frameworks, instead reporting exploratory findings, for instance using data analysis (unsupervised machine learning methods such as principal component analysis) to identify prominent themes and observe the influence of their use on patients' caregivers' perceived quality of communication.

Similar to the variety of aspects investigated, the large set of frameworks used for reference or in the assessment reported in A.9 makes it difficult to compare the results of the studies and integrate them into a meta interpretation.

3.3. Paralinguistic and non-verbal communication

While the semantic aspect of the interaction has been frequently investigated, partly automated in the frame of this review but also more globally in observational studies of the clinician-patient interaction, non-semantic analysis of communication during consultations has been less studied. From the studies retrieved in this review, a number of aspects can be identified as promising.

Visual cues constitute the most frequent modality investigated. *Facial features* of the patient during communication has been used to detect facial expressions of different affects (happiness, social smiles, sadness, fear, anger, disgust, contempt) [40] in relation to signs of illness. Beyond the scope of this review, facial features were also used for the detection of illness. Barzilay et al. [2] classified patients' affect using Face Action Recognition, noting the potential of the method as a clinician-supporting tool to detect schizophrenia. Joshi et al. [19] extracted generic facial spatio-temporal descriptors — Local Binary Patterns on Three Orthogonal Planes (LBP-TOP) and Space-Time Interest Points (STIP) — as part of a multimodal classification model of depression (speech and video features), demonstrating the capacity of automated analysis to classify patients, but using extreme cases of the DSM-IV scale.

Focusing on *gaze and eye contact*, Pearce et al. [38] limited its use to detect computer activity while Porhet et al. [39] investigated gaze as elements of patterns of interaction (cues leading to cues in reaction) in verbal and non-verbal communication during consultations. Gaze was present in detected rules alongside other visual cues (nods, hand movements), however with low confidence scores for the strength of the observed patterns. Visual elements of bodily actions in time, *gestures and movements* have also been investigated.

The *posture* was investigated by Chakraborty et al. [7] to quantify symptoms of schizophrenia, finding a negative correlation between motor movements and negative symptoms.

Using a small number of interactions ($n = 10$) Mase et al. [27] analysed gestures as part of more abstracted interactional patterns. They did not analyse the gestures in themselves however, and their use was only as elements of sequential patterns for the interpretation of the interaction as a whole. At the smaller scale of motions realised during the interaction, Hart et al. [18] looked at interpersonal motions - synchrony and mutual-followership - between two communication styles in acted scenarios (disengaged, engaged). While their corpus is larger, investigation of real interactions would be required to validate these findings.

Finally, a few studies used a combination of visual cues. Porhet et al. [39] extracted head movements, posture, gaze, eyebrow, hand gesture, and smile to identify of cues leading to patients' feedback in the form of rules ($X \implies Y$, e.g. $\text{doctor}_{\text{head nod}} \implies \text{patient}_{\text{head nod}}$). They assessed the confidence of an extracted rule by computing the proportion of cases verifying the rule. While patterns of interactions were identified, low confidence (the confidence scores of the top 11 rules are between 0.36 and 0.12) and the acted nature of the data limits generalisability.

Finally, speech related investigations are mostly focused on *silences and pauses*. Identified as a significant component of the medical consultation, notably by Byrne and Heath [5], the therapeutic use of silences described in theoretical models can be detected using a systematic approach, while evidence of more complex usage and functions of pauses and silences is reported. Durieux et al. [12] investigated connectional silences: pauses between clinician's and patient's turns identified as potential markers of shared understanding and presence. They demonstrated the capacity of machine learning to detect connectional silences (recall 0.58, precision 1 compared with human coders) and support the annotation by human coders (human annotation without automation took 61% more time) but did not proceed to their analysis as a part of the communication beside a quantification over 32 samples. Conversational pauses are an element of the dynamics of the interaction (as a marker of engagement, power distribution, turn-taking, listening, connection, politeness, etc.). Manukyan et al. [26] investigated the performance of automated methods for the identification of conversational pauses, on its own (they report an accuracy of 94.4%) and as a supporting tool for manual coders (the annotation of one hour of audio took between

113 and 156 minutes for human coders, whereas the automated classification took 1.46 minute on a standard laptop). All studies used simple definitions of pauses, usually based on the length of silences ($t_{duration} > 3s$), and simple definition of pauses, i.e. not characterising types of pauses.

3.4. Methodologies employed

A first overview of the assessment of the studies (see table A.10) outlines shared limitations. Concerning research implications, reviewed studies used generally novel methodologies (23 out of 27), going beyond the simple application of existing tools. Replicability was low (10 studies) or partial (17), notably due to the expected unavailability of datasets. Generalisability was globally high (seventeen studies) with only five studies using a methodology tailored for a specific setting and five studies requiring sensible work to adapt it to other contexts. Concerning the evaluation of the risks of bias, the major limitation came from feature imbalance (25 studies) associated with a lack of suitable metrics in twelve studies. Fifteen did not provide contextualised results and six did not account for overfitting. Seven did not use real life settings (e.g. features simulated interactions), and ten had rather small sample size (seven used less than 50 documents, three less than 100). Regarding other limitations, automation was only partial in twenty-three studies, and the large majority (nineteen) required the transcription of the encounters while eighteen relied on the spoken content of the interaction (the difference is explained by one study that investigated phases of the interaction [4]).

Most studies (12) used supervised learning with common classifiers (e.g. decision trees, SVM, and neural networks) to predict a type of interaction at the utterance level (e.g. coarse coding of VRCoDES in Birkett et al. [3]) or at the session level (e.g. prediction of student success in Carnell et al. [6]). Tanana et al. [49] predicted of MISC behavioural codes at the utterance and session level, with good results at session level but low performance on utterances. Venek et al. [52] used conversation dynamic features, verbal information (topic identification) and acoustic features to classify non-suicidal and suicidal patients, and a second classification of repeaters and non-repeaters. The use of clinicians' features in addition to patients' features lead to a slight accuracy improvement (90% vs 85%) in the first step but marginally reduced the performance of the second step (-1.2%). Chakraborty et al. [7] had a similar task, correlating body movement and speech with prediction of negative symptoms of schizophrenia. This approach was used to assess successful interactions [47, 30, 6], to detect connecting silences [12] and in content-based

analysis to classify topics [35], emotional valences [34], speech acts [28], and gain words [17].

Observational studies, identifying patterns from extracted features constitute another group of investigations. These studies focus on specific elements of the communication, such as semantic similarity between the patient and the physician, to find correlation between observed variations and expected dependant and independent variables. Word-based studies are common, including studies of dominance [46] anf temporal ordering of activities in [20]. Wong et al. [58] investigated word-related statistics (e.g. occurrence and co-occurrence) in relation with the perceived quality of the consultation by patients. Vrana et al. [54] searched semantic (dis-)similarities across patients and doctors of different ethnic backgrounds, observing significantly lower communication similarity from white physicians, controlling for confounders (gender of both participants). Other features were used. Rasting et al. [40] used facial display of affect to correlate patients expression with therapists emotional reactions. Porhet et al. [39] investigated sequences of multimodal behaviour elements that elicit feedback from patients. Pearce et al. [38] observed computer use behaviour. Mase et al. [27] identified points of interest in the recordings of trainings based on patterns of interactions (sequences of multimodal behaviour).

Another group of studies performed clustering to detect types of interactions (unsupervised learning, e.g. grouping clinicians by style of communication), or to distinguish between known groups (supervised learning, e.g. interaction featuring good and bad communication). For instance, Wallace et al. [56] clustered physicians based on turn-taking patterns and speech act transitions through semantics, detecting two clusters corresponding to the difference in patients' evaluation for three categories of questions investigated. Cuffy et al. [10] captured semantic aspects of communication, notably the relatedness between discourse content (however limited by the small scale of the study and the disparity between computed scores and self-reported questionnaires). Using the opposite approach (i.e. using fixed groups), Watson et al. [57] extracted word-related statistics on topics to compare speech during effective and ineffective interactions, as evaluated by experts using behavioural analysis, and found significant difference between effective and ineffective interactions in four out of five aspect of the communication. Manukyan et al. [26] evaluated the performance of automating the detection of conversational pauses with good results (accuracy=94.4%). Chiba et al. [8] investigated differences in topics found in conversations be-

tween doctors interacting with caregivers of patients who died at home or at hospital. Blomqvist et al. [4] found differences between patients with and without attention deficit and hyperactivity disorder ADHD (higher degree of non-coordination for patients with ADHD).

Finally, some studies used a combination of approaches, for instance Hart et al. [18] first conducted an exploration of motion synchrony between patients and nurses, before classifying interactions using engaged and disengaged scenarios (accuracy=0.72%)

4. Conclusion

Many of the studies identified reviewed used structured or semi-structured interviews, featuring more restricted interactions than in medical consultations. While this helps retrieving investigated cues (behaviours, emotions, gestures) more consistently, it also limits the weight of the findings of these studies as regards the less restricted range of interaction that take place during medical consultations.

A number of aspects of the patient-clinician communication, verbal and non-verbal, have been investigated using systematic approaches to facilitate objective evaluations. However, while much of the focus has been set on the semantic of the interaction, investigations using paralinguistic and non-verbal components are much less common. In fact, the analysis of non-verbal behaviour has focused more on visual aspects (face, posture, movements).

The analysis of speech is fairly common in studies seeking to discriminate impaired speech of a person (e.g. patients affected by a physical or neurological conditions). However, the characterisation of speech during the patient-clinician communication is mostly limited to the quantification of silences and pauses using simple definitions.

While some touched upon some of its elements, very few studies have investigated the structure of turns in the interaction. Turn-taking behaviours combined with the analysis of speech patterns remains an area that was not investigated, supporting and legitimating the focus of the work. The rather unexplored domain of the paralinguistic and non-verbal elements associated with the automation of the assessment of the communication happening in consultations constitutes its background [44].

Overall, the result of this review shows that the automated analysis of consultation is feasible. Numerous elements of the communication happening during medical encounters can be retrieved and analysed automatically.

The literature focuses largely on semantics, while little work exists on paralinguistic analysis. Methodologies employed in consultation analysis vary. Whereas semantic analysis often use existing frameworks as a basis, studies of non-verbal and paralinguistic communication shared little methodological common ground. Much remains to be done to standardise elements, features, and metrics for the analysis of medical consultations. While majority of studies we reviewed used automation in classification tasks, exploration and identification of patterns of interaction is also a focus of research. The results of this review shows that features of multimodal behaviour in consultations can be extracted and identified. The characterisation of these features complement existing knowledge on elements of patient-clinician communication that are new and complementary, notably relating to linguistic, paralinguistic and non-verbal behaviour.

Table 1: PICOS table.
ADHD: Attention Deficit Hyperactivity Disorder, ANOVA: Analysis of variance, BRL: Bayesian Rule Lists, CAT: Communication Accommodation Theory, C-SSRS: Columbia Suicide Severity Rating Scale, kNN: k-Nearest Neighbours, LSA: Latent Semantic Analysis, MHD: Mental Health Discussion , MISC: Motivational Interviewing Skill Code, ML: Machine Learning, OSCE: Objective Structured Clinical Examination , PCA: Principal Component Analysis, RNN: Recurrent Neural Network, SIQ-JR: Suicidal Ideation Questionnaire-Junior, SP: Standardised Patients, SVM: Support Vector Machine, TF-IDF: Term frequency-inverse document frequency, UQ: Ubiquitous Questionnaire

Study	**Participants**	**Interventions**	**Comparison groups**	**Outcomes studied**
Birkett et al. [3]	91 adult female breast cancer patients, 2 therapeutic radiographers	One-on-one consultations of patients undergoing radiotherapy	-	Classification accuracy Text-based analysis.
[4]	22 children with ADHD, 47 children without, parents, 1 dentist	Annual dental recall visit	Children with and without ADHD	Statistical difference in interactions patterns between groups.
Carnell et al. [6]	464 graduate students, AI agents (number not reported)	Student training sessions of GP consultations.	-	Predictive accuracy of interpretable classification model - BRL.

Table 1: PICOS table (continued)

Study	Participants	Interventions	Comparison groups	Outcomes studied
Chakraborty et al. [7]	46 patients and 23 healthy controls	Dedicated medical interview (not consultation).	Patients diagnosed with schizophrenia, healthy controls participants	Association between objective and clinicians' subjective evaluations of motor movement. Performance of the classification between individuals with schizophrenia and healthy individuals.
Chiba et al. [8]	18 patients at terminal stage of cancer receiving periodical medical care, 24 doctors, family caregivers.	Doctors' visits to patients.	Home death cases and hospital death cases.	No evaluation of automated processing. Difference of which topics were discussed with caregivers during doctors' visits between the two groups.
Cuffy et al. [10]	132 patients, 17 physicians	Patient–physician interactions in a primary care clinic	-	Word similarity (global: reuse of similar words in the whole interaction), responsiveness between participants (utterance-based), topic reuse. Pearson's correlations between the computed quality scores and patients' self-reported trust and satisfaction. Low linear correlation between scores of the 3 methods and patient's evaluation
Durieux et al. [12]	225 hospitalized patient referred for palliative care consultation, clinicians (number not reported)	Palliative care consultations	Comparison of semi-automatic and manual silences categorisation.	Reliability, efficiency and sensitivity of the identification
Fridman et al. [17]	208 patients diagnosed with low or intermediate-risk prostate cancers. 8 urologist, 3 radiation oncologists	Outpatient consultations about treatment options with patients diagnosed with early-stage prostate cancer	Patients choosing cancer treatment, patients choosing active surveillance.	Physician's use of gain or loss words, association between words use and patients' treatment choices. Use of loss words was associated with patient's choice of cancer treatment. Physicians' use of loss words was correlated with recommendations for cancer treatment.
Hart et al. [18]	43 recruited subjects, 1 simulated physician	Presentation of a drug to the patient and direction to apply the ointment	Two acted scenarios: disengaged and detached, engaged and suggestive.	Correlation in total kinetic energy, interpersonal motion synchrony and entrainment
Kocaballi et al. [20]	31 primary care patients, 4 primary care physicians	Medical interviews in general practice.	-	Type and flow of clinician's activities.

Table 1: PICOS table (continued)

Study	Participants	Interventions	Comparison groups	Outcomes studied
Manukyan et al. [26]	225 hospitalised patients with advanced cancer, 54 palliative care clinicians	Palliative care consultation	Human annotators, Machine learning classifier	Performance and efficiency of the classifier.
Mase et al. [27]	10* medical students (*unclear), simulated patients (number not reporter)	Simulated medical interviews with simulated patients for training interview skills of medical students	Generated summaries and actual recordings by physicians	Comparison of the evaluation between generated summaries and actual recordings (38 items). The method was able to identify points of interest in the recordings of trainings.
Mayfield et al. [28]	415 patient. 45 physicians, nurse practitioners, or physician assistants	Routine outpatient visits by people living with HIV	Manual and human evaluation	Evaluation of the performance of the automation.
Mistica et al. [30]	2 SP enacted by qualified doctors, 11 international medical graduates enrolled in a bridging course	Objective structured clinical examinations with 2 stations: sexually transmitted disease genital herpes, and bowel cancer. 1 SP per station.	-	Prediction performance on the outcome of the OSCE assessment, and analysis of communication aspects influencing it.
Park et al. [35]	255 patients (evidence-based MHD, perfunctory MHD, and no MHD), 56 physicians	Periodic health examinations	-	Classification accuracy for talk-turns. precision, recall, and F1-scores at the visit level. Sequential models had higher classification accuracy at the talk-turn level and higher precision at the visit level. Sequential information across talk-turns improves topic prediction accuracy. Best results achieve with hierarchical gated recurrent units
Park et al. [34]	350 patients, 84 physicians	Elderly patients' doctor visits	Human annotators, automated annotation	Agreement between automated classification and human ratings
Pearce et al. [38]	308 patients, 36 GPs	Routine clinical consultations (UK, Australia)	-	Proportion of triadic interactions, inclusive behaviour
Porhet et al. [39]	13 doctors, actor patients (number not reporter)	Real training sessions of doctors with simulated patients (actors) for breaking bad news scenario	-	Confidence score (frequency of valid occurrences) of extracted rules (cue $X \implies$ feedback Y)

Table 1: PICOS table (continued)

Study	Participants	Interventions	Comparison groups	Outcomes studied
Rasting et al. [41]	Therapists, 12 patients with various psychosomatic disorders	Real interviews of patients for in-patient psychotherapy	- (different degrees of alexithymia)	Correlation between categories in patients' evaluation of psychosomatic disorders and behaviours.
Sakai and Carpenter [46]	86 patients and companions dyads, physicians (number not reporter)	Clinical interview, exam and formulation of diagnostic	Patients diagnosed with and without dementia.	Differences in actual and perceived verbal dominance, differences in makers of power between groups of patients and influence on patient's evaluation.
Sen et al. [47]	122 patients with stage 3 or stage 4 advanced solid tumors, 40 oncologists	Doctor-patient conversations of late-stage cancer patients	-	Speech features related to patients' evaluation. 2 clusters of communication styles identified: several communication styles associated with higher and lower communication ratings. Poor results of machine learning for the classification of doctors with highest communication ratings.
Tanana et al. [49]	341* primary care patients at a safety-net hospital, including 76* university students with problematic drug or alcohol use (* unclear). clinicians (number not reporter)	Short motivational interviews	-	Capacity of machine leaning methods to predict MISC codes at utterance and session level.
Venek et al. [52]	60 adolescents: 30 suicidal (13 repeaters and 17 non-repeaters), 30 non-suicidal. 1 social worker	Q and A to 16 questions: Columbia Suicide Severity Rating Scale (C-SSRS version 1/14/2009), Suicidal Ideation Questionnaire-Junior (SIQ-JR version 1987 [16]), Ubiquitous Questionnaire (UQ version 2011 [1])	Suicidal (repeater / non-repeater) and non suicidal patients Classification of the patients using a two layers hierarchical classifier	Statistical significance of differences of discourse features between suicidal and non-suicidal adolescents, and between suicidal repeaters' behaviours and non-repeaters. mainly acoustic information are statistically significant to discriminate between repeaters and non-repeaters. Verbal behaviour of patients and clinicians is important to assess suicidal risk. Nonverbal behaviour, notably acoustic features, is important to assess the potential of suicidal re-attempt. Accuracy of hierarchical classification is fairly good (67.7%)

Table 1: PICOS table (continued)

Study	**Participants**	**Interventions**	**Comparison groups**	**Outcomes studied**
Vrana et al. [53]	132 low-income patients, 17 physicians	Medical appointment in a primary care clinic	-	Patient-physician communication similarity and correlation with trust levels. Results were influenced by physician's race and gender, and patient's gender. Higher communication similarity was associated with less trust in physicians before the interaction and higher after.
Wallace et al. [56]	360* patients (* unclear), 41 doctors	Physician-patient visits	-	Correlation between detected clusters and patients' evaluations.
Watson et al. [57]	8* Simulated patients or family members (* unclear), 8 clinicians.	Simulated consultation of a clinician training program to discuss adverse events in patient care	4 effective consultations, 4 ineffective consultations (set by experts using behavioural analysis)	1: Statistical comparison (t-tests) of agreement on effective/ineffective rating of the interactions, and of each part of the CAT using the mean scores of students' evaluations. 2: Interest of visualisation of concepts reuse for the discourse analysis of clinician-patient interaction. 2: In effective interactions, physicians approximated to patients more than patients approximated to physicians. Physicians engaged with the patients' conceptual contributions. The visualisation provided meaningful interpretation capacities for discourse analysis.
Wong et al. [58]	62 cases, paediatric dentists, certificated dental surgery assistants, child patients, and their care-givers (not detailed)	dental conversation with child patient and caregiver	-	Relation of themes to evaluation. Frequent use of positive reinforcement/reassurance was significantly associated with higher perceived quality of communication. Specific terms and behaviour were identified.

Table 2: Design, methodoloy and results summary.
ADHD: Attention Deficit Hyperactivity Disorder, ANOVA: Analysis of variance, BRL: Bayesian Rule Lists, CAT: Communication Accommodation Theory, C-SSRS: Columbia Suicide Severity Rating Scale, kNN: k-Nearest Neighbours, LSA: Latent Semantic Analysis, MHD: Mental Health Discussion , MISC: Motivational Interviewing Skill Code, ML: Machine Learning, OSCE: Objective Structured Clinical Examination , PCA: Principal Component Analysis, RNN: Recurrent Neural Network, SIQ-JR: Suicidal Ideation Questionnaire-Junior, SP: Standardised Patients, SVM: Support Vector Machine, TF-IDF: Term frequency-inverse document frequency, UQ: Ubiquitous Questionnaire

Study	**Design, description, aim**	**Methodology**	**Results**
Birkett et al. [3]	Automation of the coding of textual transcripts of medical interactions (VR-CoDES)	Utterance representation using bag-of-words and tf-idf, classification (naïve Bayes, logistic regression, support vector machines, decision trees)	High accuracy of the automated classification of VR-CoDES. Similar performance of the different classifiers and n-grams, TF-IDF outperformed other data representation.
Blomqvist et al. [4]	Analysis of behavioural interactions between the dentist and child patients	Quantification of the different parts of interaction using video recordings	No differences in the number of initiatives (questions), focus, and functions of verbal expressions by the dentist. Children with ADHD made significantly more initiatives, made fewer verbal responses, more frequently did not respond, and had a higher degree of avoidance of response, no-response or incongruity between the verbal and non-verbal response
Carnell et al. [6]	Investigation on practicability of ML algorithms for classification of students' success (outcome of the evaluation: pass or fail).	Classification and performance of ML over prior probability of success based on manually annotated textual content of the interaction: domain skills (medical discovery information, science reasoning) and communication skills (medical question style, dialect switching).	Machine learning using communication-based features can be used to predict success of student interaction. Interpretable classifier offers slightly lower performance than classic classifiers (0.62 vs 0.66), both slightly better than baseline (accuracy 5% over prior probability of success).
Chakraborty et al. [7]	Development of objective methods to quantify symptoms of schizophrenia	Categorisation of participants. Extraction of body posture and movements, and classification using SVM and kNN (with and without feature selection).	Multiple moderate negative correlations between objective (detected) motor movements and negative symptoms. 3 movements with $corr \leq -0.47$ and $p<0.001$, 7 with $corr \leq -0.44$ $p<0.01$, 28 movements with $corr \leq -0.29$ $p<0.05$.

Table 2: Design, methodoloy and results summary (continued).

Study	Design, description, aim	Methodology	Results
Chiba et al. [8]	Study of topics discussed by doctors and caregivers during end-of-life care to identify topics related to patients' home death	Identification of topics from recorded exchanges.	The patients' places of death is correlated with difference in the topics discussed (2 out of 3 main topics, 8 out of 15 subtopics).
Cuffy et al. [10]	Methods to capture, model and evaluate patient–physician communication using semantics.	Communication quality based on patient's evaluation (trust before, trust after, satisfaction). Clustering of interactions based on word embeddings trained on corpus and generic corpora (Wikipedia and Medline), utterances represented by centroid vector.	Patients were generally more responsive to their physician. Low linear correlation between scores of the 3 methods and patient's evaluation
Durieux et al. [12]	Identification of connectional silence in palliative care consultations using Machine learning and manual annotation.	Manual and automated extraction of silences with manual classification.	Connectional Silence can be identified using a semi-automated method with good reliability (kappa 0.62 on the found clips), efficiency (+61%) and sensitivity (No silence missed).
Fridman et al. [17]	Study of gain–loss information framing in the physicians' choice of words during risks and benefits discussion with cancer patients	Extraction of gain/loss words using a dictionnary. Logistic regression tests between word use and outcome.	Physicians recommending cancer treatment used slightly fewer words related to losses and significantly fewer words related to death. Use of loss words was associated with patient's choice of cancer treatment. Physicians' use of loss words was correlated with recommendations for cancer treatment.
Hart et al. [18]	Automated video analysis tool for non-verbal interactions	Pixel based quantification of movement in the videos of the encounters	Large differences found between scenarios. Engaged: higher motion synchrony, actor and subject follow each other's motion in turns, more equal turn-taking.
Kocaballi et al. [20]	Identification of clinical activities and their interrelationships during primary care visits.	Manual annotation of activities: type, frequency, sequence, network. Semantic analysis of transitions between activities.	Identification of temporal sequencing of activities and transitions between activities (central activity: discussion about patients' present complaint).
Manukyan et al. [26]	Automating conversation analysis in clinical settings.	Identification of contiguous intervals without voicing>2s.	Positive capacity of machine learning to automatically identify conversational pauses in inpatient serious illness conversations, while reducing coding time by two orders of magnitude.

Table 2: Design, methodoloy and results summary (continued).

Study	Design, description, aim	Methodology	Results
Mase et al. [27]	Visual summarisation method for multi-modal dialogues using pattern and motif mining.	Identification of patterns based on annotations of elements of interaction, selection of salient patterns.	39.5% features matched between summaries and manual reviews of the videos (26.3% mis-matched features, 26% unknown from summary) The method was able to identify points of interest in the recordings of trainings.
Mayfield et al. [28]	Automation of the coding of speech acts in clinical communication	Prediction of patient-reported measures of communication quality based on information-giving ratio.	Reliability is too low for the replacement of manual evaluation, but the lowered cost of the evaluation can help in exploratory research, preliminary evaluation of annotation schemes, and rapid screening of interactions.
Mistica et al. [30]	Discourse analysis of training sessions	Supervised classification of interactions based on extracted features from manual annotations.	High correlation between assessment criteria based on communication and language skills and successful outcome. Word-based feature sets were the best predictors.
Park et al. [35]	Detection of conversation topics in primary care using machine learning	Bag-of-words encoding of texts. Classification using machine learning models: single/multiple talk-turns (logistic classifiers, support vector machines, gated recurrent units), and sequential models (conditional random fields, hidden Markov models, and hierarchical gated recurrent units)	Independent models had higher recall scores at the visit level. Sequential models had higher classification accuracy at the talk-turn level and higher precision at the visit level. Sequential information across talk-turns improves topic prediction accuracy. Best results achieve with hierarchical gated recurrent units
Park et al. [34]	Evaluation of machine learning to classify emotional valence of utterances.	Classification of emotional valence (positive, negative, neutral) of utterances (bag-of-word representation) using 2 machine learning models (recurrent neural network with a hierarchical structure, logistic regression classifier).	Performance of automated emotion classification was comparable to human-human inter-rater agreement.

Table 2: Design, methodoloy and results summary (continued).

Study	Design, description, aim	Methodology	Results
Pearce et al. [38]	How computer interaction is integrated in the communication during medical consultation	Manual annotation of participants' behaviour	The way clinicians integrate the use of the computer in the interaction results in more inclusive consultations, influences patient's engagement, and is associated with more complete clinical records. 36.5% of consultations classified as inclusive. Triadic interactions during inclusive consultations are more frequent and longer.
Porhet et al. [39]	Exploration of verbal and nonverbal signals in clinician-patient communication. Identification of doctor's verbal and nonverbal cues leading to patients' feedback	Manual annotation of cues, extraction of sequences in the last five tokens preceding a feedback, extraction of rules based on sequences (sequence of type of interaction leading to specific type of feedback)	10 rules identifies, confidence score between 0.36 and 0.12, 5 rules with cs<0.2
Rasting et al. [41]	Study of affective facial expression in a dyadic therapeutic interaction in clinician-patient communication. Identification of emotional reactions of therapists to facial affect display by patients.	Manual annotation of interviews (beginning and end) and coding of facial expressions analysed by computer.	Patients with high alexithymia displayed less aggressive affects (anger, disgust, contempt). Therapists interacting with alexithymic patients tended to display negative affects: contempt using total score, and contempt, fear, and sadness using subscales of the assessment tool.
Sakai and Carpenter [46]	Investigation of markers of power in linguistic expressions during dementia diagnosis disclosure	Statistical analysis. Assessment of differences in perception of verbal dominance: ANOVA. Influenced by dementia status: t-test. Confounders: Bivariate correlation. Prediction of patient's evaluation and condition on use of markers: hierarchical regression.	Consultations were dominated by clinicians in speech duration (83%). Companions spoke more when patients had dementia. Makers of power were not predictive of patient's anxiety, depression, or satisfaction.
Sen et al. [47]	Identification of latent styles in doctor-patient communication using affective and nonaffective speech features	Extraction of speech features, sentiment analysis using Natural Language ToolKit and lexicon. Statistical analysis of features, unsupervised clustering for communication styles identification and association with outcome	Differences in numerous language features between best-rated doctors and other doctors. 2 clusters of communication styles identified: several communication styles associated with higher and lower communication ratings. Poor results of machine learning for the classification of doctors with highest communication ratings.

Table 2: Design, methodoloy and results summary (continued).

Study	Design, description, aim	Methodology	Results
Tanana et al. [49]	Automated coding of motivational interviewing using Natural Language Processing	Dependency trees with discrete sentence features (N-grams) and RNN with word embedding.	Common utterance and session level codes could be predicted, with results comparable to human reliability. Rarer codes were not well predicted.
Venek et al. [52]	Identification and assessment of suicidal risk using verbal and nonverbal responses to a questionnaire	Interviews were separated in two: interaction during UQ or not. Discourse features(conversational, verbal and acoustic) were extracted and tested individually for significance. Classification of the patients using a two layers hierarchical classifier	Significant differences found in all three types of features (22 for patients and 21 for clinicians) the classification of suicidal and non-suicidal patients. mainly acoustic information are statistically significant to discriminate between repeaters and non-repeaters. Verbal behaviour of patients and clinicians is important to assess suicidal risk. Nonverbal behaviour, notably acoustic features, is important to assess the potential of suicidal re-attempt. Accuracy of hierarchical classification is fairly good (67.7%)
Vrana et al. [53]	Characterisation of semantic similarity of the patient's and physician's language	Extraction of participants' speech in semantic space using Latent Semantic Analysis and relation to evaluation of trust in the physician before and after the interaction (General Estimating Equations regressions to correct for bias).	LSA captured individual differences during medical interactions. Significant positive relationship was found between patients' and physicians' speech. Results were influenced by physician's race and gender, and patient's gender. Higher communication similarity was associated with less trust in physicians before the interaction and higher after.
Wallace et al. [56]	Characterisation of physicians' variation in communication patterns to cluster communication styles.	sequential analysis of speech acts transitions grouped into a physician-specific vector. Clustering of physicians' vectors into 2 classes using k-means on PCA reduce matrix of physicians.	Variations between the two detected clusters are detected but are not significant in 2 of 3 categories of the patients evaluation (positive correlation with HIV-specific issues evaluation but not for Overall issues, Adherence).

Table 2: Design, methodoloy and results summary (continued).

Study	Design, description, aim	Methodology	Results
Watson et al. [57]	Accommodative communication strategies of clinician discussing the consequent patient harm following adverse events in patient care using direct evaluation and computer support tool.	Two parts study: (1) rating of CAT strategies (Overall progress, interpretability, discourse management, interpersonal control and emotional expression) by first-year psychology students, and (2) textual analysis of approximation using convergence and divergence in reuse of concepts. 2: Interest of visualisation of concepts reuse for the discourse analysis of clinician-patient interaction.	1: significant agreement on the rating of the interaction. Significant difference between effective and ineffective interactions 4 out of 5 CAT parts, while discourse management was not more highly rated in the effective recordings. 2: In effective interactions, physicians approximated to patients more than patients approximated to physicians. Physicians engaged with the patients' conceptual contributions. The visualisation provided meaningful interpretation capacities for discourse analysis.
Wong et al. [58]	Content analysis of prominent themes in the clinician-patient conversation and its relation to perceived quality of communication	Visual text analytics using word occurrence and co-occurrence statistics, dimensionality reduction using PCA followed by qualitative analysis of related conversation content.	5 themes were identified as prominent out of 13 extracted: disease treatment, treatment procedure related instructions, preparation for examination, positive reinforcement/reassurance, family/social history. Frequent use of positive reinforcement/reassurance was significantly associated with higher perceived quality of communication. Specific terms and behaviour were identified.

Acknowledgements

This research received funding from the Health Research Board, Ireland, towards the INCA project (Interaction Analytics for Automatic Assessment of Communication in Healthcare).

Conflicts of Interest

The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

[1] Alloatti, F., Bolioli, A., Bosca, A., Guadalupi, M., 2020. The RiMotivAzione dialogue corpus Analysing Medical Discourse to Model a Digital Physiotherapist, in: LREC 2020 Language Resources and Evaluation Conference 11-16 May 2020, p. 16.

[2] Barzilay, R., Israel, N., Krivoy, A., Sagy, R., Kamhi-Nesher, S., Loebstein, O., Wolf, L., Shoval, G., 2019. Predicting affect classification in mental status examination using machine learning face action recognition system: A pilot study in schizophrenia patients 10, 288.

[3] Birkett, C., Arandjelović, O., Humphris, G., 2017. Towards objective and reproducible study of patient-doctor interaction: Automatic text analysis based VR-CoDES annotation of consultation transcripts, in: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE. pp. 2638–2641.

[4] Blomqvist, M., Augustsson, M., Bertlin, C., Holmberg, K., Fernell, E., Dahllöf, G., Ek, U., 2005. How do children with attention deficit hyperactivity disorder interact in a clinical dental examination? A video analysis 113, 203–209. doi:10.1111/j.1600-0722.2005.00211.x.

[5] Byrne, P.S., Heath, C.C., . Practitioners' use of non-verbal behaviour in real consultations 30, 327–331. arXiv:7411517.

[6] Carnell, S., Lok, B., James, M.T., Su, J.K., 2019. Predicting student success in communication skills learning scenarios with virtual humans, in: Proceedings of the 9th International Conference on Learning Analytics & Knowledge, pp. 436–440.

[7] Chakraborty, D., Tahir, Y., Yang, Z., Maszczyk, T., Dauwels, J., Thalmann, D., Thalmann, N.M., Tan, B.L., Lee, J., 2017. Assessment and prediction of negative symptoms of schizophrenia from RGB+ D movement signals, in: 2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP), IEEE. pp. 1–6.

[8] Chiba, H., Ogata, T., Ito, M., Kaneko, S., 2018. Identification of Topics Explained by Home Doctors to Family Caregivers with Cancer Patients Died at Home: A Quantitative Text Analysis of Actual Speech in All Visits 245, 251–261.

[9] Cockerham, W.C., 2017. Medical Sociology. American Cancer Society. doi:10.1002/9781118410868.wbehibs548.

[10] Cuffy, C., Hagiwara, N., Vrana, S., McInnes, B.T., . Measuring the quality of patient–physician communication 112, 103589. doi:10.1016/j.jbi.2020.103589.

[11] Davis, M.S., 1968. Variations in patients' compliance with doctors' advice: An empirical analysis of patterns o communication. 58, 274–288. arXiv:5688773.

[12] Durieux, B.N., Gramling, C.J., Manukyan, V., Eppstein, M.J., Rizzo, D.M., Ross, L.M., Ryan, A.G., Niland, M.A., Clarfeld, L.A., Alexander, S.C., 2018. Identifying connectional silence in palliative care consultations: A tandem machine-learning and human coding method 21, 1755–1760.

[13] Elwyn, G., Barr, P.J., Grande, S.W., Thompson, R., Walsh, T., Ozanne, E.M., a. Developing CollaboRATE: A fast and frugal patient-reported measure of shared decision making in clinical encounters 93, 102–107. doi:10.1016/j.pec.2013.05.009.

[14] Elwyn, G., Edwards, A., Kinnersley, P., Grol, R., 2000. Shared decision making and the concept of equipoise: The competences of involving patients in healthcare choices. 50, 892–899.

[15] Elwyn, G., Grande, S.W., Barr, P., b. Observer OPTION 5 Manual.

[16] Fairhurst, K., May, C., . Knowing patients and knowledge about patients: Evidence of modes of reasoning in the consultation? 18, 501–505. doi:10.1093/fampra/18.5.501.

[17] Fridman, I., Fagerlin, A., Scherr, K.A., Scherer, L.D., Huffstetler, H., Ubel, P.A., . Gain–loss framing and patients' decisions: A linguistic examination of information framing in physician–patient conversations 44, 38–52. doi:10.1007/s10865-020-00171-0.

[18] Hart, Y., Czerniak, E., Karnieli-Miller, O., Mayo, A.E., Ziv, A., Biegon, A., Citron, A., Alon, U., . Automated Video Analysis of Non-verbal Communication in a Medical Setting 7. doi:10.3389/fpsyg.2016.01130, arXiv:27602002.

[19] Joshi, J., Goecke, R., Alghowinem, S., Dhall, A., Wagner, M., Epps, J., Parker, G., Breakspear, M., 2013. Multimodal assistive technologies for depression diagnosis and monitoring 7, 217–228.

[20] Kocaballi, A.B., Coiera, E., Tong, H.L., White, S.J., Quiroz, J.C., Rezazadegan, F., Willcock, S., Laranjo, L., 2019. A network model of activities in primary care consultations 26, 1074–1082.

[21] Kurtz, S.M., Silverman, J.D., 1998. Calgary Cambridge Guide to the Medical Interview.

[22] Laws, M.B., Beach, M.C., Lee, Y., Rogers, W.H., Saha, S., Korthuis, P.T., Sharp, V., Wilson, I.B., 2013. Provider-patient adherence dialogue in HIV care: Results of a multisite study 17, 148–159.

[23] Liu, C., Calvo, R.A., Lim, R., 2016. Improving Medical Students' Awareness of Their Non-Verbal Communication through Automated Non-Verbal Behavior Feedback 3. doi:10.3389/fict.2016.00011.

[24] Luz, S., 2009. Locating case discussion segments in recorded medical team meetings, in: Proceedings of the ACM Multimedia Workshop on Searching Spontaneous Conversational Speech (SSCS'09), ACM Press, Beijing, China. pp. 21–30.

[25] Luz, S., 2012. The nonverbal structure of patient case discussions in multidisciplinary medical team meetings. ACM Transactions on Information Systems (TOIS) 30, 17:1–17:24. doi:10.1145/2328967.2328970.

[26] Manukyan, V., Durieux, B.N., Gramling, C.J., Clarfeld, L.A., Rizzo, D.M., Eppstein, M.J., Gramling, R., 2018. Automated detection of conversational pauses from audio recordings of serious illness conversations in natural hospital settings 21, 1724–1728.

[27] Mase, K., Sawamoto, Y., Koyama, Y., Suzuki, T., Katsuyama, K., 2009. Interaction pattern and motif mining method for doctor-patient multimodal dialog analysis, in: Proceedings of the ICMI-MLMI'09 Workshop on Multimodal Sensor-Based Systems and Mobile Phones for Social Computing, pp. 1–4.

[28] Mayfield, E., Laws, M.B., Wilson, I.B., Penstein Rosé, C., 2014. Automating annotation of information-giving for analysis of clinical conversation 21, e122–e128.

[29] McKinstry, B., 2000. Do patients wish to be involved in decision making in the consultation? A cross sectional survey with video vignettes 321, 867–871.

[30] Mistica, M., Baldwin, T., Cordella, M., Musgrave, S., 2008. Applying discourse analysis and data mining methods to spoken OSCE assessments, in: Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pp. 577–584.

[31] Mitra, V., Shriberg, E., Vergyri, D., Knoth, B., Salomon, R.M., 2015. Cross-corpus depression prediction from speech, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 4769–4773.

[32] Moher, D., . Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement 151, 264. doi:10.7326/0003-4819-151-4-200908180-00135.

[33] Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., . Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement 339, b2535. doi:10.1136/bmj.b2535, arXiv:19622551.

[34] Park, J., Jindal, A., Kuo, P., Tanana, M., Lafata, J.E., Tai-Seale, M., Atkins, D.C., Imel, Z.E., Smyth, P., a. Automated rating of patient and physician emotion in primary care visits doi:10.1016/j.pec.2021.01.004.

[35] Park, J., Kotzias, D., Kuo, P., Logan IV, R.L., Merced, K., Singh, S., Tanana, M., Karra Taniskidou, E., Lafata, J.E., Atkins, D.C., Tai-Seale, M., Imel, Z.E., Smyth, P., b. Detecting conversation topics in primary care office visits from transcripts of patient-provider interactions 26, 1493–1504. doi:10.1093/jamia/ocz140.

[36] Parsons, T., 1975. The Sick Role and the Role of the Physician Reconsidered 53, 257–278. doi:10.2307/3349493, arXiv:3349493.

[37] Pearce, C., Dwan, K., Arnold, M., Phillips, C., 2006. Analysing the doctor-patient-computer relationship: The use of video data. 14.

[38] Pearce, C., Kumarapeli, P., De Lusignan, S., 2010. Getting seamless care right from the beginning-integrating computers into the human interaction., in: EFMI-STC, pp. 196–202.

[39] Porhet, C., Ochs, M., Saubesty, J., De Montcheuil, G., Bertrand, R., 2017. Mining a multimodal corpus of doctor's training for virtual patient's feedbacks, in: Proceedings of the 19th ACM International Conference on Multimodal Interaction, pp. 473–478.

[40] Rasting, M., Brosig, B., Beutel, M.E., 2005a. Alexithymic Characteristics and Patient-Therapist Interaction: A Video Analysis of Facial Affect Display 38, 105–111. doi:10.1159/000085772, arXiv:15897680.

[41] Rasting, M., Brosig, B., Beutel, M.E., 2005b. Alexithymic characteristics and patient-therapist interaction: A video analysis of facial affect display 38, 105–111.

[42] Richardson, W.S., Wilson, M.C., Nishikawa, J., Hayward, R.S., 1995. The well-built clinical question: a key to evidence-based decisions. ACP Journal Club 123, A12–A13.

[43] Roter, D.L., Larson, S., . The Roter interaction analysis system (RIAS): Utility and flexibility for analysis of medical interactions 46, 243–251. doi:10.1016/S0738-3991(02)00012-5, arXiv:11932123.

[44] Ryan, P., Luz, S., Albert, P., Vogel, C., Normand, C., Elwyn, G., . Using artificial intelligence to assess clinicians' communication skills 364, l161. doi:10.1136/bmj.l161, arXiv:30659013.

[45] Sakai, E.Y., Carpenter, B.D., 2011a. Linguistic features of power dynamics in triadic dementia diagnostic conversations 85, 295–298. doi:10.1016/j.pec.2010.09.020, arXiv:21030193.

[46] Sakai, E.Y., Carpenter, B.D., 2011b. Linguistic features of power dynamics in triadic dementia diagnostic conversations 85, 295–298.

[47] Sen, T., Ali, M.R., Hoque, M.E., Epstein, R., Duberstein, P., 2017. Modeling doctor-patient communication with affective text analysis, in:

2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE. pp. 170–177.

[48] Stubenrouch, F.E., Pieterse, A.H., Falkenberg, R., Santema, T.K.B., Stiggelbout, A.M., van der Weijden, T., Aarts, J.A.W.M., Ubbink, D.T., . OPTION5 versus OPTION12 instruments to appreciate the extent to which healthcare providers involve patients in decision-making 99, 1062–1068. doi:10.1016/j.pec.2015.12.019.

[49] Tanana, M., Hallgren, K.A., Imel, Z.E., Atkins, D.C., Srikumar, V., 2016. A comparison of natural language processing methods for automated coding of motivational interviewing 65, 43–50.

[50] Tuckett, D., 1976. An Introduction to Medical Sociology. Routledge. arXiv:kShmAgAAQBAJ.

[51] Varul, M.Z., . Talcott Parsons, the Sick Role and Chronic Illness 16, 72–94. doi:10.1177/1357034X10364766.

[52] Venek, V., Scherer, S., Morency, L.P., Pestian, J., 2017. Adolescent suicidal risk assessment in clinician-patient interaction 8, 204–215.

[53] Vrana, S.R., Vrana, D.T., Penner, L.A., Eggly, S., Slatcher, R.B., Hagiwara, N., a. Latent Semantic Analysis: A new measure of patient-physician communication 198, 22–26. doi:10.1016/j.socscimed.2017.12.021.

[54] Vrana, S.R., Vrana, D.T., Penner, L.A., Eggly, S., Slatcher, R.B., Hagiwara, N., b. Latent Semantic Analysis: A new measure of patient-physician communication 198, 22–26. doi:10.1016/j.socscimed.2017.12.021.

[55] Walker, N., Cedergren, J.H., Trofimovich, P., Gatbonton, E., Mikhail, E., 2008. Someone to talk to: A virtual patient for medical history interview training in a second language , 1–9.

[56] Wallace, B., Dahabreh, I., Trikalinos, T., Laws, M.B., Wilson, I., Charniak, E., . Identifying differences in physician communication styles with a log-linear transition component model, in: Proceedings of the AAAI Conference on Artificial Intelligence.

[57] Watson, B.M., Angus, D., Gore, L., Farmer, J., 2015. Communication in open disclosure conversations about adverse events in hospitals 41, 57–70.

[58] Wong, H.M., Bridges, S.M., McGrath, C.P., Yiu, C.K.Y., Zayts, O.A., Au, T.K.F., 2017. Impact of prominent themes in clinician-patient conversations on caregiver's perceived quality of communication with paediatric dental visits 12, e0169059.

[59] Zimmermann, C., Del Piccolo, L., Bensing, J., Bergvik, S., De Haes, H., Eide, H., Fletcher, I., Goss, C., Heaven, C., Humphris, G., 2011. Coding patient emotional cues and concerns in medical consultations: The Verona coding definitions of emotional sequences (VR-CoDES) 82, 141–148.

Appendix A. Supplemental tables

Abbreviation	Full name
ANOVA	analyses of variance
BoW	bag of words
BTS	Bartlett's test of sphericity
CFR	conditional random fields
CNN	Counterpropagation Neural Networks
DT	Decision Tree
GRU	gated recurrent units
GT	golden truth
HMM	hidden Markov models
HGRU	hierarchical gated recurrent units
JSD	Jensen-Shannon Divergence
KLD	Kullback-Leibler Divergence
KMO	Kaiser-Meyer-Olkin
kNN	k-nearest neighbors
LIWC	Linguistic Inquiry and Word Count
LR	logistic regression
LSM	Language style matching
NB	Naïve Bayes
RNN	recurrent neural network
SVM	Support Vector Machine
t-tests	Student's t-tests
tf-idf	term frequency-inverse document frequency

Table A.3: List of abbreviation for methods and terms used in studies reported in the review.

Abbreviation	Full name
DCSQ	Dementia Care Satisfaction Questionnaire
DPFC	16-item Dental Patient Feedback on Consultation skills
GDS	Geriatric Depression Scale
MISC	Motivational Interviewing Skill Code
NSA	Negative Symptom Assessment
STAI	State-Trait Anxiety Inventory
TAS	Toronto Alexithymia Scale (TAS-26)

Table A.4: List of questionnaires (top) and medical scales (bottom) used in studies reported in the review.

Table A.5: Tasks table.

Study	Framework used	Material	Task performed	Performance	Datatset
[3]	Verona coding definitions of emotional sequences (VR-CoDES)	semi-structured textual transcripts	utterances-based classification of VR-CoDES	Manual: Kappa=0.67 (2 annotators on 5% of the corpora). Automated: F-score=0.72, Kappa=0.45	200 audio recordings of consultations
[4]	Ad-hoc (interaction phase, sequence, elements)	video, audio	topic identification, sequence annotation	Weighted kappa=0.98 on 5 documents	69 video recordings of the introduction phase (1-2 minutes)
[6]	Linguistic Inquiry and Word Count (LIWC), language style matching (LSM), distinction between biomedical or psychosocial utterances from RIAS.	Transcripts and manual annotations (Domain Topics for 3 stages).	Classification of student success for topic discovery (binary class - threshold on amount of information retrieved) in the 3 stages.	Held-out test folds. Classifier results between 5% and 10% (BRL) above baseline (always True classifier) for communication skill in one domain. No difference in the other two.	464 transcripts of student interactions with 6 Virtual Patients
[7]	Negative Symptoms Assessment (NSA-16) scale	Video (tracking of limbs) and audio recording. Linear and angular speeds of skeleton joints. Annotations of behaviour by psychologists.	Identification of postures correlated with NSA-16 items. Detection of NSA-16 items. Classification of participants (healthy, schizophrenic).	Leave-one-out cross-validation. Prediction of subjective ratings - 61-78% accuracy. Classification of patient - 74-87% accuracy.	69 medical semi-structured interviews by 1 trained psychometrician. (34 hours)
[8]	None, ad-hoc list of topics	Transcripts from speech annotated for number of occurrences of topics.	Topic identification (3 topics, 15 sub-topics) from parts of speech tagging and dictionary of terms.	Not reported	227 visits to 18 patients, interactions recorded during medical examinations and conversations with family caregivers.

Table A.5: Tasks table (continued).

Study	Framework used	Material	Task performed	Performance	Datatset
[10]	-	transcripts	Evaluation of communication quality (global, utterance, topic)	Pearson's correlation scores (Patient's Satisfaction): global=0.14, utterance=-0.07, topic= 0.08	132 video recorded patient–physician interactions in a primary care clinic
[12]	Connectional Silences taxonomy for the context of palliative care: emotional, compassionate, invitational.	audio recording	extraction of conversational pauses (1000 clips), manual annotation of type of silence	Automated extraction misidentified 41.3% of the clips as silences while none was missed. Manual annotation required 61% more time than the semi-automated method.	354 audio-recorded inpatient palliative care consultations
[17]	Ad-hoc dictionary of gain-loss terms	transcripts	word coding (2 classes: gain, loss)	manual annotation Krippendorf α=0.93 (50 documents). Automated annotation Krippendorff Alpha coefficient = 0.97.	286 audio-recorded face-to-face consultations (1 or 2 per patient)
[18]	-	Video recordings	Automated classification of interpersonal motion in video recordings using synchrony and mutual-followership indicators.	Accuracy: 0.72	43 videos of simulated encounters (22 disengaged, 21 engaged).
[20]	Ad-hoc scheme adapted from Waitzkin's framework	Video, verbatim transcripts.	Coding of clinician's activities, chart of performed tasks.	370 activities detected.	31 consultations: audio, video, computer screen video capture, notes from an observer.
[26]	Adapted definition of conversational pauses, Acoustic features (MFCC + zero-crossing rate, energy, energy entropy, and spectral entropy)	audio, manual annotation of conversational pauses (reference, 60 conversations)	Identification of conversational pauses (random forest classifier, 50 trees), extraction and aggregation of acoustic features.	Sensitivity=90.5%, specificity=94.5%, accuracy=94.4%, positive predictive value=30%	354 audio recordings of real-world serious illness conversations (9770 minutes)

Table A.5: Tasks table (continued).

Study	Framework used	Material	Task performed	Performance	Datatset
[27]	-	video, audio, manually annotated dialogue (speech, gaze, gesture).	Patterns extraction, motifs extraction (sequence of patterns)	1569 patterns observed, 18 patterns selected (covering 45% of the interview times)	10 Videotaped simulated medical interview (Performance evaluated by a medical doctor)
[28]	Generalized Medical Interaction Analysis System (GMIAS), Comprehensive Analysis of the Structure of Encounters System (CASES)	Manually annotated speech acts, aggregated in 3 classes: information-giving and requesting, other. Patients' ratings of provider communication.	(1) Text analysis: speech acts classification (logistic regression) during 3 types of interaction across the consultation (Presentation, Information, Resolution). (2) Prediction of communication quality.	(1) Accuracy=71.2%, κ=0.57 (full corpus). (2) 80% correct evaluation (corpus of 5 documents).	40 transcripts of routine outpatient visits
[30]	-	transcripts, manual annotation of turns and pauses. 38 features grouped in 11 feature sets.	Prediction of passed or failed examination (binary classification) based on combinations of features	Best results on all features and separating dataset by station. Baseline (majority vote): F-score=0.871, IB1 algorithm: F-score=0.882	22 video-recorded consultations of candidates enacting medical consultation scenarios (1 per station)
[35]	Multi-Dimensional Interaction Analysis coding system (modified, 39 topic labels)	transcripts, evaluations scores	Classification of talk turns	Talk-turn level accuracy: Hier-GRU=61.77%. Visit level accuracy: Windowed SVM=78.37	279 audio recorded primary care visits
[34]	-	transcripts	emotion recognition	Pearson correlation coefficients: Human (one vs rest)=0.60, RNN=0.60. R-precision(positive class) Human (OVR)=0.47, RNN=0.58. R-precision(negative class) Human (OVR)=0.44, RNN=0.45	353 video recorded patient–physician interactions in a primary care clinic

Table A.5: Tasks table (continued).

Study	Framework used	Material	Task performed	Performance	Datatset
[38]	[37] method for video analysis of doctor-patient-computer relationship.	audio, video, text of the medical notes, manual annotation of gaze	Classification of computer activity (doctor only, shared with patient): occurrences and length	Not reported	308 consultations with associated generated notes in the computer medical record
[39]	Verbal Cues: Enriched Orthographical Transcription, MarsaTag. Visual Cues: ad-hoc annotation set	Audio, segmented transcripts, Part of Speech tags	Manual annotation: head movements, posture, gaze, eyebrow, hand gesture, smile. Extraction of multimodal sequences leading to feedbacks	manual annotation k=0.63	13 videos of patient-doctor interaction (119 minutes)
[40]	Emotionally relevant movements in the face (EmFACS). Ad hoc: happiness, social smiles, sadness, fear, anger, disgust, contempt, social smiles, different affects	Video	Manual annotation of expressions. Affect recognition and pattern recognition (clinician reaction to patient display of affect).	Not reported	12 videos of patient-doctor interaction
[46]	Patient's evaluation: 20-item State-Trait Anxiety Inventory, 15-item Geriatric Depression Scale, Dementia Care Satisfaction Questionnaire	Transcripts, outcome questionnaires (anxiety, depressive symptoms, satisfaction)	Verbal dominance (number of words), ratio of first person singular and plural pronouns relative to all words	Not reported	86 videotaped sessions of physician-patient-companion triads

Table A.5: Tasks table (continued).

Study	Framework used	Material	Task performed	Performance	Datatset
[47]	14 word features. sentiment analysis: Valence Aware Dictionary for sEntiment Reasoning. Patient evaluation: 5 points Likert-type scales	transcripts, audio recordings, patient surveys	number of spoken/unique words, average positive/negative sentiment expressed, number of questions asked. Clustering of conversation features into "styles". Prediction of doctor-patient interaction rating.	Rating prediction: 71% accuracy	122 audio recordings of patient's visits
[49]	Motivational Interviewing Skill Code V2.1	transcripts, parsing of utterance, ratings	Prediction of behavioural codes (utterances) and summary elements (session)	Best performing: DSF. Utterance: $\kappa > 0.60$ open and closed questions, affirm, giving information, and follow/neutral. Session level: Intraclass correlation (ICC) > 0.75: inter-rater agreement, affirm, facilitate, giving information, follow/neutral, simple reflections, and open and closed questions all were in the excellent range. $0.60 <$ ICC < 0.75: sustain talk. ICC < 0.40: confront, structure, and advise	341 psychotherapy sessions in 6 MI clinical trial (78,977 clinician and client talk turns)
[53]	-	transcripts, trust scores	text-level extraction of semantic space by participant	- (mean patient-physician communication similarity correlation=0.142)	132 video recorded interactions

Table A.5: Tasks table (continued).

Study	Framework used	Material	Task performed	Performance	Datatset
[52]	speech features (9): COVAREP toolbox (v.1.2.0), linguistic features: LIWC	Segmented transcripts, audio	Statistical significance of conversation dynamic features, verbal information (topic identification) and acoustic features. Classification of suicidal risk (2 classes), classification of repeater's behaviour (2 classes), and combined hierarchical classification	Hierarchical classification accuracy=71.7%. Repeaters/non-repeaters accuracy =67.7%. Suicidal/non-suicidal accuracy=88.3%.	60 audio-recorded dyadic clinician-patient interviews.
[56]	General Medical Interaction Analysis System (GMIAS - 10 classes)	transcripts, patients' evaluation of physician communication.	topic annotation	Interannotator kappa=0.81 to 0.95.	360 physician-patient visits
[57]	Communication accommodation theory (CAT), Discursis	Audio and video, visualisation generated from transcripts	Categorisation, annotation of communication strategies, topic tracking: immediate topic repetition (ITR), topic consistency other (TCO), and topic consistency self (TCS).	1: significant difference in categorisation between effective/ineffective interactions ($p<0.001$), 2: ITR>0 in effective vs ITR<0 in ineffective interactions. extreme TCS values indicative of effective interactions. TCO<0 in 3/4 of effective and 1/4 of ineffective interactions.	8 audio video recordings of interaction of trainings with simulated patient
[58]	feedback: 16-items Dental Patient Feedback on Consultation	transcripts, patients' demographic information, caregiver perceived quality of communication	Topic identification using word occurrence and co-occurrence statistics	13 themes, grouped in 5 using PCA explaining 60.2% of the total variance (15.3%, 14.4%, 11.9%, 9.9%, 8.8%).	162 video recordings of clinician-patient conversations: appointments for consultation, oral examination, dental treatment, and follow-up.

Table A.6: Patients information. (SEI = socioeconomic information; $\bar{x}$ = mean)

ID	Population	Condition	Location	Age	Sex	Ethnicity	SEI
[3]	91	breast cancer	Edinburgh, Scotland	28 to 85 ($\bar{x}$ = 58, sd = 11.3)	female	-	-
[4]	69 p: 22, c: 47	ADHD	Sigtuna, Sweden	10-11	p: 18m, 4f c: 18m, 29f	21% background from a foreign country (main study, pop=555)	-
[6]	6 (virtual avatars+AI agent)	speech language pathology	University of Florida/University of Auckland/University of Queensland, USA/New Zealand/Australia	38-73 ($\bar{x}$ = 55)	2m, 6f	Black: 3, White: 3	
[17]	208	early-stage prostate cancer, low or intermediate-risk (Gleason score 6 or 7)	Midwest, USA	$\bar{x}$ = 62	male	88% white	64% college-educated
[7]	69: 46 patients, 23 controls	schizophrenia (BAC+NSA-16)	Singapore	p: 20-52 (mean 31.2), c: 19-47 (28.4)	p: 23m 23f, c: 11m 12f	p: 38 Chinese 5 Malay 3 Indian, c: 19C 3M 1I	p: education: university 6, diploma/vocational 25, high school 15, c: 3U 14D/V 6HS
[8]	18	end-stage of cancer	Hokkaido / Kanto / Tokai / Kinki / Tyugoku / Kyusyu, Japan	> 20, $\bar{x}$ = 71.9, sd = 12.4	16m, 2f	(Japanese) -	-

m: male, f: female, c: control group, p: patients, -: not reported.

Table A.6: Patients information (continued).

ID	Population	Condition	Location	Age	Sex	Ethnicity	SEI
[10]	132	(primary care)	large city, Midwest, USA	18–82 ($\bar{x}$ = 43.8, sd = 14.0)	32m, 100f	Black/African American	low income
[12]	231	hospitalised patients with advanced cancer	New York/San Francisco, USA	<55:27, 55-70: 45, >70: 28	51m 49f	Black: 13, Hispanic/Latino: 8, Either Black or Latino: 20, No Black/Latino: 80	Education: ≥Bachelors 29, HS/some college 55, <High school 16. Financial security: secure 38, partially secure 28, insecure 33
[18]	43	volunteers for analgesic ointment evaluation	Israel	$\bar{x}$ = 24 yo(18-39)	34m, 9f	-	Education: $\bar{x}$ = 14 years(12-18)
[20]	40	primary care visits	Australia	-	-	-	-
[26]	231	hospitalised patients with advanced cancer	New York/San Francisco, USA	<55:27, 55-70: 45, >70: 28	51m 49f	Black: 13, Hispanic/Latino: 8, Either Black or Latino: 20, No Black/Latino: 80	Education: ≥Bachelors 29, HS/some college 55, <High school 16. Financial security: secure 38, partially secure 28, insecure 33
[27]	(10)	medical interview (simulated patients)	Nagoya University Hospital, Japan	-	-	-	-
[28]	415	HIV-infected	USA	-	-	-	-

m: male, f: female, c: control group, p: patients, -: not reported.

Table A.6: Patients information (continued).

ID	Population	Condition	Location	Age	Sex	Ethnicity	SEI
[30]	2	SP acted by qualified doctors	Australia	-	1f 1m	-	-
[35]	279	patients due for a colorectal cancer screening	Michigan, USA	range 50-80 (parent study: $\bar{x}$ = 59.6)	- (parent study: f=63%)	- (parent study: white=66%)	High school/GED or higher=95.7%
[34]	350	primary care office visits	USA	$\bar{x}$ = 62	65.6% female	66.2% white	-
[38]	308	GP consultations	Australia: 141 , UK: 167	-	Australia: f=81, m=60 , UK: f=105, m=62	-	-
[39]	<13	breaking bad news situations (virtual patients, doctors and actors)	France	-	-	-	-
[40]	12	in-patients with various psychosomatic disorders (8 anxiety disorders, 5 depressive and adjustment disorders, 5 somatoform disorders)	Germany	$\bar{x}$ = 32.9 (SD = 8.1)	f=9, m=3	-	single=8, divorced=2, married=2. 9 employed full-time, 2 in school, 1 housekeeping

m: male, f: female, c: control group, p: patients, -: not reported.

Table A.6: Patients information (continued).

ID	Population	Condition	Location	Age	Sex	Ethnicity	SEI
[46]	86 patients, 86 companions	diagnosed with dementia	USA	p: $\bar{x}$ = 72.93 (sd = 8.10), comp: $\bar{x}$ = 62.46 (sd = 13.72)	p: f=52, comp: f=60	p: black=8, white=78, comp: black=5, white=77	Education (years) p: $\bar{x}$ = 14.52 (sd = 3.48) comp: $\bar{x}$ = 15.1 (sd = 2.84)
[47]	122	cancer patients, stage 3 or stage 4 (late stage) advanced solid tumours	USA	-	-	-	-
[49]	(341)	6 studies: -,10 first year college students with indication of drinking related problems, 20 students intending to drink during their upcoming spring break trip, 41 alcohol intervention for students turning twenty-one, 70 adults presenting at primary care clinics who indicate drug use, 7 college students with some indication of marijuana-related problems	USA	-	-	-	78 students
[52]	60	adolescents. 30 suicidal (13 repeaters, 17 non-repeaters), 30 non-suicidal.	Cincinnati Children's Hospital Medical Center Emergency Department, USA	13 < age < 18	30m, 30f	-	-
[54]	132	primary care	USA	$\bar{x}$ = 43.8, range=18–82	76% women	Black / African American	low-income
[56]	360	Patients with HIV	USA	-	-	-	-

m: male, f: female, c: control group, p: patients, -: not reported.

Table A.6: Patients information (continued).

ID	Population	Condition	Location	Age	Sex	Ethnicity	SEI
[57]	1 professional actor	open disclosure (simulated)	Brisbane, Australia	-	-	(Australian)	-
[58]	162 (not unique)	dental visit	Hong Kong	4-5: 47, 6-10: 85, 11-16: 30	m: 71, f:91	(Chinese)	-

m: male, f: female, c: control group, p: patients, -: not reported.

Table A.7: Clinicians information. (SEI = socioeconomic information)

ID	Population	Speciality	Experience	Location	Age	Sex	Ethnicity	SEI
[3]	2	therapeutic radiographers	-	Edinburgh, Scotland	-	-	-	-
[4]	1	dentist		Sigtuna, Sweden	-	-	-	-
[6]	464	speech language pathology	students	University of Florida/University of Auckland/University of Queensland, USA	-	-	-	-
[7]	1	psychometrician	-	Singapore	-	-	-	-
[8]	24	specialists in home medical care	years of experience: $\bar{x}$ = 18.4, sd = 8.5. years of home care experience: $\bar{x}$ = 5.5 sd = 4.6	Hokkaido: 14, Kanto: , Tokai 2, Kinki 3, Tyugoku 1 - Japan	-	20m, 4f	-	-
[10]	17	GPs	physicians, or medical residents in training	large city, Midwest, USA	26–35 ($\bar{x}$ = 27.1)	8m, 9f	variety of ethnic groups	low income
[12]	54	palliative care clinicians: 52 physicians, 11 nurse practitioners, 26 physician fellow, 6 nurses, 3 social workers, 2 chaplain	-	New York/San Francisco, USA	-	46m 54f	-	-
[17]	11	8 urologist, 3 radiation oncologist (in 1/3 of the interactions: 10 nurse practitioners, 34 residents, 4 medical students)	-	Midwest, USA	$\bar{x}$ = 62	-	-	-

m: male, f: female, c: control group, p: patients, -: not reported, (values in italic): assumed from content.

Table A.7: Clinicians information (continued).

ID	Population	Speciality	Experience	Location	Age	Sex	Ethnicity	SEI
[18]	1	doctor	actor	Israel	- (est. 40-50)	male	white	-
[20]	4	primary care physicians	>5 years of clinical experience	Australia	-	2 male, 2 female	-	-
[26]	54	palliative care clinicians: 52 physicians, 11 nurse practitioners, 26 physician fellow, 6 nurses, 3 social workers, 2 chaplain	-	New York/San Francisco, USA	-	46m 54f	-	-
[27]	(10)	-	medical students	Nagoya University Hospital, Japan	-	-	-	-
[28]	(45)	physicians, nurse practitioners (NPs), or physician assistants (30 with a second provider, an NP, or fellow)	-	USA	-	-	-	-
[30]	11	-	medical students	Melbourne, Australia	-	-	-	-
[35]	59	physicians	-	Michigan, USA	$X = 49.4$	male=41.5%	-	-
[34]	84	physicians	-	USA	-	-	-	-
[38]	36	GPs	-	Australia: 25, UK: 11	-	Australia: f=7, m=13 , UK: f=4, m=12	-	-
[39]	13	students	France	-	-	-	-	-
[40]	-	therapists	Germany	-	-	-	-	-
[46]	-	physician	USA	-	-	-	-	-
[47]	40	oncologists	USA	-	-	-	-	-

m: male, f: female, c: control group, p: patients, -: not reported, (values in italic): assumed from content.

Table A.7: Clinicians information (continued).

ID	Population	Speciality	Experience	Location	Age	Sex	Ethnicity	SEI
[49]	-	graduate or undergraduate students, clinic-based social workers	not reported	USA	-	-	-	-
[52]	1	trained social worker	-	Cincinnati Children's Hospital Medical Center Emergency Department, USA	-	-	-	-
[54]	17	physicians	second or third year medical residents	USA	$\bar{x}$ = 27.1, f: 53% range=26–35		8 India / Pakistan (5f). 6 Asia, other (3f). 2 white (2m). 1 Black (1f)	
[56]	41	physicians	-	USA	-	-	-	-
[57]	-	clinicians	-	Brisbane, Australia	-	-	-	-
[58]	-	paediatric dentists, certificated dental surgery assistants	-	Hong Kong	-	-	-	-

m: male, f: female, c: control group, p: patients, -: not reported, (values in italic): assumed from content.

Table A.8: Interactions.

ID	Type of interaction	Medical interaction
[3]	dyadic	breast cancer consultations
[4]	triadic, analysis: dyadic only	introduction phase of dentist visits
[6]	dyadic	GP consultations
[7]	dyadic	semi-structured interviews
[8]	triadic	medical examinations and conversations with family caregivers
[10]	dyadic	patient–physician interactions in a primary care clinic
[12]	dyadic	palliative care consultations
[17]	dyadic	medical interview (discuss treatment options). 1/3 visits included a discussion with a nurse practitioner or resident before the interview
[18]	dyadic	presentation and instructions of a drug for an evaluation study (simulated). Engaged scenario: opened questions. Disengaged scenario closed questions
[20]	dyadic	GP consultation
[26]	dyadic	palliative care consultations
[27]	dyadic	Simulated medical interview
[28]	dyadic (36 with a second clinician)	routine outpatient visits of HIV patients
[30]	dyadic	OSCE examinations: sexually transmitted disease – genital herpes and bowel cancer (scripted, uncued, free from)
[35]	dyadic, small fraction triadic (nurse, family member)	primary care visits (preventive health discussions.)
[34]	dyadic	elderly patients primary care office visits
[38]	dyadic	GP consultations
[39]	dyadic	breaking bad news, real training sessions
[40]	dyadic	intake interviews for in-patient psychotherapy
[46]	triadic	dementia diagnosis disclosure sessions with the patient and companion
[47]	triadic (family caregiver)	physician-patient visits
[49]	dyadic	Motivational Interviewing (substance use disorders)

OSCE: Objective structured clinical examination

Table A.8: Interactions (continued).

ID	**Type of interaction**	**Medical interaction**
[52]	dyadic	interviews, 16 questions (Columbia Suicide Severity Rating Scale (C-SSRS version 1/14/2009), Suicidal Ideation Questionnaire-Junior (SIQ-JR version 1987), Ubiquitous Questionnaire (UQ version 2011).
[54]	dyadic	primary care
[56]	dyadic	adherence dialogue in HIV care
[57]	dyadic	discussions about adverse events in patient care, open disclosure interactions scenarios taken from actual adverse events
[58]	triadic	examination: 60, treatment: 71, consultation: 31

OSCE: Objective structured clinical examination

Table A.9: Data analysis.

ID	Preprocessing	Feature extraction	Task/method	Evaluation	Results
[3]	transcription (professional, manual), stopwords removal. Representation: Term abstraction: unigram, unordered bigram, ordered bigram Term set representations: binary bag, tf-idf (95% most frequent words, n_{words}=300)	Coarse level coding of VR-CoDES.	Supervised learning. classification: NB, LR, SVM (Gaussian kernel, scale 0:25 $\sqrt{n_p}$ where n_p: number of predictors), boosted ensemble DT	CV: 5-fold cross-validation. T: no	Best representation: unordered bigram,little difference across classifiers. Results - BoW, Gaussian kernel SVM. precision=0.93, recall=0.86, AUC=0.75, F-score=0.72, κ=0.45. Results - tf-idf higher that BoW, more consistent: mean acc=0.94.
[4]	-	sequences of interactional elements	manual identification of sequences of interaction and comparison between the 2 patient's groups	comparison. PM: Students t-test adjusted for gender using LR. other type of/unclear focus: Fishers' exact test	ADHD: more initiatives (P=0.002), focus of the initiative was most frequently unclear (P=0.018). fewer verbal responses (P=0.090) and more frequent missing responses (verbal and non-verbal) (P=0.080). Higher degree of missing responses (P=0.061). Higher degree of non-coordination - avoidance of response, no-response or incongruity between the verbal and non-verbal response (P=0.072).

B: baseline, PM: performance metric, CV: cross-validation technique used, T: test set held out and its size, -: not reported. ANOVA: analysis of variance, BoW: Bag of Words, BRL: Bayesian Rule Lists, CART: Classification and regression trees, CNN: Convolutional Neural Network, GEE: General Estimating Equations, GRU: Gated Recurrent Unit, HMM: Hidden Markov Model, ICC: intraclass correlations, kNN: k-nearest neighbours, LOO: leave one out, LR: Linear regression, ML: Machine Learning, NB: Naive Bayes, PCA: Principal Component analysis, PCC: Pearson's Correlation Coefficient, SVM: support vector machine, tf-idf: Term frequency-inverse document frequency

Table A.9: Data analysis (continued).

ID	Preprocessing	Feature extraction	Task/method	Evaluation	Results
[6]	transcription (-), manual annotations	doctor's word frequency: (LIWC) - 9 categories (topics). Participants' word frequency (LSM): adverbs, articles, auxiliary verbs, conjunctions, indefinite pronouns, negations, personal pronouns, prepositions, and quantifiers + average. Type of utterance (RIAS' psychosocial or biomedical, topic direct correspondence); 6 combined features (linguistic). topic data reduction (13 annotated domain topics). feature binning ([0, 1]: 3 classes, [-1, 1]: 6 classes)	Prediction of student success (passing the evaluation) in three stages - binary classification of student's success (success, failure): NB, KNN, LR, SVM, CART, BRL.	**B:** prior probability (acc=0.58), PM: accuracy, CV: 4-fold	Accuracy per phase. Diet and Eating Habits, Medical History: no improvement over B($acc_D EH = 0.64, acc_M H = 0.87$). History of Present Illness: CART, KNN, NB > 5% over B (acc=0.53). BRL LR >10%. Global LR, NB, SVM >5% (B=0.58)

B: baseline, PM: performance metric, CV: cross-validation technique used, T: test set held out and its size, -: not reported. ANOVA: analysis of variance, BoW: Bag of Words, BRL: Bayesian Rule Lists, CART: Classification and regression trees, CNN: Convolutional Neural Network, GEE: General Estimating Equations, GRU: Gated Recurrent Unit, HMM: Hidden Markov Model, ICC: intraclass correlations, kNN: k-nearest neighbours, LOO: leave one out, LR: Linear regression, ML: Machine Learning, NB: Naive Bayes, PCA: Principal Component analysis, PCC: Pearson's Correlation Coefficient, SVM: support vector machine, tf-idf: Term frequency-inverse document frequency

Table A.9: Data analysis (continued).

ID	Preprocessing	Feature extraction	Task/method	Evaluation	Results
[7]	3D position of the skeleton and coordinates of 20 joints, audio: - (task 3 only)	global movement of upper body limbs (head, elbows, wrists, hands): linear/angular speed/acceleration, Head movement (linear/angular speed), gestures: angular difference in elbow and wrist (top 0.1 percent). Mean+SD for each. 41 non-verbal speech signals (task 3 only)	Supervised learning. Task 1: linear correlation between movements and NSA-16 items. Task 2: prediction of subjective ratings (binary from multiclass: unobservable/observable): Linear SVM with Stochastic Gradient Descent and kNN. Task 3 classification of participants (binary: patient, control): same as task 2	Task 1 PM: p-value<0.05. Task 2 and 3 **B:** no (manual GT), PM: precision, recall, F-score, ROC area, accuracy, CV: LOO.	Task 1: Negative correlation between body movement signals and Reduced Expressive Gestures (8 features/14) and speech items of NSA-16 (Restricted Speech Quantity (10/14) and Prolonged Time to Respond(4/14)). Task 2: Restricted Speech Quantity 78.26% SVM, Reduced Expressive Gestures 73.91% SVM, Impoverished Speech Content 67.39% SVM, Affect Reduced Modulation of Intensity 63.04% kNN, Prolonged Time to Respond 60.87% SVM. Task 3: body only: acc=73.91% SVM, body+speech: acc=86.76% multilayer perceptron

B: baseline, PM: performance metric, CV: cross-validation technique used, T: test set held out and its size, -: not reported. ANOVA: analysis of variance, BoW: Bag of Words, BRL: Bayesian Rule Lists, CART: Classification and regression trees, CNN: Convolutional Neural Network, GEE: General Estimating Equations, GRU: Gated Recurrent Unit, HMM: Hidden Markov Model, ICC: intraclass correlations, kNN: k-nearest neighbours, LOO: leave one out, LR: Linear regression, ML: Machine Learning, NB: Naive Bayes, PCA: Principal Component analysis, PCC: Pearson's Correlation Coefficient, SVM: support vector machine, tf-idf: Term frequency-inverse document frequency

Table A.9: Data analysis (continued).

ID	Preprocessing	Feature extraction	Task/method	Evaluation	Results
[8]	transcription (-, doctor only)	Part of Speech tagging (KH-Coder version 2). Frequency of nouns, adjectives, and verbs. Topic extraction (15 sub-topics, 3 main)	Comparison of occurrence frequency of topics discussed between patients' dying at home and at hospital. Chi-square test	not applicable	difference in the occurrence frequency of topics between the two groups: 8 sub-topics more discussed at home. p¡0.01: Visiting 24 hours and 365 days (76.9% vs. 23.1%), Predicted sudden deterioration pattern (84.6% vs. 15.4%), Ease of contacting or consulting with doctors (88.5% vs. 38.5%), Current life expectancy (46.2% vs. 7.7%), Decay and death caused by ageing (76.9% vs. 7.7%). p¡0.05: Calling home care doctors instead of an ambulance (61.5% vs. 15.4%), Home care service based on a long-term care insurance system (76.9% vs. 38.5%), Medical insurance system and payment (61.5% vs. 15.4%).

B: baseline, PM: performance metric, CV: cross-validation technique used, T: test set held out and its size, -: not reported. ANOVA: analysis of variance, BoW: Bag of Words, BRL: Bayesian Rule Lists, CART: Classification and regression trees, CNN: Convolutional Neural Network, GEE: General Estimating Equations, GRU: Gated Recurrent Unit, HMM: Hidden Markov Model, ICC: intraclass correlations, kNN: k-nearest neighbours, LOO: leave one out, LR: Linear regression, ML: Machine Learning, NB: Naive Bayes, PCA: Principal Component analysis, PCC: Pearson's Correlation Coefficient, SVM: support vector machine, tf-idf: Term frequency-inverse document frequency

Table A.9: Data analysis (continued).

ID	Preprocessing	Feature extraction	Task/method	Evaluation	Results
[10]	transcription (professional), anonymisation	text analysis: word embeddings (word to word co-occurrence matrix) per participant. Continuous bag of words (CBOW, probability of a word given a context) and skip-gram (probability of the context given a word). Centroid for each utterance (average of embeddings). Extraction of topics: k number of discriminative clusters based on content per participant over the whole interaction.	Capture of patient–physician relatedness between discourse content. Word embeddings (word2vec) trained with corpus/"2015 Medline abstracts and titles" / "2017 Wikipedia". Comparison per participants of 3 models: all to all (average of all utterances)/utterance-based (resident-to-patient, patient-to-resident)/topic-based (interaction's word clustering). All-to-all: global use of similar/related words. Utterance-to-utterance: participant's responsiveness. Topic-to-topic: topics used by each participant and how related are they. Comparison of mean cosine similarity (centroids).	(1) Resident-Patient Interaction evaluation. **B:** no. PM: comparison of computed quality of communication scores (QCS). Statistical significance (Fisher's R-to-Z transformation, p¡0.05) (2) Physician conversation quality (QCS averaged per physician). **B:** no. PM: Significance of variations of QCS. (3) Resident-Patient questionnaires evaluation **B:** patient self-reported metrics, PM: PCC between B and QCS	(1) Interactions were ranked similarly using all-to-all and clustering methods. Utterance-based methods rank interactions in a similar manner. (2) No statistical significance of variations of quality score (p¿0.05) for all methods (3) very low overall linear correlation (-0.16 to 0.15) among all methods

B: baseline, PM: performance metric, CV: cross-validation technique used, T: test set held out and its size, -: not reported. ANOVA: analysis of variance, BoW: Bag of Words, BRL: Bayesian Rule Lists, CART: Classification and regression trees, CNN: Convolutional Neural Network, GEE: General Estimating Equations, GRU: Gated Recurrent Unit, HMM: Hidden Markov Model, ICC: intraclass correlations, kNN: k-nearest neighbours, LOO: leave one out, LR: Linear regression, ML: Machine Learning, NB: Naive Bayes, PCA: Principal Component analysis, PCC: Pearson's Correlation Coefficient, SVM: support vector machine, tf-idf: Term frequency-inverse document frequency

Table A.9: Data analysis (continued).

ID	Preprocessing	Feature extraction	Task/method	Evaluation	Results
[12]	-	85 audio features / 0.5s nonoverlapping intervals. 17 audio features: 13 Mel Frequency Cepstral Coefficients, zero-crossing rate, energy, energy entropy, and spectral entropy. 50 ms intervals (with 25 ms overlap), 5 statistical aggregators: minimum, maximum, mean, median, SD.	Supervised machine learning, 2-steps. Binary classification (speech, silence), contiguous (>2s). Classifier: RF, 50 decision trees. (1) ML binary classifiers: RF, SVM, CNN. Raw feature vectors and PCA. (2) ML+Human Coders vs HC alone	(1) **B:** human coders, PM: accuracy, sensitivity, specificity, CV: 10-fold T: no (2) **B:** manual annotations (golden truth set by 2 coders), PM: Cohen's Kappa of HC over detected silences, task time, sensitivity CV: no T: no	(1) $accuracy_RF = 0.98$ (2) . task time HCvsHM+ML: +61%. Sensitivity: 100%

B: baseline, PM: performance metric, CV: cross-validation technique used, T: test set held out and its size, -: not reported. ANOVA: analysis of variance, BoW: Bag of Words, BRL: Bayesian Rule Lists, CART: Classification and regression trees, CNN: Convolutional Neural Network, GEE: General Estimating Equations, GRU: Gated Recurrent Unit, HMM: Hidden Markov Model, ICC: intraclass correlations, kNN: k-nearest neighbours, LOO: leave one out, LR: Linear regression, ML: Machine Learning, NB: Naive Bayes, PCA: Principal Component analysis, PCC: Pearson's Correlation Coefficient, SVM: support vector machine, tf-idf: Term frequency-inverse document frequency

Table A.9: Data analysis (continued).

ID	Preprocessing	Feature extraction	Task/method	Evaluation	Results
[17]	transcription, clean-up (deletion of punctuation, description of noises), lemmatisation, stemming. Manual dictionary of terms related to gains or losses (3 reviewers).	3 word counts per healthcare provider: total, terms related to gains/to losses (dictionary based)	(1) word coding: Linguistic Inquiry and Word Count (LIWC) software, context extraction Contextualizer software (coding of negations: negation word in a 30 words window) (2) Framing: association between physician recommendation and physicians' words, LR (3) relation with patients' choice	(1) **B:** manual coding PM: accuracy, CV: no T: no (2) not evaluated (3) not evaluated	(1) gain words: 100% loss words: 80%, Krippendorff's $\alpha = 0.9$ (2) physicians recommending cancer treatment used fewer loss words ($p = .097$). No significant association was found over gain words. Words associated with death were related to physicians' recommendation (treatment: 43%, active surveillance: 58%, both: 50%) (3) Association between patients' choice of cancer treatment and loss words in the first clinical consultation ($p = .05$).

B: baseline, PM: performance metric, CV: cross-validation technique used, T: test set held out and its size, -: not reported. ANOVA: analysis of variance, BoW: Bag of Words, BRL: Bayesian Rule Lists, CART: Classification and regression trees, CNN: Convolutional Neural Network, GEE: General Estimating Equations, GRU: Gated Recurrent Unit, HMM: Hidden Markov Model, ICC: intraclass correlations, kNN: k-nearest neighbours, LOO: leave one out, LR: Linear regression, ML: Machine Learning, NB: Naive Bayes, PCA: Principal Component analysis, PCC: Pearson's Correlation Coefficient, SVM: support vector machine, tf-idf: Term frequency-inverse document frequency

Table A.9: Data analysis (continued).

ID	Preprocessing	Feature extraction	Task/method	Evaluation	Results
[18]	separation into subject/doctor parts of the frame	velocity of each pixel in each frame (optical flow algorithm), total energy (sum of squared pixel velocities) per person. cross-correlation between persons' energies, motion synchrony, kinetic energy cross-correlation at zero lag, total / instantaneous entrainment and leading/-following behaviour, power spectrum of the motion, jitter	(1) Exploration: correlation of motions, jitter as a marker of followership, dominance (cross correlation over 20s). (2) Classification of interactions: scenario (engaged/disengaged) LR using synchrony and asymmetry	(1) Significance. Motions: cross-correlation at lag $-5s$¡τ¡$5s$. Jitter and dominance: Mann–Whitney test (2) **B:** none PM: accuracy CV: no, T: no	Motion synchrony higher in engaged scenario (p¡0.001). disengaged scenario: participants follow each other's motion in turns, engaged: one-way followership (patient follows doctor). Jitter higher in engaged scenario. Dominance duration ratio higher in engaged scenario (p¡0.03) (2) Classification $accuracy = 0.72$
[20]	transcription (research team).	extraction of activity networks, temporal sequencing of activities and transitions	Visualisations. Temporal ordering of activities: heatmap. Network diagram: Fruchterman Reingold centrality	not evaluated	highly interactive, fragmented, and nonlinear process. Central cluster: discussion about patients' presenting complaint was the most central activity and was highly connected to medical history taking, physical examination, and assessment. Remaining activities were more peripheral and less connected.

B: baseline, PM: performance metric, CV: cross-validation technique used, T: test set held out and its size, -: not reported. ANOVA: analysis of variance, BoW: Bag of Words, BRL: Bayesian Rule Lists, CART: Classification and regression trees, CNN: Convolutional Neural Network, GEE: General Estimating Equations, GRU: Gated Recurrent Unit, HMM: Hidden Markov Model, ICC: intraclass correlations, kNN: k-nearest neighbours, LOO: leave one out, LR: Linear regression, ML: Machine Learning, NB: Naive Bayes, PCA: Principal Component analysis, PCC: Pearson's Correlation Coefficient, SVM: support vector machine, tf-idf: Term frequency-inverse document frequency

Table A.9: Data analysis (continued).

ID	Preprocessing	Feature extraction	Task/method	Evaluation	Results
[26]	-	detection of conversational pauses. 85 audio features/0.5s nonoverlapping intervals. 17 audio features: 13 Mel Frequency Cepstral Coefficients, zero-crossing rate, energy, energy entropy, and spectral entropy. 50 ms intervals (with 25 ms overlap), 5 statistical aggregators: minimum, maximum, mean, median, SD.	Supervised machine learning, 2-steps. Binary classification (speech, silence), contiguous (>2s). Classifier: RF, 50 decision trees. ML binary classifiers: RF, SVM, Counter-propagation Neural Networks. Raw feature vectors and PCA.	**B:** manual annotations (3 annotators, 261mins), PM: accuracy, sensitivity, specificity, CV: 10-fold T: 6 consultations (260.5mins)	accuracy=94.4%, sensitivity=90.4, specificity=94.5.

B: baseline, PM: performance metric, CV: cross-validation technique used, T: test set held out and its size, -: not reported. ANOVA: analysis of variance, BoW: Bag of Words, BRL: Bayesian Rule Lists, CART: Classification and regression trees, CNN: Convolutional Neural Network, GEE: General Estimating Equations, GRU: Gated Recurrent Unit, HMM: Hidden Markov Model, ICC: intraclass correlations, kNN: k-nearest neighbours, LOO: leave one out, LR: Linear regression, ML: Machine Learning, NB: Naive Bayes, PCA: Principal Component analysis, PCC: Pearson's Correlation Coefficient, SVM: support vector machine, tf-idf: Term frequency-inverse document frequency

Table A.9: Data analysis (continued).

ID	Preprocessing	Feature extraction	Task/method	Evaluation	Results
[27]	digitisation of videos, manual annotation (dialogue primitives, e.g. gaze, speech, 0.1 s precision)	Patterns of interaction	Extraction concurrence of primitives, motifs, pattern clusters from annotations. Feature reduction: JSD derived from KLD. Pattern evaluated with: basic pattern measure and characteristic pattern measure. Motifs evaluated with normalised expected/actual occurrence. Pattern clusters: distance between patterns. Clustering: Ward method. Reduction: thresholding.	**B:** evaluation from video PM: Comparison of human evaluation of communication from video recording and corresponding extracted patterns (matched features, mis-matched features,unknown, other). CVT/T: no	out of 38 items: 39.5% matched features, 26.3% mis-matched, 26% unknown, 13.2% other.
[28]	transcription (professional or research assistant), coding (GMIAS: all, CASES: 50). Mapping from GMIAS codes to information-giving	Part-of-speech tagging. Unigrams (BoW), bigrams, part-of-speech bigrams, role-specialized N-grams, adjacent speech - content similarity, adjacent speech - hypothesis label	Speech acts classification (text analysis): trained on 40 conversations, L2-regularised LR	**B:** manually coded transcripts, PM: Cohen's κ. Reproduction of the Information-Giving Ratios. T: 375 conversations, overfitting: selection of quality indicators prior to experiments.	κ = 0.573, accuracy=71.2%, Information-Giving Ratio, $r = 0.96$. Performances increased with the size of the training set, gain was logarithmic. Automated annotation did not significantly correlated with outcomes.

B: baseline, PM: performance metric, CV: cross-validation technique used, T: test set held out and its size, -: not reported. ANOVA: analysis of variance, BoW: Bag of Words, BRL: Bayesian Rule Lists, CART: Classification and regression trees, CNN: Convolutional Neural Network, GEE: General Estimating Equations, GRU: Gated Recurrent Unit, HMM: Hidden Markov Model, ICC: intraclass correlations, kNN: k-nearest neighbours, LOO: leave one out, LR: Linear regression, ML: Machine Learning, NB: Naive Bayes, PCA: Principal Component analysis, PCC: Pearson's Correlation Coefficient, SVM: support vector machine, tf-idf: Term frequency-inverse document frequency

Table A.9: Data analysis (continued).

ID	Preprocessing	Feature extraction	Task/method	Evaluation	Results
[30]	transcripts (manual verbatim, research team, ELAN)	text analysis: 38 features (overall word count, length of interaction, number of turns, number of uh and ah, number of unfinished words, number of overlapping words, length of overlap (time), transition pauses, within turn pauses, all time-based features, all turn-based features, number of turns, longest turn, single word responses, number of introduced content words by each speaker, number of times speaker uses word introduced by other, number of words in dialogue, longest number of words in a turn)	Outcome prediction (binary: fail, success). supervised classifier: IB1 (lazy learner) using 11 feature sets (different clustering of extracted features).	**B:** zero-R (majority vote) PM: accuracy, precision, recall, F-score CV: 10-fold stratified and LOO	10-fold - best word based=.916, all=.872 LOO - best word based=.919, all=.872

B: baseline, PM: performance metric, CV: cross-validation technique used, T: test set held out and its size, -: not reported. ANOVA: analysis of variance, BoW: Bag of Words, BRL: Bayesian Rule Lists, CART: Classification and regression trees, CNN: Convolutional Neural Network, GEE: General Estimating Equations, GRU: Gated Recurrent Unit, HMM: Hidden Markov Model, ICC: intraclass correlations, kNN: k-nearest neighbours, LOO: leave one out, LR: Linear regression, ML: Machine Learning, NB: Naive Bayes, PCA: Principal Component analysis, PCC: Pearson's Correlation Coefficient, SVM: support vector machine, tf-idf: Term frequency-inverse document frequency

Table A.9: Data analysis (continued).

ID	Preprocessing	Feature extraction	Task/method	Evaluation	Results
[34]	transcripts (human-generated)	utterances extraction (punctuation based)	2 ML to recognise emotional valence of utterances (3 classes: positive, negative, neutral): RNN with a hierarchical structure, LR (bag-of-words) objective function : minimising the log-loss (cross-entropy) using gradient-based search in an end-to-end fashion	**B:** LR , LR. PM: Average and One-versus-Rest (OvR) human valence rating. Pearson correlation coefficient, R-precision CV: 10-fold	Pearson correlation coefficients: human OvR=0.60, RNN=0.60, LR=0.55. R-precisions (positive class): 0.47, 0.58, 0.53 (same order). R-precisions (negative class): 0.44, 0.45, 0.42 (same order). RNN consistently better than LR. RNN similar to human OvR.
[35]	removal of potentially identifiable information. word tokenisation: Natural Language ToolKit (NLTK) tokenizer. Stopwords removal (except for NN models)	binary word vectors (vocabulary size=14800) of each talk-turn aggregated into a single talk-turn vector (bag-of-words, tf-idf weights). Except for NN: embedding layer (GloVe vectors) and bidirectional set of gated recurrent units (size=128), resulting in talk-turn vectors (size=256).	Classification of talk-turn topic labels: independent (LR, SVM, GRU), window-based (Windowed LR, Windowed SVM), sequential (CRF, HMM-LR, HMM-SVM, HMM-GRU, Hier-GRU).	**B:** prediction of most common topics. PM: Turn level: accuracy. Visit level (aggregated): precision recall F1 (human golden truth). Significance: dependent t tests for paired samples	Turn: Hier-GRU accuracy=61.7% sequential models are more accurate than others (P<.01). Visit: Windowed SVM F1=78.37%. Lower gains in performance between models. Semantic similarity of discussion topics can be a significant contributor to prediction error

B: baseline, PM: performance metric, CV: cross-validation technique used, T: test set held out and its size, -: not reported. ANOVA: analysis of variance, BoW: Bag of Words, BRL: Bayesian Rule Lists, CART: Classification and regression trees, CNN: Convolutional Neural Network, GEE: General Estimating Equations, GRU: Gated Recurrent Unit, HMM: Hidden Markov Model, ICC: intraclass correlations, kNN: k-nearest neighbours, LOO: leave one out, LR: Linear regression, ML: Machine Learning, NB: Naive Bayes, PCA: Principal Component analysis, PCC: Pearson's Correlation Coefficient, SVM: support vector machine, tf-idf: Term frequency-inverse document frequency

Table A.9: Data analysis (continued).

ID	Preprocessing	Feature extraction	Task/method	Evaluation	Results
[38]	verbal and body language clues, clinician style (inclusive or not)	Automated extraction of computer use. Unsure: gaze, detailed computer use	observation techniques and simple descriptive statistics	Extraction not evaluated. Significance of observations: Chi square, NPAR Man Whitney	20% of consultations without computer use. Computer shapes the beginning actively (7 %), passively (10%). 23% of consultations were patient initiated. inclusive consultations: patient looked more at the computer screen (number of times, duration). Triadic (doctor, patient, computer screen) interactions were more common.
[39]	transcription (manual), annotation of non-verbal behaviours, audio segmented into Inter-Pausal Units	part-of speech (POS), automated segmentation and extraction of sequences (SPPAS).	automatic extraction of multimodal sequences leading to feedbacks. Extraction of significant rules (types of sequence X leading to specific feedback Y, $X \implies Y$)	**B:** none, PM: Confidence score of extracted rules ($freq(X \cup Y)/freq(X)$).	10 rules identified, confidence comprised between 0.36 and 0.12, 5 rules with cs<0.2

B: baseline, PM: performance metric, CV: cross-validation technique used, T: test set held out and its size, -: not reported. ANOVA: analysis of variance, BoW: Bag of Words, BRL: Bayesian Rule Lists, CART: Classification and regression trees, CNN: Convolutional Neural Network, GEE: General Estimating Equations, GRU: Gated Recurrent Unit, HMM: Hidden Markov Model, ICC: intraclass correlations, kNN: k-nearest neighbours, LOO: leave one out, LR: Linear regression, ML: Machine Learning, NB: Naive Bayes, PCA: Principal Component analysis, PCC: Pearson's Correlation Coefficient, SVM: support vector machine, tf-idf: Term frequency-inverse document frequency

Table A.9: Data analysis (continued).

ID	Preprocessing	Feature extraction	Task/method	Evaluation	Results
[40]	compositing of the two video streams into one. analysis of 15 min of each interview (first 10min, last 5min): emotionally relevant movements in the face (EmFACS), TAS-26 scale	facial expressions of happiness, social smiles, sadness, fear, anger, disgust, contempt, different affects	congruence of codings with the facial expressions of happiness, social smiles, sadness, fear, anger, disgust, contempt, social smiles, different affects. Evaluation of facial affect display and corresponding emotional reactions of the therapists. Relation between Patients' Facial Affective Display and TAS	Extraction not evaluated. Correlations: correlation coefficient	Correlations TAS with categories of facial affect display (patients): significant negative correlation between the total score of the TAS-26/the first TAS-26 subscale 'problems in identifying feelings' and the facial display of aggressive affects (anger, disgust, contempt). Correlations TAS with facial affect display (patients): Anger, Contempt (p¡0.05), Anger, Contempt, Blends, Surprise, Disgust ($p<0.1$). Correlations TAS with facial affect display (therapists): Contempt, Fear(p¡0.05) Contempt, Fear, Sadness (p¡0.1).

B: baseline, PM: performance metric, CV: cross-validation technique used, T: test set held out and its size, -: not reported. ANOVA: analysis of variance, BoW: Bag of Words, BRL: Bayesian Rule Lists, CART: Classification and regression trees, CNN: Convolutional Neural Network, GEE: General Estimating Equations, GRU: Gated Recurrent Unit, HMM: Hidden Markov Model, ICC: intraclass correlations, kNN: k-nearest neighbours, LOO: leave one out, LR: Linear regression, ML: Machine Learning, NB: Naive Bayes, PCA: Principal Component analysis, PCC: Pearson's Correlation Coefficient, SVM: support vector machine, tf-idf: Term frequency-inverse document frequency

Table A.9: Data analysis (continued).

ID	Preprocessing	Feature extraction	Task/method	Evaluation	Results
[46]	transcripts	text analysis (Linguistic Inquiry and Word Count)	Measures of verbal dominance and use of pronoun	Observation of differences in actual and perceived verbal dominance (ANOVA). influenced of dementia status on verbal dominance (Independent samples t-tests). differences in pronoun use (ANOVA). Not evaluated. Bivariate correlations between observations and characteristics, hierarchical regressions between observations and outcomes.	Physicians dominated the conversation (83% of the total words). Patients 10%, companions 6%. Significant difference in the use of first person pronouns across participants (p<.001). Physicians used fewer singular pronouns, Companions used fewer singular pronouns than patients. Physicians used more plural pronouns. Power indices did not predict outcomes.

B: baseline, PM: performance metric, CV: cross-validation technique used, T: test set held out and its size, -: not reported. ANOVA: analysis of variance, BoW: Bag of Words, BRL: Bayesian Rule Lists, CART: Classification and regression trees, CNN: Convolutional Neural Network, GEE: General Estimating Equations, GRU: Gated Recurrent Unit, HMM: Hidden Markov Model, ICC: intraclass correlations, kNN: k-nearest neighbours, LOO: leave one out, LR: Linear regression, ML: Machine Learning, NB: Naive Bayes, PCA: Principal Component analysis, PCC: Pearson's Correlation Coefficient, SVM: support vector machine, tf-idf: Term frequency-inverse document frequency

Table A.9: Data analysis (continued).

ID	Preprocessing	Feature extraction	Task/method	Evaluation	Results
[47]	transcription	Speech Features by speaker: number of words spoken, number of questions asked, word diversity (unique word count). Affective Features: sentiments expressed (positive, negative, neutral). Valence Aware Dictionary for sEntiment Reasoning (VADER) + Natural Language ToolKit (NLTK)	(1) Comparison of feature averages between best rated and other interactions. (2) Classification of: LR, kNN (k=13). with and without L1 regularisation validation (3) unsupervised clustering of features (k-means) (4) Linguistic Inquiry Word Count analysis between best rated doctors group and others doctor group	(1) PM: t-test p-value, Bonferroni corrected. Effect size: Cohen's d (2) **B:** survey responses PM: accuracy CV: 5-fold (3) PM: Silhouette coefficient, Student's t-test comparison (4) PM: t-test	(1) p¡0.05: number of words spoken by doctor, Doctor unique word count (2) acc = 71%. (3) 4 styles identified (% words spoken by doctor and Doctor positive sentiment, % words spoken by patient and Patient positive sentiment, Doctor Unique Word Count and Patient Unique Word Count, number of unique words spoken by doctor and Doctor positive sentiment). Not statistically significant (4) Largest effect: *You*, *I*, and Personal Pronoun categories

B: baseline, PM: performance metric, CV: cross-validation technique used, T: test set held out and its size, -: not reported. ANOVA: analysis of variance, BoW: Bag of Words, BRL: Bayesian Rule Lists, CART: Classification and regression trees, CNN: Convolutional Neural Network, GEE: General Estimating Equations, GRU: Gated Recurrent Unit, HMM: Hidden Markov Model, ICC: intraclass correlations, kNN: k-nearest neighbours, LOO: leave one out, LR: Linear regression, ML: Machine Learning, NB: Naive Bayes, PCA: Principal Component analysis, PCC: Pearson's Correlation Coefficient, SVM: support vector machine, tf-idf: Term frequency-inverse document frequency

Table A.9: Data analysis (continued).

ID	Preprocessing	Feature extraction	Task/method	Evaluation	Results
[49]	transcripts	part of speech. Text analysis (n-grams, word embedding)	Prediction of MISC behavioural codes (utterance) and session-level MISC summary indices. SL ML, 2 dependency trees methods. Discrete Sentence Feature (DSF): dependency parse tree and N-grams, RNN Model: dependency parse tree and word embedding, multinomial regression	**B:** human inter-rater agreement (n=63) PM: utterances: Cohen's kappa. sessions: two-way, absolute-agreement, single-measures ICC. CV: 10 fold, T: yes (n=109, 30%)	Utterances: Varied performance (better than chance except advise with permission, advise without permission, and confront). lowest performance on low frequency categories. $\kappa > 0.50$: open and closed questions, facilitate, giving information, affirm and follow/neutral. $0.30 > \kappa > 0.50$: simple and complex reflections. DSF performed better than RNN (.055 to .113). Session: DSF outperformed RNN, $ICC > 0.75$: affirm, facilitate, giving information, follow/neutral, simple reflections, and open and closed questions. $0.60 ¡ ICC ¡ 0.74$: sustain talk

B: baseline, PM: performance metric, CV: cross-validation technique used, T: test set held out and its size, -: not reported. ANOVA: analysis of variance, BoW: Bag of Words, BRL: Bayesian Rule Lists, CART: Classification and regression trees, CNN: Convolutional Neural Network, GEE: General Estimating Equations, GRU: Gated Recurrent Unit, HMM: Hidden Markov Model, ICC: intraclass correlations, kNN: k-nearest neighbours, LOO: leave one out, LR: Linear regression, ML: Machine Learning, NB: Naive Bayes, PCA: Principal Component analysis, PCC: Pearson's Correlation Coefficient, SVM: support vector machine, tf-idf: Term frequency-inverse document frequency

Table A.9: Data analysis (continued).

ID	Preprocessing	Feature extraction	Task/method	Evaluation	Results
[52]	transcription	conversation dynamic features (clinicians, patients): speech / pausetime percentages, words per second, overlap rates. Verbal information: LIWC word category scales (80: linguistic class, positive emotion and negative emotion, nonfluencies, assent words). Acoustic information features: Fundamental frequency;Normalized Amplitude Quotient, Quasi-Open Quotient, Parabolic Spectral Parameter, Maxima Dispersion Quotient, Peak Slope, Liljencrants-Fant model parameter Rd:, Formants (F1, F2)	(1) Observational during Ubiquitous Questionnaire / others. (2) Classification (2 steps binary: step 1 suicidal / non suicidal, repeater / non-repeater): SVM using statistically significant features of 1 (37 features: 6 conversational, 14 verbal, 17 acoustic features), radial basis function kernel, step 2 AdaBoostM1 (20 features: 1 conversational, 19 acoustic)	(1) PM: ANOVA (p¡0.05). (2) PM: acc, F1 CV: LOO	(2) Using patients' features: 56.7% step 1 acc=85%. step 2 acc=34.5% Using patients' and clinician: step 1 acc = 90% step 2 acc = 33.3% F1 Ubiquitous Questionnaire: Non-Suicidal 0.88 Non repeater 0.5 repeater 0.37, resp. 0.84 0.68 0.4 others

B: baseline, PM: performance metric, CV: cross-validation technique used, T: test set held out and its size, -: not reported. ANOVA: analysis of variance, BoW: Bag of Words, BRL: Bayesian Rule Lists, CART: Classification and regression trees, CNN: Convolutional Neural Network, GEE: General Estimating Equations, GRU: Gated Recurrent Unit, HMM: Hidden Markov Model, ICC: intraclass correlations, kNN: k-nearest neighbours, LOO: leave one out, LR: Linear regression, ML: Machine Learning, NB: Naive Bayes, PCA: Principal Component analysis, PCC: Pearson's Correlation Coefficient, SVM: support vector machine, tf-idf: Term frequency-inverse document frequency

Table A.9: Data analysis (continued).

ID	Preprocessing	Feature extraction	Task/method	Evaluation	Results
[54]	transcripts: separated into doctor / patient raw text files, cleaned of special characters and formatting	text analysis: Latent Semantic Analysis (LSA)	detection of patient-physician communication similarity using LSA trained of whole corpus.	**B:** none, PM: similarity correlation, GEE regressions	patient-physician communication similarity correlation: $\bar{x}$ = 0.142, median=0.150, sd = 0.185. Physicians differed significantly in patient-physician communication similarity. White physicians exhibited significantly lower semantic similarity with their patient's speech than Indian/Pakistani or other Asian, resp (mean $r = 0.028$, $SE = 0.0325$ / mean $r = 0.179$, $SE = 0.024$ / $r = 0.185$, $SE = 0.025$). Female physicians had marginally greater semantic similarity ($p = .082$). Female patients' speech exhibited greater semantic similarity ($p = .017$). Greater communication similarity was associated with less trust in physicians in general ($p = .002$) and greater trust in their own physician (p¡.010).

B: baseline, PM: performance metric, CV: cross-validation technique used, T: test set held out and its size, -: not reported. ANOVA: analysis of variance, BoW: Bag of Words, BRL: Bayesian Rule Lists, CART: Classification and regression trees, CNN: Convolutional Neural Network, GEE: General Estimating Equations, GRU: Gated Recurrent Unit, HMM: Hidden Markov Model, ICC: intraclass correlations, kNN: k-nearest neighbours, LOO: leave one out, LR: Linear regression, ML: Machine Learning, NB: Naive Bayes, PCA: Principal Component analysis, PCC: Pearson's Correlation Coefficient, SVM: support vector machine, tf-idf: Term frequency-inverse document frequency

Table A.9: Data analysis (continued).

ID	Preprocessing	Feature extraction	Task/method	Evaluation	Results
[56]	transcription	turn-taking patterns, relative frequencies of speech act transitions	Clustering of physicians by their communication: parameters estimation to capture physician-level communication characteristics (speech act usage, speech act transitions). Relation with rating over 12 questions. Gradient descent optimisation, feature-space reduction (PCA). Probability of speech act conditioned on the preceding speech act, the speaker pattern and the participating physician	**B:** none. PM: cluster coefficient estimates, t-test between group values	2 clusters of physicians, significant difference (p¡.05) for questions regarding communication around HIV-specific issues, suggestive for the two other sets of questions. Three reported significant patterns: physician issuing commissives within a single turn, physician issuing directives within a single turn (positive association with evaluation), directive following questions (negative association). i.e. advising or making decisions without patient input and patients appreciate instruction when solicited, not when unsolicited.
[57]	transcription (-)	Discursis visual output, immediate topic repetition, topic consistency other, and topic consistency self (automated extraction). Features converted to z-scores relative to the mean and SD of aggregated values	Comparison of features between the 2 types of interactions	no evaluation (z-value)	-

B: baseline, PM: performance metric, CV: cross-validation technique used, T: test set held out and its size, -: not reported. ANOVA: analysis of variance, BoW: Bag of Words, BRL: Bayesian Rule Lists, CART: Classification and regression trees, CNN: Convolutional Neural Network, GEE: General Estimating Equations, GRU: Gated Recurrent Unit, HMM: Hidden Markov Model, ICC: intraclass correlations, kNN: k-nearest neighbours, LOO: leave one out, LR: Linear regression, ML: Machine Learning, NB: Naive Bayes, PCA: Principal Component analysis, PCC: Pearson's Correlation Coefficient, SVM: support vector machine, tf-idf: Term frequency-inverse document frequency

Table A.9: Data analysis (continued).

ID	Preprocessing	Feature extraction	Task/method	Evaluation	Results
[58]	transcription: speaker of utterances and texts (research team)	word occurrence and co-occurrence statistics. feature reduction: PCA on 13 themes extracted from word occurrence and co-occurrence to obtain 5 themes (KMO=0.536, BTS: p¡0.001), PC1 to PC5 (*Disease / treatment*, *Treatment procedure related instructions*, *Preparation for examination*, *Positive reinforcement / reassurance* and *Family / social history*)	Relation of six variables on each theme with perceived quality of communication. t-tests or one-way ANOVA	**B:** patient's evaluation. PM: association between Dental Patient Feedback on Consultation skills (DPFC) score and extracted variables for each themes (p-value).	p¡0.05: *Percentage of related utterances in total number of utterances*, *Percentage of time spent in total time duration*. p¡0.01: *Number of related words*, *Percentage of related words in total number of words*, *Number of related utterances*.

B: baseline, PM: performance metric, CV: cross-validation technique used, T: test set held out and its size, -: not reported. ANOVA: analysis of variance, BoW: Bag of Words, BRL: Bayesian Rule Lists, CART: Classification and regression trees, CNN: Convolutional Neural Network, GEE: General Estimating Equations, GRU: Gated Recurrent Unit, HMM: Hidden Markov Model, ICC: intraclass correlations, kNN: k-nearest neighbours, LOO: leave one out, LR: Linear regression, ML: Machine Learning, NB: Naive Bayes, PCA: Principal Component analysis, PCC: Pearson's Correlation Coefficient, SVM: support vector machine, tf-idf: Term frequency-inverse document frequency

*

Table A.10: Study assessment.

ID	Research implications	Risk of bias	Strengths/Limitations
[3]	**Novelty:** Yes, **Replicability:** partial, **Generalisability:** high	**RL:** partial (structured transcripts), **FB:** no, **SM:** no, **CR:** manual annotation of 5% $\kappa = 0.67$, 95% CI=0.58-0.75; Spearman's $\rho = 0 : 98$, p¡0 : 001 Overfitting: HO but no CV, S: ≤50	**Spontaneous speech:** yes, **Conversational speech:** yes, **Automation:** partial, **Transcription-free:** no, **Content-independence:** yes.
[4]	**Novelty:** Yes, **Replicability:** partial, **Generalisability:** medium	**RL:** yes, **FB:** no, **SM:** n/a, **CR:** manual annotation of 21 phases $\kappa = 0.98$, 21 sequences $\kappa = 0.98$, 21 interaction elements $\kappa = 0.95$. Overfitting: n/a, S: ≤100	**Spontaneous speech:** yes, **Conversational speech:** yes, **Automation:** no, **Transcription-free:** yes, **Content-independence:** no.
[6]	**Novelty:** Yes, **Replicability:** partial, **Generalisability:** medium	**RL:** no, **FB:** no, **SM:** no, **CR:** yes (prior probability of success). Overfitting: CV, S: ≥100	**Spontaneous speech:** yes, **Conversational speech:** yes, **Automation:** partial, **Transcription-free:** no, **Content-independence:** no.
[7]	**Novelty:** Yes, **Replicability:** partial, **Generalisability:** low	**RL:** yes, **FB:** no, **SM:** yes, **CR:** no. Overfitting: n/a, S: ≤50	**Spontaneous speech:** yes, **Conversational speech:** yes, **Automation:** partial, **Transcription-free:** no, **Content-independence:** no.
[8]	**Novelty:** No, **Replicability:** partial, **Generalisability:** low	**RL:** yes, **FB:** no, **SM:** yes, **CR:** no. Overfitting: LOO CV, S: ≤100	**Spontaneous speech:** yes, **Conversational speech:** yes, **Automation:** partial, **Transcription-free:** video: yes, audio: unclear, **Content-independence:** video: yes, audio: unclear.

-: not reported. RL: Real-life interactions, FB: Feature balance, SM: Suitable metrics, CR: Contextualised results. CI: Confidence Interval.

Table A.10: Study assessment (continued).

ID	Research implications	Risk of bias	Strengths/Limitations
[10]	**Novelty:** Yes, **Replicability:** partial, **Generalisability:** medium	**RL:** yes, **FB:** no, **SM:** yes, **CR:** partial (patient evaluation). Overfitting: -, S: ≥100	**Spontaneous speech:** yes, **Conversational speech:** yes, **Automation:** yes, **Transcription-free:** no, **Content-independence:** no.
[12]	**Novelty:** No, **Replicability:** partial, **Generalisability:** high	**RL:** yes, **FB:** no, **SM:** yes, **CR:** no (baseline not assessed). Overfitting: -, S: ≥100	**Spontaneous speech:** yes, **Conversational speech:** yes, **Automation:** partial, **Transcription-free:** yes, **Content-independence:** yes.
[17]	**Novelty:** Yes, **Replicability:** low, **Generalisability:** low	**RL:** yes, **FB:** partial, **SM:** yes, **CR:** no. Overfitting: no, S: ≥100	**Spontaneous speech:** yes, **Conversational speech:** yes, **Automation:** partial, **Transcription-free:** no, **Content-independence:** no.
[18]	**Novelty:** Yes, **Replicability:** partial, **Generalisability:** high	**RL:** no, **FB:** no, **SM:** no, **CR:** no. Overfitting: no, S: ≤50	**Spontaneous speech:** yes, **Conversational speech:** yes, **Automation:** yes, **Transcription-free:** yes, **Content-independence:** yes.
[20]	**Novelty:** Yes, **Replicability:** low, **Generalisability:** high	**RL:** yes, **FB:** no, **SM:** no, **CR:** no. Overfitting: n/a, S: ≤50	**Spontaneous speech:** yes, **Conversational speech:** yes, **Automation:** partial, **Transcription-free:** no, **Content-independence:** no.
[26]	**Novelty:** No, **Replicability:** partial, **Generalisability:** high	**RL:** yes, **FB:** no, **SM:** yes, **CR:** yes. Overfitting: yes, S: ≥100	**Spontaneous speech:** yes, **Conversational speech:** yes, **Automation:** yes, **Transcription-free:** yes, **Content-independence:** yes.

-: not reported. RL: Real-life interactions, FB: Feature balance, SM: Suitable metrics, CR: Contextualised results. CI: Confidence Interval.

Table A.10: Study assessment (continued).

ID	Research implications	Risk of bias	Strengths/Limitations
[27]	**Novelty:** Yes, **Replicability:** low, **Generalisability:** high	**RL:** no, **FB:** no, **SM:** no, **CR:** yes. Overfitting: n/a, S: ≤50	**Spontaneous speech:** yes, **Conversational speech:** yes, **Automation:** partial, **Transcription-free:** yes, **Content-independence:** yes.
[28]	**Novelty:** Yes, **Replicability:** low, **Generalisability:** high	**RL:** yes, **FB:** no, **SM:** yes, **CR:** yes. Overfitting: no, S: ≥100	**Spontaneous speech:** yes, **Conversational speech:** yes, **Automation:** partial, **Transcription-free:** no, **Content-independence:** no.
[30]	**Novelty:** Yes, **Replicability:** low, **Generalisability:** high	**RL:** no, **FB:** no, **SM:** yes, **CR:** yes. Overfitting: yes, S: ≤50	**Spontaneous speech:** yes, **Conversational speech:** yes, **Automation:** partial, **Transcription-free:** no, **Content-independence:** no.
[34]	**Novelty:** Yes, **Replicability:** low, **Generalisability:** high	**RL:** yes, **FB:** no, **SM:** yes, **CR:** yes. Overfitting: yes, S: ≥100	**Spontaneous speech:** yes, **Conversational speech:** yes, **Automation:** partial, **Transcription-free:** no, **Content-independence:** no.
[35]	**Novelty:** Yes, **Replicability:** partial, **Generalisability:** high	**RL:** yes, **FB:** no, **SM:** yes, **CR:** yes. Overfitting: yes, S: ≥100	**Spontaneous speech:** yes, **Conversational speech:** yes, **Automation:** partial, **Transcription-free:** no, **Content-independence:** no.
[38]	**Novelty:** Yes, **Replicability:** low, **Generalisability:** medium	**RL:** yes, **FB:** no, **SM:** no, **CR:** no. Overfitting: no, S: ≥100	**Spontaneous speech:** yes, **Conversational speech:** yes, **Automation:** partial, **Transcription-free:** yes, **Content-independence:** yes.

-: not reported. RL: Real-life interactions, FB: Feature balance, SM: Suitable metrics, CR: Contextualised results. CI: Confidence Interval.

Table A.10: Study assessment (continued).

ID	Research implications	Risk of bias	Strengths/Limitations
[39]	**Novelty:** Yes, **Replicability:** low, **Generalisability:** high	**RL:** no, **FB:** no, **SM:** no, **CR:** no. Overfitting: n/a, S: ≤50	**Spontaneous speech:** yes, **Conversational speech:** yes, **Automation:** partial, **Transcription-free:** no, **Content-independence:** no.
[40]	**Novelty:** Yes, **Replicability:** low, **Generalisability:** low	**RL:** yes, **FB:** no, **SM:** no, **CR:** no. Overfitting: n/a, S: ≤50	**Spontaneous speech:** yes, **Conversational speech:** no, **Automation:** partial, **Transcription-free:** yes, **Content-independence:** yes.
[46]	**Novelty:** Yes, **Replicability:** partial, **Generalisability:** medium	**RL:** yes, **FB:** no, **SM:** no, **CR:** no. Overfitting: n/a, S: ≤100	**Spontaneous speech:** yes, **Conversational speech:** yes, **Automation:** partial, **Transcription-free:** no, **Content-independence:** yes.
[47]	**Novelty:** Yes, **Replicability:** partial, **Generalisability:** high	**RL:** yes, **FB:** no, **SM:** no, **CR:** no. Overfitting: yes, S: ≥100	**Spontaneous speech:** yes, **Conversational speech:** yes, **Automation:** partial, **Transcription-free:** no, **Content-independence:** no.
[49]	**Novelty:** Yes, **Replicability:** low, **Generalisability:** high	**RL:** yes, **FB:** no, **SM:** no, **CR:** yes. Overfitting: yes, S: ≥100	**Spontaneous speech:** yes, **Conversational speech:** yes, **Automation:** partial, **Transcription-free:** no, **Content-independence:** no.
[52]	**Novelty:** Yes, **Replicability:** partial, **Generalisability:** low	**RL:** yes, **FB:** no, **SM:** yes, **CR:** no. Overfitting: yes, S: ≤100	**Spontaneous speech:** yes, **Conversational speech:** yes, **Automation:** partial, **Transcription-free:** no, **Content-independence:** no.

-: not reported. RL: Real-life interactions, FB: Feature balance, SM: Suitable metrics, CR: Contextualised results. CI: Confidence Interval.

Table A.10: Study assessment (continued).

ID	Research implications	Risk of bias	Strengths/Limitations
[54]	**Novelty:** Yes, **Replicability:** partial, **Generalisability:** high	**RL:** yes, **FB:** no, **SM:** no, **CR:** no. Overfitting: no, S: ≥100	**Spontaneous speech:** yes, **Conversational speech:** yes, **Automation:** partial, **Transcription-free:** no, **Content-independence:** no.
[56]	**Novelty:** Yes, **Replicability:** partial, **Generalisability:** high	**RL:** yes, **FB:** no, **SM:** no, **CR:** no. Overfitting: no, S: ≥100	**Spontaneous speech:** yes, **Conversational speech:** yes, **Automation:** partial, **Transcription-free:** no, **Content-independence:** no.
[57]	**Novelty:** No, **Replicability:** partial, **Generalisability:** high	**RL:** no, **FB:** no, **SM:** no, **CR:** n/a, Overfitting: n/a, S: ≤50	**Spontaneous speech:** yes, **Conversational speech:** yes, **Automation:** partial, **Transcription-free:** no, **Content-independence:** no.
[58]	**Novelty:** Yes, **Replicability:** partial, **Generalisability:** high	**RL:** yes, **FB:** partial, **SM:** yes, **CR:** no, Overfitting: n/a, S: ≥100	**Spontaneous speech:** yes, **Conversational speech:** yes, **Automation:** partial, **Transcription-free:** no, **Content-independence:** no.

-: not reported. RL: Real-life interactions, FB: Feature balance, SM: Suitable metrics, CR: Contextualised results. CI: Confidence Interval.

Table A.11: Datasets assessment.

ID	Data set/Subset size	Data type	Data annotation	Data balance	Data availability	Language
[3]	200 consultations. D: 2m13s-26m19s, $\bar{x}$ = 8m46s, sd = 4m55s	audio, textual transcripts: professionally produced (manual), Demographic (age), treatment (radiotherapy type (3), chemotherapy (yes/no), degree of cancer recurrence), fears before study (16 point scale), self-reported rating of general health state (1–100 scale), living situation (alone or not), consultation number (1-4), consultation duration. **type**: semi-structured	Behavioural coding: VR-CoDES (6 cues, 1 concern), manual	a:No, g:No (by design), s:No	Available: no	English
[4]	69 interactions (1-2 minutes)	audio, video **type**: conversational	Interaction phase (i.e. theme). Interaction sequence (1 initiative-response). Interaction elements: syntax, source, focus, type, response, source of response, manual	a: no (by design), g: no, s: -	Available: no	(Swedish

D: duration, -: not reported. a: age, g: gender, s: socio-professional class. SD: standard deviation

Table A.11: Datasets assessment (continued).

ID	Data set/Subset size	Data type	Data annotation	Data balance	Data availability	Language
[6]	464 GP consultations (student, virtual patient)	textual transcripts (-)	discovery segments (segment containing information critical to the diagnosis); stages (Diet and Eating Habits (DEH), History of Present Illness (HPI), and Medical History (MH)); discovery proficiency prediction (success, failure); overall discovery proficiency $(n_{(studentdiscoveries)}/n_{(totalnumberofdiscoveries)})$. no evaluation	a: no , g: no, s: no	Available: no	English
[7]	69 clinical interview (34 hours, mean duration=30mins)	Videos. **type**: semi-structured.	subjective ratings of symptoms of schizophrenia (NSA-16 scale) No evaluation.	a: yes , g: yes, s: yes	Available: no	English (non native)
[8]	227 patients visits (home: $\bar{x} = 12.9$, sd = 6.7. Hospital $\bar{x} = 10.0$, sd = 4.2). 5415 words	Audio. **type**: conversational.	None.	a: no , g: no, s: -	Available: no	English (non native)
[10]	132 GP consultations	transcripts (professionally), patient questionnaires: general trust in the medical system (prior), trust in the consultation resident (after), satisfaction with the consultation (after). **type**: conversational.	not annotated	a: p=yes/d=no , g: p=no/d=yes, s: no	Available: no	English

D: duration, -: not reported. a: age, g: gender, s: socio-professional class. SD: standard deviation

Table A.11: Datasets assessment (continued).

ID	Data set/Subset size	Data type	Data annotation	Data balance	Data availability	Language
[12]	587 clips from 354 palliative care consultations (9770 minutes)	audio. **type**: conversational	Type of connectional silences (emotional, compassionate, invitational), linguistic features (pre/post speakers, pause length, temporal reference (past, present, future) preceding the pause)	a: no, g: yes, s: yes	Available: no	English
[17]	286 transcripts (words: $\bar{x}$ = 5348, median=4880, sd = 2921)	audio, transcripts **type**: conversational	providers' roles in the conversation, patient and physician IDs, consultation order (1 or 2), speciality of the attending physician urologist, radiation oncologist), patients' choices. No evaluation	a: no, g: no (by design), s: no	Available: no	English
[18]	43 videos (duration: $\bar{x}$ = 210s, sd = 49s)	video **type**: structured	Not annotated	a: yes, g: no, s: yes	Available: no	-
[20]	40 (pilot=7, study=31)	audio, video of electronic health records screens, annotations: times when tools were used (papers, notebooks, websites), occurrence of phone calls, interruptions, new or a regular patient, patient alone or accompanied **type**: conversational	consultation activity (ad-hoc, adapted from Waitzkin, 1989)	a: -, g: -, s: -	Available: no	English
[26]	587 clips from 354 palliative care consultations (9770 minutes)	audio. **type**: conversational	Connectional silences and speech	a: no, g: yes, s: yes	Available: no	English

D: duration, -: not reported. a: age, g: gender, s: socio-professional class. SD: standard deviation

Table A.11: Datasets assessment (continued).

ID	Data set/Subset size	Data type	Data annotation	Data balance	Data availability	Language
[27]	10 videotaped interactions(about 10 minutes each)	video, audio **type**: conversational	25 primitives (12 shared per participants+memo-taking for doctors): speak, gaze to human, gaze to memo, head nod, rhythm, and touching self, memo-taking, "major nonverbal behaviours in communication psychology research literatures"	a: -, g: -, s: -	Available: no	(Japanese)
[28]	415 recordings	audio, transcripts, patients' quality indicators: communication quality (overall), provider decision-making, participatory decision-making, interpersonal style, interpersonal trust. **type**: conversational	Speech acts: 118 287 Giving Information, 28 576 Requesting Information, 92 448 Other. not evaluated	a: -, g: -, s: -	Available: no	English
[30]	22 (<8mins)	video, time-aligned transcripts, examination result: OSCE marking scheme (17 pass, 5 fail) **type**: conversational	a: -, g: -, s: -	Available: no	English	

D: duration, -: not reported. a: age, g: gender, s: socio-professional class. SD: standard deviation

Table A.11: Datasets assessment (continued).

ID	Data set/Subset size	Data type	Data annotation	Data balance	Data availability	Language
[34]	353 interactions (210k utterances)	transcripts (Mental Health Discussion study by Tai-Seale et al., Assessment of Doctor-Elderly Patient Transactions (ADEPT) study by Teresi et al.) **type**: conversational	emotional valence of utterances. Scale: -3 (very negative) to +3 (very positive). 14 raters (students), 4 discarded (distributions of assigned ratings significantly different from the other raters). Each utterance was rated by 2.3 raters. Evaluation: Intraclass Correlation Coefficient (ICC), two-way random effects model ICC: 0.90.	a: -, g: no, s: -	Available: no	English
[35]	279 interactions (122 083 talk-turns, median=408, $\bar{x}$ = 438, upper/lower quartiles=312/522)). (subset differs from reported parent study)	transcripts **type**: conversational	topic label (27: modified MDIA coding system. 3 most frequent topics (BiomedHistory, PreventiveCare, and MusSkePain)>50% of the corpus. Not evaluated	a: -, g: -, s: -	Available: no	English
[38]	308 consultations	multi-channel video, screen capture, key strokes, mouse coordinates	gaze, computer use, detailed use of the computer, verbal and body language clues, clinician style (inclusive or not). Not evaluated	a: -, g: no, s: -	Available: no	English

D: duration, -: not reported. a: age, g: gender, s: socio-professional class. SD: standard deviation

Table A.11: Datasets assessment (continued).

ID	Data set/Subset size	Data type	Data annotation	Data balance	Data availability	Language
[39]	13 videotaped interactions (total= 119 minutes, $\bar{x} = 15$)	audio, video, transcripts (manual) **type**: conversational	transcripts (Transcription Orthographique Enrichie / Enriched Orthographical Transcription), part-of speech (MarsaTag), Visual cues (Head movements: nod, shake, tilt, bottom, up, side. Posture change: forward, backward, other. Gaze: oneself, interlocutor, other direction, closed eyes. Eyebrow expression: frown, raise. Hand gesture, Smile). Evaluation. Visual Cues, 5% of the corpus $\kappa = 0.63$	a: -, g: -, s: -	Available: no	French
[40]	12 interactions (180 min)	videos of each participant **type**: semi-structured	EmFACS: emotionally relevant movements in the face, based on the earlier Facial Action Coding System (FACS). Evaluation: test of the coder, reliability r¿0.80.	a: -, g: -, s: -	Available: no	German

D: duration, -: not reported. a: age, g: gender, s: socio-professional class. SD: standard deviation

Table A.11: Datasets assessment (continued).

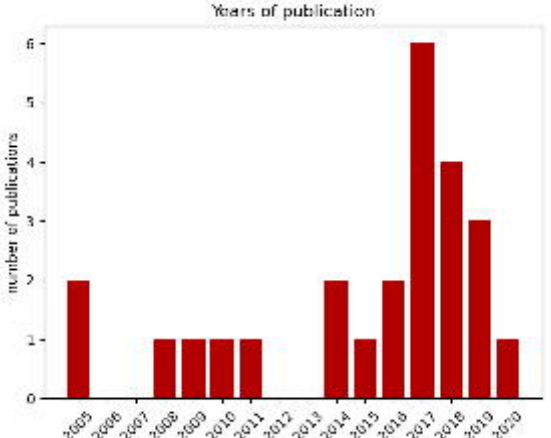
ID	Data set/Subset size	Data type	Data annotation	Data balance	Data availability	Language
[46]	86 videotaped interactions	audio, video, transcripts, patient and companion questionnaires: anxiety (20-item State-Trait Anxiety Inventory), depression (15-item Geriatric Depression Scale), satisfaction with their appointment (Dementia Care Satisfaction Questionnaire) 2-3 days after the session **type**: conversational	not annotated	a: -, g: -, s: -	Available: no	English
[47]	122	transcripts (professionals), audio recordings of the interactions, patient surveys: wellbeing (Likert-type scales), physician's communication skill (5 questions) **type**: conversational	not annotated	a: -, g: -, s: -	Available: no	English
[49]	341 (1.7 million words, 175,000 utterances, and 79,000 talk turns)	transcripts (human raters) **type**: conversational	modified Motivational Interviewing Skill Code (MISC version 2.1): single, categorical behavioural code to each client and clinician utterance. MISC behavioural codes and session-level MISC summary indices only, no global ratings.	a: -, g: -, s: -	Available: no	English

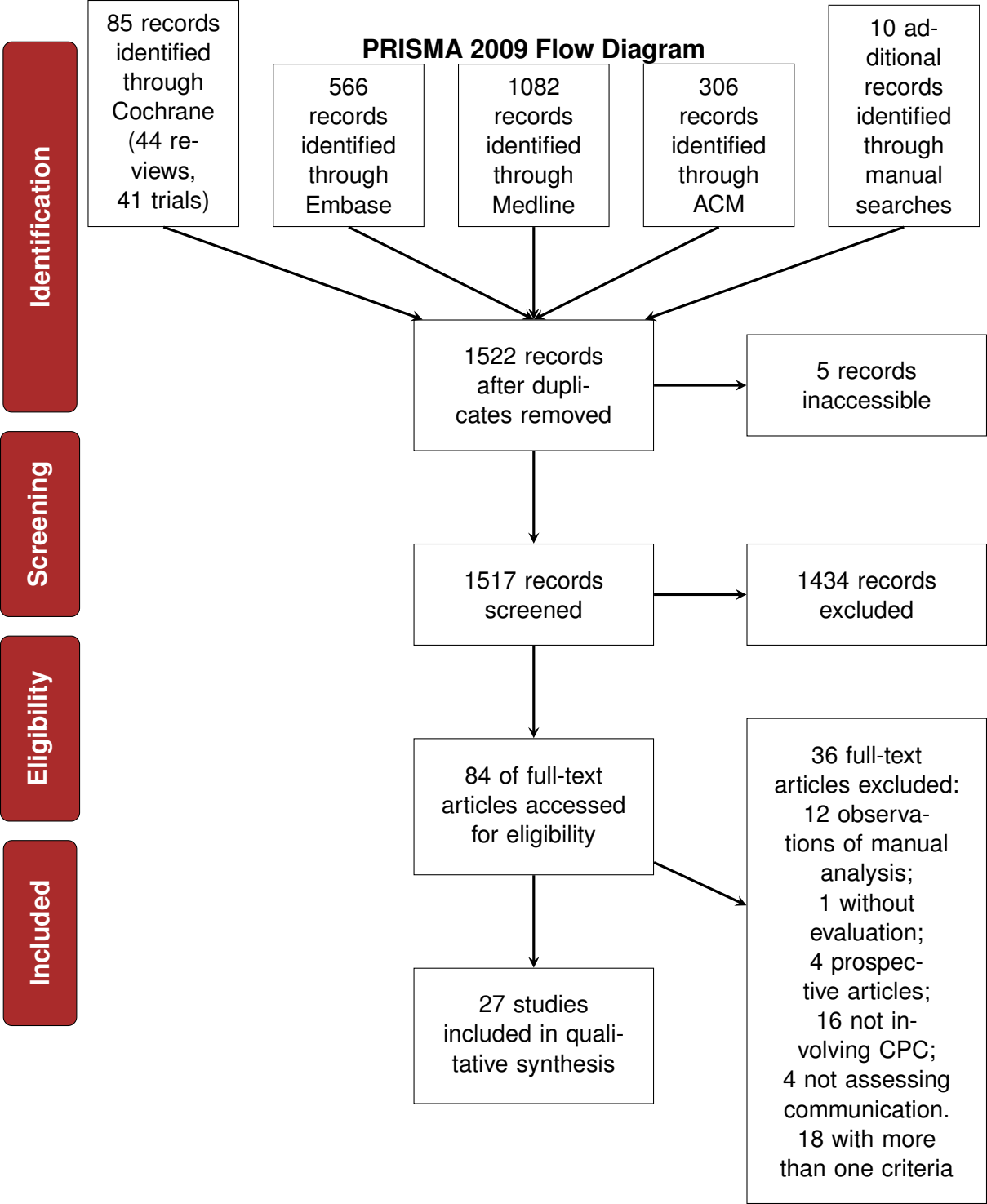
D: duration, -: not reported. a: age, g: gender, s: socio-professional class. SD: standard deviation

Table A.11: Datasets assessment (continued).

ID	Data set/Subset size	Data type	Data annotation	Data balance	Data availability	Language
[52]	60 interactions (mean duration: suicidal: 869s, non-suicidal: 490s)	audio recordings (mono, SNR=17.2 dB), transcripts, speech segments (software: ELAN) **type**: semi-structured	not annotated	a: -(by design), g: yes, s: -	Available: no	English
[54]	132 video recorded interactions	video, transcripts (professional) questionnaire: previous history with medical interactions, trust questionnaire (n=65) **type**: semi-structured	not annotated	a: no, g: no, s: no	Available: no	English
[56]	360 interactions (median length=605 utterances)	transcripts (manual), patient's questionnaire (physician communication, Likert scale)	General Medical Interaction Analysis System (GMIAS), kappa= 0.81 to 0.95	a: no, g: no, s: no	English	
[57]	8 interactions. Duration: - (segment≥3min)	Audio, video, transcripts (-). Type: conversational	effectiveness of interaction (effective, ineffective). No evaluation	a: -, g: -, s: -	Available: no	(English)
[58]	162 interactions (2-4 participants). Duration: -	Audio, video, transcripts (research team, manual), questionnaires (DPFC). Type: conversational	6: Number of related words in the grouped themes, number of related utterances containing the related words, time spent on related utterances in a record and percentages of these three variables in total number of words, utterances and time duration of a record. No evaluation	a: yes (children), g: yes, s:-	Available: no	(Chinese and/or English)

D: duration, -: not reported. a: age, g: gender, s: socio-professional class. SD: standard deviation


Years of publication
number of publications
0
1
2
3
4
5
6
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020


PRISMA 2009 Flow Diagram
Identification
Screening
Eligibility
Included
85 records identified through Cochrane (44 reviews, 41 trials)
566 records identified through Embase
1082 records identified through Medline
306 records identified through ACM
10 additional records identified through manual searches
1522 records after duplicates removed
5 records inaccessible
1517 records screened
1434 records excluded
84 of full-text articles accessed for eligibility
36 full-text articles excluded: 12 observations of manual analysis; 1 without evaluation; 4 prospective articles; 16 not involving CPC; 4 not assessing communication. 18 with more than one criteria
27 studies included in qualitative synthesis