Natural Language Processing for Mimicking Clinical Trial Recruitment in Critical Care: A Semi-automated Simulation Based on the LeoPARDS Trial

Hegler C. Tissot, Anoop D. Shah, David Brealey, Steve Harris, Ruth Agbakoba, Amos Folarin, Luis Romao, Lukasz Roguski, Richard Dobson, and Folkert W. Asselbergs

***Abstract*—Clinical trials often fail to recruit an adequate number of appropriate patients. Identifying eligible trial participants is resource-intensive when relying on manual review of clinical notes, particularly in critical care settings where the time window is short. Automated review of electronic health records (EHR) may help, but much of the information is in free text rather than a computable form. We applied natural language processing (NLP) to free text EHR data using the CogStack platform to simulate recruitment into the LeoPARDS study, a clinical trial aiming to reduce organ dysfunction in septic shock. We applied an algorithm to identify eligible patients using a moving 1-hour time window, and compared patients identified by our approach with those actually screened and recruited for the trial, for the time period that data were available. We manually reviewed records of a random sample of patients identified by the algorithm but not screened in the original trial. Our method identified 376 patients, including 34 patients with EHR data available who were actually recruited to LeoPARDS in our centre. The sensitivity of CogStack for identifying patients screened was 90% (95% CI 85%, 93%). Of the 203 patients identified by both manual screening and CogStack, the index date matched in 95 (47%) and CogStack was earlier in 94 (47%). In conclusion, analysis of EHR data using NLP could effectively replicate recruitment in a critical care trial, and identify some eligible patients at an earlier stage, potentially improving trial recruitment if implemented in real time.**

I. Introduction

RANDOMISED clinical trials can provide robust evidence of the effectiveness of medicines and other treatments, but are expensive to conduct and may fail to recruit a sufficient number of appropriate patients to have adequate statistical power [1]. Clinical trials units try to use a variety of techniques to increase patient recruitment, such as increasing the awareness amongst patients and clinicians [2]. However, identification of suitable patients can be resource-intensive, often relying on manual review of clinical notes to identify potentially eligible patients, where the information may be split over different systems. This can be particularly difficult in emergency settings and intensive care units (ICU), where it is important to identify eligible patients early, so that the window of opportunity is not missed [3]. Staff shortages and inconvenient timing can potentially lead to eligible participants being missed [4].

Electronic heath records (EHRs) are increasingly used for research [5] and have been proposed as a way of improving trial recruitment, either via a patient-centric approach or in the form of decision support for clinicians, such as point-of-care alerts [6]. Algorithms to identify trial participants may reduce the human resource needed for identifying patients earlier. Patient characteristics extracted from EHR databases can be mapped to trial information derived from study eligibility criteria [7], [8]. However, much of the information in EHRs is unstructured, in the form of free text, rather than in a structured form. Natural language processing (NLP) techniques can extract relevant information from free text, but cannot be relied upon to be completely accurate because of typographical errors and nuances of human language. However, NLP may be used to pre-screen potential trial participants, reducing the number of patient records that need manual review [9]–[11].

Algorithms incorporating NLP have been tested for their ability to identify patients eligible for clinical trials [9], [12]–[14]. Previous studies have tended to evaluate patients at a single time-point only, but critical care patients may have rapidly changing physiology, and trials may have narrow time windows for recruitment. One such trial was the LeoPARDS trial [15], which tested a drug for improving outcomes in life-threatening infections in ICU patients. We aimed to test whether NLP in combination with electronic structured data could assist in recruiting patients to a critical care trial such as this. The simulation was conducted within one of the LeoPARDS trial sites, University College London Hospitals NHS Foundation Trust (UCLH). UCLH is is a teaching hospital and part of a National Institute for Health Research (NIHR) Biomedical Research Centre (BRC), and is leading a collaboration across multiple BRCs to curate a critical care research database within the NIHR Health Informatics Collaboration (CCHIC) [16]. We demonstrate that even simple NLP techniques combined with a multi-contextualized searchable

H. C. Tissot, A. D. Shah, D. Brealey, S. Harris, R. Agbakoba, L. Romao and L. Roguski are with the Institute of Health Informatics, University College London, London, UK; Health Data Research UK London, University College London, London, UK; and University College London Hospitals, London, UK

A. Folarin, and R. Dobson are with the Institute of Health Informatics, University College London, London, UK; Health Data Research UK London, University College London, London, UK; and the Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK

F. W. Asselbergs is with the Institute of Health Informatics, University College London, London, UK; Health Data Research UK London, University College London, London, UK; and Department of Cardiology, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands

platform are able to support automatic or semi-automatic clinical trial screening.

II. METHODS

A. Data Sources and Informatics Infrastructure

The Critical Care Health Informatics Collaboration (CCHIC) is a research platform comprising EHR patient data from critical care units at five large BRCs (Cambridge, Guys/Kings/St Thomas', Imperial, Oxford, and UCLH) [16]. Data are available from 2014 onwards, extracted from a diverse range of EHR systems in a standardised format. Data are curated into a research-ready database which has been approved by an NHS Research Ethics Committee (14/LO/1031). The CCHIC dataset includes 108 hospital, unit, patient and episode descriptors (recorded once per admission) and 154 time-varying variables including physiological measurements, laboratory tests, nursing activities and drug administration. CCHIC includes structured data only; unstructured data such as images and free text are not included.

For this study we combined structured UCLH data from CCHIC (comprising admission and discharge dates, physiological measurements and laboratory results) with unstructured narrative notes (free text) entered by clinicians, which may include patient histories, examination findings, diagnoses, past medical history, suspected conditions and care plans. Free text was extracted from the UCLH critical care EHR (the IntelliVue Clinical Information Portfolio (ICIP) by Phillips) recorded in the following fields: problem lists, event timeline, reason for admission, admission history, past medical history, and pre-admission medication.

Stuctured and free text data from the EHR were combined into a searchable indexed repository using the CogStack [17] platform for document processing and distributed analysis. CogStack comprises a set of open source services coordinated by a batch processing framework that provides multiple interfaces for NLP and document processing tools (such as image to text conversion). It enables structured information from relational databases to be combined with free text documents in a configurable, searchable, patient-oriented representation. Full text searching is possible using Elasticsearch [18].

Bio-Yodie [19] is a named entity linking system developed as part of the KConnect Horizon 2020 project. It finds mentions in the text that correspond to Unified Medical Language System (UMLS) [20] concepts, and uses various knowledge sources to choose the best interpretation when more than one possible match is found. The pipeline utilizes a gazetteer to locate words or phrases that may indicate an entity mention. After using a stop list to remove low precision terms, these mentions are then used to retrieve all the possible candidates for that term, and a number of scores is used to pick the most likely matching UMLS concept. The system returns the text annotated with matched UMLS Concept Unique Identifiers (CUI) and other relevant information from the UMLS.

Finally, SemEHR [21] is a CogStack tool that enables concepts retrieved by an information extraction system such as Bio-Yodie to be contextualised. It uses the ConText algorithm [22], a rule-based algorithm to determine experiencer (patient or other), affirmation status (affirmed, negative or hypothetical) and temporality (past or recent). NLP annotations are then assembled at the patient or document level to generate a timeline view, used to provide semantic data via ontology-based search and analytic interfaces. We only used affirmative UMLS concepts that were experienced by the patient in this study.

We developed an application to mimic the trial screening process using CogStack. We converted the trial eligibility criteria to a computable algorithm comprising: logical conditions based on numerical structured data, clinical concepts extracted from text by a contextualized search function, and logical compounding functions to group multiple specific conditions into higher level selection criteria. We compared potentially eligible patients identified by CogStack with those screened or recruited in the original LeoPARDS trial for the intersection of time periods between trial recruitment and the CCHIC data. All analysis on the EHR data was carried out by researchers blinded to the trial recruitment log, with no involvement in the original trial.

B. The LeoPARDS Trial

The LeoPARDS trial (Levosimendan for the Prevention of Acute oRgan Dysfunction in Sepsis) investigated whether a 24-hour infusion of levosimendan improved organ dysfunction in septic shock [15]. The trial screened 2,382 patients in 2014–2015 across 31 centres and recruited 526 patients, of whom 47 were from UCLH, recruited between June 2014 and December 2015. The primary outcome was the mean daily Sequential Organ Failure Assessment (SOFA) score, which is used to track the evolution of organ dysfunction. The study showed no significant difference between levosimendan and placebo (mean difference in SOFA score, 0.61; 95%CI, -0.07 to 1.29).

Recruitment into LeoPARDS required patients with new onset septic shock to be identified within 24 hours, so that they could be randomised to the study drug or placebo. Eligible patients were identified by dedicated research nurses who spent 4 to 6 hours each day reviewing the notes of all new ICU admissions, to identify those who should be selected for screening. The selection criteria are shown in Table I. The inclusion criteria aimed to identify adult patients ($\geq$ 18 years) with septic shock [23], and the exclusion criteria aimed to exclude patients in whom the trial therapy was inappropriate or unsafe, and patients with medical conditions which might make the outcome of the trial difficult to interpret.

C. Simulation of Patient Identification Using CogStack

We simulated the review of all patients in the ICU every hour, looking back on clinical data collected during the previous 24 hours to classify whether the patient had new onset septic shock; see Fig. 1. The patient was marked as eligible at the earliest timepoint that they fulfilled the selection criteria, and they were not considered eligible for future time points.

Septic shock was defined in EHR data as concurrent existence of systemic inflammatory response syndrome (SIRS) due to known or suspected infection, and use of vasopressor drugs.

TABLE I
SELECTION CRITERIA FOR THE LEOPARDS TRIAL [15].

Inclusion Criteria	Exclusion Criteria
(A) Fulfil at least 2/4 of the criteria of the systemic inflammatory response syndrome (SIRS) due to known or suspected infection within the previous 24 hours. The SIRS criteria are: 1. fever ($> 38°C$) or hypothermia ($< 36°C$), 2. tachycardia (heart rate > 90 beats per minute), 3. tachypnoea (respiratory rate > 20 breaths per minute or $PaCO_2 < 4.3$ kPa) or need for mechanical ventilation, and 4. abnormal leukocyte count ($> 12{,}000$ cells/mm^3, < 4000 cells/mm^3, or $> 10\%$ immature [band] forms). (B) Hypotension, despite adequate intravenous fluid resuscitation, requiring treatment with a vasopressor infusion (e.g. noradrenaline / adrenaline / vasopressin analogue) for at least four hours and still having an ongoing vasopressor requirement at the time of randomisation.	(A) more than 24 hours since meeting all the inclusion criteria; (B) end-stage renal failure at presentation (previously dialysis-dependent); (C) severe chronic hepatic impairment (Child-Pugh class C) [24]; (D) a history of torsades de pointes; (E) known significant mechanical obstructions affecting ventricular filling or outflow or both; (F) treatment limitation decision in place (e.g. 'Do Not Resuscitate' or not for ventilation/dialysis); (G) known or estimated weight > 135 kg; (H) known to be pregnant; (I) previous treatment with levosimendan within 30 days; (J) known hypersensitivity to levosimendan or any of the excipients; (K) known to have received another investigational medicinal product within 30 days or currently in another interventional trial that might interact with the study drug – potential co-enrolment into other studies would be considered on an individual study basis.

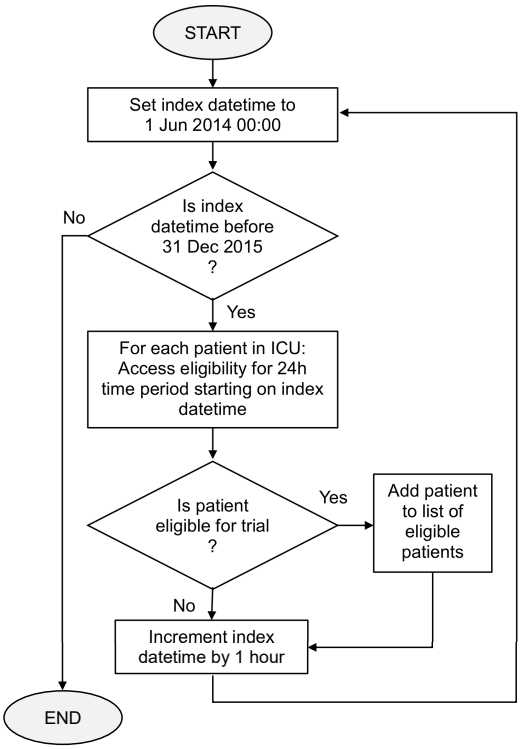

Fig. 1. Flowchart showing iteration through screening times.

SIRS was defined on the basis of meeting physiological threshold values for at least two of the following four parameters: body temperature (fever or hypothermia), ventilation (high respiratory rate or requirement for mechanical ventilation), heart rate, and white cell count. We did not assess hypotension or measures of fluid resuscitation, but assumed that patients prescribed vasopressor medication had refractory hypotension (Fig. 2).

We operationalised 'known or suspected infection' as a recent diagnosis of infection from SemEHR (with specific types and sites of infection listed as search terms; see Table IV).

We then applied the LeoPARDS exclusion criteria using structured and unstructured data as follows: end stage renal failure, dialysis, torsades de pointes, mechanical obstruction (mitral stenosis or aortic stenosis) or severe hepatic impairment (using either *recent* or *past* temporal context provided by SemEHR), or pregnancy (using only *recent* temporal context). Relevant UMLS concepts are listed in Table V. We additionally identified patients with severe hepatic impairment by the presence of any two of bilirubin ≥ 34.2 micromol/L (CCHIC structured data), ascites or encephalopathy. This is an approximation of Child-Pugh class C [24], assuming encephalopathy is severe, ascites is moderate, and the international normalised ratio and albumin are in the middle of the scoring ranges. We were unable to apply the exclusion criteria of previous treatment with levosimendan, hypersensitivity to levosimendan or enrolment into another interventional trial (items I, J, and K from Table I) as this information was not entered in a structured way or in a form of text that could extracted as UMLS concepts. Fig. 2 presents the overall workflow for identifying patients eligible for LeoPARDS trial.

D. Technical Implementation of Eligibility Criteria

Our approach was designed to support an unlimited recursive nested set of conditional clauses connected by grouping logical operators. Partial matches and temporal constraints were also required as part of the formal criteria specification.

The selection criteria for the LeoPARDS trial were designed by following an inner hierarchical structure of conditional components. The inclusion criteria comprised the default mandatory component for defining patient eligibility, requiring at least an inner logical group or an inner logical specification. The exclusion criteria were a complementary component

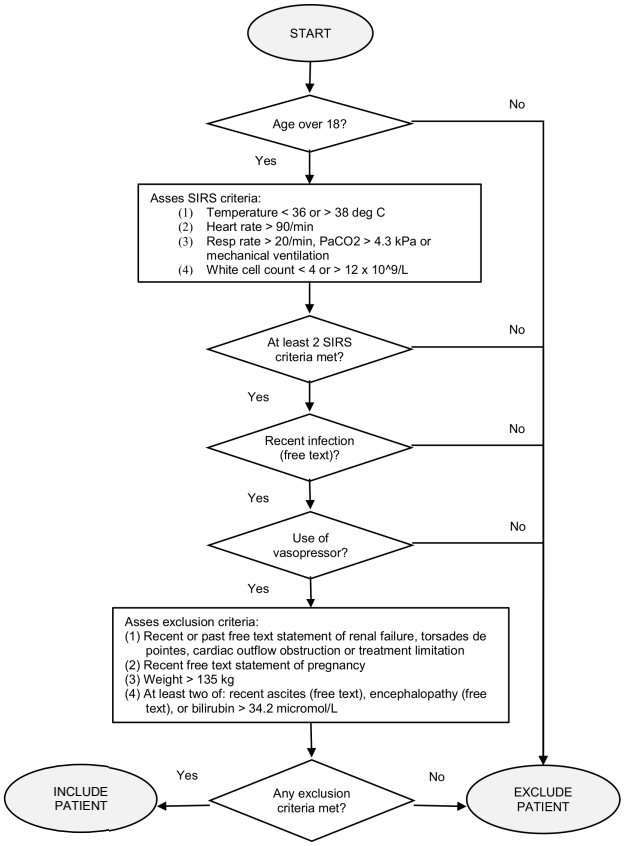

Fig. 2. Workflow for identifying patients eligible for LeoPARDS trial using electronic health records.

comprising an inner logical group of conditions, specifying the set of patients to be subtracted (or flagged) from the cohort matching the inclusion conditions.

In order to formally describe the inclusion and exclusion criteria for the LeoPARDS trial, we defined a set of logical compounding functions (LCF). LCFs group a set of logical conditions that are individually evaluated and logically combined to give an overall *True* or *False* output. LCF results can be hierarchically combined to specify more complex logical operations. The proposed LCFs are described below – n is a numerical constraint parameter and L is the set of logical conditions to be evaluated (all LCFs return $False$ when $L = \emptyset$):

- $\texttt{MIN}(n, L)$: each logical condition $c \in L$ is logically evaluated, resulting $True$ when at least n conditions from L result $True$;
- $\texttt{MAX}(n, L)$: each logical condition $c \in L$ is logically evaluated, resulting $True$ when no more than n conditions from L result $True$;
- $\texttt{ALL}(L)$: results $True$ if, and only if, there is no condition $c \in L$ logically evaluated resulting $False$;
- $\texttt{ANY}(L)$: results $True$ when there is at least one logical condition $c \in L$ that is logically evaluated resulting $True$ – equivalent to: $MIN(1, L)$;
- $\texttt{ONE}(L)$: results $True$ if there is only one condition $c \in L$ that is logically evaluated resulting $True$, all the other conditions resulting $False$ – equivalent to: $MAX(1, L)$;
- $\texttt{NOT}(L)$: results $True$ if, and only if, there is no condition $c \in L$ logically evaluated resulting $True$ – equivalent to: $MAX(0, L)$.

In addition to the LCFs described above, we defined a textual contextualised function that searches specific UMLS concepts resulting from SemEHR and stored in an Elasticsearch repository:

- $\texttt{UMLSearch}(umls, temporality)$: is a textual contextualised search condition that matches annotated documents (free text notes) against one or more UMLS concepts (*umls* parameter) in a given time constraint (*temporality* parameter) – *temporality* can be set as *past* or *recent*, from which *recent* takes into account any UMLS concepts mentioned in any documents dated up to the last 72 hours from the reference screening date (index *date/time* in Fig. 1) set as "recent" by SemEHR, whereas *past* considers any historical occurrences of the given UMLS concepts. When $temporality$ is not given, $UMLSearch$ searches for any mention of the given UMLS identifiers that have been experienced by the patient. For example, the inclusion criteria $\texttt{UMLSearch(UMLS\{infection\},recent)}$ results a list of all *affirmative* annotations (excludes negations) corresponding to the UMLS codes for *infection* (Table IV) found in *recent* documents (last 72 hours in which the patient is set as the experiencer (not a family member).

We started by using LCFs to design the primary filters required to match patients according to the inclusion criteria. Primary conditions are supported by structured data points available in CCHIC. Table II provides a formal description of some of the filters used to design the selection criteria in terms of logical conditions coupling variables, logical operators, and grouping LCFs.

E. Comparison of Automated and Manual Screening

We used the CogStack algorithm to identify the start of septic shock episodes, and compared patients identified with this method to those actually screened or recruited to the trial in UCLH. In order to make the comparison valid, we restricted the comparison to the set of patients who could potentially have been selected by either method. This was the set of patients who were admitted to UCLH ICU between July 2014 and June 2015, or October 2015 to December 2015. This was a period of time when LeoPARDS was actively recruiting and CCHIC data were available. We excluded the time period July to September 2015, when UCLH ICU audit activities were suspended because of staffing shortages, and CCHIC data were therefore incomplete. We included only the first ICU admission per patient. All English citizens have an NHS number, which was used to link free text on the ICIP system with structured data in CCHIC. We excluded patients without NHS numbers (such as foreign patients), in whom the datasets could not be linked by this method.

We compared the set of patients identified as eligible for LeoPARDS by the CogStack algorithm with the screening logs for the original trial. For patients detected as eligible by CogStack but not screened in the original trial, we carried out

TABLE II
EXAMPLES OF ELIGIBILITY CRITERIA CONDITIONS DESIGNED WITH STRUCTURED LOGICAL CONDITIONS AND LOGICAL COMPOUNDING FUNCTIONS (LCFS).

Condition	Logical description
fever	$body\_temperature > 38.0$
hypothermia	$body\_temperature < 36.0$

tachycardia	ANY {
	$heart\_rate > 90$,
	$heart\_rhythm > 90$
	}

tachypnoea	ANY {
	$resp\_rate > 20$,
	$PaCO_2 > 4.3$,
	$mechanical\_ventilation > 0$
	}

[1] abnormal leukocyte	ANY {
	$white\_cell\_count > 12000.0$,
	$white\_cell\_count < 4000.0$
	}

[2] hypotension	ANY {
	$noradrenaline > 0$,
	$vasopressin > 0$,
	$terlipressin > 0$,
	$dopamine > 0$,
	$dobutamine > 0$,
	$adrenaline > 0$
	}

overall inclusion criteria	ALL {
	$UMLSearch(UMLS\{infection\}^3, recent)$
	,
	$MIN(2)$ {
	ANY $\{fever, hypothermia\}$,
	$tachycardia$,
	$tachypnoea$,
	$abnormal\_leukocite$
	}
	}

[1] The CCHIC dataset does not include structured variables for leukocyte morphology, so it was not possible to extract the criterion "immature [band] forms". However, we assumed that in the majority of cases the leukocyte criterion would be met based on absolute numbers.

[2] We used vasopressor treatment rather than blood pressure for the definition of "hypotension", because the criterion required that the patient had persistent hypotension despite adequate fluid resuscitation and required vasopressor treatment, and we assumed that standard clinical practice of ensuring adequate fluid resuscitation would have been followed.

[3] See Table IV for UMLS concepts defining $infection$.

a manual case note review of a random sample. Two clinicians reviewed the original EHR case notes on the ICIP system to ascertain whether the algorithm correctly applied the eligibility criteria, and why the patient was not included in the screening log.

III. RESULTS

For the actual LeoPARDS trial in UCLH, 315 ICU admissions (303 patients) were assessed as being potentially eligible on screening, and 47 patients were recruited. Our CogStack algorithm identified 407 ICU admissions (395 patients) that met the eligibility criteria for LeoPARDS. Fig. 3 shows the numbers of screened and recruited patients by month from June 2014 to December 2015.

We restricted the comparison to patients with NHS numbers admitted to ICU during the period when CCHIC data was non-missing (excluding July to September 2015), and all subsequent results are based on this subset. There were 2571 ICU patients with a total of 2862 admissions. Of these, 376 patients were identified as eligible by CogStack and 226 were selected for screening by the manual process. Thus using CogStack to pre-select patients for screening could potentially reduce the number of patients for manual review by 85% (2195/2571; 95% CI 84%, 87%).

The sensitivity of CogStack for identifying patients screened for LeoPARDS was 90% (95% CI 85%, 93%). Of the 203 overlapping patients identified by both manual screening and CogStack, 74% (151/203, 95% CI 68%, 80%) had a screening date which matched within one day. All 34 patients who were actually recruited to LeoPARDS were detected by CogStack (see Table III). The 173 additional patients detected by CogStack had a similar sex distribution (56.1% male, 97/173) to the overlapping patients (55.7% male, 113/203, P value for comparison = 1). The age distribution was also similar (Fig. 4, P value for comparison by Wilcoxon test 0.66).

We also analysed the ability of CogStack to identify eligible patients earlier than the original UCLH screening log. From the 203 overlapping patients, 95 (47%; 95% CI 40%, 54%) were found by CogStack the same day as screening, 48 patients (24%; 95% CI 18%, 30%) were detected one day earlier, and 46 patients (23%; 95% CI 17%, 29%) were detected two or more days earlier (Fig. 5). Where CogStack was not able to identify patients as early as the manual screening log, this was because they had been matched to the same patient in an earlier or later ICU admission.

Among the 173 patients detected by Cogstack but not screened in the original trial, we manually reviewed the clinical notes of a random sample of 20 (11.6%). We found only 2 patients (10%) who could potentially have been enrolled, and one of these was not screened because it was during the New Year holiday period when trial staff were not working. Of the remainder, 4 (20%) had a Child-Pugh score of 1 according to CogStack, i.e. mild liver dysfunction not meeting the exclusion criteria, but on manual screening they were excluded because of liver impairment (3 patients) and pancreatitis instead of sepsis being the cause of SIRS (1 patient). Eight patients strictly met the inclusion and exclusion

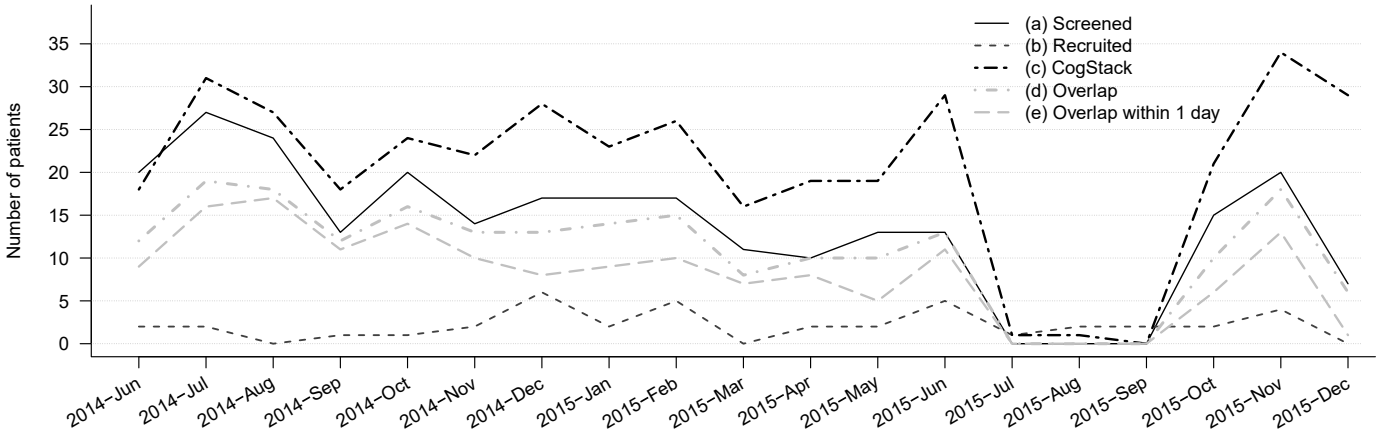

Fig. 3. (a) number of patients in the original UCLH screening log; (b) number of patients recruited in UCLH; (c) number of patients found by CogStack; (d) number of patients overlapping between CogStack and the original UCLH screening log; (e) number of patients overlapping with no more than 1 day difference between the screening date from UCLH screening log and CogStack. (Note that CogStack was not able to find patients matching the selection criteria between July and September 2015 due to a known lack of data in the CCHIC dataset.)

TABLE III
COMPARISON OF NUMBER OF PATIENTS IDENTIFIED BY COGSTACK AS ELIGIBLE FOR LEOPARDS TRIAL, COMPARED TO THE GOLD STANDARDS OF THOSE ACTUALLY SCREENED OR ACTUALLY RECRUITED.

CogStack	**Actual screening**		**Actual recruitment**	
	Screened	Not screened	Recruited	Not recruited
Detected and included	203	173	34	342
Detected and excluded*	4	0	0	4
Not detected	19	2172	0	2191
Recall (sensitivity)	0.898 (95% CI 0.851, 0.934)		1.000 (95% CI 0.897, 1.000)	
Precision	0.540 (95% CI 0.488, 0.591)		0.090 (95% CI 0.063, 0.124)	
Specificity	0.926 (95% CI 0.915, 0.936)		0.865 (95% CI 0.851, 0.878)	
F1-score	0.674		0.166	

*Patient matches at least one of the exclusion criteria

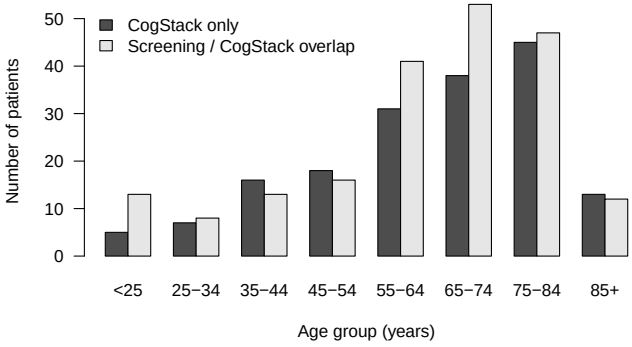

Fig. 4. Age distribution of patients identified by CogStack as being eligible for the trial, according to whether or not they were included in the actual trial screening.

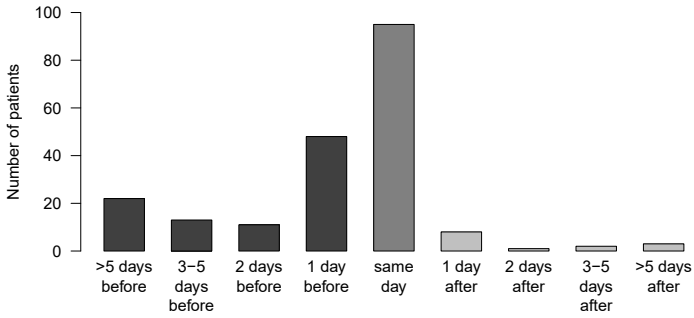

Fig. 5. Timing of eligibility identified by CogStack compared to manual screening.

criteria, but were considered clinically unlikely to benefit from an experimental therapy because they were either too sick or dying (5 patients) or at the least severe end of the spectrum (3 patients). Six patients (30%) had an alternative explanation for the combination of antibiotic treatment and physiological parameters that suggested sepsis, that was not programmed into the CogStack algorithm. A typical example was a post-surgical patient on prophylactic antibiotics, with raised respiratory rate and heart rate (possibly due to pain), requiring inotropes for hypotension due to an epidural.

Finally, in terms of processing time, CogStack demonstrated potential to serve as a near real-time search and filtering tool in order to facilitate the pre-screening process. The full process of screening approximately 11,500 24-hour-sliding windows (during the time period of the study) was performed in less then 15 minutes (890 seconds), corresponding to less than 0.1 second per window screening. Experiments were performed using a 64-bit Linux server with 8 core Intel® Xeon®CPU (E5-2680 v4 2.40GHz) and 64GB RAM.

IV. Discussion

This study showed that an electronic algorithm incorporating NLP could successfully match patients against the selection criteria for a clinical trial in the critical care setting with a time-sensitive recruitment window. Use of the algorithm for pre-screening could potentially reduce the number of patient notes that need manual review by 85%, a considerable saving in time and labour. Although taking into account only a subset of free text notes, CogStack was able to identify all the patients actually recruited and almost all that were screened. Besides being faster, CogStack was also able to identify that patients were eligible at least 2 days earlier in 20% of cases, suggesting that it may be able to help prevent failure of recruitment due to missing a timeline.

Results from the manual check showed that strict application of the criteria resulted in some patients being identified who would not be included based on clinical judgement (if they were not sick enough to risk an experimental treatment, or if they were so sick that any intervention was likely to be futile). This suggests that trial inclusion and exclusion criteria need to be more explicit if they are to be accurately applied by computer algorithms while truly capturing the desired patient population. Very few additional eligible patients were detected by the CogStack algorithm, which shows that the manual processes for participant identification were thorough, albeit resource-intensive.

A. Improving efficiency of clinical trials

Randomised controlled trials (RCTs) are considered the gold standard to assess the effects of medical interventions such as pharmacological treatments [25]. However, they are time consuming and expensive, and the populations included in RCTs often do not resemble real-world patient populations [2], [6], [8]. There is interest in using EHR databases to conduct clinical trials, with randomised treatment allocation as per conventional trials, but recruitment and follow-up managed through the EHR rather than with bespoke visits [26]–[28]. Such trials may recruit a more representative population than conventional trials, and the results may better reflect effectiveness in routine clinical practice [26], [29]. Another advantage of EHR-based automatic patient selection is that the algorithms can be modified and re-applied to test different patient selection criteria, making it easier to design future trials [30].

Clinical trials need to recruit participants according to the eligibility criteria defined in the protocol in order to accurately answer the study question. Despite effort and expense, and a long time period allocated for patient numbers to accrue, attainment of enrolment goals seems elusive in many studies [1], [31]. Among randomised controlled trials funded by the NIHR Health Technology Assessment programme, the final recruitment target sample size was achieved in only 56% [32]. This can have major impact on the feasibility, power and validity of the trials.

Electronic health records provide the potential to identify trial participants more efficiently [7], [8], but there are a number of challenges. Mapping the selection criteria to logical conditions can be difficult, as eligibility criteria are described using natural language designed for human rather than computer interpretation. There has been interest in using NLP to develop computable algorithms from free text trial descriptions [33]–[36], and an 'eligibility criteria representation language' has been proposed [37]. However, unless EHR data sources are standardised, it is a major task to enable complex queries to run on disparate data sources [38]. Representation of time constraints also needs to be taken into account [39]. Temporal references can be described in diverse ways with varying degrees of precision (e.g. "within the previous 24 hours", "previous treatment within 30 days", "for at least four hours") [40], [41].

B. Natural language processing

Detailed information on patient characteristics that are relevant to trial inclusion and exclusion criteria may not be included in the structured data, and only available in the free text of EHRs. Although narrative text is a valuable asset for improving healthcare [42], it is usually inaccessible due to its lack of structure, hence the need for NLP applications to extract information in a structured form.

Many of the NLP approaches to date have a fairly narrow focus using simple rule-based approaches (e.g. regular expression patterns) in order to address specific information extraction tasks, but they require extensive human intervention for application to new tasks [43]. There are a number of open source tools that can annotate clinical text using UMLS or another terminology; examples include cTakes [44] (Mayo Clinic), Freetext Matching Algorithm [45] and Bio-Yodie [19], which was used in this project.

Machine learning NLP approaches have been growing in popularity, as they are more flexible in enabling the system to 'learn' from training data [46]. This is typically done in a 'supervised' manner using manually annotated training data, but semi-supervised [47] and unsupervised methods have also been developed. An unsupervised approach incorporating an ontology could accurately identify arrhythmia events from Italian medical reports [48].

Text analytics platforms such as SemEHR (built on CogStack) [17], [21] and GATE [49] are increasingly being used across large document repositories, and can incorporate a range of NLP tools.

C. Strengths and limitations

The main strength of this study was the demonstration of algorithms combining structured EHR data and NLP to assist participant recruitment in a simulation of a real clinical trial. The LeoPARDS trial had particular recruitment challenges – the time-sensitive nature of the task, and the severity of the patients' condition.

A limitation was that our algorithm attempted to identify a diagnosis of sepsis which may be difficult even for experienced clinicians. Hence application of the strict inclusion and exclusion criteria identified patients who were not eligible because they had an alternative explanation for their physiological state that was not sepsis; this was apparent to clinician reviewers but

not to the algorithm because it was not programmed in. This highlighted the need for much more explicit trial inclusion algorithms if they are to be interpreted automatically, and it may be difficult to plan for all such nuances in advance.

Another limitation was that the algorithm included only key portions of the free text rather than the entire clinical record, and the identification of some criteria was not possible (such as white cell morphology). We were limited to the single site which had free text available for NLP, but the method could potentially be scaled to many sites and adapted for different studies.

D. Clinical and research implications

This study has demonstrated the feasibility of this approach in a critical care trial. Future work should apply this method at other sites and for other studies, and to develop a method for a current clinical trial in order to evaluate its utility and performance for real-time patient screening and recruitment. UCLH is currently building an 'Experimental Medical Application Platform' which will combine a rich research repository of EHR data with text analytics (CogStack) and real-time data feeds from the operational EHR, to enable the rapid development of this type of research application.

The algorithm could also be tuned by testing out different thresholds for inclusion and exclusion, in order to achieve a combination of sensitivity and specificity which best suits its use in combination with manual review in a trial recruitment scenario. Use of EHR data with NLP could also be used to extract participant data for the trial case report forms. This will save even more time by avoiding the need for duplicate data entry, and enable the use of more detailed measures of health status, such as continuous monitoring of physiological parameters rather than a single measurement in a case report form. However, it also introduces new challenges such as ensuring validity, completeness and accuracy of the data [50], [51], and harmonising heterogenous data across institutes.

V. Conclusions

Electronic health record data may potentially be used in computer algorithms to help identify trial participants and increase recruitment in clinical trials, but much of the detailed clinical information is available only in the form of free text. We simulated screening and recruitment for the LeoPARDS trial in critical care, using the Cogstack platform with rule-based natural language processing tools to process electronic health record data. CogStack was able to identify the majority of patients originally screened, including all those recruited, and in many cases able to identify patients as eligible one or two days before the actual manual screening process. This approach could be implemented in real time to facilitate clinical trial recruitment, and reduce the burden of time-consuming manual case note review.

Acknowledgements

This study was supported by the National Institute for Health Research (NIHR) University College London Hospitals (UCLH) Biomedical Research Centre (BRC) Clinical and Research Informatics Unit (CRIU), NIHR Health Informatics Collaborative (HIC), and by awards establishing the Institute of Health Informatics at University College London (UCL).

Patient data analysed for this project was extracted from the Critical Care Health Informatics Collaboration research database, which has had National Research Ethics Service approval (14/LO/1031). Individual participant consent was not required, as section 251 exemption was granted by the Confidentiality Advisory Group of the Health Research Authority.

This study was funded Health Data Research UK (grant No. LOND1), which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome Trust. ADS is supported by a postdoctoral fellowship from THIS Institute.

RD is supported by: (a) the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and Kings College London. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health; (b) Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome Trust; (c) The BigData@Heart Consortium, funded by the Innovative Medicines Initiative-2 Joint Undertaking under grant agreement No. 116074. This Joint Undertaking receives support from the European Unions Horizon 2020 research and innovation programme and EFPIA; it is chaired, by DE Grobbee and SD Anker, partnering with 20 academic and industry partners and ESC; (d) The National Institute for Health Research University College London Hospitals Biomedical Research Centre.

FWA is supported by UCL Hospitals NIHR Biomedical Research Centre.

DB is partially funded by the Division of Critical Care, University College Hospital and NIHR University College London Hospitals Biomedical Research Centre.

TABLE IV
UMLS RELEVANT CONCEPTS FOR "INFECTION".

Infection	UMLS IDs
cellulitis	C0007642, C0267568, C0742022, C0554110, C0263145, C0343024
cholangitis	C0311273, C0267922
cholecystitis	C0008325, C0149520, C0267841
chronic sinusitis	C0149516
colitis	C0277524, C0343386
cystitis	C0010692
diverticulitis	C0012813
empyema	C0014009
endocarditis	C0014118, C0014122, C0014121, C0155685, C0746604
epididymitis	C0014534
epiglottitis	C0014541
folliculitis	C0016436
gastroenteritis	C0017160
herpes encephalitis	C0276226
infection (generic)	C3714514, C0009450, C0876973, C0037278, C0151317, C0850034, C0262655, C0275518, C0877046, C0242172, C0744926, C1699561, C0022729, C0238990, C0439633, C1112209, C0749769, C1698666, C0035243, C0442886, C0745687, C0860239, C0851989, C0149778, C0038941, C0042029
ludwig angina	C0024081
mastitis	C0024894
maxillary sinusitis	C0024959
mediastinitis	C0025064
meningitis	C0025289, C0085437, C0085436
myocarditis	C0027059
myopericarditis	C0854532
necrotizing pancreatitis	C0267941
necrotizing fascitis	C0238124
osteomyelitis	C0029443, C0564832
peritonitis	C0031154, C0473119, C0341503, C0275551
pharyngitis	C0031350
pneumonia	C0032310, C0577702, C0339961, C0032300, C0702135, C0155870, C0155862, C0032290, C0264383, C0004626, C0949083, C1142578, C0519030, C1701940, C0747651, C0694549, C0585104, C0585105
prostatitis	C0033581
pyelonephritis	C0034186
pyonephrosis	C0034216
recurrent bronchitis	C0741796
sepsis	C0243026, C0036690, C0684256, C0152965, C1142182, C0877153, C1141926, C1719672, C0036685
sinusitis	C0037199
tonsillitis	C0040425

TABLE V
OTHER UMLS CONCEPTS USED TO COMPOUND THE LEOPARDS SELECTION CRITERIA.

Concept	UMLS IDs
torsades de pointes	C0040479, C1960156, C1963250, C3150851, C4510938, C4510799, C4511461
renal failure	C0011946, C0015354, C0019004, C0019014, C0022661, C0031139, C0041612, C0191116, C0200017, C0206075, C0264654, C0268810, C0271932, C0398312, C0398338, C0398340, C0398343, C0398344, C0403462, C0403463, C0403464, C0403465, C0419061, C0419062, C0455667, C0558708, C0565539, C0748315, C1561829, C3494724, C3531744, C3536572, C3649547, C3697607, C4038741, C4047993
mechanical obstruction	C0003492, C0003499, C0003507, C0024164, C0026269, C0151241, C0152417, C0155567, C0158618, C0264766, C0264772, C0275846, C0332886, C0340335, C0340361, C0340371, C0340372, C0340373, C0340375, C0344401, C0345086, C0345087, C0349073, C0349075, C0349516, C0406810, C0700637, C1290389, C1306822, C1850635, C1868705, C1960800, C3532372, C3532376, C3839320, C3839383, C3839635
treatment limitation	C0582114, C3472262, C4305111
liver impairment	C0019147, C0019212, C0085605, C0162557, C0274386, C0400927, C0400928, C0400929, C0745744, C1619727, C2936476, C4039103
pregnant	C0026751, C0032979, C0032980, C0032981, C0032995, C0033150, C0041747, C0149973, C0232989, C0232990, C0232991, C0232992, C0232993, C0232994, C0242786, C0269675, C0278056, C0404831, C0404842, C0425965, C0425979, C0425983, C0425984, C0425985, C0425986, C0425987, C0549206, C0585066, C0860096, C1291689, C2586154
ascites	C0003962, C0008732, C0019086, C0025184, C0031144, C0220656, C0267772, C0267773, C0267774, C0267776, C0269720, C0275919, C0341525, C0401037, C0401038, C0437001, C0585187, C0741244, C1285291, C3532188, C3665480, C4038874, C4038944
encephalopathy	C0019147, C0019151, C3266165

REFERENCES

[1] M. Cuggia, P. Besana, and D. Glasspool, "Comparing semi-automatic systems for recruitment of patients to clinical trials," *Int. J. Med. Inform.*, vol. 80, no. 6, pp. 371–388, Jun. 2011. [Online]. Available: http://dx.doi.org/10.1016/j.ijmedinf.2011.02.003

[2] T. Kennedy-Martin, S. Curtis, D. Faries, S. Robinson, and J. Johnston, "A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results," *Trials*, vol. 16, no. 1, p. 495, Nov 2015. [Online]. Available: https://doi.org/10.1186/s13063-015-1023-4

[3] K. E. A. Burns, C. Zubrinich, W. Tan, S. Raptis, W. Xiong, O. Smith, E. McDonald, J. C. Marshall, R. Saginur, R. Heslegrave, G. Rubenfeld, and D. J. Cook, "Research recruitment practices and critically ill patients. a multicenter, cross-sectional study (the consent study)," *American journal of respiratory and critical care medicine*, vol. 187, no. 11, pp. 1212–1218, 2013. [Online]. Available: https://app.dimensions.ai/details/publication/pub.1040741281

[4] N. Pattison, N. A. S. Humphreys, and T. Walsh, "Exploring obstacles to critical care trials in the uk: A qualitative investigation," *Journal of the Intensive Care Society*, vol. 18, no. 1, pp. 36–46, 2017.

[5] H.-U. Prokosch and T. Ganslandt, "Perspectives for medical informatics. reusing the electronic medical record for clinical research." *Methods of information in medicine*, vol. 48 1, pp. 38–44, 2009.

[6] P. M. Rothwell, "Commentary: External validity of results of randomized trials: disentangling a complex concept," *International Journal of Epidemiology*, vol. 39, no. 1, pp. 94–96, 2010. [Online]. Available: http://dx.doi.org/10.1093/ije/dyp305

[7] C. Rowlands, L. Rooshenas, K. Fairhurst, J. Rees, C. Gamble, and J. M. Blazeby, "Detailed systematic analysis of recruitment strategies in randomised controlled trials in patients with an unscheduled admission to hospital," *BMJ Open*, vol. 8, no. 2, 2018. [Online]. Available: https://bmjopen.bmj.com/content/8/2/e018581

[8] E. Gray, S. Norris, S. Schmitz, and A. OLeary, "Do disparities between populations in randomized controlled trials and the real world lead

to differences in outcomes?" *Journal of Comparative Effectiveness Research*, vol. 6, no. 1, pp. 65–82, 2017.

[9] Y. Ni, J. Wright, J. Perentesis, T. Lingren, L. Deléger, M. Kaiser, I. S. Kohane, and I. Solti, "Increasing the efficiency of trial-patient matching: automated clinical trial eligibility pre-screening for pediatric oncology patients," *BMC Med. Inf. & Decision Making*, vol. 15, p. 28, 2015. [Online]. Available: https://doi.org/10.1186/s12911-015-0149-3

[10] N. Sager, M. Lyman, C. Bucknall, N. Nhan, and L. J. Tick, "Natural language processing and the representation of clinical data," *Journal of the American Medical Informatics Association*, vol. 1, no. 2, Mar/Apr 1994.

[11] S. R. Jonnalagadda, A. K. Adupa, R. P. Garg, J. Corona-Cox, and S. J. Shah, "Text mining of the electronic health record: An information extraction approach for automated identification and subphenotyping of hfpef patients for clinical trials," *Journal of Cardiovascular Translational Research*, vol. 10, no. 3, pp. 313–321, Jun 2017. [Online]. Available: https://doi.org/10.1007/s12265-017-9752-2

[12] L. Li, H. S. Chase, C. O. Patel, C. Friedman, and C. Weng, "Comparing ICD9-encoded diagnoses and NLP-processed discharge summaries for clinical trials pre-screening: a case study," *AMIA Annu. Symp. Proc.*, pp. 404–408, Nov. 2008. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/18999285

[13] Y. Ni, S. Kennebeck, J. W. Dexheimer, C. M. McAneney, H. Tang, T. Lingren, Q. Li, H. Zhai, and I. Solti, "Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department," *J. Am. Med. Inform. Assoc.*, vol. 22, no. 1, pp. 166–178, Jan. 2015. [Online]. Available: http://dx.doi.org/10.1136/amiajnl-2014-002887

[14] S. V. Pakhomov, J. Buntrock, and C. G. Chute, "Prospective recruitment of patients with congestive heart failure using an ad-hoc binary classifier," *J. Biomed. Inform.*, vol. 38, no. 2, pp. 145–153, Apr. 2005. [Online]. Available: http://dx.doi.org/10.1016/j.jbi.2004.11.016

[15] A. C. Gordon, G. D. Perkins, M. Singer, D. F. McAuley, R. M. Orme, S. Santhakumaran, A. J. Mason, M. Cross, F. Al-Beidh, J. Best-Lane, D. Brealey, C. L. Nutt, J. J. McNamee, H. Reschreiter, A. Breen, K. D. Liu, and D. Ashby, "Levosimendan for the prevention of acute organ dysfunction in sepsis," *New England Journal of Medicine*, vol. 375, no. 17, pp. 1638–1648, 2016, pMID: 27705084.

[16] S. Harris, S. Shi, D. Brealey, N. S. MacCallum, S. Denaxas, D. Perez-Suarez, A. Ercole, P. Watkinson, A. Jones, S. Ashworth, R. Beale, D. Young, S. Brett, and M. Singer, "Critical care health informatics collaborative (cchic): Data, tools and methods for reproducible research: A multi-centre uk intensive care database," *International Journal of Medical Informatics*, vol. 112, no. 1, pp. 82–89, 2018.

[17] R. Jackson, I. Kartoglu, C. Stringer, G. Gorrell, A. Roberts, X. Song, H. Wu, A. Agrawal, K. Lui, T. Groza, D. Lewsley, D. Northwood, A. Folarin, R. Stewart, and R. Dobson, "Cogstack - experiences of deploying integrated information retrieval and extraction services in a large national health service foundation trust hospital," *BMC Medical Informatics and Decision Making*, vol. 18, no. 1, p. 47, Jun 2018. [Online]. Available: https://doi.org/10.1186/s12911-018-0623-9

[18] "Open Source Search: Elastic," https://www.elastic.co/, accessed: 27 Jan 2020. [Online]. Available: https://www.elastic.co/

[19] G. Gorrell, X. Song, and A. Roberts, "Bio-yodie: A named entity linking system for biomedical text," *CoRR*, vol. abs/1811.04860, 2018. [Online]. Available: http://arxiv.org/abs/1811.04860

[20] National Library of Medicine (U.S.), *UMLS Knowledge Sources: Metathesaurus, Semantic Network, [and] SPECIALIST Lexicon*. U.S. Department of Health and Human Services, National Institutes of Health, National Library of Medicine, 2003. [Online]. Available: https://books.google.co.uk/books?id=xTtrAAAAMAAJ

[21] H. Wu, G. Toti, K. I. Morley, Z. M. Ibrahim, A. Folarin, R. Jackson, I. Kartoglu, A. Agrawal, C. Stringer, D. Gale, G. Gorrell, A. Roberts, M. Broadbent, R. Stewart, and R. J. Dobson, "Semehr: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research*," *Journal of the American Medical Informatics Association*, vol. 25, no. 5, pp. 530–537, 2018. [Online]. Available: http://dx.doi.org/10.1093/jamia/ocx160

[22] H. Harkema, J. N. Dowling, T. Thornblade, and W. W. Chapman, "ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports," *J. Biomed. Inform.*, vol. 42, no. 5, pp. 839–851, Oct. 2009. [Online]. Available: http://dx.doi.org/10.1016/j.jbi.2009.05.002

[23] M. M. Levy, M. P. Fink, J. C. Marshall, E. Abraham, D. Angus, D. Cook, J. Cohen, S. M. Opal, J.-L. Vincent, G. Ramsay, and for the International Sepsis Definitions Conference, "2001 sccm/esicm/accp/ats/sis international sepsis definitions conference," *Intensive Care Medicine*, vol. 29, no. 4, pp. 530–538, Apr 2003. [Online]. Available: https://doi.org/10.1007/s00134-003-1662-x

[24] R. N. Pugh, I. M. Murray-Lyon, J. L. Dawson, M. C. Pietroni, and R. Williams, "Transection of the oesophagus for bleeding oesophageal varices," *Br. J. Surg.*, vol. 60, no. 8, pp. 646–649, Aug. 1973. [Online]. Available: http://dx.doi.org/10.1002/bjs.1800600817

[25] P. M. Spieth, A. S. Kubasch, A. I. Penzlin, B. M.-W. Illigens, K. Barlinn, and T. Siepmann, "Randomized controlled trials: a matter of design," in *Neuropsychiatric disease and treatment*, 2016.

[26] T. P. van Staa, L. Dyson, G. McCann, S. Padmanabhan, R. Belatri, B. Goldacre, J. Cassell, M. Pirmohamed, D. Torgerson, S. Ronaldson, J. Adamson, A. Taweel, B. Delaney, S. Mahmood, S. Baracaia, T. Round, R. Fox, T. Hunter, M. Gulliford, and L. Smeeth, "The opportunities and challenges of pragmatic point-of-care randomised trials using routinely collected electronic records: evaluations of two exemplar trials," *Health Technol Assess*, vol. 18, no. 43, pp. 1–146, 2014.

[27] M. C. Gulliford, T. P. van Staa, L. McDermott, G. McCann, J. Charlton, and A. Dregan, "Cluster randomized trials utilizing primary care electronic health records: methodological issues in design, conduct, and analysis (ecrt study)," *Trials*, vol. 15, p. 220, 2014.

[28] S. James, S. V. Rao, and C. B. Granger, "Registry-based randomized clinical trials—a new clinical trial paradigm," *Nature Reviews Cardiology*, vol. 12, pp. 312 EP –, 03 2015. [Online]. Available: https://doi.org/10.1038/nrcardio.2015.33

[29] K. Loudon, S. Treweek, F. Sullivan, P. Donnan, K. E. Thorpe, and M. Zwarenstein, "The precis-2 tool: designing trials that are fit for purpose," *BMJ*, vol. 350, 2015. [Online]. Available: https://www.bmj.com/content/350/bmj.h2147

[30] C. G. Walsh and K. B. Johnson, "Observational cohort studies and the challenges of in silico experiments," *JAMA Oncology*, vol. 3, no. 1, pp. 55–57, 2017. [Online]. Available: +http://dx.doi.org/10.1001/jamaoncol.2016.3478

[31] J. Kremidas, "Recruitment roles," *Applied Clinical Trials*, vol. 20, no. 9, pp. 32–33, Sep 2011.

[32] S. J. Walters, I. Bonacho dos Anjos Henriques-Cadby, O. Bortolami, L. Flight, D. Hind, R. M. Jacques, C. Knox, B. Nadin, J. Rothwell, M. Surtees, and S. A. Julious, "Recruitment and retention of participants in randomised controlled trials: a review of trials funded and published by the united kingdom health technology assessment programme," *BMJ Open*, vol. 7, no. 3, 2017. [Online]. Available: https://bmjopen.bmj.com/content/7/3/e015276

[33] S. W. Tu, M. Peleg, S. Carini, M. Bobak, J. Ross, D. Rubin, and I. Sim, "A practical method for transforming free-text eligibility criteria into computable criteria," *J. Biomed. Inform.*, vol. 44, no. 2, pp. 239–250, Apr. 2011. [Online]. Available: http://dx.doi.org/10.1016/j.jbi.2010.09.007

[34] C. Patel, J. Cimino, J. Dolby, A. Fokoue, A. Kalyanpur, A. Kershenbaum, L. Ma, E. Schonberg, and K. Srinivas, "Matching patient records to clinical trials using ontologies," in *The Semantic Web*, vol. 4825. Springer, Berlin, Heidelberg, 01 2007, pp. 816–829.

[35] B. Olasov and I. Sim, "RuleEd, a web-based semantic network interface for constructing and revising computable eligibility rules," *AMIA Annu. Symp. Proc.*, p. 1051, 2006. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/17238670

[36] D. W. Lonsdale, C. Tustison, C. G. Parker, and D. W. Embley, "Assessing clinical trial eligibility with logic expression queries," *Data Knowl. Eng.*, vol. 66, no. 1, pp. 3–17, Jul. 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0169023X07001528

[37] C. O. Patel and C. Weng, "ECRL: an eligibility criteria representation language based on the UMLS Semantic Network," *AMIA Annu. Symp. Proc.*, p. 1084, Nov. 2008. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/18999200

[38] H. Tissot and R. Dobson, "Combining string and phonetic similarity matching to identify misspelt names of drugs in medical records written in portuguese," *Journal of Biomedical Semantics*, vol. 10, no. 1, p. 17, Nov 2019. [Online]. Available: https://doi.org/10.1186/s13326-019-0216-2

[39] H. Tissot, A. Roberts, L. Derczynski, G. Gorrell, and M. Didonet Del Fabro, "Analysis of temporal expressions annotated in clinical notes," in *Proceedings of 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*. London, UK: ACL, 2015, pp. 93–102.

[40] H. Tissot, M. D. Del Fabro, L. Derczynski, and A. Roberts, "Normalisation of imprecise temporal expressions extracted from text," *Knowledge and Information Systems*, Feb 2019. [Online]. Available: https://doi.org/10.1007/s10115-019-01338-1

[41] N. Viani, H. Tissot, A. Bernardino, and S. Velupillai, "Annotating temporal information in clinical notes for timeline reconstruction: Towards the definition of calendar expressions," in *Proceedings of the 18th BioNLP Workshop and Shared Task*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 201–210. [Online]. Available: https://www.aclweb.org/anthology/W19-5021

[42] F. W. Asselbergs, F. L. Visseren, M. L. Bots, G. J. de Borst, M. P. Buijsrogge, J. M. Dieleman, B. G. van Dinther, P. A. Doevendans, I. E. Hoefer, M. Hollander, P. A. de Jong, S. V. Koenen, G. Pasterkamp, Y. M. Ruigrok, Y. T. van der Schouw, M. C. Verhaar, and D. E. Grobbee, "Uniform data collection in routine clinical practice in cardiovascular patients for optimal care, quality control and research: The utrecht cardiovascular cohort," *European Journal of Preventive Cardiology*, vol. 24, no. 8, pp. 840–847, 2017, pMID: 28128643. [Online]. Available: https://doi.org/10.1177/2047487317690284

[43] K. Kreimeyer, M. Foster, A. Pandey, N. Arya, G. Halford, S. F. Jones, R. Forshee, M. Walderhaug, and T. Botsis, "Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review," *Journal of biomedical informatics*, vol. 73, pp. 14–29, 2017.

[44] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, "Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications," *Journal of the American Medical Informatics Association*, 2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/20819853

[45] A. D. Shah, E. Bailey, T. Williams, S. Denaxas, R. Dobson, and H. Hemingway, "Natural language processing for disease phenotyping in uk primary care records for research: a pilot study in myocardial infarction and death," *Journal of Biomedical Semantics*, vol. 10, 2019. [Online]. Available: https://doi.org/10.1186/s13326-019-0214-4

[46] S. Gehrmann, F. Dernoncourt, Y. Li, E. T. Carlson, J. T. Wu, J. Welt, J. Foote, John, E. T. Moseley, D. W. Grant, P. D. Tyler, and L. A. Celi, "Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives," *PLoS One*, vol. 13, p. e0192360, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29447188

[47] Z. Wang, A. D. Shah, A. R. Tate, S. Denaxas, J. Shawe-Taylor, and H. Hemingway, "Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning." *PLoS One*, vol. 7, p. e30412, 2012.

[48] N. Viani, C. Larizza, V. Tibollo, C. Napolitano, S. G. Priori, R. Bellazzi, and L. Sacchi, "Information extraction from italian medical reports: An ontology-driven approach," *International Journal of Medical Informatics*, vol. 111, pp. 140–148, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1386505617304586

[49] H. Cunningham, V. Tablan, A. Roberts, and K. Bontcheva, "Getting more out of biomedical documents with gate's full lifecycle open source text analytics," *PLOS Computational Biology*, 2013. [Online]. Available: http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1002854

[50] M. S. Lauer and R. B. D'Agostino, "The randomized registry trial: The next disruptive technology in clinical research?" *New England Journal of Medicine*, vol. 369, no. 17, pp. 1579–1581, 2013, pMID: 23991657. [Online]. Available: https://doi.org/10.1056/NEJMp1310102

[51] L. Xiao, N. Lv, L. G. Rosas, D. Au, and J. Ma, "Validation of clinic weights from electronic health records against standardized weight measurements in weight loss trials," *Obesity*, vol. 25, no. 2, pp. 363–369, 2017. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/oby.21737