Published in final edited form as:

Methods Inf Med. 2016 May 17; 55(3): 266–275. doi:10.3414/ME15-01-0112.

Valx: A system for extracting and structuring numeric lab test comparison statements from text

Tianyong Hao, PhD^{a,b}, Hongfang Liu, PhD^c, and Chunhua Weng, PhD^{a,*}

^aDepartment of Biomedical Informatics, Columbia University, New York, NY, United States

^bKey Lab of Language Engineering and Computing of Guangdong Province, Guangdong University of Foreign Studies, Guangzhou, China

^cDepartment of Health Sciences Research, Rochester, MN, United States

Abstract

Objectives—To develop an automated method for extracting and structuring numeric lab test comparison statements from text and evaluate the method using clinical trial eligibility criteria text.

Methods—Leveraging semantic knowledge from the Unified Medical Language System (UMLS) and domain knowledge acquired from the Internet, Valx takes 7 steps to extract and normalize numeric lab test expressions: 1) text preprocessing, 2) numeric, unit, and comparison operator extraction, 3) variable identification using hybrid knowledge, 4) variable - numeric association, 5) context-based association filtering, 6) measurement unit normalization, and 7) heuristic rule-based comparison statements verification. Our reference standard was the consensus-based annotation among three raters for all comparison statements for two variables, i.e., HbA1c and glucose, identified from all of Type 1 and Type 2 diabetes trials in ClinicalTrials.gov.

Results—The precision, recall, and F-measure for structuring HbA1c comparison statements were 99.6%, 98.1%, 98.8% for Type 1 diabetes trials, and 98.8%, 96.9%, 97.8% for Type 2 Diabetes trials, respectively. The precision, recall, and F-measure for structuring glucose comparison statements were 97.3%, 94.8%, 96.1% for Type 1 diabetes trials, and 92.3%, 92.3%, 92.3% for Type 2 diabetes trials, respectively.

Conclusions—Valx is effective at extracting and structuring free-text lab test comparison statements in clinical trial summaries. Future studies are warranted to test its generalizability beyond eligibility criteria text. The open-source Valx enables its further evaluation and continued improvement among the collaborative scientific community.

Keywords

Medical Informatics; Natural Language Processing; Patient Selection; Clinical Trial; Comparison Statement

*Corresponding Author: Chunhua Weng, Ph.D., Department of Biomedical Informatics, Columbia University, New York City, 622 W 168th Street, PH-20, New York, NY 10032, chunhua@columbia.edu.

1. Introduction

Both the application of a clinical practice guideline on a patient and the recruitment of a research volunteer into a clinical study need to first assess if the patient or the volunteer meets the clinical care or research eligibility criteria, which exist largely as free text in clinical practice guidelines or clinical trial protocols [1–6]. Anecdotally over 40% of free-text eligibility criteria contain numeric comparison statements, e.g., “*HbA1c superior or equal to 7.5%*” and “*age eligibility for study: 18 years and older.*” Comparison statements are important for assessing how quantifiable clinical phenotypes such as lab tests, clinical attributes, or demographic variables are compared to specified thresholds or value ranges according to particular measures. Automatic extraction and normalization of such comparison statements are indispensable for enabling efficient electronic eligibility determination for executing computerized clinical practice guidelines [7], electronically screening patients for clinical trial studies [8, 9], and automating systematic reviews or meta-analysis of clinical trial evidence in published studies at scale [10, 11].

Extraction of a numeric comparison statement involves the extraction of four information constructs, i.e., numeric variable, comparison operator, threshold or value range, and measurement unit. This task faces several challenges. First, multiple numeric comparison statements can co-exist in the same sentence (e.g., “QRS $\geq$ 120 ms and LVEF $\leq$ 35%” in clinical trial NCT02018029 in [ClinicalTrials.gov](https://clinicaltrials.gov)) and hence accurate co-reference resolution is necessary to identify correct associations among the information constructs. Second, many statements contain implicit or incomplete specification such as measurement units, making normalization difficult. In addition, there exist heterogeneous representations of logic operators including symbols or English phrases, with or without negation, and sometimes involving implicit contextual information. For example, in “Transaminases $>$ 1.5X ULN” in NCT00365794, it is unclear what “ULN” stands for without contextual knowledge. Moreover, some sentences contain typos and logical errors that should be corrected before parsing. For example, “*HbA1c = 7.5% and = 10%*” in NCT00117780 has an intention of “*HbA1c $\geq$ 7.5% and $\leq$ 10%*” but omits “ $\geq$ ” and “ $\leq$ ”.

Methods for extracting numeric comparison statements are available but have various limitations. Tu et al. [12] developed a rule-based method for numeric expression extraction and translated seven numeric types into Arabic numeric. However, this method did not support automatic variable identification and unit extraction. Lonsdale et al. [13] extracted the logical expressions in clinical trial eligibility criteria text. In their method, a link grammar parser first labeled word dependencies, which were then used to extract semantics using a rule-based parser. The evaluation on 12 clinical trials achieved precision values of 70% and 65% for arguments and recall values of 61% and 63% for predicates, respectively, where argument and predicate are logic representational constructs of criteria [13]. Tu et al. also developed a method called ERGO for transforming free-text eligibility criteria into computable formats [14]. Using manual annotations of atomic statements from clinical trial summaries as the reference standard, ERGO obtained a recall of 67%, including 47.8% of exact matches and 19.5% of partial matches. This performance may not be adequate for reliably extracting numeric comparison statements from eligibility criteria. Demen et al. developed PASTEL to enable clinicians to manually create eligibility criteria with the

support of an ontology [15], but the method was labor-intensive and costly for formalizing the large number of existing trials. Murata et al. [16] developed a text mining system to extract numerical and named entities, such as person, location, and organization from newspapers. The highest and average accuracy were 85% and 48%, respectively. Due to the limited performance of the above methods for this task, we identify a real-world need for a practical method that can extract numeric comparison statements automatically from clinical eligibility criteria text at scale.

In order to reliably automate numeric comparison statement extraction, in this paper, we present a knowledge-based automatic system - Valx for extracting and normalizing numeric lab test comparison statements in clinical trial eligibility criteria texts. We evaluated Valx against the clinical trial summaries by manually creating a reference standard for example variables HbA1c and glucose in all Type 1 and Type 2 diabetes trials in ClinicalTrials.gov [17]. In the evaluation Valx achieved high precision and recall, both above 96% for HbA1c and glucose. We also evaluated Valx's performance on other variables such as body mass index (BMI) using a reference standard created for 50 randomly selected clinical trials and achieved equivalent performance. We further provided a proof-of-concept demonstration of an application enabled by Valx for aggregate analysis of all the included Type 1 and Type 2 diabetes clinical trials in ClinicalTrials.gov.

2. Material and methods

A complete numeric comparison statement includes four consecutive constructs: (1) a lab test variable; (2) a comparison logic operator; (3) a threshold or a value range; and (4) a measurement unit. For example, as for text "Body mass index must between 20–40kg/m", the extracted numeric comparison statements are (a) "BMI, greater equal, 20, kg/m" and (b) "BMI, lower equal, 40, kg/m", where BMI is a lab test variable, "greater equal" and "lower equal" (corresponding to 'between') are two logic operators, "20" and "40" are value thresholds, and "kg/m" is a measurement unit.

Each numeric comparison statement is represented using Extensible Markup Language (XML), which follows the design initiative of a well-adopted temporal markup language TimeML [3, 18, 19]. TimeML is a robust specification language for representing temporal expressions in XML format [20]. For example, the expression "2 months" is annotated as "<TIMEX3 tid='t1' type='Duration' value='P2M'>2 months</TIMEX3>" using one of TimeML versions - TimeX3.

Our representation, different from TimeML, contains two types of markup labels for representing variables and comparison statements. A variable includes three constructs, i.e., identified term, mapped concept, and mapping reference, and hence is represented in the format of "<VL Label=[*concept*] Source=[*reference for identifying the concept*]>[*identified terms*]</VL>". Similarly, a statement can be represented as (numeric, logic operator, unit) in the format of "<VML Logic=[*the relation of a variable to the numeric*] Unit=[*the unit of the numeric*]>[*a numeric*]</VML>". We therefore can associate the variable and the statement to obtain a complete annotation using our representations. For instance, text "BMI value must between 20–40kg/m" is annotated as "<VL Label = 'BMI' Source = 'DK'>BMI

combined to recognize composite measurement units, e.g., “mg/dl”. Rules are defined to detect unknown or incomplete units. For example, supposing the unit “kg / m²” is unknown, it can be identified through a connected identified numeric using the rule “[*numeric*] [*string*]/[*string*]”, “[*atomic unit*]/[*string*]”, or “[*string*]/[*atomic unit*]”. The rules are also designed to extend a unit into a more complete representation. For instance, the unit “mg” is extended to “mg per day”. Furthermore, rules are created to complete missing measurement units according to their context since a missing unit is a common problem (Appendix 3). For example, in “*fasting blood glucose between 70 and 250 mg/dl*” the missing unit “mg/dl” is added after “70” using the rules. When there is no contextual information, a variable-specific default unit as available is inserted after variable identification and association.

The identification of comparison operator representations (e.g., “not higher than or equal to”) utilizes pre-defined operator features, negation, context, and rules (Appendix 4). Many expressions are highly contextual. For example, a commonly used representation “HbA1c value between 6.5–10%” contains no comparison operators but clearly indicates “>= 6.5% and <=10%”. Some comparison operators need transformation, e.g., “HbA1c = 7% at most” should be transformed as “HbA1c <= 7%”. Afterwards, comparison operators are combined logically. For example, “=” and “>=” in “HbA1c = or >= 7%” are combined as “>=”. We also define the relations among the comparison operator, thus a comparison operator can be easily converted to its opposite. For example, “> 7%” in exclusion criteria text can be easily converted to “<= 7%” as an inclusive criterion since “larger than” is the opposite of “lower than and equal to”.

Step 3: Variable identification using hybrid knowledge—Our variable extraction uses hybrid information, including contextual knowledge, domain knowledge, the Unified Medical Language System (UMLS) Metathesaurus [21], and n-Gram co-occurrence information, represented as Context, DK, UMLS, and n-Gram, respectively. The purpose of combining different knowledge is to detect both known and unknown variables.

We first use contextual information to detect unknown variables in certain formats. For example, according to a typical format “<Comparison operator> [Variable] <Comparison operator>”, “IBW” can be detected from “18.5 <= IBW < 25”. We then apply domain knowledge (DK), which defines frequently used variable representations, legitimate units, illegitimate units, preferred unit for normalization, maximum value range, and value range reference source. For example, for the variable “HbA1c”, a knowledge item includes different representations as “*glycated hemoglobin, glycated hemoglobin, glycosylated hemoglobin, hemoglobin A1c, HemoglobinA1c, hgbA1c, hga1c, hba1c, ha1c, balc, a1c, hb1c*”, the legitimate units as “%, percent”, the illegitimate units as “all the other units”, the preferred unit as “%”, and the maximum value range as [4, 12].

The UMLS is used to recognize and normalize concepts with different semantic representations. It potentially helps find the same variable in different representations, identifies variables from texts, and normalizes variables with a more standard representation. However, UMLS is only integrated into offline version of Valx due to the large size of UMLS and the requirement of real-time processing efficiency for online version. As Valx is

designed for extracting lab test comparison statements, users can add the concept representations of target variables into DK to enhance Valx's ability to recognize the target terms. Particularly, users can add additional representations beyond UMLS, e.g., commonly seen typos of the variables, e.g., "hb1c" for variable HbA1c.

Then we generate n-gram strings using all identified comparison statements as separators. N-grams frequently co-occurring with numeric (i.e., higher than a threshold, 10, in this paper) are identified. These n-grams are further matched with DK and the UMLS after removing stop words and ordering by length. An n-gram is normalized as an eligibility variable once a match is detected. In case no variable is identified through contextual knowledge, DK and UMLS matching, the longest n-gram in front of a numeric value is selected as a variable. For instance, "cancer other than basal cell skin cancer" is selected as a variable in "cancer other than basal cell skin cancer within 5 years" in trial NCT00784511.

Step 4: Variable - Numeric Association—We firstly detect certain pre-defined structures (Appendix 5) and associate the related variables and values by referencing the structures. For example, a typical structure is "<value 1> (and/to/-) <value 2> (of/for) <variable>". Sentence "45 and 70 years of age" matches with the structure. Accordingly, the value expressions "45 years" and "70 years" are associated with the variable "age" and the association is formalized as "[age, greater or equal than, 45, years]" and "[age, lower or equal than, 70, years]".

The associations between variable and values are far more complex than just strict structures. Typical cases in eligibility criteria text are summarized as follows: 1) explicit connectors, e.g., "of", "between ... and...", "from... to...", etc., 2) connected logic operators, e.g., "hbA1c less than or equal to 8.5%" in NCT00442767, 3) disconnected logic operators, e.g., "HbA1c (glycosylated hemoglobin A1c) within the range of 7.5% to 10.0%" in NCT00669864, and 4) implicit logic operators, e.g., "hemoglobin a1c 5–10%" in NCT01653210. Cases 2), 3) and 4) do not contain connectors and case 4) does not contain both connectors and logic operators in text. Particularly, a sentence may contain more than one variable, e.g., "fpg 80, 140 mg/dl and glycated hemoglobin (a1c) > 6.5% and < or = 10.0%" in NCT00747006.

Valx utilizes both structural and distributional information for co-reference resolution, i.e., association of the detected variables with numeric comparison statement. Structure-based method is firstly used to detect certain pre-defined structures (Appendix 5) and associate the related variables and values. The structure-based method includes the processing of associations to switch the order of variables and values in case that the values occur before variables. On this basis, we use word sequence to associate variables and values within different sub-strings. As to the example "FPG at screening visit >= 100 mg/dl or A1C >= 5.8%", two substrings "FPG ... >= 100 mg/dl" and "A1C >= 5.8%" are extracted. For each substring, the variables and values are checked with the sequence "<variable>...<value 1>...<value n>" and all the values from 1 to n were associated with the variable even if they might not be adjacent to the variable.

2.2 Phase 2 – Association verification and filtering

Step 5: Context-based Filtering—Candidate associations are filtered with two types of exceptions using contextual information at sentence level: 1) variable-specified unit matching exceptions, and 2) comparison operator matching exceptions. All the associated expressions are first checked by comparing their units either in front of or behind the expressions with the variable-specified units defined in DK. An association is considered an exception if the associated unit exists in the illegitimate unit list of the variable. For example, “kg” is an illegitimate unit for the variable “HbA1c”, so the expressions containing “HbA1c” with unit “kg” are removed. We then detect surrounding contextual information to identify pre-defined special operators, such as “+/-”, “|±”, or “+ -” (Appendix 6). All numeric expressions modified directly by a special operator are identified as non-comparison statements and kept as they are. In the example “hemoglobin a1c (HbA1c) of >= 7% and < 10% (+/- 0.1%)”, “+/- 0.1%” is recognized as non-comparison statements and preserved as a semantic unit without parsing.

Step 6: Measurement Unit Normalization—To normalize measurement units, we firstly complete empty units (missing unit mentions) using the closest units detected within the same sentences. For example, the sentence “HbA1c greater than 7.0 and less than 12.0%” misses a unit for “7.0”. Valx automatically detects the missing unit after “7.0” and thus adds “%” based on the identification of “12.0%”. In case there is no contextual information, the preferred unit of the associated variable defined in DK is assigned. We further detect unit format errors and rectify them based on the variable-specified legitimate units defined in DK. For instance, “mg/dl”, “g/l”, “mmol/l”, and “millimoles/liter” are all legitimate units for variable “glucose”. According to this lookup list, unit formatted as “mmol” is recognized as incorrect due to missing “/l” and is rectified to become “mmol/l”. Afterwards Valx convert all the units into a preferred unit defined in our DK and normalize the corresponding values accordingly. The conversion rules are from public resources on the Internet, e.g., [22]. For example, the expression “lower than 250 mg/dl” associated with the variable “Glucose” is mapped to “lower than 13.89 mmol/l” by mapping “mg/dl” to “mmol/l”. Valx needs pre-defined unit conversion rules to specify how the units can be converted in advance.

Step 7: Heuristic rule-based comparison statement verification—The normalized expressions still need further processing because of the following reasons: 1) values without any unit mentioning are originally represented in special units rather than commonly used units; 2) expressions are incorrectly associated with a variable due to either incomplete DK or the units being mistakenly recognized as a legitimate unit. For example, the normalized expression “<VL Label=BMI>stable BMI</VL><VML Logic=lower Unit=%> 5 </VML>” for criterion “stable BMI (varied less than 5%)” contains an incorrect association because our DK does not specify that “%” is not the legitimate measurement unit for BMI so that Valx mistakenly considers “%” as the measurement unit for variable BMI.

A list of heuristic rules is defined (Appendix 7) to verify and filter these incorrect statement associations based on value range enlargement. The purpose of the enlargement is to avoid the occasional removal of outliers, as outliers are commonly used for clinical trial. The

purpose of filtering based on range enlargement is to remove extremely unlikely associations beyond normal outliers. Two enlargement strategies are applied: (1) maximum value range-based extension, and (2) average value-based extension. When there is a specific maximum value range defined in DK, strategy (1) is used to enlarge the maximum value range to $[\min_value/threshold1, \max_value*threshold1]$ directly. For example, the enlarged range is [2, 24] for HbA1C supposing the defined maximum value range is [4, 12]. Once there is no pre-defined value range available, Valx calculates the average value associated with the same variable across statements. The average value-based range enlargement is calculated as $[average_value/threshold2, average_value*threshold2]$. The threshold1 is set as 2 and the threshold2 is set as 8 from our empirical experiments on randomly selected statement samples. Afterwards, the statements are further verified to judge if their values exceed the enlarged value ranges. For example, the maximum value range for “Glucose” variable is “between 4 to 10 mmol/l” according to [21]. The enlarged value range is [2, 20] by maximum value range-based extension strategy. An example of association “glucose” with “25 mmol/l” is thus filtered out as incorrect association.

2.3 Evaluation design

We used the eligibility criteria text from all 149,048 clinical trials on ClinicalTrials.gov [17] retrieved as of 07/15/2013. After excluding trials where the eligibility criteria section was missing or contained only the phrase “Please contact site for information”, 148,381 trials were retained. We identified 7,714 trials for diabetes, including 3,331 trials for Type 2 diabetes and 1,052 trials for Type 1 diabetes, by searching for “diabetes Type 1” or “diabetes Type 2” in the “condition” field in the online clinical trial search form in ClinicalTrials.gov. All eligibility criteria texts of these trial summaries were used to evaluate Valx. Inclusion and exclusion criteria were processed separately.

To generate a reference standard for evaluation, we selected two frequently used variables in both Type 1 and Type 2 diabetes trials [23], i.e., HbA1c and glucose. The two variables have a large number of statements for generating the reference standard for evaluation. Three informatics researchers manually annotated all numeric comparison statements sequentially. The first annotator did the initial annotation. The second annotator verified the annotation results. The third annotator reviewed and approved the annotations as the reference standard. From 7,714 trials, the first annotator labeled 2,053 trials containing HbA1c comparison statements and 709 trials containing glucose comparison statements. The second annotator corrected inaccurate annotations and found new statements (110 for HbA1c and 46 for glucose). The final reference standard dataset included 3,466 statements for HbA1c from 2,052 trials and 1,142 statements for glucose from 706 trials. We further separated the reference standard into two disease-specific sub datasets (Type 1 and Type 2), where the dataset for Type 1 contains 469 HbA1c and 52 Glucose statements and Type 2 contains 2,020 HbA1c and 385 Glucose statements. Each trial was further separated as inclusion and exclusion text. For example, the dataset for Type 2 diabetes trials contained 1,934 HbA1c comparison statements in inclusion criteria and 186 in exclusion criteria.

Performance was evaluated using precision, recall, and F1 score as standard evaluation measures for information extraction systems [24]. The precision was defined as the

proportion of correctly extracted statements by Valx among all the statements extracted by

Valx as $\frac{|True\ positives|}{|True\ positives| + |False\ positives|}$. The recall was the proportion of correctly extracted statements by Valx among all annotated statements in the reference standard as

$\frac{|True\ positives|}{|True\ positives| + |False\ negatives|}$. F1 score was calculated as $2 \times \frac{precision \times recall}{precision + recall}$. To determine if a comparison statement was correct, we compared it with the reference standard. A statement is correct if and only if all statement components are correct at the same time along with correct trial ID and text section, to which the statement belonged. Any mismatch of the components in a statement causes the statement to be regarded as incorrect.

3. Results

Valx is an open source tool as part of http://www.ohnlp.org/index.php/OHNLP_Tool_List. It can be used both offline and online for extracting numeric comparison statements for different variables. Fig. 2 shows a screenshot of the online Valx demo available at <http://columbiaelixer.appspot.com/valx>. A user can either input a specific clinical trial ID or paste a block of text to extract a list of variables and their associated statements, where the statements are presented in the order from “Inclusion” to “Exclusion” and each of them contains markup representation in four colors. For example, “bmi 25.0–39.9” is formalized as “[BMI] [greater than or equal to] [25.00] [kg/m2]” and “[BMI] [lower than or equal to] [39.9] [kg/m2]”, where yellow, orange, light blue, and blue colors are for variable, comparison operator, value, and unit, respectively. The parsing results can be downloaded in CSV format.

Table 1 shows the evaluation results using the HbA1c comparison statements in Type 2 diabetes dataset. Valx extracted 2079 statements (1895 for inclusion and 184 for exclusion). Compared with the reference standard, 2054 statements (1877 for inclusion and 177 for exclusion) were identified as correct, leading to an overall precision of 98.8%, a recall of 96.9%, and a F1 of 97.8%. Similarly, for Type 1 diabetes dataset, Valx achieved an overall precision of 99.6% with 99.7% precision for inclusion and 98.5% for exclusion. The overall recall for Type 1 dataset was 98.1%, slightly higher than that for Type 2 dataset. The F1 scores for both Type 2 and Type 1 diabetes datasets were higher than 97%. Similarly, for variable glucose, Valx correctly extracted 365 among 375 identified statements on diabetes Type 2 diabetes dataset, leading to a precision of 97.3%, a recall of 94.8%, and a F1 of 96.1%. The performance for Type 1 diabetes dataset included a precision of 96.7%, a recall of 94.5%, and a F1 of 95.6%. We also evaluated the performance of Valx on other lab test variables, such as body mass index (BMI). From the 3,331 diabetes Type 2 trials, Valx extracted a total of 1,643 statements. According to human verification on the extracted statements from 50 randomly selected clinical trials, Valx achieved a precision of 96.3% for the statement extraction for the BMI variable.

4. Discussion

4.1. Error analysis

The evaluation of Valx indicates its robust reliability for structuring free-text numeric comparison statements. We reviewed 102 incorrectly parsed statements for HbA1c or glucose, including 27 false positives and 75 false negatives. Seven major error categories were identified according to the differences of error characteristics. We further investigated the causes of errors and classified them into sub-categories. Fig. 3 shows the results.

Incorrect parsing of the semantic constructs causes 46.2% of all errors, which can be further broken up into 81% in parsing comparison operators, 4.8% in parsing measurement units, and 14.2% in parsing other representations. For example, “HbA1c = 7.5% and = 10%” in NCT00117780 was incorrectly parsed as “HbA1c = 7.5 %, HbA1c = 10 %”, but it should be “HbA1c=7.5% or HbA1c=10%”. Similarly, “increased A1c level of more than 2%” in NCT00728403 was incorrectly parsed as “HbA1c > 2 %”. Further evaluation on extracted statements for BMI variable is provided in Appendix 9, which also demonstrates the incorrect parsing of semantic constructs such as “BMI >=85th percentile.”

The second top category of error falls in incorrect interpretation of contextual information (16.5%), including the sub-categories: missing context for reference values (73.3%) and multiple value ranges associated with constraints (26.7%). Other top error types include the incorrect association of variables and values (13.2%), text parsing errors (9.9%), incorrect variable identification (7.7%), irregular text coding (3.3%), and incorrect numerical representations (3.3%). Appendix 8 lists the error types and examples.

The currently unavailable functionalities in Valx contribute to about 34% of all the errors. One limitation is the lack of identification of constraints for conditional value range identification. Some variable specifications may be associated with multiple conditional value ranges. For example, the sentence, “the patient who has been taking oral hypoglycemic agent since 3 months with HbA1c 6.5 to 9% at screening test or who is drug-naïve or stopped taking oral hypoglycemic agent more than 3 months with HbA1c 7 to 10% at screening test” in NCT01001611, defines two constraints for the two numeric comparison statements. Valx can recognize the comparison statements but it fails to associate the conditions associated with the statements, e.g., the condition “first visit” in “HbA1c <8.5% measured at first visit”. The second limitation is that Valx is only evaluated on clinical trial summaries. Future work is warranted to test Valx for structuring numeric comparison statements in other text.

The inaccuracy in clinical trial summaries is another significant contributor to parsing errors. Data quality issues such as incomplete specification and typos, e.g., the typo “egal” in “HbA1c superior or egal to 7.5%” (NCT01144728), have contributed to 65.9% of all errors. Our analysis indicates that many comparison statements are inaccurate or vague. A common type of vagueness is the use of a reference standard. For instance, in “glycosylated hemoglobin (hba1c) <1.15 times the upper limit of normal” (NCT00191282), Valx incorrectly identifies “times the upper limit of normal” as a unit but this “upper limit” depends on what “normal” is used by this protocol. The second type of inaccuracy is the

incorrect numeric representations, e.g., “7, 5” within “recent HbA1c determination (between 7, 5 and 10%)” (NCT01867502) should be “7.5”. The third type of inaccuracy is the inaccurate comparison operator representations, e.g., “7.0 % = HbA1c < 10.0 %” (NCT00154414) is incorrectly written as “=” but should be “<”. The fourth type of inaccuracy is the inconsistency between inclusion and exclusion criteria. For example, “HbA1c < 10.5%” (NCT00231530) in the inclusion section but “HbA1c of > 10.5%” in the exclusion section, leaving “= 10.5%” unclear to be included or excluded. Other difficulty for parsing is the use of complex threshold specification, e.g., “HbA1c ≤ 130% of upper limit of normal of local hospital lab” requires recognition of “normal of local hospital lab” as one entity, but Valx is not capable of processing it. As discovered by Bhattacharyaa et al., the eligibility criteria text in ClinicalTrials.gov are condensed with lesser information than full clinical trial protocols [25].

4.2. Application for Aggregate Analysis of Clinical Trials

Despite the aforementioned limitations, Valx enables potential applications for facilitating electronic screening of eligible patients for clinical research opportunities, guideline-based medical care, and scalable meta-analyses of study population characteristics for clinical trials. The effort required to find target candidates for participation in clinical trials is significant for modern medical research [8, 25]. Valx can enable automatic eligibility determination using criteria with numeric comparison statements. Moreover, to systematically aggregate study populations of multiple related clinical trials, it is necessary to extract quantitative study population characteristics from free text clinical trial eligibility criteria. Valx can assist with precise extraction and structuring of numeric comparison statements for the analysis of clinical study population characteristics and assessment of their population representativeness during early design. Valx has been successfully used to enable a visualization system called VITTA (<http://is.gd/VITTA>) for visualizing aggregated clinical trial study populations using all numeric expressions for all medical conditions in ClinicalTrials.gov [11] and for comparing clinical study populations to real-world patient populations using numeric traits [10].

To illustrate how it works for the purpose of scalable aggregate analysis across trials, here we use Valx to facilitate the analysis of the distributions of collective study populations of multiple clinical trials for two example variables, HbA1c and glucose. Fig. 4 shows the distributions of trials covering a list of value points, where red solid line is for Type 1 diabetes trials and blue dashed line for Type 2 diabetes trials. X-axis represents the allowable value range for HbA1c (Fig 4-A) or glucose (Fig 4-B). The special value points (e.g., 0% of HbA1c) are caused by incomplete specifications. Most of all Type 2 trials (91%) recruited patients whose HbA1c was 8, but 20% and 15% trials did not specify lower and upper boundaries, respectively. The portion of Type 1 trials lacking lower boundary was much higher, 50%. The distribution for glucose (Fig 4-B) showed 43% of trials lacked lower boundary for Type 2 trials, while 45% trials lacked upper boundary for Type 1 trials.

5. Conclusions

We presented a validated system Valx for automatically extracting and normalizing numeric comparison statements for quantifiable lab test variables from the eligibility criteria text of clinical trials. Our evaluation results demonstrated its high precision and recall for two example variables, HbA1c and glucose. More studies are warranted to test its wider application. The open source Valx has been made available through www.OHNLP.org to interested researchers for conducting additional evaluations and for collaborating on its continual improvement.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This project was supported by National Library of Medicine grant **R01LM009886** (PI: Weng), National Institute of General Medical Sciences grant **R01GM102282** (PI: Liu), and National Center for Advancing Translational Sciences grant **UL1TR000040** (PI: Ginsberg). The first author was later funded by National Natural Science Foundation of China grant **NO 61403088** (PI: Hao) to complete the manuscript.

References

1. Boland MR, Miotto R, Gao J, Weng C. Feasibility of feature-based indexing, clustering, and search of clinical trials. A case study of breast cancer trials from ClinicalTrials.gov. *Methods Inf Med.* 2013; 52(5):382–94. [PubMed: 23666475]
2. Miotto R, Jiang S, Weng C. eTACTS: a method for dynamically filtering clinical trial search results. *J Biomed Inform.* 2013; 46(6):1060–7. [PubMed: 23916863]
3. Tianyong Hao AR, Weng Chunhua. Extracting and Normalizing Temporal Expressions in Clinical Data Requests from Researchers. *Lecture Notes in Computer Science.* 2013; 8040:10.
4. Hao T, Rusanov A, Boland MR, Weng C. Clustering clinical trials with similar eligibility criteria features. *J Biomed Inform.* 2014; 52:112–20. [PubMed: 24496068]
5. Hao T, Weng C. Adaptive semantic tag mining from heterogeneous clinical research texts. *Methods Inf Med.* 2015; 54(2):164–70. [PubMed: 25327613]
6. Bache R, Taweel A, Miles S, Delaney BC. An eligibility criteria query language for heterogeneous data warehouses. *Methods Inf Med.* 2015; 54(1):41–4. [PubMed: 24985949]
7. Weng C, Payne PRO, Velez M, Johnson SB, Bakken S. Towards Symbiosis in Knowledge Representation and Natural Language Processing for Structuring Clinical Practice Guidelines. *Studies in Health Technology and Informatics.* 2013; 201:461–469.
8. Thadani SR, Weng CH, Bigger JT, Ennever JF, Wajngurt D. Electronic Screening Improves Efficiency in Clinical Trial Recruitment. *Journal of the American Medical Informatics Association.* 2009; 16(6):869–873. [PubMed: 19717797]
9. Miotto R, Weng C. Case-based reasoning using electronic health records efficiently identifies eligible patients for clinical trials. *Journal of the American Medical Informatics Association.* 2015 available online.
10. Weng C, Li Y, Ryan P, Zhang Y, Liu F, Gao J, Bigger JT, Hripcsak G. A Distribution-based Method for Assessing The Differences between Clinical Trial Target Populations and Patient Populations in Electronic Health Records. *Applied Clinical Informatics.* 2014; 5(2):463–479. [PubMed: 25024761]
11. He Z, Carini S, Sim I, Weng C. Visual aggregate analysis of eligibility features of clinical trials. *Journal of Biomedical Informatics.* 2015; 54(0):241–255. [PubMed: 25615940]

12. Tu M, Z Y, Zong CA. Universal Approach to Translating Numerical and Time Expressions. 9th International Workshop on Spoken Language Translation. 2012:209–216.
13. Lonsdale DW, Tustison C, Parker CG, Embley DW. Assessing clinical trial eligibility with logic expression queries. *Data & Knowledge Engineering*. 2008; 66(1):3–17.
14. Tu SW, Peleg M, Carini S, Bobak M, Ross J, Rubin D, Sim I. A practical method for transforming free-text eligibility criteria into computable criteria. *Journal of Biomedical Informatics*. 2011; 44(2):239–250. [PubMed: 20851207]
15. Damen D, L K, Hellebaut G, Bulcke TVD. PASTEL: A Semantic Platform for Assisted Clinical Trial Patient Recruitment. *International Conference on Healthcare Informatics*. 2013:269–276.
16. Murata M, S T, Torisawa K, Iwatate M, Ichii K, Ma Q, Kanamaru T. Sophisticated Text Mining System for Extracting and Visualizing Numerical and Named Entity Information from a Large Number of Documents. *NTCIR-7 Workshop Meeting*. 2008:555–562.
17. US National Institutes of Health. ClinicalTrials.gov. [cited 2014; Available from: <http://www.clinicaltrials.gov/>]
18. Pustejovsky, J., I, B., Sauri, R., Castano, J., Littman, J., Gaizauskas, R., et al. *The Language of Time: A Reader*. Oxford University Press; 2005. The Specification Language TimeML; p. 545-557.
19. Boguraev, B., A, RK. 19th international joint conference on Artificial intelligence. Morgan Kaufmann Publishers Inc.; 2005. TimeML - Compliant Text Analysis for Temporal Reasoning; p. 997-1003.
20. Pustejovsky, James, C, JM., Ingria, Robert, Sauri, Roser, Gaizauskas, Robert J., Setzer, Andrea, Katz, Graham, Radev, Dragomir R. *New Directions in Question Answering*. AAAI Press; 2003. TimeML: Robust Specification of Event and Temporal Expressions in Text; p. 28-34.
21. National Library of Medicine. Unified Medical Language System Glossary. [cited 2014; Available from: http://www.nlm.nih.gov/research/umls/new_users/glossary.html]
22. Units Conversion. [cited 2014; Available from: http://www.globalrph.com/conv_si.htm]
23. Gillett MJ. International Expert Committee report on the role of the A1c assay in the diagnosis of diabetes: *Diabetes Care* 2009; 32(7): 1327–1334. *Clin Biochem Rev*. 2009; 30(4):197–200. [PubMed: 20011212]
24. Manning, Christopher D., R, P., Schütze, Hinrich. *Introduction to information retrieval*. Cambridge University Press; 2009.
25. Parker CG, Embley DW. Generating medical logic modules for clinical trial eligibility criteria. *AMIA Annu Symp Proc*. 2003:964. [PubMed: 14728468]

Fig. 1.

The framework of Valx for extracting and normalizing numeric lab test comparison statements

Valx - Numerical Expression Extraction and Normalization

Extract and normalize quantifiable variables including variable name and values from eligibility criteria text into computer-interpretable representations. Input a trial ID or a eligibility criteria text for processing.

Trial ID: (e.g., NCT00784511). Set empty when use eligibility criteria text only.

Eligibility criteria text:

Inclusion Criteria:

- African-American by self designation
- Glucose intolerance defined as FPG $\geq$ 100 mg/dl or A1c $\geq$ 5.8%
- BMI 25.0-39.9
- Age 40 or older

Exclusion Criteria:

Medical Conditions

Variable option: ☒ Detect all variables

Identified variables: ☐ Age_and_Gender ☐ Glucose ☐ HBA1C ☐ BMI ☐ Age ☐ renal stone ☐ cancer other than basal cell skin cancer ☐ pregnancy ☐ menopause onset ☐ AST ☐ ALT ☐ estimated glomerular filtration rate ☐ Creatinine ratio ☐ abnormal SERUM CALCIUM ☐ Hematocrit ☐ consumption ☐ corresponds to a 24-hour urinary calcium excretion

[Click to view the trial on ClinicalTrials.gov](#)

Variable: Structured AGE & GENDER

Gender information: both

Age information: [Minimum:40 years, Maximum:]

Text section: Inclusion

Sentence: glucose intolerance defined as fpg $\geq$ 100 mg/dl or a1c $\geq$ 5.8%

Representation: glucose intolerance defined as <VL Label=Glucose Source=DK>fpg</VL> <VML Logic=greater_equal Unit=mg/dl>100</VML> or <VL Label=HBA1C Source=DK>a1c</VL> <VML Logic=greater_equal Unit=%>5.8</VML>

Normalized variables and values: **Glucose** greater than or equal to 5.56 mmol/l; **HBA1C** greater than or equal to 5.80 %;

Text section: Inclusion

Sentence: bmi 25.0-39.9

Representation: <VL Label=BMI Source=DK>bmi</VL> <VML Logic=greater_equal Unit=>25.0</VML> - <VML Logic=lower_equal Unit=>39.9</VML>

Normalized variables and values: **BMI** greater than or equal to 25.00 kg/m2; **BMI** lower than or equal to 39.90 kg/m2;

Text section: Inclusion

Sentence: age 40 or older

Representation: <VL Label=Age Source=DK>age</VL> <VML Logic=greater_equal Unit=>40</VML>

Normalized variables and values: **Age** greater than or equal to 40.00 years;

Text section: Exclusion

Sentence: diabetes potentially requiring pharmacotherapy, defined as a1c $>$ 7%

Representation: diabetes potentially requiring pharmacotherapy, defined as <VL Label=HBA1C Source=DK>a1c</VL> <VML Logic=greater Unit=%>7</VML>

Normalized variables and values: **HBA1C** greater than 7.00 %;

Fig. 2.

A screenshot of the graphical user interface demonstrating how the online Valx works

Fig. 3.
The summary of error analysis in top categories (middle chart) and sub-categories

Fig. 4.
A and B The percentile value distribution of all the diabetes Type 2 (blue dashed lines) and Type 1 (red solid lines) trials containing variable “HbA1c” (left) and “glucose” (right)

Table 1

The evaluation of Valx on Diabetes Type 2 and Type 1 diabetes trials using variable HbA1c compared with human-based reference standard dataset

Dataset	Text section	# by human	# by Valx	# Correct	Precision	Recall	F1
Diabetes Type 2	Inclusion	1934	1895	1877	99.1%	97.1%	98.0%
	Exclusion	186	184	177	96.2%	95.2%	95.7%
	Overall	2120	2079	2054	98.8%	96.9%	97.8%
Diabetes Type 1	Inclusion	403	397	396	99.7%	98.3%	99.0%
	Exclusion	66	65	64	98.5%	97.0%	97.7%
	Overall	469	462	460	99.6%	98.1%	98.8%
Both	Overall	2589	2541	2514	98.9%	97.1%	98.0%