Improving the accuracy of therapy descriptions in clinical trials using a bottom-up approach

Maurine Tong, BS[1], Ricky K. Taira, PhD[1]

[1]University of California, Los Angeles, CA

Abstract

Randomized clinical trial (RCT) reports commonly have complicated therapy descriptions that are written in free-text. Drug therapy is difficult to describe due to the dynamic nature of how protocols change and the many ways drugs can be administered. Details regarding protocol changes and drug administration must be explained clearly for reproducibility and reliability. A process model supplemented with concept ontologies can clarify the dynamics of how therapies change and make knowledge more explicit. We demonstrated the process to develop a representation model to reveal specific context concerning drug therapies within clinical trial report literature. A PubMed search was conducted to identify RCTs on non-small-cell lung cancer (NSCLC) pertaining to epithelial growth factor receptor (EGFR) mutations. Twenty-seven clinical trials were used to develop the model using a bottom-up approach. This representation describes drug dosage, administration details, and drug cycles within different experimental arms and control groups. We then presented preliminary evaluation of the clarity and understandability of the representation.

Introduction

Knowledge from RCT studies/published literatures provide answers on the effectiveness of particular therapies. Elucidating ambiguous treatment protocols improve the quality of scientific research and enhance physician performance. When designing a drug therapy, it is important to clearly define the course of administration with minimal change to the protocol. Despite efforts to control for changes, RCTs often do not follow the initial therapies planned. Discrepancies in interventions have implications on outcome results and statistical analyses. Unexpected events can take place during the conduction of the study, resulting in differences in patients' treatment interventions. Discontinuities can occur in treatments as well as in individualized care from the clinical team. It is important to foresee all possible changes and have a clearly defined protocol for handling each case. In addition, another hindrance that makes therapies unclear occurs at the reporting level. The prevalence of incomplete protocol reporting is high[1], and trialists seem generally unaware of the implications of not reporting all outcomes and protocol changes[1]. Ambiguities arising from the complicated nature of treatments and incomplete reporting motivate the need for a common standard representation model. The utilization of a common and standard representation model for treatments helps detect missing data and/or errors, and promotes interchange and replication of treatments leading to better interpretation of patient results. The research focus of this paper is to document the course of interventions in a precise and accurate manner. We first conducted a systematic review of RCTs for a specific disease. Within these RCTs, we looked at the free-text statements, diagrams and tables documenting treatment, then constructed a representation model to store this information. We built a representation model of treatment protocols using a bottom-up development method, concentrating on the domain of non-small-cell lung cancer (NSCLC) and looking specifically at clinical trial papers dealing with epithelial growth factor receptor (EGFR) mutations.

Background

Most NSCLC patients, if left untreated, have a median survival of 4-5 months after diagnosis and <10% chance of 1 year survival[2]. In particular because of their association with malignant proliferation, the EGFR pathway is critical to some lung adenocarcinoma cells. Members of the ErbB receptor tyrosine kinase family, which includes EGFR, are often deregulated by cancer cells and are validated targets for anticancer therapies. Small molecule reversible inhibitors specific for EGFR have great potential for clinical benefit[3]. Unfortunately, the clinical benefit of the EGFR-tyrosine kinase inhibitors (TKIs) is limited by both primary and acquired resistance. Patients who initially respond to EGFR TKIs develop acquired resistance after a median of 12 months[4]. One way to understand the reasons for primary and required resistance is by looking at drug administration methods. However, drug therapy descriptions are mainly written in an unstructured, free-text form, hence making it manually intensive to compare between therapies.

Efforts and motivation for structuring and synthesizing treatment protocols have been long researched by the Consolidated Standards of Reporting Trials (CONSORT). CONSORT is a 21-point checklist of required items describing what researchers did during trial such as methods, results, and analysis that aims to improve the critical appraisal and completeness of RCT reports[5,6]. It requires that interventions for each testing group be explained in sufficient detail to allow for replication, including how and when the interventions were administered. The description allows fellow researchers to know exactly how to administer the intervention that was evaluated in a particular trial. Specific to treatments and interventions, there is a checklist of characteristics which consists of drug name, dose, method of administration, timing and duration of administration, conditions under which interventions are withheld, and titration regimen. While CONSORT gives a detailed checklist for the necessary information a clinical trial needs to include, its representation is not standardized and there is no criteria for how clearly and completely this information is conveyed.

Other groups have worked on defining and structuring information related to clinical trials as a whole, including treatment protocols. The Ontology of Clinical Research (OCRe) is a formal ontology for describing human studies that can join external information standards (e.g. BRIDG[7], CDISC[8]) and clinical terminologies (e.g. SNOMED CT)[9]. OCRe is an extension of the RCT Schema, which captures concepts related to a trial's design, basic intervention description, execution, administration, and results. The Ontology for Biomedical Investigations (OBI) project developed an integrated ontology for the description of biological and medical experiments and investigations[12]. This ontology aims to model the design of an investigation, including protocols, instrumentation, materials, and data. The Ontology of Scientific Experiments (EXPO) standardizes organization, execution, and analysis of a scientific experiment[13]. The above ontologies formalize descriptions of experimental protocols in varying degrees of granularity. However, our representation model differs from previous efforts because it combines a process model with an ontology to encompass a domain that is not covered by any of the above representations.

The idea of process modeling has been well studied in business process re-engineering (BPR) and proves to be beneficial for representing steps in treatment protocols. The act of representing a sequence of tasks has the potential to identify inefficiencies and opportunities for cost reduction[14]. Payne et al.'s efforts on workflow modeling focuses on abstracting individual participant-related events from a corpus of documents and performing hierarchical cluster analysis[15]. While Payne's work studies a series of typical participant events, it lacks a standard representation model to organize the information. De Carvalho et al. models operational workflow for clinical trials in Unified Modeling Language (UML)[16]. While their application of the concepts of process modeling is specific for the application of operational workflow, their application can be extended to describe treatment interventions.

Methods

Due to a lack of standardized knowledge for describing drug therapies, we used a bottom-up approach to learn the typical language for writing treatment protocols. A bottom-up approach gathers information and generalizes it into a concept description[17,18]. Users pinpoint typical information before generalizing them into a concept for the ontology, which provides a standard way to express concepts. This model provides an infrastructure for analysis and comparison. The steps for creating a representation model for RCT research papers on NSCLC involving EGFR mutations are as follows: (I) Search of relevant literature; (II-IV) Bottom-up approach for data/process modeling for treatment descriptions; and (V) Evaluation of the representation model.

Step I. Search Strategy: This search strategy encompassed a PubMed search for relevant publications pertaining to NSCLC and EGFR mutation. We refined and modified search criteria to identify relevant articles. Our final search criteria combined keywords from RCTs (*"clinical trial"* AND *"Phase I"*) and lung cancer with EGFR mutations(*"lung cancer"* AND *"non-small cell"* AND *"EGFR"*). We then systematically reviewed each article that matched our search.

The search strategy yielded 36 papers, which we retrieved and manually reviewed. Review and case-study papers were excluded, as were papers without access to full text in English. The 27 remaining articles were included in the corpus of NSCLC RCTs for ontology development.

Step II. Representation Model Development – Domain and Scope: Our representation model is able to answer the following competency questions:

- What is a drug regimen?
- What are typical courses of treatment issued to groups of patients?
- Assuming no changes take place, what is the expected progression of events?
- What are the stopping conditions and what are the stopping protocols for a trial design?

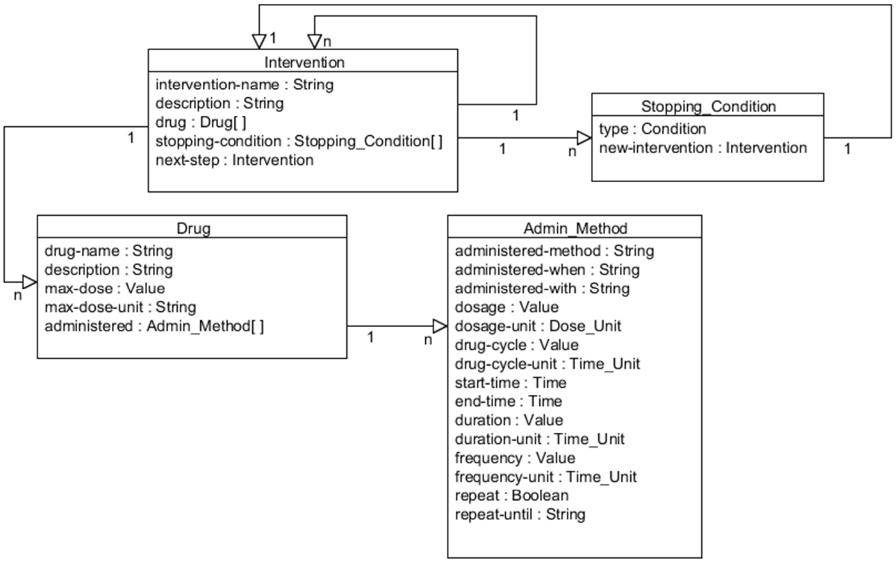
Step III. Representation Model Development – Important Terms: Using the corpus of text on treatments, we identified terms important to the description of drug therapy and to conditions that can affect the course of treatment. As a running example, we will use Price et al. 2010[3]:

> *"After obtaining informed consent, patients were treated with gefitinib 250 mg daily and everolimus 5 mg daily as determined in our earlier phase I study. Dose reduction of everolimus to 2.5 mg daily was allowed for toxicity not managed by optimal supportive care. Dose reduction of gefitinib to 250 mg every other day was allowed for side effects attributable to gefitinib. Dose interruption of both everolimus and gefitinib for grade 3 or 4 toxicities was allowed until resolution of the toxicity (≤ grade 1). For grade 3 or 4 skin toxicity, dose interruption of gefitinib only was allowed with continuation of everolimus unless the toxicity did not resolve within 1 week. For grade 3 or 4 dyslipidemia, dose interruption of everolimus only was permitted. Patients with grade 3 or 4 toxicities that did not resolve in 2 weeks were removed from the study."*

The process for identifying terms was split into two parts: (1) Identification of specific drug entities and their modifiers. In the first sentence, two drug entities, *"gefitinib"* and *"everolimus,"* were identified along with their dosage, *"250 mg daily"* and *"5 mg daily"* respectively. (2) Identification of stopping conditions. In the sixth sentence, the stopping condition entities, *"grade 3-4 dyslipidemia,"* were identified along with the intervention they stop, the *"everolimus"* drug regimen. The fourth sentence contains two stopping conditions: *"grade 3 or 4 skin toxicities"* and *"toxicity lasting more than 1 week."* The resulting intervention depends on both stopping conditions. Our representation model demonstrates the important terms relating to treatments and stopping conditions.

Step IV. Representation Model Development – Classes, Class Hierarchies and Properties: We organized the classes using a bottom-up ontology development approach. Here we illustrate our representation model (Figure 1). Each block in the model represents a class, that can have multiple properties. Each of which are filled with values or other classes. Our process model contains four classes and 27 properties and builds off existing ontologies[12] by including predefined types, which include `Dose_Unit`, `Time_Unit`, and `Condition`. Values for these classes are drawn from a controlled vocabulary.

Figure 1. Representation Model



The `Drug` Class describes all information needed to replicate administration of a particular drug, including administration method, dosage, drug cycles, duration, and frequency. The `Drug` class can contain multiple instances of administration methods, `Admin_Method`, capturing the various ways a drug can be administered. For instance, a drug can be administered daily, weekly, or monthly, as well as in different dosages. Each instance of the `Admin_Method` class describes one type of administration (Table 1).

The `Stopping_Condition` class describes reasons for changing the intervention. Typically, drug regimens have complicated protocols to discontinue or change the use of a drug. Each `Stopping_Condition` class requires an entry from the `Condition` class, where a standardized list of stopping conditions can be found. Table 2 contains examples of text entered into the `Condition` class. When capturing intervention changes, both the stopping condition and the new intervention need to be captured. For example, sentence 5 of Price et al. 2010 identifies a change in protocol for participants who showed grade 3 toxicities: the condition is *"grade 3 or 4 skin toxicity"* and the new-intervention is *"dose interruption of gefitinib"*. The appropriate stopping condition is selected from a pre-defined list (Table 2), and the protocol changes are described as a new intervention.

The `Intervention` class describes the events in a drug regimen, represents how therapies change, and is labeled with a name and description. Because a drug regimen can contain more than one drug, the `Intervention` class is allowed to have multiple instances of the `Drug` class. Each instance of the `Drug` class corresponds to one drug. The dynamic nature of drug therapies is modeled in the `Intervention` class. Within the `Intervention` class, there are properties to describe stopping conditions and subsequent changes in the protocol. Using the stopping-condition or next-step property, the user can input the next step in the protocol.

Table 1. Properties of the `Admin_Method` Class and example entries (from Price et al. 2010).

Property Name	Description	Example Entry
administered-method	Method of drug delivery	*"Orally", "IV"*
administered-when	Additional information describing when the drug was administered	*"Before breakfast"*
administered-with	Co-delivery agents; can include other drugs or inactive ingredients.	*"250 mL saline"*
dosage	Dosage of the drug.	250
dosage-unit		`Dose_Unit` object containing information *"mg"*
drug-cycle	Length of a drug cycle, as defined by trialists. This is different than the frequency property. For example, drug can be administered every day; however, the drug cycle can be defined for 2 weeks.	2
drug-cycle-unit		`Time_Unit` object containing information *"week"*
duration	Duration of drug infusion. This is useful to describe intravenous (IV) drugs, and is usually null for orally administered drugs.	90
duration-unit		`Time_Unit` object containing information *"min"*
frequency	Frequency with which the drug was administered(daily, weekly, etc).	1
frequency-unit		`Time_Unit` object containing information *"day"*
repeat	Answers the question: Was this drug pattern repeated? Allows for the entry for the number of treatment cycles	TRUE
repeat-until		*"6 cycles"*

Table 2. List of Stopping Conditions for Price et al. 2010

Text Abstracted for type in Stopping_Conditions
"toxicity not managed by optimal supportive care"
"side effects attributable to gefitinib"
"until resolution of the toxicity (≤ grade 1)"
"grade 3 or 4 skin toxicity"
"grade 3 or 4 dyslipidemia"

Step V. Evaluation of Representation Model: We evaluated the representation model for its usefulness in explaining a therapy within our corpus and its ability to extend and capture information from a broader search. The representation model can be evaluated using a descriptive method which assesses its ability to sum up free-text information[19]. We first tested the representation with our running example of text from Price et al. 2010. Then, we tested the validity of the representation by applying it to an NSCLC clinical trial not specific to EGFR mutations, Johnson et al. 2004[20], in order to reaffirm that the classes within our representation comprehensively cover the essential components of treatment descriptions.

The following text from Johnson et al. 2004, which presents chemotherapy treatments for NSCLC not specific to the EGFR mutation, was used to evaluate our representation model:

> *"Patients received up to six cycles of carboplatin/paclitaxel. Paclitaxel (200 mg/m2) was administered over 3 hours every 3 weeks. Carboplatin dosing was based on the Calvert formula14 with a target area under the curve of 6 mg/mL x min and glomerular filtration rate (GFR) estimated for males as GFR = (140-age) x weight/72 x (serum creatinine). For females, a correction factor of 0.85 was used. Carboplatin was administered over 15 to 30 minutes, beginning 60 minutes after completion of the paclitaxel. Dose reductions were permitted for febrile neutropenia or absolute neutrophil count less than 1,000/µL for ≥ 5 days, any clinically significant bleeding (≥ grade 2), grade 3 nausea/vomiting not controlled by antiemetic medication, evidence of hepatic (AST > 5 x ULN or bilirubin > 3 x ULN), cardiovascular (symptomatic arrhythmia, hypotension [< 90/60 mmHg or fluid replacement] or chest pain), neurologic (≥ grade 2) or other grade 3 or 4 toxicity. Bevacizumab was administered by intravenous infusion over 90 minutes, 1 hour after each cycle of chemotherapy. If the initial infusion was well tolerated, subsequent infusion times were shortened to 30 to 60 minutes. The bevacizumab dose was not modified during this study. On completing the planned chemotherapy, nonprogressing patients were allowed to continue on bevacizumab at the same dose and schedule for up to a maximum of 18 doses. Patients in the control arm were permitted to receive bevacizumab (15 mg/kg) on disease progression."*

Results

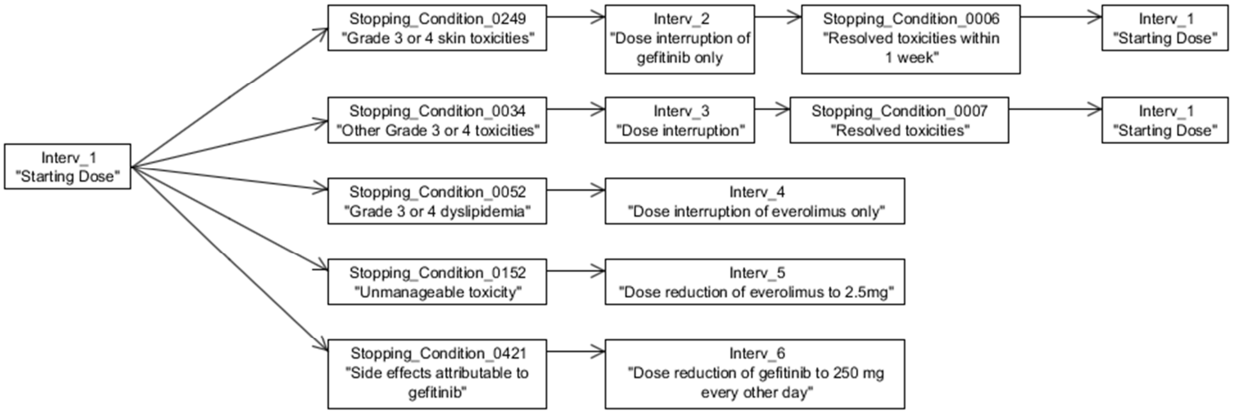
Clinical Trial Features

A total of nine papers out of the 36 that met our PubMed search criteria were excluded from this study; among these, six papers were reviews or case studies, one paper was an update of a previously published report, one paper did not allow free access to the full text, and one paper was written in Chinese without an English translation. In the end, twenty seven papers met our inclusion criteria, including six papers from Journal of Clinical Oncology; four papers from Journal of Thoracic Oncology; three papers from Annals of Oncology; three papers from Clinical Cancer Research; two papers from Cancer; and two papers from Investigational New Drugs. Our research also included one paper from each of the following: Anticancer Research, Bull Cancer, Cancer Chemotherapy and Pharmacology, The Lancet Oncology, Lung Cancer, Medical Oncology, and Oncologist. The research papers included in our study encompassed sixteen unique drug therapies; among these, twelve (44%) of the trials used more than one drug, fifteen (56%) used a combination of two drugs, and no trials used a combination of three or more drugs. The most common drug for EGFR used to treat NSCLC patients was erlotinib, used by nine trials. We observed several new stopping conditions, the most common of which were disease progression and grade 3 or 4 toxicities. The most common protocol change action was dose reduction.

Evaluation of Representation Model

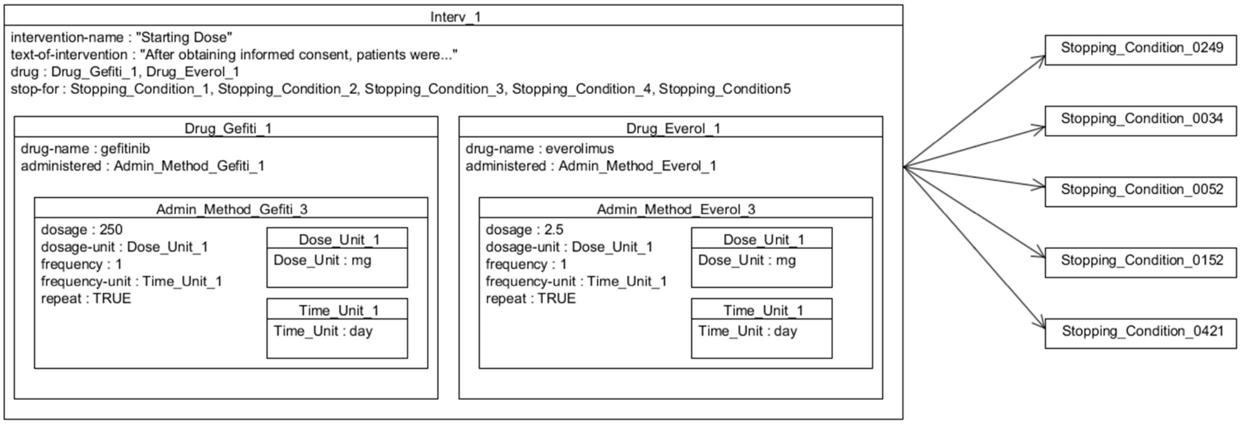
Our representation model captured drug therapy information from a NSCLC clinical trial specific to EGFR mutation, Price et al. 2010 (Figure 2). The representation can be broken down sentence-by-sentence to show the details of our abstraction. Recall the first sentence in our running example: *"patients were treated with gefitinib 250 mg daily and everolimus 5 mg daily as determined in our earlier phase I study."* We started by creating an instance of `Intervention` for the starting drug therapy, named *`Interv_1`*. *`Interv_1`* leads to 5 different stopping conditions, each of which leads to a different and revised intervention. An advantage of our representation model is the ability to reuse blocks. After *`Interv_2`* there is a stopping condition, after which the original intervention is reinstated. The same happens after *`Interv_3`*. By portraying interventions as a process of events, we were better able to identify errors in the protocol. For example, we would see an error if the intervention after a stopping condition matches the intervention before the stopping condition. In Figure 2, we noticed that each intervention used by Price et al. is unique, as denoted by a unique identifier.

Figure 2. Representation Model for Price et al. 2010



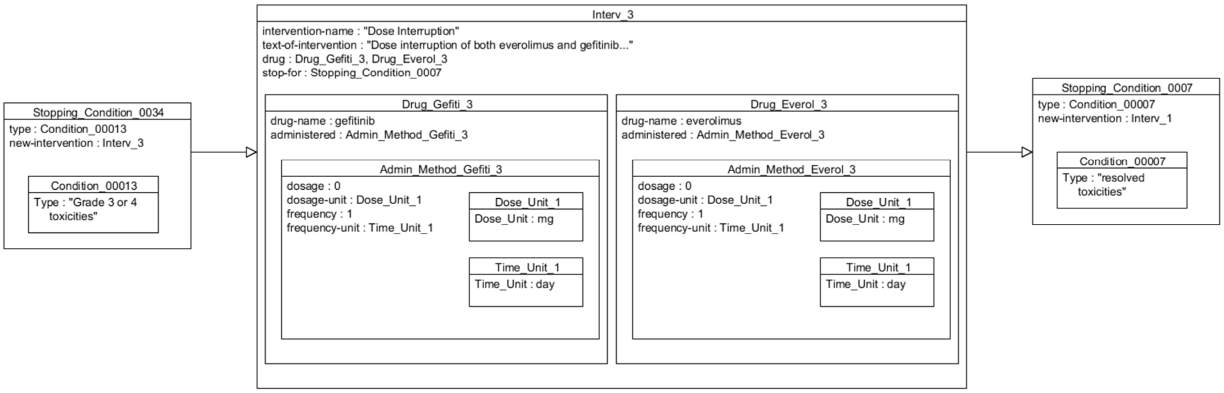
The drugs are described in detail within the starting intervention, *`Interv_1`*. Sentence 1 mentions two drugs, gefitinib and everolimus. Hence, we created two instances of Drug class, *`Drug_gefiti_1`* and *`Drug_everol_1`*. We then created an instance of the `Admin_Method` class for each drug to account for all administration details. For visualization purposes, we drew subclasses within each class to represent the hierarchy of classes (Figure 3).

Figure 3. Abstraction of sentence 1 from Price et al. 2010: *"patients were treated with gefitinib 250 mg daily and everolimus 5 mg daily as determined in our earlier phase I study."* The first box is the starting intervention; the remaining 5 boxes show the five different stopping conditions that can occur.


Interv_1
intervention-name : "Starting Dose"
text-of-intervention : "After obtaining informed consent, patients were..."
drug : Drug_Gefiti_1, Drug_Everol_1
stop-for : Stopping_Condition_1, Stopping_Condition_2, Stopping_Condition_3, Stopping_Condition_4, Stopping_Condition5
Drug_Gefiti_1
drug-name : gefitinib
administered : Admin_Method_Gefiti_1
Admin_Method_Gefiti_3
dosage : 250
dosage-unit : Dose_Unit_1
frequency : 1
frequency-unit : Time_Unit_1
repeat : TRUE
Dose_Unit_1
Dose_Unit : mg
Time_Unit_1
Time_Unit : day
Drug_Everol_1
drug-name : everolimus
administered : Admin_Method_Everol_1
Admin_Method_Everol_3
dosage : 2.5
dosage-unit : Dose_Unit_1
frequency : 1
frequency-unit : Time_Unit_1
repeat : TRUE
Dose_Unit_1
Dose_Unit : mg
Time_Unit_1
Time_Unit : day
Stopping_Condition_0249
Stopping_Condition_0034
Stopping_Condition_0052
Stopping_Condition_0152
Stopping_Condition_0421

To illustrate how the stopping conditions work, we abstracted sentence 4 from Price et al. 2010: *"Dose interruption of both everolimus and gefitinib for grade 3 or 4 toxicities was allowed until resolution of the toxicity (≤ grade 1)."* This sentence contains a two-step stopping condition. In the first step, the stopping condition is the appearance of grade 3 or 4 toxicities. In the second step, the stopping conditions can be lifted if the toxicities resolve to grade 1 or better. Focusing on the first step, we created an instance of `Stopping_Condition` called *`Stopping_Condition_0034`* and filled in the property values for this class. *"Grade 3 or 4 toxicities"* is filled in for the type property. In the new-intervention property, we created a new instance of `Intervention` called `Interv_3` (Figure 4) which we populated in the same manner as `Interv_1` (Figure 3). In the second step (resolved toxicities), the dose interruption stops and the original treatment continues. Note that the `Stopping_Condition` class includes stopping conditions, but can also be generalized to any changes in patient status, such as the resolution of toxicities. We created an instance of `Stopping_Condition` called *`Stopping_Condition_0007`*. The type property is *"resolved toxicities"*, and the new-intervention property is `Interv_1`, which corresponds to the original intervention. Thus, using our representation model, we can accurately and precisely characterize interventions with their corresponding stopping conditions.

Figure 4. Abstraction of sentence 4 from Price et al. 2010: *"Dose interruption of both everolimus and gefitinib for grade 3 or 4 toxicities was allowed until resolution of the toxicity (≤ grade 1)."* The first box contains the stopping condition, *"grade 3 or 4 toxicities,"* and the new resulting intervention. The second box contains the new resulting intervention, *"dose interruption of everolimus and gefitinib"*. The third box contains the stopping condition, *"resolution of toxicity,"* and the new resulting intervention.


Stopping_Condition_0034
type : Condition_00013
new-intervention : Interv_3
Condition_00013
Type : "Grade 3 or 4 toxicities"
Interv_3
intervention-name : "Dose Interruption"
text-of-intervention : "Dose interruption of both everolimus and gefitinib..."
drug : Drug_Gefiti_3, Drug_Everol_3
stop-for : Stopping_Condition_0007
Drug_Gefiti_3
drug-name : gefitinib
administered : Admin_Method_Gefiti_3
Admin_Method_Gefiti_3
dosage : 0
dosage-unit : Dose_Unit_1
frequency : 1
frequency-unit : Time_Unit_1
Dose_Unit_1
Dose_Unit : mg
Time_Unit_1
Time_Unit : day
Drug_Everol_3
drug-name : everolimus
administered : Admin_Method_Everol_3
Admin_Method_Everol_3
dosage : 0
dosage-unit : Dose_Unit_1
frequency : 1
frequency-unit : Time_Unit_1
Dose_Unit_1
Dose_Unit : mg
Time_Unit_1
Time_Unit : day
Stopping_Condition_0007
type : Condition_00007
new-intervention : Interv_1
Condition_00007
Type : "resolved toxicities"

After verifying the representation with Price et al., a text from our corpus, we then used Johnson et al. 2004 to test our model. We were able to identify relevant text referring to treatment protocols and identify the appropriate class and properties to extract dosing information of all drugs (carboplatin, paclitaxel, and bevacizumab), as well as information about the number of cycles, duration of dose, and other administration details (Table 3). This RCT consisted of two arms, one experimental arm and one control arm; we instantiated two `Intervention` classes, one for each arm. For the experimental arm, we sorted the sentences according to which drug they pertained to, then examined each sentence to determine the information we can extract. We repeated this procedure for the control arm (Table 4).

In addition, we were able to capture a number of stopping conditions (Table 5), which, for this RCT, consisted of reasons why participants were unable to continue with the treatment intervention. Reasons for discontinuation ranged from grade 3-4 toxicities to specific organ problems.

Table 3. Organization of the text from Johnson et al. 2004 broken down by sentence and sorted by administered drug for the experimental group. Each sentence is listed with the property that stores the extractible administration details.

Carboplatin	*"Patients received up to six cycles of carboplatin/paclitaxel."*—Repeat *"Carboplatin dosing was based on the Calvert formula14 with a target area under the curve of 6 mg/mL x min and glomerular filtration rate (GFR) estimated for males as GFR = (140-age) x weight/72 x (serum creatinine). For females, a correction factor of 0.85 was used. Carboplatin was administered over 15 to 30 minutes, beginning 60 minutes after completion of the paclitaxel."*—Administered-method, Administered-when, Dosage, Duration
Paclitaxel	*"Paclitaxel (200 mg/m2) was administered over 3 hours every 3 weeks."* –Dosage, Frequency, Duration *"Patients received up to six cycles of carboplatin/paclitaxel."*—Repeat
Bevacizumab	*"Bevacizumab was administered by intravenous infusion over 90 minutes, 1 hour after each cycle of chemotherapy."*—When administered, Duration, Frequency

Table 4. Organization of the text from Johnson et al. 2004 by administered drug for the control group

Carboplatin	NA
Paclitaxel	NA
Bevacizumab	*"Patients in the control arm were permitted to receive bevacizumab (15 mg/kg) on disease progression"*—Dosage

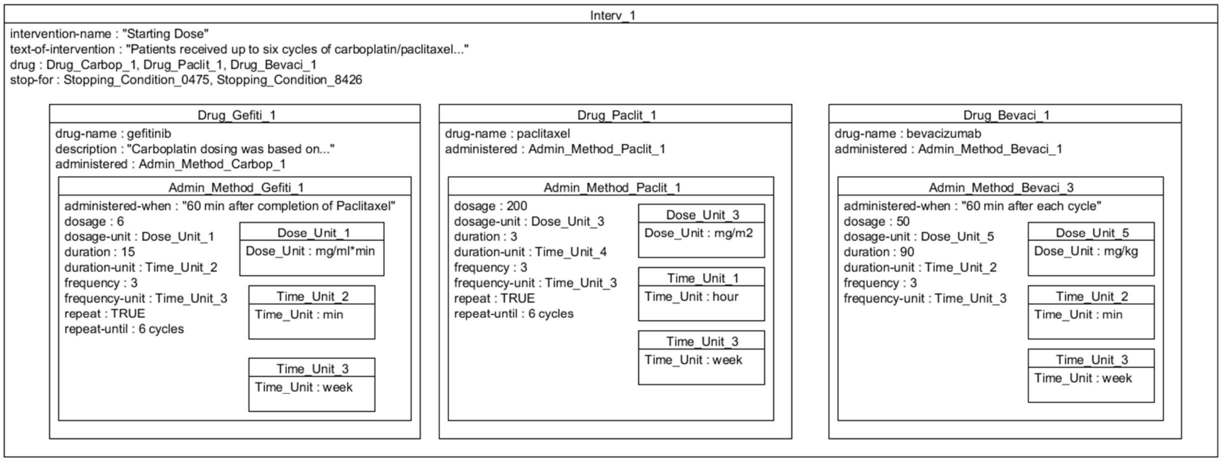
Table 5. List of Stopping Conditions for Johnson et al. 2004

Text Abstracted for type in Stopping_Condition	Text Abstracted for new-interventions
"febrile neutropenia or absolute neutrophil count less than 1,000/μL	Unspecified dose reduction

for ≥ 5 days"	
"any clinically significant bleeding (≥ grade 2)"	Unspecified dose reduction
"grade 3 nausea/vomiting not controlled by antiemetic medication"	Unspecified dose reduction
"evidence of hepatic (AST > 5 x ULN or bilirubin > 3 x ULN)"	Unspecified dose reduction
"cardiovascular (symptomatic arrhythmia, hypotension [< 90/60 mmHg or fluid replacement] or chest pain)"	Unspecified dose reduction
"neurologic (≥ grade 2) or other grade 3 or 4 toxicity"	Unspecified dose reduction
"initial infusion was well tolerated"	*"infusion times were shortened to 30 to 60 minutes"*

We started our representation for Johnson et al. 2004 by creating an instance of `Intervention` consisting of 3 drugs: carboplatin, paclitaxel, and bevacizumab. For each drug, we filled in the values for each property within the `Admin_Method` class (Figure 5).

Figure 5. Starting intervention for the clinical trial by Johnson et al. 2004


Interv_1
intervention-name : "Starting Dose"
text-of-intervention : "Patients received up to six cycles of carboplatin/paclitaxel..."
drug : Drug_Carbop_1, Drug_Paclit_1, Drug_Bevaci_1
stop-for : Stopping_Condition_0475, Stopping_Condition_8426
Drug_Gefiti_1
drug-name : gefitinib
description : "Carboplatin dosing was based on..."
administered : Admin_Method_Carbop_1
Admin_Method_Gefiti_1
administered-when : "60 min after completion of Paclitaxel"
dosage : 6
dosage-unit : Dose_Unit_1
duration : 15
duration-unit : Time_Unit_2
frequency : 3
frequency-unit : Time_Unit_3
repeat : TRUE
repeat-until : 6 cycles
Dose_Unit_1
Dose_Unit : mg/ml*min
Time_Unit_2
Time_Unit : min
Time_Unit_3
Time_Unit : week
Drug_Paclit_1
drug-name : paclitaxel
administered : Admin_Method_Paclit_1
Admin_Method_Paclit_1
dosage : 200
dosage-unit : Dose_Unit_3
duration : 3
duration-unit : Time_Unit_4
frequency : 3
frequency-unit : Time_Unit_3
repeat : TRUE
repeat-until : 6 cycles
Dose_Unit_3
Dose_Unit : mg/m2
Time_Unit_1
Time_Unit : hour
Time_Unit_3
Time_Unit : week
Drug_Bevaci_1
drug-name : bevacizumab
administered : Admin_Method_Bevaci_1
Admin_Method_Bevaci_3
administered-when : "60 min after each cycle"
dosage : 50
dosage-unit : Dose_Unit_5
duration : 90
duration-unit : Time_Unit_2
frequency : 3
frequency-unit : Time_Unit_3
Dose_Unit_5
Dose_Unit : mg/kg
Time_Unit_2
Time_Unit : min
Time_Unit_3
Time_Unit : week

The process model more clearly illustrates how treatment interventions were described in free text and enables us to pinpoint items missing from the protocol. In this trial, we can see that after a stopping condition occurs the resulting dose reductions were not specified (Table 5).

This representation model for Johnson et al. 2004 was limited in its ability to effectively describe dosing requirements. While our representation model was able to capture the information on `Intervention` as separate entities, it was unable to explain interventions in relation to each other. For example, in sentence 3 of Johnson et al. 2004, we see the text: *"For females, a correction factor of 0.85 was used."* Instead storing information about a correction factor, our representation model abstracts this information by creating another instance of `Intervention`. Yet another limitation is the inability of our model to capture a range of time. This is because each drug is assumed to be administer instantaneously. For example, sentence 4 of Johnson et al. 2004 states that the drug was administered *"over 15 to 30 minutes."* This information is not captured in our representation model.

Conclusion

Treatment descriptions within primary RCT papers are not well modeled. Drug therapies are difficult to administer and complications requiring change of protocol are common. The traditional method of presenting treatments has many limiting factors; for example, it lacks flexibility in modeling dynamic changes in therapy administration protocol. A common solution is to use free-text descriptions; however, free-text is at times vague and must be translated to a computer-understandable format to store and analyze data. Our research addresses the standardization of treatments written in a set of clinical trial reports for an example disease.

In this paper we established a way to build a representation model specific for describing drug therapies in a set of clinical trial papers. We built a representation model using a bottom-up approach and demonstrated its utility representing treatment protocols. The representation model characterizes a single event and treatments are modeled as a sequential list of individual events. By understanding the drug therapy temporal order, a researcher or biostatistician can unambiguously follow the study protocol for replication or understand the assumptions of the experimental design. Information structured in this way can eventually be used to evaluate the efficacy and tolerability of treatments, as well as for executing systematic reviews and meta-analysis of randomized trials.

We discovered that despite the different types of therapies, each therapy can be modeled as a sequence of events and stopping conditions, with each event and stopping conditions standardized by their own controlled vocabularies. While stopping conditions were initially modeled for protocol changes as a result of diminishing participant health, we found that they work well to model any change in participants' status. For example, a stopping condition can be *"resolution of toxicities,"* in which case the original intervention is continued, or *"well tolerated"* treatments, in which case treatments are readjusted. These conditions can be incorporated in our representation model to allow for ease of synthesis and disambiguation. Some studies allow for dose reduction due to sudden toxicities; however, the protocol for dose reduction is not specified. Details needed include how much to reduce a patient's dose, how long the reduction should last, and what happens after a patient returns to his/her previous status.

One limitation of our model is that we did not capture all possible diseases and more research is required before the model can be considered comprehensive. Our corpus of papers looked specifically at treatments for NSCLC patients with an EGFR mutation. However, our method for developing the representation model can be extended to include other types of lung cancer in the short term, eventually encompassing all cancer in the long term. Another limitation is restricted computer readability due to the model's use of free-text for some properties; for example, administration-method, administration-with, and administration-when all store information as a string. Part of our on-going work is abstracting string components to a more computer-readable format. Our representation model makes the assumption that drugs used in combination drug therapies are administered simultaneously, and is not able to capture drugs that were not administered at the same point. Another limitation of our work is the constriction that we placed on our method of evaluation: we evaluated the representation model based on how well it can abstract a new piece of text. However, other methods of evaluation (e.g. compared with data on the domain(s) of interest, assessed by pre-defined criteria by use[19]) may provide better assessment on the utility and purpose of the representation model.

Moving forward, we will use the bottom-up method to improve the robustness of our representation model by extending its content to include a variety of therapies for patients with other types of cancer; and we will eventually broaden our efforts to build on existing ontologies for other sections of a clinical trial paper.

References

1. Smyth RM, Kirkham JJ, Jacoby A, Altman DG, Gamble C, Williamson PR. Frequency and reasons for outcome reporting bias in clinical trials: interviews with trialists. BMJ. 2011 Jan 6;342:c7153.
2. Sharma SV, Bell DW, Settleman J, Harber D. Epidermal growth factor receptor mutations in lung cancer. Nature Vol 7. March 2007
3. Price KA, Azzoli CG, Krug LM, Pietanza MC, Rizvi NA, Pao W, Kris MG, Riely GJ, Heelan RT, Arcila ME, Miller VA. Phase II trial of gefitinib and everolimus in advanced non-small cell lung cancer. J Thorac Oncol. 2010 Oct;5(10):1623-9.
4. Yap TA, Vidal L, Adam J, Stephens P, Spicer J, Shaw H, Ang J, Temple G, Bell S, Shahidi M, Uttenreuther-Fischer M, Stopfer P, Futreal A, Calvert H, de Bono JS, Plummer R. Phase I trial of the irreversible EGFR and HER2 kinase inhibitor BIBW 2992 in patients with advanced solid tumors. J Clin Oncol. 2010 Sep 1;28(25):3965-72. Epub 2010 Aug 2.
5. Moher D, Jones A, Lepage L; CONSORT Group (Consolitdated Standards for Reporting of Trials). Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. JAMA. 2001 Apr 18;285(15):1992-5.
6. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, Elbourne D, Egger M, Altman DG. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ. 2010 Mar 23;340:c869. doi: 10.1136/bmj.c869.
7. The Biomedical Research Integrated Domain Group (BRIDG) Model. <http://www.bridgmodel.org/>
8. Clinical Data Interchange Standards Consortium. 2012. <http://www.cdisc.org/>
9. Sim I, Carini S, Tu S, Wynden R, Pollock BH, Mollah SA, Gabriel D, Hagler HK, Scheuermann RH, Lehmann HP, Wittkowski KM, Nahm M, Bakken S. The human studies database project: federating human studies design data using the ontology of clinical research. AMIA Summits Transl Sci Proc. 2010 Mar 1;2010:51-5.

10. Sim I, Olasov B, and Carini S. The Trial Bank System: Capturing Randomized Trials for Evidence-Based Medicine. AMIA Annu Symp Proc. 2003; 2003: 1076.
11. Sim I, Owens DK, Lavori PW, Rennels GD. Electronic trial banks: a complementary method for reporting randomized trials. Med Decis Making. 2000 Oct-Dec;20(4):440-50.
12. Brinkman RR, Courtot M, Derom D, Fostel JM, He Y, Lord P, Malone J, Parkinson H, Peters B, Rocca-Serra P, Ruttenberg A, Sansone SA, Soldatova LN, Stoeckert CJ Jr, Turner JA, Zheng J; OBI consortium. Modeling biomedical experimental processes with OBI. J Biomed Semantics. 2010 Jun 22;1 Suppl 1:S7.
13. Soldatova LN, King RD. An ontology of scientific experiments. J R Soc Interface. 2006 Dec 22;3(11):795-803.
14. Neill PO, Sohal AS (1999) Business Process Reengineering A Review of Recent Literature. Technovation 19: 571–581.
15. Payne PRO, Eneida AM, Justin BS (2007) Modeling Participant-Related Clinical Research Events Using Conceptual Knowledge Acquisition Techniques. AMIA Annu Symp Proc 593–597
16. de Carvalho ECA, Jayanti MK, Batilana AP, Kozan ZMO, Rodrigues MJ, Shah J, Loures MR, Pietrobon R. Standardizing clinical trials workflow representation in UML for international site comparison. PLoS ONE. 5(11):e13893, 2010.
17. Van der Vet PE, Mars NJI. Bottom-up construction of ontologies. IEEE Jul/Aug 1998 Vol 10, Issue 4, 513-526. DOI: 10.1109/69.706054
18. Guarino N, Oberle D, Staab S. What is an Ontology? International Handbooks on Information Systems, 2009, Part 1, 1-17.
19. Brank J, Groberlnik M, Mladenic D. A survey of ontology evaluation techniques. SIKDD 2005 at multiconference IS 2005, 17 Oct 2005, Ljubljana, Slovenia.
20. Johnson DH, Fehrenbacher L, Novotny WF, Herbst RS, Nemunaitis JJ, Jablons DM, Langer CJ, DeVore RF 3rd, Gaudreault J, Damico LA, Holmgren E, Kabbinavar F. Randomized phase II trial comparing bevacizumab plus carboplatin and paclitaxel with carboplatin and paclitaxel alone in previously untreated locally advanced or metastatic non-small-cell lung cancer. J Clin Oncol. 2004 Jun 1;22(11):2184-91.