

Journey to the Centre of the Protocol

Explorations in text mining for study eCRF development

Abstract

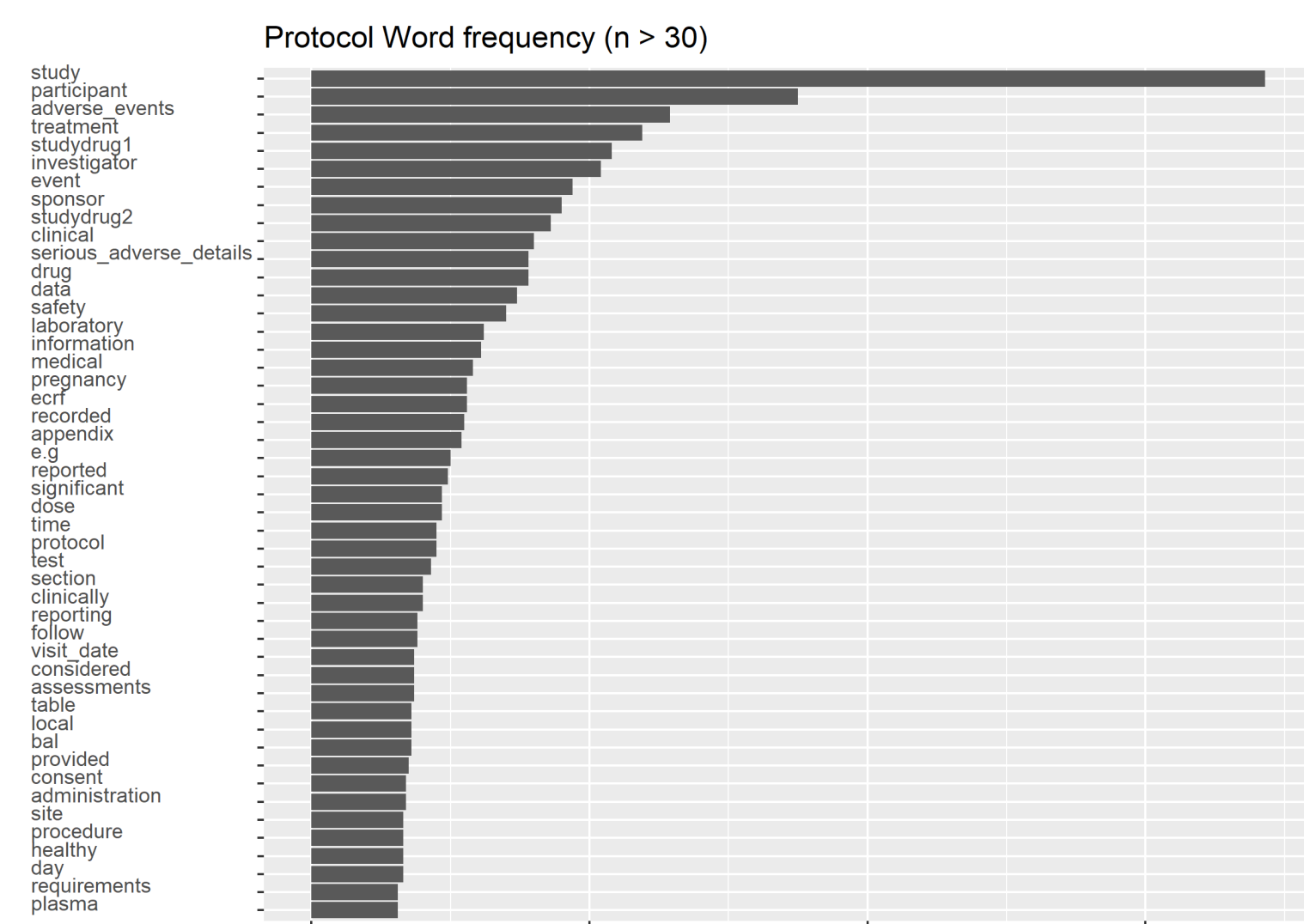
CRF development for clinical trial database build should be driven by the study protocol. Currently, a manual review of the protocol takes place to identify the data collection needs. Tasks include reviewing study assessments and procedures to translate them into the standard library equivalents. A well-defined set of data standards, protocol templates and review guidelines greatly expedites the process. However, this may be time consuming and lack consistency.

This poster explores the use of text mining techniques applied to a protocol, in regards to clinical data collection. The aim being to highlight which standard data collection forms are applicable and where possible, what datapoints are to be collected.

rodrick.bower@roche.com

Explore

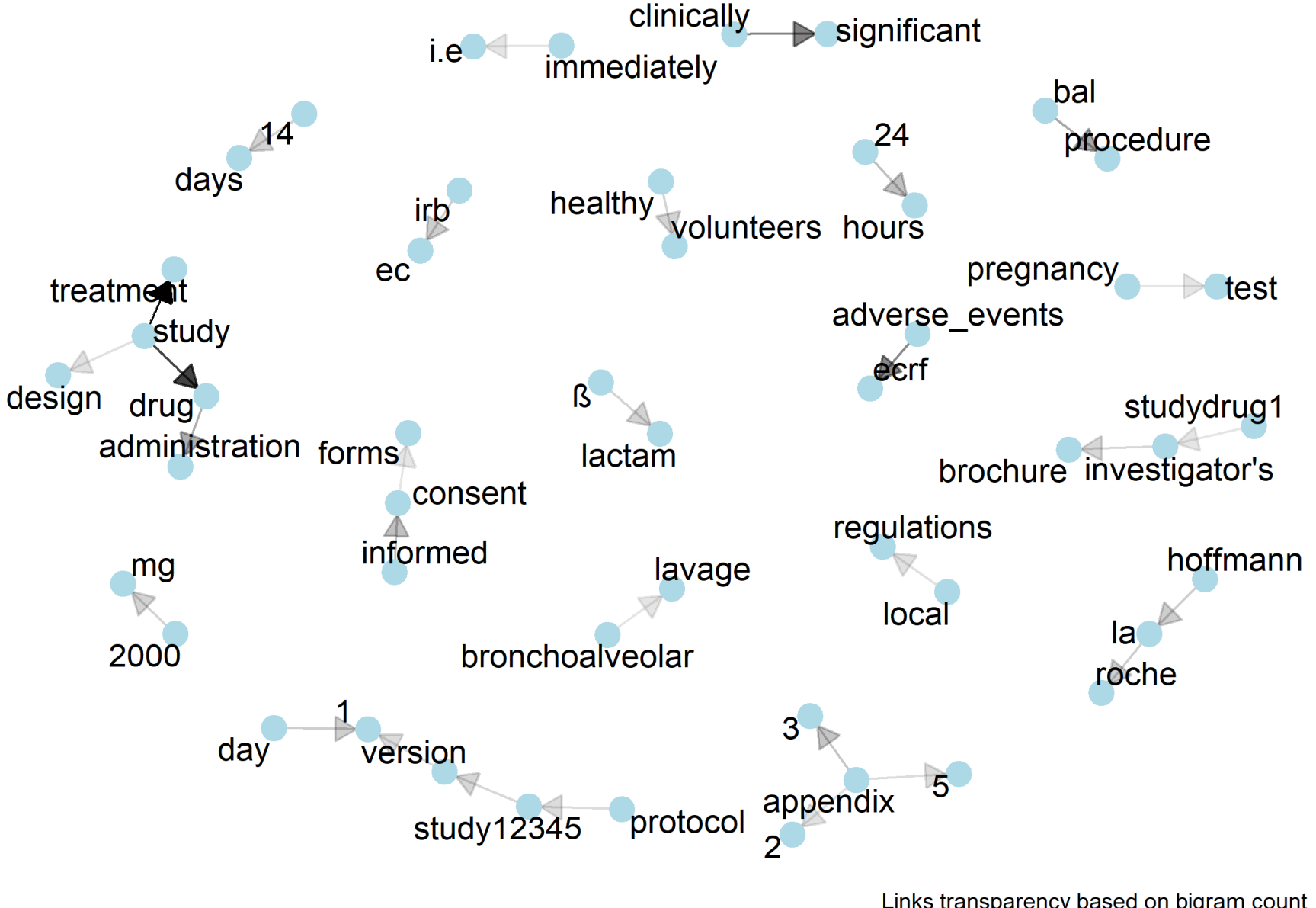
Term Frequency



- Important/meaningful words occur frequently
- Pre-process: Text separated into single words/terms
- Count and visualize as a bar chart

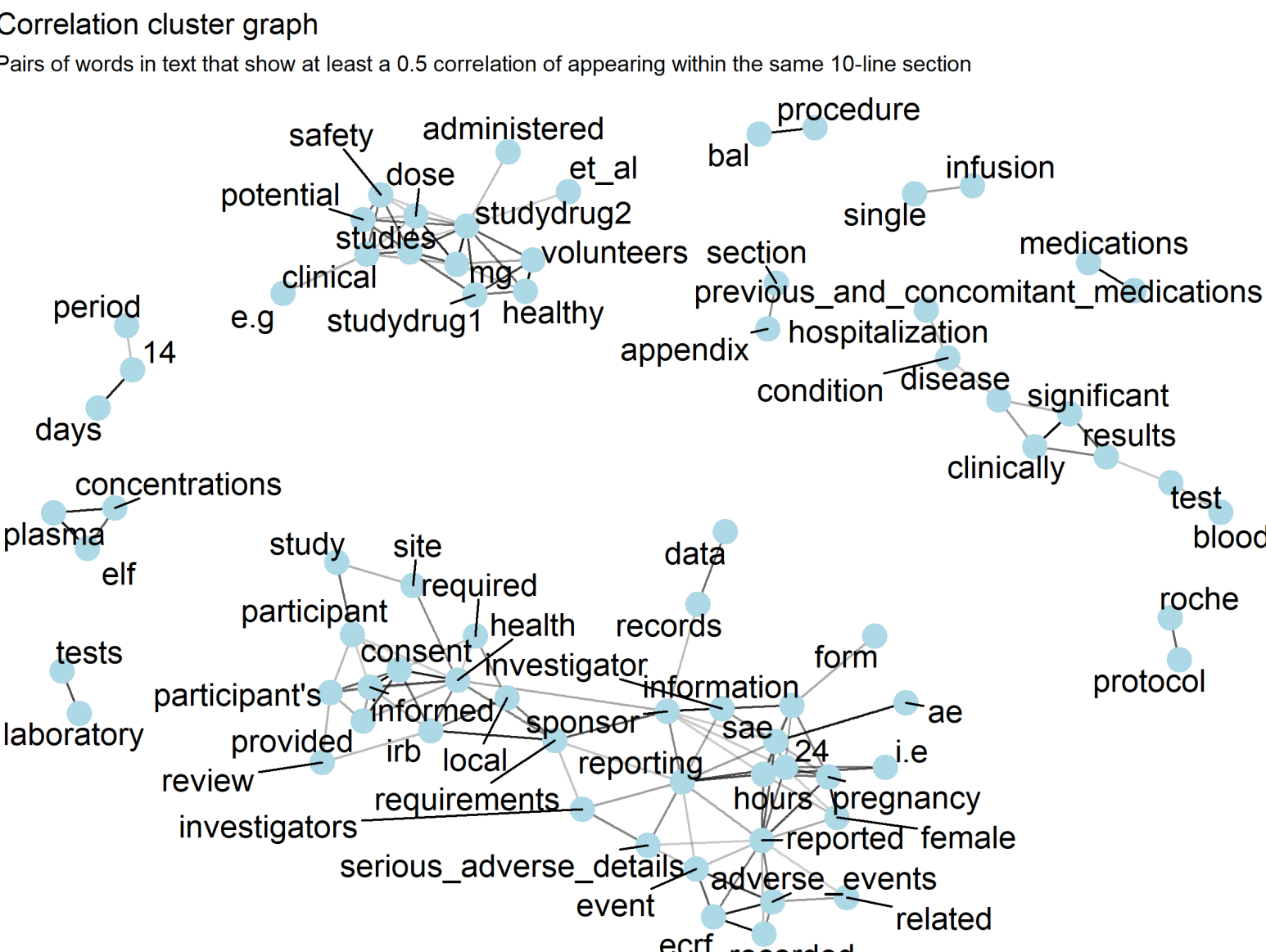
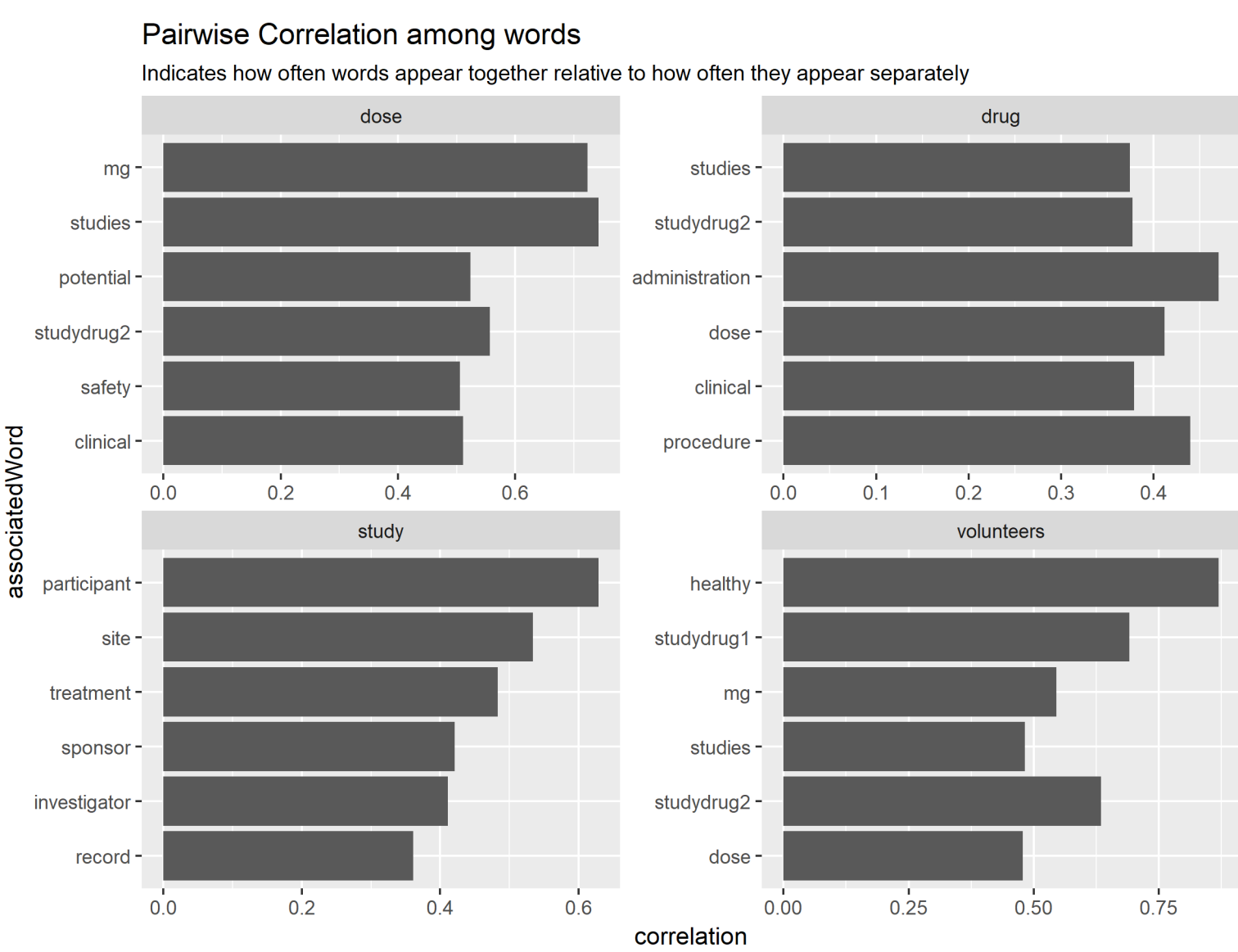
Term Relationship

Bigram Graph
Pairword combinations > 10



- Highlight words that tend to follow each other to indicate meaning
- Pre-process: Text divided into pairs of adjacent words (bigrams)
- Visualize common protocol bigrams as a network

Term Correlation



- Analyse words that occur in the same section of text
- Pre-process: Protocol text divided into rows then grouped
- Filter terms of interest and highlight most common co-occurring words
- Show correlation among words (occurrence frequency relative to how often they appear separately – phi coefficient / Pearson correlation)
- Output helpful for exploration. For example, we could find the words most associated with words of interest like "dose", "drug", "study" etc.
- Visualize target terms or most associated words as a network to show connections

Implement

Mission One: Identify Standard Assessments in the Protocol

Match standard library form names in the protocol text to identify those required for the study build.

Compile a dictionary of Library assessments. Use to make the protocol text and library terms uniform.

- Combine multiple word terms (e.g. the words 'Subject' and 'Characteristics' remain together when consecutive).
- Replace alternative names/abbreviations (e.g. ECG/ Electrocardiogram).
- Indirect match (height/weight indicates anthropometric measurements)

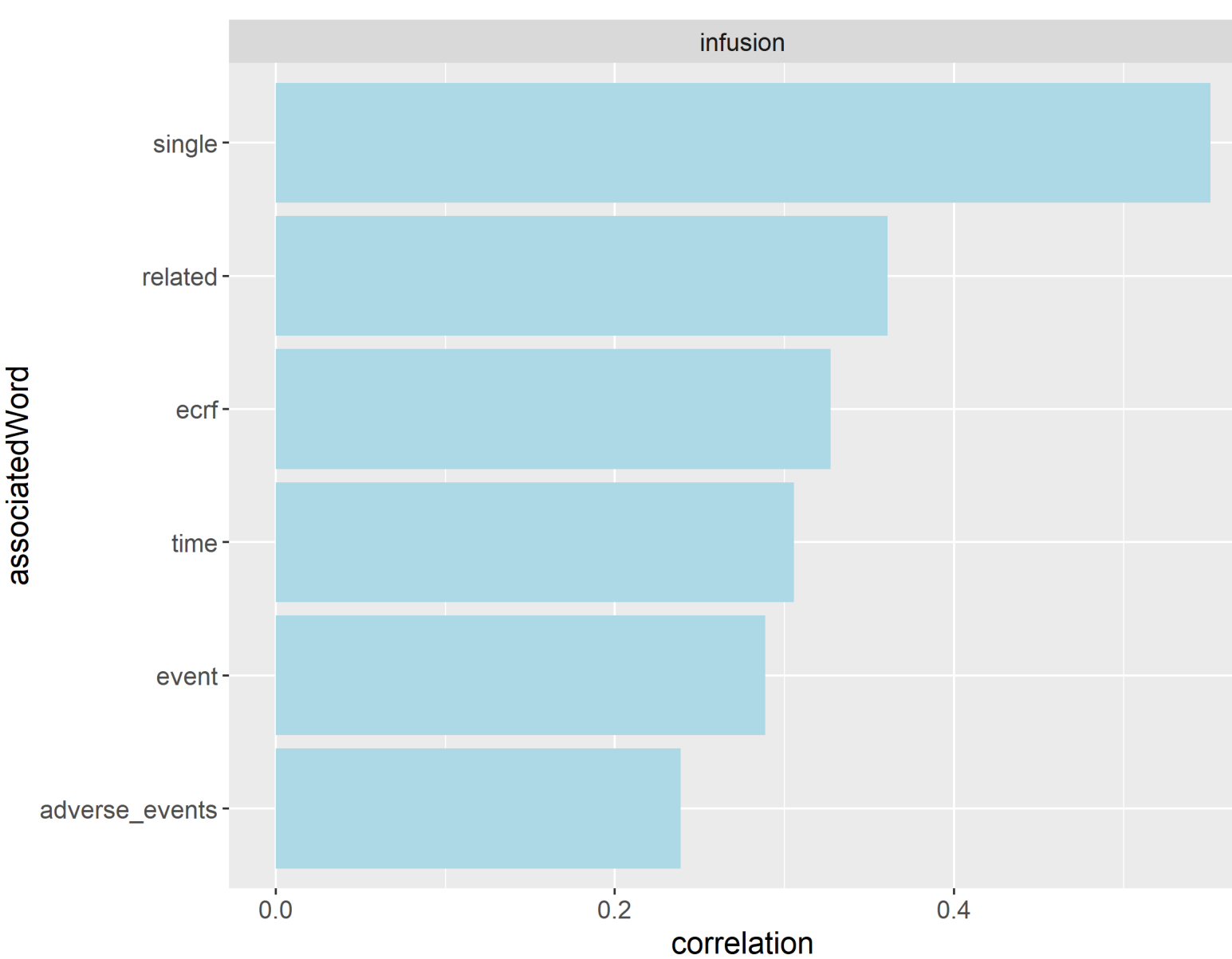
A join of the protocol and library is used to highlight the standard assessments required by the study.

1	adverse_events	129
2	pregnancy	56
3	visit_date	37
4	subject_disposition	10
5	vital_signs	10
6	medical_history	8
7	electrocardiogram	6
8	enrollment	6
9	hematology	4
10	blood_chemistry	3
11	urinalysis	3
12	anthropometric_measurements	3
13	risk_factors	2
14	serology	2
15	surgery_and_procedure_history	2
16	demography	1
17	subject_characteristics	1

Mission Two: Identify Version of Form Required

For certain assessments, the library contains multiple versions. E.g. the study drug admin form has two versions, one for oral admin another for infusions/injections.

Use the parsing/correlation technique to determine which applies.



Mission Three: Fields within a Form

Identify individual data points to be captured within an assessment.

Example: Lab Chemistry – which parameters are to be collected?

Working within form standards this can be achieved using term matching. Use of a common dictionary is beneficial to increase the accuracy of terms identified using a join.

word	n
<chr>	<int>
1 urea	10
2 alanine_aminotransferase	8
3 aspartate_aminotransferase	8
4 repeat	6
5 potassium	5
6 alkaline_phosphatase	3
7 sodium	3
8 total_bilirubin	3
9 chloride	2
10 glucose	2
11 albumin	1
12 bicarbonate	1
13 calcium	1
14 creatinine	1
15 direct_bilirubin	1
16 lactate_dehydrogenase	1
17 phosphate	1

Protocol Terms

Library Terms
(form = Chemistry)