I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Amarendra Das, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Daniel Rubin, Co-Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Lawrence Fagan

Approved for the Stanford University Committee on Graduate Studies.

Patricia J. Gumport, Vice Provost Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format. An original signed hard copy of the signature page is on file in University Archives.
ABSTRACT

This dissertation describes the Rule-Based Response Assessment Framework, a goal-driven, context-driven, and knowledge-based temporal interpretation method that computationally performs treatment response assessment. The goal of this work is to enable consistent application of standardized response criteria to clinical trial data analysis, and to enable research and development of response criteria. Treatment response assessment is an important clinical and research task in medicine. Standardized treatment response criteria have been developed to aid researchers conducting clinical trials compare the efficacy and toxicity of novel treatments within and across studies. With these criteria, the clinical state and response state of each patient in the cohort can be consistently assessed and then further aggregated into a summary statistic that describes the cohort’s response to a treatment. Multiple domain-specific response criteria have been developed, that range in complexity and specificity such that consistent application of response criteria within and across clinical trials can be challenging. Furthermore, there is no common framework to represent the generic and domain specific components of the criteria to enable application, comparison and evaluation of response criteria. Yet response criteria have common features and tasks within and across clinical domains that present an opportunity for knowledge-based modeling approaches.

In this work, I have developed the Rule-Based Response Assessment Framework, which defines the dimensions of response assessment including the knowledge and tasks required to apply response criteria to patient data. The Framework consists of several knowledge models and several sub-methods that make up the Response Assessment Method. The methods take as input knowledge and data encoded in the models and generate as output new knowledge that is asserted back into the knowledge models. The knowledge models in the Framework include the Protocol, Clinical Data and Response Criteria Models and the Temporal Data and Temporal Restrictions Models that they share. The Response Assessment Method consists of the Context Restriction Method and the Context-Based Interpretation Methods. The Interpretation Methods are further categorized as classification or statistical methods based on whether they generate
qualitative or quantitative interpretations respectively. The interpretation methods generate interpretations at the finding level, patient level and cohort level.

The Rule-Based Response Assessment Framework is implemented using ontologies and rules using the Semantic Web formalisms of OWL and SWRL respectively. The framework is applied to two response criteria in the oncology domain and one response criterion for rheumatoid arthritis. The implementation of the framework in OWL and SWRL was sufficiently expressive to generate clinically accurate response interpretations compared to human interpretation. In addition, we demonstrate that approximately 87% of the criteria knowledge is shared between the two oncology criteria, and that the framework enables a comparison of the outputs of the two criteria applied to the same patient data set. We have demonstrated that the implementation of the Rule-Based Response Assessment Framework is extensible, reusable and enables consistent application of response criteria to patient data and comparison of response criteria within a clinical domain.

This dissertation makes several informatics and clinical contributions. Informatics contributions include a formal definition of the response assessment task as a type of temporal interpretation method. In addition, we have developed the Response Assessment Method as a goal-driven, context-driven, temporal interpretation method that extends the range of interpretation tasks for temporal interpretation methods to include classification and statistical interpretation tasks across single and multiple-contexts, for single and multiple parameters and subjects. We have also introduced the rule template approach to domain knowledge acquisition and implemented the framework using knowledge representation formalism of OWL and SWRL that can be inspected, extended, and transported.

Clinical contributions include the development of a formal model of the dimensions of response criteria and complete decision-logics for three domain response criteria. In addition, automated application of response criteria could enable more consistent application within and across clinical trials. Finally, the framework enables comparison of response criteria knowledge and results, essential for the ongoing development of response criteria. Future research directions will incorporate the Rule-
Based Response Assessment Framework into a larger informatics infrastructure for application and development of oncology response criteria.
ACKNOWLEDGEMENTS

I would like to dedicate this dissertation in memory of my mother Lenore Levy who passed away from complications of breast cancer during the second year of my PhD program. My mother’s cancer was diagnosed during my first year of medical school and inspired me to devote my professional life to the advancement of cancer medicine. The example of her life has instilled in me a strong sense of curiosity and work ethic. I am so very thankful for my mother and father Dr. Lawrence Levy’s support and encouragement through many years of graduate and post-graduate training.

My husband Jonathan Holt has been my most cherished intellectual and spiritual support since the end of medical school and throughout my Internal Medicine Residency, Medical Oncology Fellowship, and PhD program. Jonathan continues to help me discover my inner philosopher. I could not have accomplished so much in such a condensed period of time without his continued support. Our daughter Dana Holt continues to bring joy and wonder to our world that is otherwise too often overrun by work. She has been very patient with her mother’s long working hours.

I first became interested in developing informatics methods to support cancer treatment response assessment during the Biomedical Informatics Foundations Project Course BMI 212. I would like to acknowledge my LesionViewer collaborators Yael Garten, Ankit Garg, and Aaron Tam for their contributions to this first pilot project exploring the challenges and opportunities for cancer treatment response assessment. I would also like to thank Dr. Ellie Guardino for her mentorship in cancer clinical trial design and cohort response assessment.

My co-advisors Dr. Daniel Rubin and Dr. Amar Das have helped me to refine and articulate the work in this dissertation. They helped me to understand the nature of informatics methods as a scientific and philosophical discipline. As members of the dissertation committee, Dr. Larry Fagan and Dr. Russ Altman have contributed both intellectual and general academic support during my PhD program.
The technical implementation of this work would not have been possible without the contributions of Martin O’Connor. Martin’s work to extend the protégé SWRL tab rule language implementation to include set based and abox reasoning enabled me to completely encode the Framework in a knowledge based formalism.

This work has been supported by several training grants including: Stanford’s National Library of Medicine Fellowship training grant, the Veterans Administration Informatics Fellowship program at the Palo Alto Veteran’s Hospital, and Stanford’s Medical Oncology Fellowship Program.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>iv</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>vii</td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>vix</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>xiv</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>xv</td>
</tr>
<tr>
<td><strong>CHAPTER 1: INTRODUCTION</strong></td>
<td>1</td>
</tr>
<tr>
<td>1.1: Response Assessment</td>
<td>1</td>
</tr>
<tr>
<td>1.2: Treatment Response Assessment in Medicine</td>
<td>3</td>
</tr>
<tr>
<td>1.3: Definition of Treatment Response Assessment</td>
<td>7</td>
</tr>
<tr>
<td>1.4: Overview of Dissertation Document by Chapter</td>
<td>12</td>
</tr>
<tr>
<td><strong>CHAPTER 2: ANALYSIS OF RESPONSE ASSESSMENT CRITERIA</strong></td>
<td>13</td>
</tr>
<tr>
<td>2.1: Domain Specific Response Assessment Criteria for Clinical Trials</td>
<td>13</td>
</tr>
<tr>
<td>2.1.1: Oncology Response Criteria</td>
<td>13</td>
</tr>
<tr>
<td>2.1.1.1: Application of RECIST</td>
<td>17</td>
</tr>
<tr>
<td>2.1.2: HIV Response Criteria</td>
<td>24</td>
</tr>
<tr>
<td>2.1.3: Rheumatology Response Criteria</td>
<td>25</td>
</tr>
<tr>
<td>2.2: Common Features of Response Criteria</td>
<td>27</td>
</tr>
<tr>
<td>2.2.1: Defining the Clinical State in Response Criteria</td>
<td>28</td>
</tr>
<tr>
<td>2.2.1.1: Assessment Modality</td>
<td>28</td>
</tr>
<tr>
<td>2.2.1.2: Measurement Technique</td>
<td>29</td>
</tr>
<tr>
<td>2.2.1.3: Measurement Data Type</td>
<td>29</td>
</tr>
<tr>
<td>2.2.1.4: Timing of Response Assessment</td>
<td>30</td>
</tr>
<tr>
<td>2.2.1.5: Classification and Calculation of Findings</td>
<td>32</td>
</tr>
<tr>
<td>2.2.1.6: Quantitative Estimate of the Clinical State</td>
<td>34</td>
</tr>
<tr>
<td>2.2.1.7: Qualitative Estimate of the Clinical State</td>
<td>35</td>
</tr>
<tr>
<td>2.2.2: Defining the Response State in Response Criteria</td>
<td>36</td>
</tr>
<tr>
<td>2.2.2.1: Quantitative Assessment of Response to Treatment</td>
<td>36</td>
</tr>
</tbody>
</table>
4.2.2: Temporal Restrictions Model .................................................. 81
4.2.3: Clinical Data Model .............................................................. 84
4.2.4: Temporal Data Model ............................................................ 85
4.2.5: Context Restriction Method ................................................... 86
  4.2.5.1: Anchor Generation Method ............................................... 86
  4.2.5.2: Context Generation Method ........................................... 86
  4.2.5.3: Context Classification Method ....................................... 86
  4.2.5.4: Composite Context Generation Method ....................... 87
4.3: Knowledge and Methods for Context-Based Interpretation ................. 87
  4.3.1: Response Criteria Knowledge Model .................................. 90
    4.3.1.1: Response Parameters ................................................ 91
    4.3.1.2: Interpretation Task Specification .............................. 92
      4.3.1.2.1: Single Context Interpretation Tasks .............. 93
      4.3.1.2.2: Multiple Context Interpretation Tasks .......... 94
    4.3.1.2.3: Specification of Classification Tasks ............ 94
    4.3.1.2.4: Specification of Statistical Tasks .................. 95
  4.3.2: Context Based Interpretation Methods ................................ 95
    4.3.2.1: Single Context Classification Methods .................. 95
    4.3.2.2: Multiple Context Classification Methods ........... 99
    4.3.2.3: Single Context Statistical Methods ..................... 101
    4.3.2.4: Multiple Context Statistical Methods ................ 102

CHAPTER 5: IMPLEMENTATION OF THE RULE-BASED RESPONSE
ASSESSMENT FRAMEWORK ............................................................. 104
5.1: Knowledge Representation Implementation Choice .......................... 104
5.2: Ontologies and Rules for the Context Restriction Procedure ............. 107
5.3: Ontologies and Rules for the Context Based Interpretation Procedure .... 110
  5.3.1: Response Criteria Parameter Ontology ................................ 111
  5.3.2: Interpretation Task Specification Ontology ..................... 112
  5.3.3: Context-Based Interpretation Rule Tempaltes .................. 114
CHAPTER 6: DOMAIN APPLICATIONS ......................................................... 118
6.1: Application of the Framework to Solid Tumor Response Criteria .......... 118
   6.1.1: Solid Tumor Clinical Trial Data ............................................. 118
      6.1.1.1: Image Annotations ..................................................... 119
   6.1.2: Annotation Image Markup Ontology and Data Rules ...................... 119
   6.1.3: Instantiation of the Response Evaluation Criteria in Solid Tumors ... 121
      6.1.3.1: Instantiation of Study Protocol Specification ....................... 122
      6.1.3.2: Importing Image Annotation Data to Clinical Data Ontology ...... 122
      6.1.3.3: Domain Extension of Response Parameter Ontology ............... 123
      6.1.3.4: Instantiation of Domain Specific Interpretation Tasks .......... 124
6.2: Application of the Framework to Lymphoma Response Criteria .......... 126
   6.2.1: Lymphoma Clinical Trial Data .............................................. 126
   6.2.2: Instantiation of the International Harmonization Criteria .............. 127
6.3: Application of the Framework to Rheumatoid Arthritis Response Criteria . 129

CHAPTER 7: EVALUATION ................................................................. 134
7.1: Evaluation of Expressivity of Framework ........................................ 134
7.2: Evaluation of Automated Outcomes Compared to Human Interpretation ... 135
7.3: Evaluation of Reuse of Implementation ......................................... 138
7.4: Evaluation of Extensibility of Framework ....................................... 139
7.5: Comparison of Response Criteria ................................................. 140

CHAPTER 8: DISCUSSION ............................................................... 144
8.1: Informatics Contributions .......................................................... 144
   8.1.1: Definition of response assessment task .................................... 144
   8.1.2: Contributions to temporal interpretation methods ....................... 145
   8.1.3: Knowledge-based implementation ......................................... 146
   8.1.4: Rule Templates for domain knowledge acquisition .................... 147
   8.1.5: Evaluation of SWRL extensions ............................................ 148
   8.1.6: Reuse and integration of existing ontologies ............................ 149
8.2: Informatics Limitations ............................................................ 149
   8.2.1: Trade-offs of rule template implementation ............................. 149
8.2.2: Missing data ................................................................. 150
8.2.3: Lack of truth maintenance ............................................. 150
8.3: Clinical Contributions ...................................................... 151
  8.3.1: Formal model of dimensions of response criteria ................. 151
  8.3.2: Complete decision-logic for three domain response criteria .... 151
  8.3.3: Automated response assessment ...................................... 153
  8.3.4: Comparison of response criteria knowledge and results .......... 153
8.4: Clinical Limitations ....................................................... 154
  8.4.1: Lack of decision support for input data set ....................... 154
  8.4.2: Limited modeling of assessment modalities ...................... 154
  8.4.3: Retrospective versus prospective data acquisition and analysis .... 155
  8.4.4: Limited to protocol-driven care .................................... 156

CHAPTER 9: FUTURE RESEARCH DIRECTIONS .......................... 158
9.1: Proposed system for application of response criteria ................ 158
  9.1.1: Data Acquisition Tools ................................................ 159
  9.1.2: Data Analysis Tools .................................................. 161
  9.1.3: Data Visualization Tools ............................................. 161
  9.1.4: Treatment Decision Support tools ................................... 163
  9.1.5: Tools for reporting response assessment outcomes ............ 163
  9.1.6: Standards for data storage & transfer .......................... 164
9.2: Proposed system for development of response criteria ............. 164
  9.2.1: Data repository ...................................................... 164
  9.2.2: Data acquisition tools .............................................. 165
  9.2.3: Knowledge acquisition tools ...................................... 166
  9.2.4: Tools to enable experiments on response criteria .............. 166
9.3 Conclusions .................................................................. 167

REFERENCES ........................................................................ 169
LIST OF TABLES

Table                                           page
1.1: Definition of terms for Response Assessment 1
1.2: Definition of terms for Clinical Response Assessment 3
2.1: Dimensions of Domain Response Criteria 30
2.2: Finding and Clinical State Parameters for Response Criteria 32
2.3: Finding level reasoning tasks 33
2.4: Clinical state level reasoning tasks 35
2.5: Response State and Summary Statistic Parameters 36
2.6: Response state level reasoning tasks 37
2.7: Percentage of target lesions identified and longest diameter recorded 42
2.8: Percentage of CT scans with sufficient information to calculate response 42
3.1: Summary of systems that perform the response assessment task 51
3.2: Summary of data driven vs. goal driven temporal interpretation methods 60
3.3: Summary of Methods that perform Quantitative Classification 67
3.4: Summary of Methods that perform Qualitative Classification 69
3.5: Summary of Method that perform Statistical methods 72
4.1: Classification of Interpretation Task Specifications 93
5.1: Classification of Interpretation Task Specifications for Single Context Classification with a Single Input Parameter 112
5.2: Interpretation rule templates created to implement RECIST 115
6.1: Categories of Interpretation Task Specifications for RECIST 124
6.2: Ratio of rule templates to interpretation task specifications for RECIST 125
6.3: Shared Rules and Interpretation Task Specifications for RECIST and IHC 128
6.4: Number of Interpretation Tasks for Each Type of Response Parameter 132
6.5: Comparison of rule templates between RECIST and EURAL 132
6.6: Ratio of rule templates to interpretation task specifications for EURAL 132
7.1: Results of Comparison of Human and Automated Interpretation 135
7.2: Comparison of Lymphoma and RECIST criteria outcomes 142
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>The General and Clinical Response Assessment Task</td>
<td>2</td>
</tr>
<tr>
<td>1.2</td>
<td>Aggregation of cohort response into a summary statistic</td>
<td>4</td>
</tr>
<tr>
<td>1.3</td>
<td>Response Assessment Interpretation Task</td>
<td>5</td>
</tr>
<tr>
<td>1.4</td>
<td>Example of the clinical tasks related to treatment selection and assessment of clinical response</td>
<td>8</td>
</tr>
<tr>
<td>1.5</td>
<td>Response Assessment vs. Response Prediction</td>
<td>10</td>
</tr>
<tr>
<td>1.6</td>
<td>Response Assessment Method</td>
<td>11</td>
</tr>
<tr>
<td>2.1</td>
<td>Oncology Response Assessment Task</td>
<td>14</td>
</tr>
<tr>
<td>2.2</td>
<td>Timeline of publication of Oncology Response Assessment Criteria</td>
<td>16</td>
</tr>
<tr>
<td>2.3</td>
<td>Protocol of assessment events</td>
<td>17</td>
</tr>
<tr>
<td>2.4</td>
<td>Radiographic imaging study used to measure tumor lesions</td>
<td>18</td>
</tr>
<tr>
<td>2.5</td>
<td>RECIST flow sheet</td>
<td>19</td>
</tr>
<tr>
<td>2.6</td>
<td>Decision logic for baseline classification of tumor lesions</td>
<td>20</td>
</tr>
<tr>
<td>2.7</td>
<td>Calculation and Classification of Target Lesion Response</td>
<td>21</td>
</tr>
<tr>
<td>2.8</td>
<td>Classification of Follow-up Non-Target Lesion Response</td>
<td>22</td>
</tr>
<tr>
<td>2.9</td>
<td>Combined follow-up response classification</td>
<td>23</td>
</tr>
<tr>
<td>2.10</td>
<td>Waterfall plot of Best Response Rate</td>
<td>24</td>
</tr>
<tr>
<td>2.11</td>
<td>Disease Activity Score (DAS) calculation</td>
<td>26</td>
</tr>
<tr>
<td>2.12</td>
<td>EURAL thresholds for response classification</td>
<td>27</td>
</tr>
<tr>
<td>2.13</td>
<td>Examples of a clinical rule and mathematical formula for the RECIST</td>
<td>28</td>
</tr>
<tr>
<td>2.14</td>
<td>Common Features of Response Criteria</td>
<td>29</td>
</tr>
<tr>
<td>2.15</td>
<td>Interpretation Contexts for Response Assessment</td>
<td>31</td>
</tr>
<tr>
<td>2.16</td>
<td>Formulas to calculate the quantitative response state</td>
<td>38</td>
</tr>
<tr>
<td>2.17</td>
<td>Types of Qualitative Assessments of Treatment Response</td>
<td>39</td>
</tr>
<tr>
<td>3.1</td>
<td>Dimensions of Temporal Interpretation Methods</td>
<td>49</td>
</tr>
<tr>
<td>3.2</td>
<td>Shahar’s Knowledge-Based Temporal Abstraction (KBTA) Method</td>
<td>59</td>
</tr>
<tr>
<td>3.3</td>
<td>Example of a Data Driven Temporal Interpretation Task</td>
<td>62</td>
</tr>
<tr>
<td>3.4</td>
<td>Limitations of the KBTA method for context based reasoning</td>
<td>64</td>
</tr>
<tr>
<td>3.5</td>
<td>Inputs and output requirements of the response assessment method</td>
<td>65</td>
</tr>
<tr>
<td>3.6</td>
<td>Categories of Classification and Statistical Temporal Interpretation Methods</td>
<td>66</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
<td></td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>4.1: Rule-Based Response Assessment Framework</td>
<td>77</td>
<td></td>
</tr>
<tr>
<td>4.2: The Context Restriction Procedure</td>
<td>79</td>
<td></td>
</tr>
<tr>
<td>4.3: The Protocol Knowledge Model</td>
<td>81</td>
<td></td>
</tr>
<tr>
<td>4.4: Temporal Restriction Model</td>
<td>82</td>
<td></td>
</tr>
<tr>
<td>4.5: Context Specification</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td>4.6: Composite Context Specification</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td>4.7: The Clinical Data Model mapped to the Protocol Model</td>
<td>84</td>
<td></td>
</tr>
<tr>
<td>4.8: Temporal Data Model mapping to Clinical Data Model</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>4.9: Context Based Interpretation Procedure</td>
<td>88</td>
<td></td>
</tr>
<tr>
<td>4.10: Common interpretation rule patterns</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td>4.11: A rule template and its interpretation task specification</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td>4.12: Response Criteria Knowledge Model</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>4.13: Response Criteria Parameter Model</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>5.1: Ontology Reuse, Extension and Development for Context Restriction</td>
<td>108</td>
<td></td>
</tr>
<tr>
<td>5.2: Temporal Ontologies for Response Assessment Framework</td>
<td>109</td>
<td></td>
</tr>
<tr>
<td>5.3: Start Treatment Anchor Generation Rule implemented in SWRL</td>
<td>110</td>
<td></td>
</tr>
<tr>
<td>5.4: Ontology Development for the Context Based Interpretation Procedure</td>
<td>111</td>
<td></td>
</tr>
<tr>
<td>5.5: Domain independent Ontologies for the Context-Based Interpretation Task Procedure</td>
<td>111</td>
<td></td>
</tr>
<tr>
<td>5.6: Single Context Single Parameter Classification Task Specification Ontology</td>
<td>113</td>
<td></td>
</tr>
<tr>
<td>5.7: Parameter Restriction Ontology</td>
<td>114</td>
<td></td>
</tr>
<tr>
<td>5.8: Rule Template for Single Context Single Parameter Quantitative Property Relation Restriction Classification Task</td>
<td>116</td>
<td></td>
</tr>
<tr>
<td>5.9: Rule Template for Single Context Single Parameter Set Aggregation Statistical Task</td>
<td>117</td>
<td></td>
</tr>
<tr>
<td>6.1: Image Annotation Data for Response Assessment</td>
<td>120</td>
<td></td>
</tr>
<tr>
<td>6.2: Instantiation of RECIST</td>
<td>121</td>
<td></td>
</tr>
<tr>
<td>6.3: Extension of the Parameter Ontology for RECIST</td>
<td>123</td>
<td></td>
</tr>
<tr>
<td>6.4: Instantiation of the International Harmonization criteria</td>
<td>127</td>
<td></td>
</tr>
<tr>
<td>6.5: Instantiation of the EURAL Rheumatoid Arthritis criteria</td>
<td>129</td>
<td></td>
</tr>
<tr>
<td>6.6: Extensions to the Data Parameter Ontology for EURAL instantiation</td>
<td>130</td>
<td></td>
</tr>
<tr>
<td>6.7: Extensions to the Abstract Parameter Ontology for EURAL instantiation</td>
<td>131</td>
<td></td>
</tr>
<tr>
<td>7.1: Reuse of the implementation</td>
<td>138</td>
<td></td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
<td></td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>8.1: Decision-logic for classification of a follow-up Lymph Node Lesion</td>
<td>152</td>
<td></td>
</tr>
<tr>
<td>9.1: Systems to Support Application of Response Criteria</td>
<td>159</td>
<td></td>
</tr>
<tr>
<td>9.2: Anatomic Visualization of Temporal Changes in Lesion Features</td>
<td>162</td>
<td></td>
</tr>
<tr>
<td>9.3: Graphical Summary of Temporal Changes in Cancer Lesion Features</td>
<td>163</td>
<td></td>
</tr>
<tr>
<td>9.4: Systems to Support the Development of Response Criteria</td>
<td>165</td>
<td></td>
</tr>
</tbody>
</table>
CHAPTER 1
INTRODUCTION

The goal of this chapter is to briefly introduce the content of the dissertation document. This consists of an overview of the clinical activity of treatment response assessment and the motivation for an extensible and reusable rule-based approach for automated response assessment.

1.1: Response Assessment

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response Assessment Activity</td>
<td>Act of assessing how the state of an observed entity changes in response to an intervention</td>
</tr>
<tr>
<td>Observed Entity</td>
<td>Entity whose attributes can be observed</td>
</tr>
<tr>
<td>State of Observed Entity</td>
<td>Attributes of an observed entity that can be described</td>
</tr>
<tr>
<td>Intervention</td>
<td>Goal directed action with the intention to modify the state of the observed entity</td>
</tr>
<tr>
<td>Response Biomarker</td>
<td>Attribute of an entity used to assess a state change after an intervention</td>
</tr>
<tr>
<td>Baseline Assessment Period</td>
<td>Time prior to the start of intervention when state of entity is assessed</td>
</tr>
<tr>
<td>Follow-up Assessment Period</td>
<td>Time after the start of intervention when state of entity is assessed</td>
</tr>
<tr>
<td>Response Classification</td>
<td>Description of the change in state after an intervention</td>
</tr>
<tr>
<td>Response Protocol</td>
<td>Formal specification of the procedures for assessing an entities response to intervention</td>
</tr>
</tbody>
</table>

Response assessment in the most general form consists of assessing how the state of an observed entity changes in response to an intervention (Figure 1.1) (Table 1.1). The observed entity may be an organism, the environment, or even an abstract entity such as the economy. The state of the observed entity can be described through a set of observations, also called response biomarkers. An intervention is a goal directed action with the intention to modify the state of the observed entity. The state of the observed entity is assessed at some time point prior to the start of the intervention, referred to as the baseline assessment period, and then again using the same assessment technique sometime after the start of the intervention, referred to as
the follow-up assessment period. The change in the state of the observed entity from the baseline assessment to the follow-up assessment period describes the entity’s response to the intervention. The change in the state of the entity is classified based on thresholds that define responders and non-responders.

Response assessment occurs in many aspects of life, but is a more formal process when using the scientific method. Interventionsal scientific experiments in particular specify the dimensions of response assessment as part of their protocol for collecting and analyzing response biomarkers about the observed entities. The formal response assessment protocol enables consistent assessment of the response biomarkers from individual entities so that they may be aggregated with cohorts, and compared across interventions. Human interventional studies in particular require a rigorous procedure for response assessment to ensure the safety of the participants. In addition, therapeutic interventional studies in medicine require formal procedures for assessing the efficacy of a novel therapy. Formal response assessment criteria have been developed in medicine to enable consistent analysis of response to treatment within and across studies in a particular clinical domain. However, the dimensions of
response assessment in medicine can be quite complex, at times resulting in inconsistent application of the criteria. This challenge presents an opportunity for automated response assessment methods to enable consistent application of response criteria. While response assessment is a general activity, given this important clinical problem, we focus the scope of this dissertation on treatment response assessment in the medical domain.

1.2: Treatment Response Assessment in Medicine

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical Response Assessment</td>
<td>Act of assessing the clinical state of a patient changes in response to treatment</td>
</tr>
<tr>
<td>Patient</td>
<td>Observed entity whose attributes of health or disease can be observed</td>
</tr>
<tr>
<td>Clinical State</td>
<td>Quantitative and qualitative attributes of a patient’s health or disease that can be described</td>
</tr>
<tr>
<td>Treatment</td>
<td>Goal directed action with the intention to improve a patient’s morbidity or mortality</td>
</tr>
<tr>
<td>Clinical Response Biomarker</td>
<td>Patient attribute used to assess a state change after treatment</td>
</tr>
<tr>
<td>Baseline Assessment Period</td>
<td>Time prior to the start of treatment when the clinical state is assessed</td>
</tr>
<tr>
<td>Follow-up Assessment Period</td>
<td>Time after the start of treatment when the clinical state is assessed</td>
</tr>
<tr>
<td>Classification of Response to Treatment</td>
<td>Description of the change in state after a treatment that often guides therapeutic decisions</td>
</tr>
<tr>
<td>Clinical Response Criteria</td>
<td>Formal specifications of the procedure for treatment response assessment in medicine</td>
</tr>
</tbody>
</table>

Treatment response assessment is an important clinical and research activity in medicine. In medicine the observed entity is the patient, the clinical state of the patient is described with respect to some clinical response biomarkers, and the intervention is a therapeutic action with the goal of modifying the clinical state so as to improve morbidity or mortality (Figure 1.1) (Table 1.2). The clinical state is assessed at baseline prior to the start of treatment and again at follow-up some biologically appropriate time point after the start of the treatment. The change in the clinical state from the baseline to follow-up period is assessed with respect to biologically relevant
thresholds that are used to classify the nature of the response, which typically guides the next set of therapeutic decisions.

Therapeutic clinical trials are medical experiments that evaluate the efficacy and toxicity of an investigational therapy. Response criteria have been developed for clinical trial research as formal specifications of the procedure for treatment response assessment. They define the types of assessment procedures and timing of assessment that should be used to generate the response biomarkers, and define the clinical state and response state with specialized mathematical formulas and classification rules. Response criteria have three functions within clinical trial protocols: 1) response assessments are incorporated into protocol based treatment decision algorithms, 2) response assessments are aggregated into a response summary statistic that describes a cohort’s response to treatment, and 3) the response summary statistic is used to compare treatment response within and across trials (Figure 1.2). Response criteria have been developed for many clinical domains including Oncology (1,2), Infectious Diseases (3,4), Rheumatology (5), and Cardiology (6).

![Figure 1.2: Aggregation of cohort response into a summary statistic (green) that enables comparison of response outcomes within and across trials. B = Baseline Assessment Period, F1 = First Follow-up Assessment Period, F2 = Second Follow-up Assessment Period, etc. P1 = Patient 1, P2 = Patient 2, etc. The clinical state (blue) of each patient in the cohort is assessed at the baseline and follow-up time period. The response state (yellow) is described for each patient over that period of time and can be aggregated into a summary statistic (green) that describes the response of the entire cohort.](image)
Response criteria are domain specific and designed to be applied across multiple clinical trial protocols. This enables comparison of trial results both within and across clinical trials. This is particularly important for single arm trials where there is no control arm for direct comparison and a comparison to historical controls is used to assess the potential efficacy of a novel therapy. Response criteria thus define intermediate trial endpoints that can be used in trial designs to assess the efficacy of a novel therapy sooner than survival endpoints. In a recent conference at the Brookings Institution, Dr. David Kessler, former Commissioner of the Food and Drug Administration, identified the development of the HIV viral load serologic test and the validation of response criteria based on the HIV viral load (3,4) in the 1990’s as a key factor that enabled more rapid evaluation of novel HIV therapies compared to survival endpoints (7). The theory of response assessment for this surrogate endpoint is that if the viral count decreased because of a new drug, it was possible to validate improved survival. This enabled more rapid government approval for novel therapies than would have been possible if they had to wait for the survival endpoint to occur. Thus, potentially life prolonging therapies where brought to market more quickly.

Figure 1.3: Response Assessment. The response assessment interpretation method is central to the activities of response data acquisition, treatment decision support and cohort result reporting.
Consistent application of response criteria for response assessment in clinical research is central to the procedures for response data acquisition, treatment decision support and cohort result reporting (Figure 1.3). However, application of response criteria in clinical trials can be highly inconsistent. This can be seen when differences arise in the interpretation of response outcomes between the local providers managing patients in the trial and central reviewers (8). The FDA reported a 24-29% rate of discordance in response assessment at the patient level in oncology clinical trials between local investigator and central reviewer assessment of progression status (9,10). These differences at the patient level result in an average of 41% decrease in the cohort mean tumor response rate that is reported by central reviewers compared to local investigators in a review of 9 clinical trials (8). Even when the central review process is used, discordance exists between readers. One study in particular reported a rate of discordance for progression status up to 38% between multiple central reviewers (11). This lack of concordance is related in part to the complexity, lack of precision and lack of specification of response criteria (8), and demonstrates the need for automated methods to consistently apply response criteria within clinical trials.

This inconsistent application of response criteria is not only observed within trials, but across trials in clinical domains, making it difficult to compare trial results. For example, Tonkin in (12) reviewed 62 solid tumor clinical trials and found a wide variation in the reported technique for response assessment between trial designs. They also observed a wide variation in the reported rates of tumor response in colon cancer clinical trials all using the same 5-Fluorouracil treatment protocol ranging from 8% - 80%, with similarly wide ranges for head and neck and lung cancer trials. The authors conducted an analysis of the response criteria used in these trials and concluded that the use of different response criteria in clinical trials of the same disease with the same treatment contribute to the wide variation in reported rates of tumor response. The use of different response criteria often occurs when new assessment modalities and measurement techniques emerge which have not yet been incorporated into response criteria, yet are used as part of the clinical trial response assessment protocol. This poses a significant problem when trying to compare trial
results, and demonstrates the need for automated methods to consistently apply response criteria across trials.

At the same time, this variance in clinical trial response assessment protocols demonstrates the need for response criteria to evolve over time as new assessment techniques are introduced, and the need to validate these new response criteria with respect to survival endpoints. However, there is no common framework to represent the dimensions of response criteria that enables their consistent application within and across clinical trials. Yet response criteria have common features and methods within and across clinical domains that present an opportunity for modeling approaches. As such, a common framework for treatment response assessment should enable consistent application of response criteria for patients within trials, across trials within clinical domains, and experimentatation with new criteria. This dissertation describes a framework for treatment response assessment that describes the dimensions of treatment response assessment in medicine. The framework also enables automated assessment of response over patient data with the goal improving the consistency with which response is assessed for patient cohorts.

1.3: Definition of Treatment Response Assessment

Response assessment is a distinct clinical activity within the treatment decision-making process. Figure 1.4 shows an example of the flow of clinical activities related to treatment selection and response assessment for a blood clot. **Diagnosis** consists of classifying the nature of the abnormality; in this case a blood clot. **Treatment selection** consists of choosing an appropriate therapy for blood clots is chosen from amongst treatment options; in this case the blood thinner warfarin is selected. Response criteria have been developed that specify the procedure for assessing how a patient with a particular diagnosis responds to a particular therapy. Response criteria include a specification of the response assessment strategy including the assessment modality; in this case the serum INR measures the effect of warfarin on blood coagulation. They also specify the frequency of response assessment as it relates to the natural course of the disease and the expected kinetics of the treatment.
In this case, the serum INR should be rechecked two to three (13) days after the start of warfarin therapy given the relatively rapid onset of this drug’s action and narrow therapeutic window. In addition, response criteria specify how to assess treatment efficacy or toxicity. In this case, the therapeutic goal for anti-coagulation using warfarin for the treatment of blood clots is an INR in the range of 2.0 to 3.0 (13). Response criteria selection consists of selecting the appropriate response criteria based on a given diagnosis, therapy and therapeutic goal.

Figure 1.4: Example of the clinical activities related to treatment selection and assessment of clinical response. In this example, the diagnosis of a blood clot is made and warfarin anticoagulation therapy is selected. The response criteria selection activity consists of defining the modality of assessment (serum INR), the therapeutic goal (INR between 2.0 and 3.0) and the frequency of assessment (every 2 days). Response prediction estimates that a daily dose of 5 mg of warfarin will result in an INR within the therapeutic range and thus this dose is selected. However, in the actual follow-up assessment, the actual INR is supra therapeutic when compared to the goal state. This results in a reassessment of the therapy and dose selection.
Once the therapy and response criteria have been selected, the starting dose must be selected. *Dose selection* consists of *response prediction* and *plan selection*. *Response prediction* evaluates the current state of the patient, the proposed treatment, and utilizes a nomogram to predict the future state of the patient. Several nomograms have been developed to predict the effect of warfarin given various patient features, such as patient’s age, gender, weight, or genome (14). In *Plan selection* the clinician chooses the plan that will most likely achieve the therapeutic goal. In this case, the nomogram predicts that an initial warfarin dose of 5 mg daily will result in an INR of 2.4. In this example, the 5 mg dose is selected since it is predicted to achieve the therapeutic goal.

After the treatment has started, the response is assessed according to the procedure specified in the response criteria. In this case, two days after the start of warfarin at 5 mg daily the INR is checked and found to be 3.3. *Response assessment* classifies the actual follow-up state with respect to the therapeutic goal. In this case, the INR demonstrates that a daily dose of 5 mg of warfarin results in a supra-therapeutic state for this patient. Since the treatment goal is not achieved, the cyclical process of dose selection followed by response assessment continues until the goal is achieved. If on the other hand it is found that the chosen therapy is not capable of achieving the therapeutic goal, then a new therapy is selected.

From a computational perspective, we define response assessment as a type of goal-driven temporal reasoning that creates interpretations of time-oriented data. Temporal reasoning includes several classes of generic methods including interpretation, projection, forecasting and planning (15). *Interpretation* involves reasoning only about past and present data, and not about the future. This is in contrast with *Projection* and *Forecasting* that predict the future values of parameters given past values. *Planning* methods on the other hand, produce a recommended sequence of actions that are anticipated will achieve the goal state given the initial state (16). From this perspective, *response assessment* is a type of interpretation method in that it reasons only about the past up to the present (Figure 1.5). *Diagnosis* is also an interpretation method that explains a set of findings, but is distinct from
response assessment in that it does not explain the finding with respect to the treatment events or treatment goals. These interpretation methods are also distinct from response prediction, a type of projection method where the future value of the clinical state is predicted from various patient features and the type of treatment. Furthermore, the response assessment is also distinct from treatment selection, dose selection, and response criteria selection methods, which are types of planning methods (Figure 1.4).

Within the medical domain, temporal interpretation methods can be further classified as abstraction, monitoring or diagnosis methods. Temporal abstraction has traditionally been defined as a procedure that transforms data from low-level quantitative form to high-level qualitative descriptions (17). This definition is limited with respect to response assessment in that it may take as input qualitative or quantitative data and generate qualitative or quantitative interpretations. The traditional conceptualization of clinical monitoring involves the interpretation of high frequency physiologic data streams in the ICU setting (17) that often lack a treatment data stream. The treatment interpretation context is central to response assessment, and as such we distinguish response assessment from the monitoring. The treatment

Figure 1.5: Response Assessment vs. Response Prediction. Response assessment is an Interpretation method that reasons over past clinical data up to the present. This is distinct from Response Prediction, a Projection method that tries to predict the future state from past states and other patient features. Colored boxes represent historical clinical events while the uncolored Clinical State box represents a future predicted state that has yet to be observed.
interpretation context further specializes response assessment as a type of context-based interpretation method where the reasoning is explicitly restricted to specific contexts as opposed to implied. Lastly, response assessment requires knowledge of the response criteria to perform the interpretation, and generates interpretations about the patient’s clinical state and response state, as well as aggregated summary statistics about the patient cohort.

We therefore frame response assessment as a type of goal-driven, knowledge-driven, and context-based interpretation method most similar to temporal abstraction, however requiring extensions to the traditional conceptualization. These extensions include a focus on procedure-oriented reasoning, the addition of statistical interpretation techniques to the traditional classification methods, as well as cohort level interpretations. Response assessment as applied to patient cohorts in clinical trials takes as input time stamped patient data, the protocol for response assessment for the clinical trial, and knowledge of the response criteria, and generates as output qualitative and quantitative interpretations regarding the clinical state, response state, and cohort summary statistic (Figure 1.6).

I thus hypothesize that a rule-based framework for response assessment will enable an extensible, reusable method for representing, applying and comparing response criteria.

![Figure 1.6: Response Assessment Method](image)

Figure 1.6: Response Assessment Method. The Response Assessment Method takes as input time stamped patient data, the clinical trial response assessment protocol, and knowledge of the Response Criteria, and generates as output time oriented qualitative and quantitative interpretations.
1.4: Overview of Dissertation Document by Chapter

Chapter 2 presents an analysis of the common features of clinical trial response criteria, an extended example of the application of response criteria in the oncology domain, current approaches and limitations to the application and development of response criteria, and the desiderata for an automated response assessment framework. Chapter 3 discusses and critiques prior work in temporal interpretation methods from the perspective of the response assessment desiderata.

Chapter 4 defines the Rule-Based Response Assessment Framework including the temporal, data and knowledge models required to encode response criteria and the reasoning methods required to apply response criteria to patient data. Chapter 5 describes the knowledge representation implementation of the Rule-Based Response Assessment Framework using ontologies and rules. Chapter 6 describes the application of the framework to two response criteria in the oncology domain and one response criterion in the rheumatoid arthritis domain.

Chapter 7 describes the evaluation of the framework for its 1) expressivity, 2) clinical accuracy, 3) reusability for multiple data sets, 4) extensibility as applied to multiple response criteria, and 5) ability to enable comparison of the knowledge level and outcomes of two response criteria in the same domain. Chapter 8 summarizes the informatics and clinical contributions and limitations of this work, and Chapter 9 discusses future research directions and conclusions.
CHAPTER 2
ANALYSIS OF RESPONSE ASSESSMENT CRITERIA

This chapter presents a review of response criteria from several clinical domains, an analysis of the common features of response criteria, limitations of the current clinical research workflow and information artifacts associated with the development and application of response criteria, and the desiderata for an automated response assessment method.

2.1: Domain Specific Response Assessment Criteria for Clinical Trials

2.1.1: Oncology Response Criteria

Objective criteria have been developed to standardize tumor and host response to treatment in cancer clinical trials. Although a complete history of the evolution of oncology response criteria is beyond the scope of this dissertation, a brief historical overview is useful for illustrating their significance and complexity. Due to the relative complexity of oncology response criteria that assess tumor response to treatment, they serve as a good example of the range of interpretation tasks required for response assessment. As such, it is useful to present a detailed example of the application of an oncology response criterion to a clinical trial patient and cohort. This example will serve as the primary descriptive use case throughout the remainder of this dissertation.

Novel cancer therapies are often first evaluated in the metastatic or advanced disease setting for their anti-tumor activity. The primary therapeutic goal for many anti-tumor treatments in the advanced disease setting is delay in tumor growth and ideally tumor shrinkage. Delay in tumor growth can correlate with improved quality of life, morbidity, and mortality (18,19). Historically, tumor shrinkage has been the hallmark of anti-tumor activity for cytotoxic therapies, which cause tumor cell death and thus have the potential to shrink tumor masses. Tumor shrinkage in a proportion of patients in phase II studies of cytotoxic drugs has been shown to be predictive of improvement in survival in phase III studies (19-21). Non-cytotoxic therapies on the other hand are typically cytostatic, and may not cause tumor shrinkage but rather
tumor stability. Several non-cytotoxic therapies have also demonstrated improvement in overall survival in randomized trials (22,23). For such cases, delay in tumor growth can also be used as evidence of anti-tumor activity (24). The time to objective tumor growth is referred to as the time to progression (TTP) and is often a primary or secondary endpoint for phase II and III oncology clinical trials.

Figure 2.1: Overview of Response Assessment in Oncology. The disease state is the tumor burden, the goal of the intervention is to delay tumor growth, and response is classified with respect to that goal.

In oncology response assessment (Figure 2.1), the clinical state is an estimate of the tumor burden, the intervention is some form of systemic or local therapy, and response is assessed with respect to the degree of tumor shrinkage or growth. Oncology response criteria have been developed to standardize the approach for estimating tumor burden, defining quantitative and qualitative changes in tumor burden, and classifying tumor response to treatment in clinical trial cohorts.

The first oncology response criterion called the World Health Organization (WHO) criteria was published in 1981 (25). The goal of the WHO criteria was to develop a response criterion that could objectively assess tumor response to treatment as compared to previous approaches that utilized clinical symptoms to assess treatment
response. While clinical symptoms of disease are an important aspect of assessing individual patient responses, they tend to be subjective assessments and not all cancer patients have such symptoms. The WHO criteria thus recommended evaluating tumor lesions with radiographic modalities, taking bi-dimensional tumor measurements of observed cancer lesions in the images, and taking the sum of the products to produce a quantitative estimate of tumor burden. This was done at baseline and again at follow-up. The quantitative response was calculated as the percent change in the tumor burden (sum of products) taking as reference the baseline assessment.

\[
\text{Response Rate} = \frac{(\text{Tumor Burden}_{\text{Follow-up}} - \text{Tumor Burden}_{\text{Baseline}})}{\text{Tumor Burden}_{\text{Baseline}}}
\]

The quantitative response was calculated at each follow-up assessment period and the minimum tumor burden observed since baseline was used to calculate the best response rate. This quantitative assessment was used to classify the nature of the patient’s response into four response categories based on heuristically defined thresholds. The response categories defined and in use to this day, are Complete Response, Partial Response, Stable Disease, and Disease Progression. While qualitative descriptions, the response categories are a form of ordinal scale. The criteria also specified observations including the appearance of any new lesions as defining the event of disease progression.

The WHO criteria were widely used in cancer clinical trial protocols. However, the designers of clinical trials began to make modifications to the criteria on an ad hoc basis to incorporate new imaging technologies and address underspecified aspects of the original document. As a result, there was a lack of standardization and it became difficult again to consistently confirm trial results and compare trial outcomes (12,26). In the mid 1990’s an international, multidisciplinary committee called the International Working Party was established to simplify and standardize the criteria. In 2000, the committee published the Response Evaluation Criteria in Solid Tumors (RECIST) guidelines (27), the next generation of solid tumor response criteria. RECIST had some significant differences from the original WHO criteria. First, it defined more specifically what was considered measurable disease both by the
anatomic location of the lesion and by a minimum size requirement. Second, up to 5 lesions per organ could be selected with a maximum of 10 lesions total to estimate tumor burden. Lastly, the criteria utilized uni-dimensional measurement of lesions, and the sum of longest diameters to estimate total anatomic tumor burden. As a result of changing from bi-dimensional to uni-dimensional measurement, the thresholds for defining response categories were also changed.

Over the last decade, RECIST has been widely utilized in cancer clinical trials and has become a “requirement” for government regulatory approval of drugs in certain contexts (2). However, just as with the original WHO criteria, after the continued application of RECIST in more settings, and the development of new imaging technologies such as metabolic imaging, it became necessary to revise the RECIST criteria. RECIST 1.1 (2) published is 2009, is so named because it is more of an evolution of the guideline than a complete departure. RECIST 1.1 further specifies and in some ways simplifies the criteria. Major changes include specific definitions for what is considered measurable with respect to lymph nodes, and a decrease in the number of target lesions used to estimate tumor burden from 10 total to 5 total with a
maximum of 2 lesions per organ. It is anticipated that RECIST 1.1 will become the new standard for response assessment for solid tumor clinical trials, as it has been endorsed by the several international cancer centers and government regulatory agencies.

However, RECIST is not an appropriate response criterion for every type of cancer (Figure 2.2). Two generations of response criteria have also been developed for lymphoma (1,29), prostate cancer (30,31), and intracranial neoplasms (32,33) to accommodate novel assessment modalities and disease characteristics. In addition, the solid tumor criteria have not been found to be sufficient for response assessment for the very rare gastrointestinal stromal tumor (GIST) (35,36) and mesothelioma (37) neoplasms. As such, response criteria have been developed for these specific diseases (32,38). Given historical trends and the continued development and validation of new response biomarkers and criteria, it is anticipated that response assessment criteria in oncology will continue to evolve.

2.1.1.1: Application of RECIST

<table>
<thead>
<tr>
<th>Assessment Modality</th>
<th>Timing of Response Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
</tr>
<tr>
<td>H&amp;P</td>
<td>X</td>
</tr>
<tr>
<td>Lab 1</td>
<td>X</td>
</tr>
<tr>
<td>Lab 2</td>
<td>X</td>
</tr>
<tr>
<td>Quality of Life Survey</td>
<td>X</td>
</tr>
<tr>
<td>CT</td>
<td>X</td>
</tr>
</tbody>
</table>

Figure 2.3 Protocol of assessment events. The rows detail the planned assessment events and the columns detail the timing of those events. An “X” indicates when each event should occur in the schedule. Relative times are used to describe the timing of events with B referring to Baseline, F1, F2, etc. referring to follow-up periods. The protocol defines the range of valid relative times related to the start of treatment for each time period.

A primary use case for the Rule-Based Response Assessment Framework in this dissertation will be assessment of tumor response to treatment for an oncology
clinical trial cohort using RECIST 1.1 (here after referred to as RECIST). As such, it is useful to describe a concrete example of how RECIST is applied to patients and cohorts during the prospective management and analysis of clinical trial data. The RECIST procedure includes the process for estimating tumor burden, estimating qualitative and quantitative changes in tumor burden, classification of response at each follow-up time point and over the entire course of treatment, and aggregate estimates of cohort response in summary statistics.

Figure 2.4: Radiographic imaging study used to measure tumor lesions at baseline and again at follow-up. The green lines define the longest diameter measurement for the cancer lesion in the liver and its perpendicular short axis.

Clinical trial research begins with the creation of a clinical trial protocol. Within the protocol are details of the eligibility criteria, treatment schedules and response assessment plan. Solid tumor oncology clinical trials use the RECIST criteria as a guide for specifying the type of assessment modalities, measurement techniques and timing of response assessment for a given protocol. This plan of assessment events is detailed in the protocol document as shown in Figure 2.3 where relative times are given for the frequency of assessment events. The timing of the assessment events is typically given in relationship to treatment events. RECIST recommends a baseline assessment no earlier than 4 weeks prior to the start of treatment, and follow-up assessments approximately every 6-8 weeks depending upon the type of solid tumor and anticipated kinetics of the experimental treatment.
RECIST recommends that solid tumors be assessed with imaging modalities, typically CT scans or MRI, PET or Chest X-ray when appropriate.

<table>
<thead>
<tr>
<th>RECIST flow sheet</th>
<th>LL, MR#123456789</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lesion ID</strong></td>
<td><strong>Description/Location</strong></td>
</tr>
<tr>
<td>Target Lesions</td>
<td>liver mass</td>
</tr>
<tr>
<td></td>
<td>liver mass</td>
</tr>
<tr>
<td>Non-target Lesions</td>
<td>liver mass</td>
</tr>
<tr>
<td></td>
<td>T4 bone lesion</td>
</tr>
<tr>
<td></td>
<td>8th rib lesion</td>
</tr>
<tr>
<td></td>
<td>Left pleural effusion</td>
</tr>
<tr>
<td>New Lesions</td>
<td>liver mass</td>
</tr>
<tr>
<td>Response</td>
<td>Sum of Diameters (SD)</td>
</tr>
<tr>
<td></td>
<td>% Change SD from Baseline</td>
</tr>
<tr>
<td></td>
<td>% Change SD from Nadir</td>
</tr>
<tr>
<td></td>
<td>Target Lesion Response</td>
</tr>
<tr>
<td></td>
<td>Non-target Lesion Response</td>
</tr>
<tr>
<td></td>
<td>Response</td>
</tr>
<tr>
<td></td>
<td>Overall Best Response</td>
</tr>
</tbody>
</table>

Figure 2.5: RECIST flow sheet: Example RECIST flow sheet for patient LL. MR# = Medical Record Number, B=Baseline, F1= First Follow-up, F2 = Second Follow-up, F3 = Third Follow-up, NA=Not Applicable, (+) = present, (-) = absent, SD = Sum of Diameters, PR = Partial Response, PD = Progressive Disease, non-PD/non-CR = Not-Progressive Disease/Not-Complete Response.

Once the patient is enrolled in the trial and prior to starting treatment, a baseline imaging study is acquired to assess the locations and sizes of identified tumor lesions. In the current clinical workflow, the oncologist orders an imaging study, the patient has the study performed and the radiologist is first to review the images. The radiologist summarizes their findings in a text report and records detailed measurements as image markups. Figure 2.4 shows a baseline CT scan of the liver of a patient with metastatic colon cancer. The green markings are an example of the image markups created by the radiologist that measures the longest diameter and its respective perpendicular (short) axis. The report is then sent to the oncologist who independently reviews the report and images. Tumor burden is assessed both quantitatively and qualitatively depending upon the location of the tumor lesion and
the type of imaging modality. At baseline, all identified tumor lesions are recorded in a RECIST flow sheet (Figure 2.5), and given a lesion identifier (Lesion ID) with a corresponding description of the type of lesion and its location. Quantitative measurements are recorded where applicable; otherwise a boolean description is made denoting the persistence or resolution of a lesion. RECIST flow sheets are often managed with paper forms or excel spreadsheets.

![Baseline Classification of Tumor Lesions](image)

Figure 2.6: Decision logic for baseline classification of tumor lesions as measurable versus non-measurable, and target versus non-target lesions

Tumor lesions are further classified based on their anatomic location and length measurements (Figure 2.6). RECIST 1.1 first distinguishes between lymph node and non-lymph node cancer lesions. Generally, cancer lesions assessed by CT are considered measurable at baseline if they are at least 10 millimeters in longest diameter, or at least 15 mm in short axis for lymph nodes. Non-measurable disease includes small lesions not meeting size criteria, bone lesions, fluid collections,
leptomeningeal disease, and lymphangitic involvement of skin or lung. A subset of the measurable lesions is selected to be included as target lesions for quantitative calculation of tumor burden. This same set of target lesions will be tracked in follow-up for quantitative calculation of the change in tumor burden. For RECIST 1.1, the largest two lesions per organ are selected as target lesions up to a total of five lesions. The remaining measurable and non-measurable lesions make up a qualitative estimate of tumor burden. Each cancer lesion is tracked at follow-up and classified according to their persistence or resolution. The quantitative tumor burden is estimated by the sum of the diameters of the target lesions, using the short axis for enlarged lymph nodes and the long axis for all other cancer lesions.

Figure 2.7: Calculation and Classification of Target Lesion Response. \( t_0 \) = baseline assessment time point, \( t_i \) = most recent follow-up assessment time point.
The quantitative and qualitative tumor burden is estimated at baseline and again at each follow-up period. Any new lesions are also noted. Quantitative changes in the tumor burden are estimated thru the calculation of the percent change in the sum of diameters from baseline, as well as the percent change from the smallest sum of diameters. These two parameters enable classification of the target lesion response according to defined quantitative thresholds (Figure 2.7). For the example case in Figure 2.5, the baseline sum of diameters is 5.5 cm. At the first follow-up the percent change from baseline is -45% indicating tumor shrinkage and giving a response classification of partial response (PR). The patient would stay on therapy at that time given demonstration of its efficacy. At the third follow-up assessment however, the percent change from the smallest sum of diameters has increased by 32%, denoting an increase in tumor burden that meets the threshold for classification of progressive disease (PD).

Figure 2.8: Classification of Follow-up Non-Target Lesion Response. CR = Complete Response, Non-CR/Non-PD = Non-Complete Response/Non-Progressive Disease, Unequivocal PD = Unequivocal Progressive Disease

The non-target lesion response is classified according to qualitative assessments of non-target lesion persistence, resolution, and progression (Figure 2.8). Figure 2.9 shows the combined response classification taking into account the
response for target and non-target lesions, and the appearance of new lesions. This patient would have also met criteria for progressive disease at the third follow-up period based on the presence of new lesions. At that point the patient would be taken off therapy given the assumption this particular treatment is no longer efficacious. The time to disease progression (TTP) is then calculated as the temporal interval between the start of the investigational therapy and the date when progressive disease was first documented. The primary study outcome, best overall response is calculated from the smallest sum of diameters achieved during the study, in this case at the second follow-up time point with an objective response rate (RR) of minus 49% and a response classification of partial response.

Figure 2.9: Combined follow-up response classification taking into account the response for target and non-target lesions, and the appearance of new lesions.

This procedure of tumor response assessment occurs for each patient in the trial cohort. The cohort’s mean response rate and median time to progression are typically compared to a historical control for single arm studies, or to the control arm in multi-arm randomized studies. These are some of the intermediate outcomes used as primary endpoints to assess the efficacy of a novel therapy. Other types of
aggregated temporal outcomes include the proportion of patients that experience disease progression or recurrence by a certain period of time, for example progression free survival at six months. Quantitative response rates can also be presented as a waterfall plot (Figure 2.10), visualizing response trends for a population cohort. Interpretation of the cohort response outcomes as a positive or negative trial result depends upon the specific disease and primary hypothesis of the trial.

The procedure just described for treatment response assessment of a clinical trial cohort using standardized response criteria is quite complex. Complex procedures often lead to errors or inconsistent interpretation as has previously been described. However, this procedure is relatively well specified through the definition of a set of response parameters, and the rules and mathematical formulae that describe them. In the following section we more briefly describe response assessment procedures for the HIV and Rheumatoid Arthritis domains. We then look for common features of response criteria across domains that inform modeling approaches.

![Response Rate - RECIST](image)

**Figure 2.10:** Waterfall plot of Best Response Rate. The best percent change in the tumor burden from baseline (y-axis) is plotted for each patient in the cohort (x-axis).

### 2.1.2: HIV Response Criteria

Just as in oncology clinical trials, HIV clinical trials have developed intermediate endpoints for assessing disease response to treatment. Several response biomarkers have been used in HIV response assessment including white blood cell counts and CD4 counts which inversely correlate with HIV disease burden. However,
these were replaced by the HIV viral load test as the primary estimate of disease burden in HIV given its strong correlation with overall survival (3,4). Compared to the Sum of Longest Diameters for RECIST, HIV viral load is a much simpler estimate of disease burden, but correlates more closely with patient outcomes in HIV. Changes in the HIV viral load over time are used to assess patient’s response to combination anti-viral regimens and have dramatically decreased the time to market for HIV therapeutics compared to the previous standard of overall survival.

After the development and validation of the HIV viral load serologic test in the 1990’s (3,4), various statistical methods were applied in HIV clinical trials to evaluate how changes in the HIV viral load with therapy predict response. Statistical methods used included absolute change of HIV-1 RNA at the end of follow-up, area-under-the-curve of HIV-1 RNA during follow-up minus baseline, HIV-1 RNA undetectability, HIV-1 RNA reduction at the end of follow-up with censoring adjustment, and mixed model analysis of HIV-1 RNA (39). The method that was best able to differentiate response in clinical trial arms was the percentage of patients with an undetectable viral load at a certain time point. Later research found that the percentage of undetectable cases at 24 weeks most closely correlated with survival endpoints. The 24-week HIV viral load value is now the standard for most HIV anti-viral trials, but variations on the approach continue to emerge.

2.1.3: Rheumatology Response Criteria

Composite measures of disease activity in rheumatoid arthritis were developed in the 1990’s to help standardize the response assessment approach for new disease modifying anti-rheumatic drugs (DMARDs) and other biologic agents. The two most common response criteria in use in clinical trials today are the American College of Rheumatology 20/50/70% improvement criteria (35) and the European League of Associations for Rheumatology (EULAR) improvement criteria (40). Like cancer, rheumatoid arthritis is a systemic disease involving multiple sites. As such disease activity cannot be expressed with a single observation and thus the Disease Activity Score (DAS) was developed combining several features of disease defined by the complex mathematical formula in Figure 2.11 (41). The Disease Activity Score
parameters include: 1) the Ritchie Articular Index (RAI), a count of the number of tender joints found on physical examination, 2) the SJC44, a swollen joint count for 44 joints, 3) the Erythrocyte Sedimentation Rate (ESR) blood test, and 4) the Global Health score, a clinical history finding describing the patient’s overall impression of their health. The DAS is calculated at baseline and at each follow-up assessment period.

\[
\text{DAS} = 0.53938 \times \sqrt{\text{RAI}} + 0.6465 \times \text{SJC44} + 0.33 \ln(\text{ESR}) + 0.00722 \times \text{GH}
\]

Figure 2.11: Disease Activity Score (DAS) calculation (41). RAI=Ritchie Articular Index, sqrt = square root, SJC44 = swollen joint count for 44 joints, ln = natural logarithm, ESR = Erythrocyte Sedimentation Rate, GH = Global Health score.

Disease response is categorized in ACR and EURAL response criteria through classification of quantitative changes in DAS from the baseline to follow-up assessment periods. In the ACR criteria, the percent change in DAS from baseline to follow-up is classified into four response categories, where a less than 20% change is a non-responder, and those with greater than 20% change are considered responders and further classified as 20%, 50% and 70% responders (42). In the EULAR criteria, response is classified based on the reached value of the DAS at the follow-up assessment period, and the absolute difference in the DAS from baseline to follow-up (40). The EURAL criterion defines thresholds for the reached DAS value and the absolute difference in DAS that define nine subsets of response (Figure 2.12). These nine subsets are further classified into three categories of response: good response, moderate response and no response. Summary statistics include a count of patients in each of the response categories.

Comparisons have been made between the outcomes of the ACR and EURAL criteria that show a high degree of concordance in their overall response classification (43). Over the years, several variations have been proposed for calculating the Disease Activity Score (DAS) used in both the ACR and EURAL response criteria. Proposed variations include simplification of the joint assessments through physical examination of 28 instead of 44 joints, the use of the C-Reactive Protein (CRP) blood
test instead of the Erythrocyte Sedimentation Rate (ESR), and the inclusion of a physician as well as a patient global health assessment score. Each variation of the DAS has its respective variation in thresholds to define the nine subsets for further response classification. Like the oncology and HIV response criteria, it is anticipated that the rheumatoid arthritis response criteria will continue to evolve with the emergence of new disease assessment modalities.

Figure 2.12: EURAL thresholds for response create nine subsets of response that are further classified into Good Response (white), Moderate Response (blue), and No Response (Yellow).

2.2: Common Features of Response Criteria

Response criteria across clinical domains specify the parameters and tasks used to estimate 1) the clinical state, 2) the response state, and 3) the summary statistics (Figure 2.14). An analysis of response criteria in the above domains demonstrates some common features with respect to defining these parameters and tasks. Parameters can be defined as attributes that have a specific range of values and a respective data type. Parameters are the inputs and outputs to the clinical rules and formulae that are chained together to define the response assessment procedure for a given response criteria. Clinical rules often take on the form of if-then statements such as shown in Figure 2.13 where a baseline cancer lesion with a long axis greater-than-or-equal-to 10 mm is classified as a measurable cancer lesion. These rules can
then be chained together as shown in the decision logic diagram in Figure 2.6 that defines a set of chained rules for classifying a baseline tumor lesion as a target or non-target lesion in the RECIST criteria. Statistical procedures are also integrated into the decision logic and include mathematical formulae such as the Disease Activity Score for rheumatoid arthritis (Figure 2.11), and the Sum of Diameters (Figure 2.13) and the Percent Change in Tumor Burden for RECIST (Figure 2.7).

The following sections describe the parameters, clinical rules and formulae that define the clinical state, response state and summary statistics for each of the domain response criteria.

Figure 2.13: Examples of a clinical rule and mathematical formula for the RECIST criteria.

2.2.1: Defining the Clinical State in Response Criteria

Response criteria specify the qualitative and quantitative features of disease to be tracked over time that make up an estimate of the clinical state for a given condition. Response criteria specify the method of disease assessment including the timing of response assessment, types of assessment modalities, measurement techniques, and measurement data types. The clinical state may be estimated qualitatively or quantitatively and is defined through a set of reasoning tasks and parameters.

2.2.1.1: Assessment Modality

Response assessment criteria specify the types of modalities that may be used to assess the clinical state. These modalities are typical of disease assessment in general clinical practice and range from history taking, physical examination, laboratory and imaging modalities. Within each category are subclasses of disease
assessment modalities such as specific types of laboratory or imaging studies. Ideally, response criteria specify a modality of disease assessment that is widely available and reproducible. Table 2.1 shows the assessment modalities that are used for each of the domain response criteria described above.

![Figure 2.14: Common Features of Response Criteria. The estimate of clinical state is used to derive the estimate of the response state and in turn the cohort aggregated summary statistic.](image)

### 2.2.1.2: Measurement Technique

For a given assessment modality there may be several types of disease measurement techniques. This is especially true of history, physical exam and imaging modalities, where multiple measurement techniques can be used to assess various features of the disease. For instance, within imaging modalities, measurement techniques of tumor lesions include length and volume assessments, or simple binary assessments of the presence or absence of a particular finding. Ideally, measurement techniques should be reproducible between observers and an accurate estimate of disease. Table 2.1 describes the assessment modality and measurement technique for each of the domain response criteria.

### 2.2.1.3: Measurement Data Type

For each type of measurement technique there is a respective data type and range of acceptable values that specifies the possible results of the measurement
activity. Quantitative techniques have either continuous, ordinal, or boolean data types. Many qualitative techniques also map to ordinal or boolean data types.

Table 2.1: Dimensions of Domain Response Criteria for Oncology (RECIST & IHC), HIV and Rheumatoid Arthritis.

<table>
<thead>
<tr>
<th>Features of Criteria</th>
<th>RECIST</th>
<th>IHC</th>
<th>HIV</th>
<th>Rheumatoid Arthritis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assessment Modality</td>
<td>Imaging (CT/MRI), Physical Exam</td>
<td>Imaging (CT, PET)</td>
<td>Laboratory (Viral PCR)</td>
<td>Joint Exam, Clinical History, Laboratory</td>
</tr>
<tr>
<td>Measurement Technique</td>
<td>Uni-dimensional Length</td>
<td>Bi-dimensional Length</td>
<td>HIV Viral Load by PCR</td>
<td>Joint physical exam for tenderness &amp; swelling, ESR</td>
</tr>
<tr>
<td>Timing of Baseline Assessment</td>
<td>Within 4 weeks start of treatment</td>
<td>Within 4 weeks start of treatment</td>
<td>not specified</td>
<td>not specified</td>
</tr>
<tr>
<td>Timing of Follow-up Assessment</td>
<td>Every 6-8 weeks</td>
<td>During Treatment: variable, minimum every 3 months After Tx Complete: minimum every 3 months for 2 years, then every 6 months for 3 years, and then annually for at least 5 years</td>
<td>24 weeks</td>
<td>Every 12 weeks</td>
</tr>
</tbody>
</table>

2.2.1.4: Timing of Response Assessment

In addition to specifying the assessment modality and measurement techniques, response criteria define the timing of response assessment before, during and after the treatment intervention. The frequency of response assessment is dependent upon the underlying biology of the disease and the anticipated change in disease with a given treatment. Response criteria may be very specific with respect to the timing requirements, or give more general recommendations that can be more loosely interpreted by the designers of a particular clinical trial protocol. Response assessment in clinical trials takes place at pre-defined time points, usually defined by a temporal range. For example, the first follow-up assessment period could take place 6-8 weeks after the start of treatment. This temporal interval is also referred to as the
temporal context in which response data should be acquired and limits the interpretation of response to pre-defined time periods.

Figure 2.15: Interpretation Contexts for Response Assessment. A: The Follow-up Assessment Context for response assessment is limited to the time between two different treatment protocols. B: Within the Baseline Assessment Context, the response data to be used to estimate the clinical state is restricted to that which occurs closest to the end of the context period. C & D: The relationship between multiple assessment periods to be used to determine the response state.

Typically, treatment events define the timing of response assessment periods. Since the goal of response assessment is to measure the effect of a treatment independent of prior or subsequent treatments, the scope of response assessment with respect to direct effects is typically limited to the time between two consecutive and distinct treatment plans (Figure 2.15A). On the other hand, late effects such as death are not limited by this constraint. Furthermore, response criteria also limit response interpretation to certain data that fall within a given context. For example, multiple assessments may occur within the baseline assessment context, but only the assessment performed closest to the end of the baseline context (i.e. closest to the start of treatment) will be used as the baseline assessment data (Figure 2.15B). In addition, some response criteria assess response across assessment periods, for instance when identifying the smallest disease burden over the course of therapy (Figure 2.15C).
Response may also be assessed with data from a single assessment period compared to data from multiple assessment periods (Figure 2.15D). Response criteria thus 1) define the timing of assessment periods with respect to treatment events, 2) define temporal restrictions on which data within an assessment period should be used for response assessment, 3) and define the relationships between multiple assessment periods with respect to determining the response state.

2.2.1.5: Classification and Calculation of Findings

Response criteria may specify several types of assessment modalities, measurement techniques and measurement data types. As such, there may be several types of quantitative and qualitative measurements of disease. Response criteria may further classify these findings for inclusion in the quantitative or the qualitative components of the composite estimate of the clinical state. Response criteria specify these finding level classification and calculation procedures in textual documents as prose that describes a set of rules and mathematical formula.

Table 2.2: Finding and Clinical State Parameters for Response Criteria in Oncology, HIV and Rheumatoid Arthritis.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Domain Response Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RECIST</td>
</tr>
<tr>
<td>Finding Level</td>
<td>Measurable/Non-</td>
</tr>
<tr>
<td></td>
<td>Measurable, Target/Non-</td>
</tr>
<tr>
<td></td>
<td>Target</td>
</tr>
<tr>
<td></td>
<td>Target</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Quantitative</td>
<td>Sum of Diameters</td>
</tr>
<tr>
<td>Clinical State</td>
<td>Target Lesions</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Qualitative</td>
<td>Lesion Persistent/</td>
</tr>
<tr>
<td>Clinical State</td>
<td>Resolution Status</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In the RECIST criteria for example, an *Image Finding* with an *ANATOMIC LOCATION* of axillary lymph node will be classified as a *Lymph Node Finding*. This is a type of simple qualitative classification where a qualitative property restriction such as a particular anatomic location is used to further classify the finding. Similarly, a Baseline *Lymph Node Finding* with a *LONGEST DIAMETER VALUE* greater-than 10 mm will be classified as *Measurable Disease*. This is a simple type of quantitative
classification utilizing the quantitative property of length measurement and the relation restriction greater-than to classify a finding.

Table 2.3: Finding level reasoning tasks: single and multiple context classification and statistical tasks

<table>
<thead>
<tr>
<th>Finding Level Reasoning</th>
<th>RECIST</th>
<th>IHC</th>
<th>HIV</th>
<th>EURAL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Set Aggregation Task</strong></td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>SW44 = Count (Set of Swollen Joints)</td>
</tr>
<tr>
<td><strong>Mathematical Formula Task</strong></td>
<td>none</td>
<td>Image Finding Product of Short and Long Axis = Length Long Axis Measurement * Length Short Axis Measurement</td>
<td>none</td>
<td>none</td>
</tr>
<tr>
<td><strong>Qualitative Restriction Classification Task</strong></td>
<td>If Baseline Normal Lymph Node and Follow-up Enlarged Lymph Node the Follow-up Lesion is New Lesion</td>
<td>If Baseline Enlarged Lymph Node and Follow-up Normal Size Lymph Node then Then Lesion Resolution Status is Resolved</td>
<td>none</td>
<td>None</td>
</tr>
<tr>
<td><strong>Mathematical Formula Task</strong></td>
<td>Absolute difference length of the longest diameter of a cancer lesion from baseline to follow-up</td>
<td>Percent change in product of the short and long axis of a lymph node lesion from baseline to follow-up</td>
<td>none</td>
<td>None</td>
</tr>
</tbody>
</table>

Response criteria also specify statistical tasks that create new quantitative finding properties or a new aggregated finding. For example, the IHC lymphoma criteria create a new cancer lesion property through the calculation of the product of the short and the long length measurements. This is a simple type of mathematical formula statistical task. The EURAL rheumatoid arthritis criteria on the other hand,
create a new aggregated finding parameter SJC44 by counting the set of swollen joints. This is a type of set aggregation statistical task.

Finding level classification and statistical tasks typically perform reasoning at a single response assessment time point. However, some finding level tasks perform reasoning over multiple response assessment time point. We refer to these as single context or multiple context reasoning tasks. For example, in the RECIST criteria, if a baseline Normal Lymph Node is found to be an Enlarged Lymph Node in the follow-up assessment period, then the follow-up lesion is classified as a New Cancer Lesion. Table 2.2 shows examples of finding level parameters and Table 2.3 shows examples of finding level interpretation tasks for each of the domain response criteria. The HIV response criterion is so simple that it does not contain any finding level parameters or reasoning tasks.

2.2.1.6: Quantitative Estimate of the Clinical State

A quantitative estimate of the clinical state has a quantitative data type. The clinical state may be estimated through the measurement of a single quantifiable clinical finding or it may be a composite of the features of several findings. In the second case, response criteria specify how these features are aggregated to create quantitative and qualitative estimates of the clinical state. Common approaches to generating an estimate of the clinical state include: 1) taking the sum of a set of quantitative features and 2) complex mathematical formulae with variable weights for each of the quantitative features. An example of the former includes the RECIST sum of diameters quantitative estimate of tumor burden. The reasoning task to generate the sum of diameters is a type of a single context statistical reasoning task that utilizes set aggregation. The EURAL Disease Activity Score (DAS) (Figure 2.11) on the other hand, is an example of a single context mathematical formula statistical reasoning task. Table 2.2 summarizes the clinical state parameters for each of the domain response criteria, while Table 2.4 gives examples of the respective reasoning tasks. Most of the reasoning tasks for the clinical state occur over a single context.
2.2.1.7: Qualitative Estimate of the Clinical State

A qualitative estimate of the clinical state has a qualitative data type. As described above, the qualitative clinical state may be a single qualitative clinical finding, or an aggregate of several findings described qualitatively. For instance, non-measurable disease can be described as persistent or resolved. In such a case, a qualitative estimate of the clinical state describes the persistence or resolution of all disease findings. Furthermore, a quantitative response biomarker may also be classified using defined thresholds to classify a quantitative clinical state as a qualitative clinical state. For instance, the HIV viral load is a quantitative clinical state that can be classified using thresholds into Detectable or Undetectable HIV viral load qualitative descriptors of the clinical state. Table 2.2 and 2.4 show examples of the qualitative estimates of clinical state and their respective reasoning tasks.

Table 2.4: Clinical state level reasoning tasks: Single and multiple context classification and statistical tasks

<table>
<thead>
<tr>
<th>Clinical State Level Reasoning</th>
<th>RECIST</th>
<th>IHC</th>
<th>HIV</th>
<th>EURAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Context Classification Task</td>
<td>Qualitative restriction classification task</td>
<td>Follow-up Target Lesion Set Contains Persistent Lesion then Follow-up Target Lesion Qualitative Response is Some Disease Persists</td>
<td>Follow-up Target Lesion Set Not Contains Persistent Lesion then Follow-up Target Lesion Qualitative Response is All Disease Resolved</td>
<td>none</td>
</tr>
<tr>
<td>Quantitative Relation restriction classification task</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>If Reached DAS Value &gt; 2.47 then Reached DAS Value is Greater Than 2.47</td>
</tr>
<tr>
<td>Statistical Task</td>
<td>Set Aggregation Task</td>
<td>Quantitative Tumor Burden = Sum Set(Target Lesion Longest Diameter)</td>
<td>Quantitative Tumor Burden = Sum Set(Target Lesion Product Short and Long Axis)</td>
<td>none</td>
</tr>
<tr>
<td>Mathematical Formula Task</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>none</td>
</tr>
<tr>
<td>Multiple Context Statistical Task</td>
<td>Set Reasoning</td>
<td>Smallest (Set Quantitative Tumor Burden at each time point)</td>
<td>Smallest (Set Quantitative Tumor Burden at each time point)</td>
<td>none</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Math Formula</th>
<th>none</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAS = 0.53938 * sqrt(RAI) + 0.6465 * SJC44 + 0.33 ln(ESR) + 0.00722 * GH</td>
<td></td>
</tr>
</tbody>
</table>
2.2.2: Defining the Response State in Response Criteria

In addition to specifying the estimate of the clinical state, response criteria specify the method for calculating and classifying how the clinical state changes over time with their definition of the response state. These methods may be quantitative or qualitative or a combination of both for estimating treatment response. Table 2.5 and 2.6 summarize the response parameters and reasoning tasks for each of the domain response criteria.

Table 2.5: Response State and Summary Statistic Parameters for Response Criteria in Oncology, HIV and Rheumatoid Arthritis.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Domain Response Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RECIST</td>
</tr>
<tr>
<td><strong>Quantitative Response State</strong></td>
<td></td>
</tr>
<tr>
<td>Percent Change Sum of Diameters from Smallest Sum of Diameters to Baseline, Percent Change Sum of Diameters Baseline, Time to Progression</td>
<td>Percent Change Sum of Products from Smallest Sum of Products to Baseline, Percent Change Sum of Products from Baseline Time to Progression</td>
</tr>
<tr>
<td><strong>Qualitative Response State</strong></td>
<td></td>
</tr>
<tr>
<td>Best Response Category</td>
<td>Best Response Category</td>
</tr>
<tr>
<td><strong>Summary Statistics</strong></td>
<td></td>
</tr>
<tr>
<td>Mean Response Rate, Median Time to Progression, Count Response Category, Percent Response Category at Time Point</td>
<td>Mean Response Rate, Median Time to Progression, Count Response Category, Percent Response Category at Time Point</td>
</tr>
</tbody>
</table>

2.2.2.1: Quantitative Assessment of Response to Treatment

Quantitative assessment of response to treatment requires a quantitative estimate of the clinical state. Response criteria specify the parameters to be used for quantitative assessment of response to treatment, the calculation method, and the data type for the quantitative response estimate. Response assessments may occur over the entire set of data, or at each follow-up response assessment period taking into account the estimates of the clinical state since the baseline assessment period. Quantitative assessments of response to treatment typically calculate a relative change from the baseline clinical state at time point \( t_0 \) to the follow-up clinical state at time point \( t_i \). Figure 2.16 shows the mathematical formulas for common quantitative response states
derived from changes in quantitative clinical states. Classes of change include absolute change, percent change, rate of change, and the difference in the rate of change before and after intervention. These formulae are a type of multiple context statistical reasoning common in many response criteria.

Table 2.6: Response state level reasoning tasks: Single and multiple context classification and statistical tasks

<table>
<thead>
<tr>
<th>Response State Level Reasoning</th>
<th>RECIST</th>
<th>IHC</th>
<th>HIV</th>
<th>EURAL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Single Context Classification Task</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Qualitative restriction classification task</td>
<td>If Target Lesion Response Category is Partial Response and Non-Target Lesion Response is Non-CR/Non-PD then Target/Non-Target Lesion Response is Partial Response</td>
<td>If Target/Non-Target Lesion Response Category is Partial Response and New Lesion Status is No New Disease Detected then Follow-up Response Category is Partial Response</td>
<td>none</td>
<td>If Change DAS Value Greater Than or Equal to 1.39 &amp; Reached DAS Value Less Than or Equal to 2.47 then Response Category is Good</td>
</tr>
<tr>
<td>Quantitative Relation restriction classification task</td>
<td>Percent Change Baseline Tumor Burden &gt; -30% then Target Lesion Quantitative Response Category is Partial Response</td>
<td>Percent Change Smallest Tumor Burden &gt; 50% then Target Lesion Quantitative Response Category is Progressive Disease</td>
<td>none</td>
<td>HIV viral load below detectable level is an Undetectable Viral Load</td>
</tr>
<tr>
<td><strong>Multiple Context Classification Task</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Qualitative Set Reasoning</td>
<td>Best Overall Response Category = Best Response Set of Categorical Response from Baseline to Follow-up</td>
<td>Best Overall Response Category = Best Response Set of Categorical Response from Baseline to Follow-up</td>
<td>none</td>
<td>Best Overall Response Category = Best Response Set of Categorical Response from Baseline to Follow-up</td>
</tr>
<tr>
<td><strong>Multiple Context Statistical Task</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mathematical Formula Task</td>
<td>Quantitative Response = Percent Change in Tumor Burden from Baseline to Follow-up</td>
<td>Quantitative Response = Percent Change in Tumor Burden from Baseline to Follow-up</td>
<td>none</td>
<td>Quantitative Response = Absolute Difference DAS from Baseline to Follow-up</td>
</tr>
</tbody>
</table>
2.2.2.2: Qualitative Assessment of Response to Treatment

There are several types of qualitative assessments of response to treatment (Figure 2.17). For a single quantitative estimate of the clinical state at a single assessment period, a quantitative threshold derived from the treatment goals can be used to define the response category for the absolute value of the clinical state irrespective of prior assessments (Figure 2.17A). For example, an HIV viral load below the detectable level has a Response Category of Complete Response. Similarly, quantitative thresholds may be applied to a single quantitative estimate of treatment response taking into account the relative change from prior assessments (Figure 2.17B). The RECIST thresholds for defining response categories from the quantitative percent change in tumor burden are one such example. Qualitative response may also be determined from a qualitative estimate of the clinical state (Figure 2.17C). For example, the qualitative non-target lesion response is derived from a qualitative assessment of non-measurable disease in RECIST. Furthermore, multiple types of qualitative assessments of treatment response may be used to estimate an overall qualitative treatment response state through a form of pattern matching over a single response period (Figure 2.17D). For example, in RECIST, the response category at each follow-up assessment period is derived from single context pattern matching of the target lesion response, the non-target lesion response, and the new lesion status.

**Figure 2.16: Formulas to calculate the quantitative response state from a quantitative clinical state. Change is measured between the baseline assessment time point \( t_0 \) and the follow-up assessment time point \( t_i \).**

<table>
<thead>
<tr>
<th>Absolute Change</th>
<th>Response State ( (t_i) = \text{Clinical State} \ (t_i) - \text{Clinical State} \ (t_0) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent Change</td>
<td>Response State ( (t_i) = \frac{[\text{Clinical State} \ (t_i) - \text{Clinical State} \ (t_0)]}{\text{Clinical State} \ (t_0)} )</td>
</tr>
<tr>
<td>Rate of Change</td>
<td>Response State ( (t_i) = \frac{[\text{Clinical State} \ (t_i) - \text{Clinical State} \ (t_0)]}{\text{Time Between} \ (t_0 - t_i)} )</td>
</tr>
<tr>
<td>Difference in Rate of Change Before &amp; After Intervention</td>
<td>Response State ( (t_i) = \text{Rate of Change After} \ (t_0) - \text{Rate of Change Before} \ (t_0) )</td>
</tr>
</tbody>
</table>
Similarly, pattern matching can be used to assess qualitative response from a single qualitative response taken over multiple response periods (Figure 2.17E). Defining the best overall response category from a series of follow-up response categories is one such type of a multiple context classification task.


2.2.2.3: Time to Response Events

Lastly, response criteria also describe several types of response events including disease and toxicity related events. Some events like death are not difficult
to define, while other events such as disease progression events are more ambiguous. Response criteria specify the method for determining when these response events have occurred, enabling calculation of the time to the event. Time to disease progression for instance is a common primary or secondary clinical trial endpoint.

2.2.3: Defining the Summary Statistic in Response Criteria

Finally, response criteria define the summary statistics that can be used as intermediate response endpoints in clinical trials for reporting how the cohort responded to the treatment. Again, the purpose of response assessment criteria is to provide a consistent approach to assessing patient response in clinical trial cohorts. This enables aggregation of response outcomes within trial arms and comparison of response outcomes across trial arms. When the same response criterion is used across clinical studies in the same domain, it also enables a method for comparing response outcomes across clinical trials. Table 2.5 lists the summary statistics used in each of the domain response criteria. Common summary statistics include the following:

- Mean of the Best Quantitative Response State
- Count of patients in each category of Best Qualitative Response State
- Percentage of patients in a particular Qualitative Response Category at a given time point (e.g. Six month progression free survival = percentage of patients who have not experienced disease progression 6 months after trial enrollment)
- Median Time to Response Event

2.3: Limitations of Current Approaches to Response Assessment

Current approaches for applying and developing response criteria have significant limitations that motivate the desiderata for the Rule-Based Response Assessment Framework. As described above in the RECIST use case, application of response criteria in analysis of patient data is a complex process. Research and development of response criteria is similarly complex. In this section we describe the current approaches used to apply and develop response criteria, and describe the limitations of these approaches that motivate the desiderata for the Rule Based Response Assessment Framework.
2.3.1: Limitations to Current Approaches for Applying Response Criteria

The predominant challenges in applying response criteria are their consistent application to patients within a particular clinical trial protocol as well as their consistent application across clinical trial designs. Consistent application of criteria within trials has been particularly difficult in the setting of imaging based approaches. This is due to several factors including data acquisition workflows for quantitative imaging data and differences in human interpretation of the data between multiple reviewers. These problems highlight the need for an automated method for consistent interpretation of response biomarkers with respect to standardize response criteria.

Acquisition of quantitative imaging data for treatment response assessment is difficult for two reasons. First, the data acquisition procedure is a complex distributed task that involves multiple providers separated in space and time, and second, quantitative measurement techniques for imaging data are imprecise. In the current clinical workflow, the oncologist orders an imaging study, the patient has the study performed and the radiologist is first to review the images. The radiologist summarizes their findings in a text report and records detailed measurements as image markups. The report is then sent to the oncologist who independently reviews the report and images. In the context of clinical trials, oncologists use flow sheets to record the quantitative aspects of the target lesions, the presence of non-target lesions, calculate the sum of diameters and the percent change from baseline to determine the response classification. Some institutions will share the flow sheets with the radiologists for evaluation of the follow-up study, however these flow sheets are not incorporated into the radiology report that composes the official medical record.

In previous work we demonstrated that the current radiologist-oncologist communication paradigm is not optimally coordinated with respect to tracking target lesions of interest, resulting in incomplete information for evaluating tumor response status being recorded in the medical record (44). Reports and images were obtained of CT scans of the chest, abdomen and pelvis from patients enrolled in clinical trials at two cancer centers where RECIST is being used to evaluate response to treatment in the metastatic setting. Target lesions were identified at baseline by the oncologists and
the radiology reports and image mark-up were evaluated for identification and measurement of these target lesions. Thirteen patients were recruited with a total of 42 CT scans (13 baseline, 29 follow-up) interpreted by 16 radiologists. 55 target lesions were identified at baseline by the oncologists (average 4 lesions per patient), and imaged a total of 167 times across studies (Table 2.7). At baseline 71% of target lesions were identified in radiology reports and 73% by image mark-up, while the longest diameter was reported 55% of the time and marked-up 50% of the time. At follow-up 38% of target lesions were identified in reports and 70% by image mark-up, while the longest diameter was reported only 28% of the time and marked-up 26% of the time. In only 26% of the studies were the report and image mark-up data sufficient to calculate the quantitative response rate as define by RECIST (Table 2.8). Current reporting of radiology results in the medical record focuses on summarizing findings, which is insufficient for consistent application of quantitative methods to evaluate response to treatment. Our results illustrated the need for new training techniques and informatics tools to improve radiologist-oncologist workflow coordination to enable quantitative evaluation of response to treatment. Our findings motivated the development of several functional requirements for informatics tools to support this workflow in application of quantitative methods for response assessment such as RECIST (44).

Table 2.7: Percentage of target lesions identified and longest diameter recorded in the radiologist report and image mark-up at baseline and follow-up assessment time points

<table>
<thead>
<tr>
<th></th>
<th>Baseline (55)</th>
<th>Follow-up (112)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Report Identified Lesion</td>
<td>71%</td>
<td>38%</td>
</tr>
<tr>
<td>Image Mark-up Lesion</td>
<td>73%</td>
<td>70%</td>
</tr>
<tr>
<td>Longest Diameter Reported</td>
<td>55%</td>
<td>28%</td>
</tr>
<tr>
<td>Longest Diameter Marked</td>
<td>50%</td>
<td>26%</td>
</tr>
</tbody>
</table>

Table 2.8: Percentage of CT scans with sufficient information to calculate the sum of longest diameters clinical state for RECIST at the baseline and follow-up assessment time point

<table>
<thead>
<tr>
<th>Number of CT Scans</th>
<th>Baseline (13)</th>
<th>Follow-up (29)</th>
<th>Total (42)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sufficient to Calculate RECIST Sum of Longest Diameters</td>
<td>54%</td>
<td>14%</td>
<td>26%</td>
</tr>
</tbody>
</table>
Limitations of the current workflow include 1) lack of communication between radiologists and oncologists as to the measurable disease to be tracked over time resulting in incomplete documentation of measurable disease, 2) lack of unambiguous identifiers for tracking measurable disease, 3) lack of education of radiologists in appropriate application of multiple oncology response criteria, and 4) publication of response criteria as underspecified text documents that are not machine computable so as to allow for automated reasoning and decision support.

In addition there is a high degree of inter-observer variation in selection and measurement of target lesions (45). This can be seen when differences arise in the interpretation of response outcomes between the local providers managing patients in the trial and central reviewers (8). This is in part due to differences in prospective vs. retrospective review of the response data, as well as differences in training and workflow of central reviewers who are given explicit instructions on how to apply the criteria to patient data and do so in a focused setting compared to and local physicians who are trying to apply complex criteria in a busy clinical setting.

This inconsistent application of response criteria is not only observed within trials, but across trials in clinical domains, making it difficult to compare trial results (12). Inconsistent application of response criteria is in part due to their natural evolutionary nature as new assessment modalities emerge over time. Within the HIV domain, criteria were inconsistently applied until the procedures for response assessment using the HIV viral load had been clinically and statistically validated with respect to their correlation with overall survival (39). A tension therefore exists between the consistent application standard response criteria in clinical trials and the need to evaluate novel response biomarkers that may better differentiate treatment effect between groups.

2.3.2: Limitations to Current Approaches to the Development of Response Criteria

Response criteria continue to evolve and require validation as new technologies for assessing disease emerge. A rational approach to the development of response criteria has recently been proposed as a guide to ongoing research in this area
(46): 1) the assessment modality should be widely available and the technique reproducible, 2) the measurement technique within the modality should be reproducible and accurate, 3) changes in the estimate of the clinical state should correlate with time to progression and overall survival, 4) new criteria should be evaluated with data from multiple clinical centers and 5) with multiple disease subtypes if applicable. There have been several recent advances in this research process, but key limitations remain that further inform the development of the Rule-Based Response Assessment Framework.

Several European and American research centers are involved in the development and testing of new oncology response assessment criteria. For the recent update of the RECIST criteria from the original version published in 2000 to the RECIST 1.1 version published in early 2009, a large retrospective database of target lesions was developed to test the impact of modifications to the criteria (47,48). Despite these recent advances in the development of large databases to conduct research on response criteria, several limitations remain. First, these data repositories currently do not use data standards for sharing imaging metadata across institutions so as to enable the collection of large imaging data set for research. Similarly, there is a lack of publicly available data sets containing baseline and follow-up imaging studies and image annotations, along with the corresponding diagnoses, therapies and clinical outcomes. A publicly available database could be used to evaluate new image processing algorithms and response criteria. In addition, the lack of a knowledge framework for encoding the parameters that define response criteria limits the ability to 1) conduct and compare experiments that modify the parameter values, and 2) deploy new criteria to be incorporated into decision support and data analysis systems.

2.4: Desiderata for an Automated Response Assessment Method

The limitations of the current approaches to application and development of response criteria motivate the desiderata for the Rule-Based Response Assessment Framework. While a comprehensive response assessment system is beyond the scope of this dissertation, an automated response assessment method is a core need for both the consistent application and development of response criteria. We thus propose the
following desiderata for an automated response assessment method. An implementation of the Framework should:

1) Be sufficiently expressive to encode response criteria and apply them to patient cohorts
2) Be clinically accurate
3) Enable reuse of code with multiple data sets for the same criteria
4) Be extensible to multiple response criteria, and
5) Enable comparison of criteria within domains

2.4.1: Sufficiently expressive

The expressivity of an implementation refers to the range of concepts and tasks that can be represented by the defined symbols and operations in the model. An implementation that is sufficiently expressive to apply response criteria to patient data must be able to represent knowledge about response entities and their properties for a variety of patients, clinical trial protocols, and domain criteria. The implementation must also be able to perform the classification and statistical tasks needed to generate the respective goal driven qualitative and quantitative interpretations at the patient and cohort level.

The above analysis of domain response criteria identified several categories of classification and statistical reasoning tasks for response assessment. These methods are typically chained together in a goal-driven procedure that defines the approach to applying response criteria to patient data. The classification and statistical reasoning methods are performed at the finding, subject and cohort level. The tasks may be performed over single or multiple contexts, and may include single or multiple input parameters. The input and output parameters may have qualitative or quantitative data types depending upon the type of reasoning task. An automated response assessment method that is sufficiently expressive must be able to handle all of these requirements.

2.4.2: Clinically Accurate

The discussion in Section 2.3.1 illustrates the various reasons for lack of consistency in application of response criteria across users. Differences in user interpretations and workflows make it difficult to establish a gold standard for accuracy of a response assessment method compared to human interpretation.
However, human generated response data flow sheets like the one in Figure 2.5 present the current standard for clinical accuracy against which an automated implementation can be compared.

2.4.3: Reusable

There are many dimensions of reuse for knowledge-based systems. Here we distinguish between reusable frameworks where aspects of the implementation can be reused, as opposed to extensible frameworks that can be applied to multiple domains. By design, response criteria are standard methods for interpreting patient data that can be applied to multiple clinical trials. As such, the implementation of the response assessment framework should be able to be reused across multiple clinical trial data sets for a single response criterion. Similarly, the evolution in the development of response criteria is such that criteria within domains have similar domain specific concepts and procedures. From this perspective, the implementation should enable reuse of domain specific criteria knowledge to generate similar criteria within the same domain.

2.5.4: Extensible

A framework that is extensible enables application to multiple domains. For response assessment, an extensible framework should enable application of response criteria in multiple clinical domains such that minimal knowledge engineering is required to instantiate new response criteria.

2.4.5: Comparison of Criteria

Each domain specific implementation of an automated reasoning system requires some degree of human inspectability so that the knowledge engineer can understand how to implement the new problem. As such there is a tension between machine interpretable and human inspectable implementations. For instance, statistical aspects of the RECIST implementation could be achieved thru the use of a spreadsheet. Spreadsheets are highly human inspectable, but lack an information model to encode the concepts in the rows and columns, nor their properties and property values. Program code on the other hand is highly machine interpretable, but is challenging for humans to comprehend. Program code that uses an information
model on the other hand can improve the readability of program code thru the use of standard terms, and also enables sharing of the inputs and outputs between programs. A knowledge-based implementation on the other hand, should enable human inspection of the knowledge required to instantiate a domain specific response criteria. Tension exists between the level of abstraction of generic methods, and the ease of inspection and authoring to enable reuse and extension. The implementation of an extensible, reusable response assessment framework should enable both human inspectability and machine interpretation of criteria knowledge. As such, an inspectable implementation should enable comparison of the knowledge level components of response criteria. This is especially useful for comparing the differences between closely related criteria in the same domain. From this perspective, the implementation should also enable comparison of the outputs of criteria to evaluate the impact of their interpretation procedures on response outcomes.

To satisfy these desiderata I hypothesized that a rule-based framework for response assessment would enable an extensible, reusable method for representing, applying and comparing response criteria. The following methodology was used to test this hypothesis:

1) Create a rule-based framework for treatment response assessment
2) Instantiate the framework for two response criteria in the same domain
   a. Instantiate the framework for a first oncology criteria
      i. Evaluate expressivity of implementation to encode criteria knowledge and apply it to patient data
      ii. Evaluate clinical accuracy of implementation compared to human interpretation in response flow sheet
      iii. Evaluate reuse of code for multiple patient data sets with same criteria
   b. Generalize implementation and evaluate for a second oncology criteria
      i. Evaluate extensibility of implementation to second criteria
      ii. Evaluate reuse of knowledge between criteria in the same domain
      iii. Evaluate the ability to compare criteria knowledge
      iv. Evaluate ability to compare response outcomes of two criteria applied to the same patient data set
3) Demonstrate generalizability of implementation to a second clinical domain

i. Evaluate extensibility of implementation to a third response criteria in the rheumatoid arthritis domain

ii. Evaluate reuse of knowledge between criteria in different clinical domains
As defined in chapter 1, response assessment is a type of temporal interpretation method that evaluates an entity’s response over time to an intervention event. Temporal interpretation methods are a class of Temporal Representation and Reasoning Methods, an area of informatics research with a long history dating back to the 1970’s that has been extensively reviewed (49) (15) (50) (17) (51). Previous reviews have taken a task-oriented view (50) (17) or a methodological view (49) (15) (51) of the temporal representation and reasoning literature. This chapter presents a task-oriented review of the literature discussing the dimensions of temporal interpretation methods within the context of response assessment.

The dimensions of temporal interpretation methods include the domain task, the temporal and non-temporal features of the data, the domain knowledge required to
perform the task, and characteristics of the reasoning method itself (Figure 3.1). Aspects of these dimensions have been extensively reviewed in prior work and will not be covered in detail in this chapter. Section 3.1 describes the types of domain tasks that are typically solved by temporal interpretation methods. A task-oriented review of temporal reasoning methods were presented in (50) (17) and included diagnosis, prognosis, and planning methods. Our review focuses on the temporal interpretation task as it is specifically related to the response assessment task and discusses prior work that has performed this task.

Another important dimension of temporal interpretation methods is the nature of the input and output data from the method. This includes both the temporal aspects of the data such modeling approached to temporal primitives and temporal data maintenance, and non-temporal aspects of the data such as quantitative or qualitative data types, data dimensionality, and data frequency. As is described in section 3.2, many aspects of these dimensions have been reviewed in prior work, and as such this review provides only a high level overview of the material as related to the response assessment task.

There are several dimensions related to the domain knowledge used for temporal interpretation methods that are also important to discuss. These include formalisms for domain knowledge representation, tools for knowledge acquisition, and approaches that enable knowledge reuse. Section 3.3 reviews the various approaches that temporal interpretation methods have taken to these dimensions of domain knowledge.

Finally, section 3.4 describes the dimensions of temporal interpretation methods as related to the characteristics of the reasoning method itself. This includes a discussion of goal-driven vs. data driven methods, context based reasoning methods, the complexity of reasoning methods, the knowledge representation of the method itself, and method reusability. The complexity of the reasoning methods is further discussed from the perspective of methods that perform classification or statistical reasoning tasks, and those that perform reasoning over a single subject or can cohort based reasoning.
3.1: Dimensions of Domain Tasks

Temporal reasoning includes several classes of generic tasks: interpretation, projection, forecasting and planning (15). Interpretation involves reasoning only about past and present data, and not about the future. This is in contrast with Projection and Forecasting tasks that predict the future values of parameters given past values. Planning tasks on the other hand, produce a recommended sequence of actions that are anticipated will achieve the goal state given the initial state (16). From this perspective, response assessment is a type of interpretation task in that it reasons only about the past up to the present.

Temporal Interpretation tasks can also be classified according to the domain tasks of monitoring, diagnosis and response assessment. Monitoring involves the creation of intermediate interpretations of high frequency physiologic data streams in the ICU setting (17) that often lack a treatment data stream. Diagnosis explains a set of findings, but is distinct from the response assessment task that explains a set of finding with respect to intervention (treatment) events and goals.

The response assessment task always includes an intervention or treatment context for interpretation. The treatment context may be explicit in the form of a treatment data set along with the response data set, or may be assumed. For example, physiologic and laboratory data parameters from a hemodialysis treatment session as in (52) are implied to be occurring during a therapeutic context. In addition, different classes of response assessment tasks exist depending upon whether the response interpretation is purely observational, or performs a comparison to a predicted response or intervention goal. Response assessment tasks that are purely observation oriented, perform an interpretation of the response data in the context of an intervention event describing increasing and decreasing trends and patterns in response data, but do not take into account a particular intervention goal. Intervention goals in medicine are typically described in terms of the disease response (efficacy) or the host response (toxicity) to treatment. For example, in oncology, the intervention goals are to maximize the efficacy of the treatment in eliminating the cancer (disease response goal) while minimizing toxicity (host response goal). A response assessment
task that takes these goals into account, can further classify the nature of the response to intervention in the context of these goals. For example, in addition to describing decreasing or increasing trends in tumor burden, the oncology response criteria classify the degree of the trend as complete response, partial response, stable disease or disease progression. Other classes of response assessment tasks are integrated with response prediction tasks whereby the predicted effect is compared to the observed effect. Deviations from the predicted effect can then be used to explain or discover the etiology for the deviation from the prediction.

Literature reviews have summarized prior work in temporal representation and reasoning related to several of these domain tasks. In particular, (50) provided a task-oriented review of the literature from the perspective of diagnosis, prognosis, and therapy planning. In this account, the authors considered the response assessment task as a type of diagnosis task and did not distinguish between systems performing one or the other type of interpretation task. In (17), the authors focused on the temporal interpretation tasks of temporal abstraction and monitoring with a specific emphasis on high frequency data streams typical of ICU settings. To date, there has not been a review that focuses on the clinical task of response assessment, likely because so few systems have implemented that specific sub-task. We review them here and summarize the findings in Table 3.1.

One of the earliest temporal interpretation methods, VM (53) (54) (55), evaluated the efficacy of various modes of mechanical ventilation. VM took as input 30 quantitative values of physiologic parameters and the treatment mode of the mechanical ventilation therapy. VM interpreted the physiologic data in the context of a particular mode of ventilation and recommended changes to the mode of mechanical ventilation to optimize the patient’s physiologic status.
Table 3.1: Summary of systems that perform the response assessment task

<table>
<thead>
<tr>
<th>Methods</th>
<th>Clinical Domain</th>
<th>Input Response Data</th>
<th>Input Treatment Data</th>
<th>Response Output Data</th>
<th>Temporal Interpretation Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>VM – Fagan (1980) (53) (54) (55)</td>
<td>Monitoring efficacy of mechanical ventilation in ICU</td>
<td>Values of 30 physiologic parameters in 2 minute intervals</td>
<td>Mode of Mechanical Ventilation</td>
<td>Intervals of adequate or inadequate mechanical ventilation based on mode of ventilation</td>
<td>Response Assessment</td>
</tr>
<tr>
<td>Kahn - ATN (1986) (56)</td>
<td>Host bone marrow toxicity response to chemotherapy</td>
<td>Laboratory Data</td>
<td>Chemotherapy Drugs and treatment contexts</td>
<td>Summary of unexpected host bone marrow response to chemotherapy</td>
<td>Response Assessment, Response Summary</td>
</tr>
<tr>
<td>Kahn - TOPAZ (1991) (49)</td>
<td>Host bone marrow toxicity response to chemotherapy</td>
<td>Laboratory Data</td>
<td>Chemotherapy Drugs and treatment contexts</td>
<td>Summary and explanation of host bone marrow response to chemotherapy</td>
<td>Response Prediction, Assessment, and Explanation</td>
</tr>
<tr>
<td>Resume - Shahar, Musen (1993) (57)</td>
<td>Host Toxicity Response to Bone Marrow Transplant</td>
<td>Laboratory Data</td>
<td>Treatment protocol for Chronic Graft vs. Host Disease</td>
<td>Grades of bone marrow toxicity and the interval of Chronic Graft versus host disease</td>
<td>Response Assessment</td>
</tr>
<tr>
<td>Resume - Shahar, Musen (1993) (57)</td>
<td>Host bone marrow toxicity response to AIDS therapy</td>
<td>Laboratory Data</td>
<td>AIDS treatment protocol</td>
<td>Intervals of various grades of anemia response to drug therapy</td>
<td>Response Assessment</td>
</tr>
<tr>
<td>Resume - Shahar (1996) (58)</td>
<td>Monitoring efficacy &amp; toxicity of insulin therapy for diabetes</td>
<td>Blood glucose levels</td>
<td>Insulin therapy, meals, physical exercise</td>
<td>Intervals and Patterns of high and low blood glucose response to insulin therapy</td>
<td>Response Assessment</td>
</tr>
<tr>
<td>CAPSUL - Chakravarty and Shahar (2001) (60)</td>
<td>Monitoring efficacy &amp; toxicity of diabetic insulin therapy</td>
<td>Blood glucose levels</td>
<td>Insulin therapy, meals, physical exercise</td>
<td>Patterns of recurring response events</td>
<td>Response Assessment</td>
</tr>
<tr>
<td>CAPSUL - Chakravarty and Shahar (2001) (60)</td>
<td>Host Toxicity Response to Bone Marrow Transplant</td>
<td>70 laboratory and physiologic parameters</td>
<td>Bone marrow therapy treatment context</td>
<td>Intervals of multi-organ toxicity. Time to recovery patterns</td>
<td>Response Assessment</td>
</tr>
<tr>
<td>Bellazzi (2005) (52)</td>
<td>Assess efficacy of hemodialysis therapy over multiple sessions</td>
<td>13 physiologic and device parameters from 4 hour hemodialysis session</td>
<td>Physician defined set of prescriptions for hemodialyzer settings and goal response</td>
<td>Pre-processing: intra-dialysis treatment efficacy; TA method: trends in quantitative parameters over time</td>
<td>Response Assessment, Knowledge Discovery</td>
</tr>
</tbody>
</table>
Several temporal interpretation methods were also developed to interface with the ONCOCIN (63) decision support system for oncology protocol driven. Kahn created several modifications to ONCOCIN’s underlying temporal model (64) (65) to enhance context sensitive querying and reasoning (66). Several temporal interpretation methods were developed with the goal of predicting, assessing, and explaining host bone marrow toxicity response to chemotherapy treatments. Initial response interpretation methods in (66) 1) analyzed each visit for deviations from the "expected" course, and then 2) reviewed multiple visits for patterns over time to create a summary of unexpected toxicity response events. Later efforts in the TOPAZ system (49), used a physiologic bone marrow response model to predict both population and patient response to chemotherapy, and then compared these predictions to observed quantitative laboratory values. When the observed response deviated from the predicted response, the system tried to create an explanation for why the results were discordant using the physiologic models. The output of the interpretation method was the abstraction of the observed findings in comparison to the expected effect and an explanation for the observed deviations when they existed.

Shahar’s Resume system that implemented the Knowledge Based Temporal Abstraction (KBTA) method (67) (58) has been used to create temporal abstractions and in particular, treatment response assessments for several clinical domains including AIDS therapy, bone marrow transplant, and diabetes insulin treatment. Initial temporal abstraction tasks with Resume (57) included identification and grading of temporal intervals of host bone marrow toxicity to AIDS therapy, and graft versus host disease following Bone Marrow Transplant. Resume was later applied to assessment of blood glucose response to insulin therapy by identifying repeating patterns of high (lack of efficacy) or low (toxicity) blood glucose response in the context of insulin administration and meal events (58). CAPSUL (60) extended the pattern detection methods of KBTA to better identify cyclical events and was similarly applied to the domains of diabetes insulin therapy and toxicity assessment after bone marrow transplant.
The RASTA (61) system later developed a parallelized implementation of the KBTA method designed to work on very large data sets and perform parallel evaluation of abstractions for improved computational efficiency. RASTA was integrated with a temporal mediator to enable querying of time-oriented biomedical data (62). The system was applied to the Stanford HIV Drug Resistance Database (68), a database containing time-stamped data on drug regimens, HIV genetic sequences, and HIV viral loads from over 2000 subjects in antiretroviral clinical trials. The temporal abstraction method created gradient response assessments such as “decreasing viral load” and pattern response assessments such as “partial drug response”. A query tool mediated accessing these abstractions.

Bellazzi applied temporal data mining methods to the domains of diabetes therapy (59) and assessment of the quality of hemodialysis therapy (52). Bellazzi’s temporal interpretation method consists of three processes for pre-processing raw time stamped data, generating qualitative temporal abstractions, and post-processing the temporal abstractions into summaries. Finally, they utilized the summaries of temporal abstractions in data mining methods to discover new knowledge about the patterns of response to therapy. In (59), Bellazzi applied this approach to the interpretation of insulin treatment efficacy and toxicity for diabetes mellitus. The method took as input time stamped response data in the form of blood glucose levels, hemoglobin AIC levels, and the presence of urine glucose. Treatment data included insulin administration events, meals and exercise. The method generated state abstractions of the type, high and low for glucose levels, trend abstractions in the form of increasing and decreasing, and complex temporal abstractions to detect patterns. In particular the authors used the complex temporal abstraction methods to identify two treatment related toxicity events, the ‘Somogyi effect’, a physiologic response to hypoglycemia while asleep that results in morning hyperglycemia without glycosuria, and ‘Metabolic instability’, a series of increasing and decreasing blood glucose trends. The post-processing step also created counts of high or low blood glucose trends over the observation period.
In (52), Bellazzi applied their three-step temporal interpretation process to the assessment of the quality of hemodialysis service for end-stage renal disease on the basis of performance indicators related to the efficacy of treatment. The method takes as input time series of 13 quantitative parameters monitored during a 4-hour hemodialysis session as well as the physician’s prescription for the hemodialysis settings and weight loss goals. In the pre-processing step, for each 4-hour hemodialysis session the median of each parameter in the time series is taken as a quantitative abstraction of that parameter for the therapy session. The quality of the therapy session is assessed by comparison of the median values to a set of reference values and a set of dialysis outcome parameters. The hemodialysis session is thus classified as a success or failure with respect to the treatment goals. In the second step, the authors apply temporal abstraction methods to median parameter values to discover the reasons of treatment failures, by deriving associations between monitoring variables and failures that may be interpreted as causal relationships.

3.2: Dimensions of Data

Central to any temporal reasoning task is the representation of the temporal primitives over which reasoning occurs and the temporal databases in which the data is stored and retrieved. These topics have been extensively reviewed (49) (15) (50) (51). Several temporal formalisms have been developed for representing time in intelligent systems including those that use time point (69) and interval based (70) temporal primitives (71,72). Work on time oriented clinical databases dates back to the 1970’s with the Time Oriented Database (73). Even today, most clinical databases utilize a point based temporal primitive for representing clinical events including treatments and findings. However, reasoning about time oriented clinical data such as with response assessment also requires a representation of temporal intervals to enable reasoning such as with Allen’s temporal intervals (70). As such it is desirable for a temporal formalism to be able to handle both types of temporal primitives. One solution is to transform time point temporal primitives into interval based temporal primitives with the same start and finish time (70).
The distinction between valid time and transaction time is also essential for reasoning with temporal databases (74). Typically most temporal reasoning in medicine occurs over valid time, while truth maintenance also requires a representation of the transaction time. Temporal representations have been further extended to specify different time granularities necessary for reasoning (75). Other types of temporal reasoning require a probabilistic approach to defining temporal truth and use a fuzzy representation of time (51).

When considering temporal interpretation methods, it is also important to consider the non-temporal characteristics of the input and output data. Non-temporal data dimensions were reviewed extensively in (17) and include issues related to the data dimensionality, data frequency, the number of input and output parameters, and the data types that can be handled by the temporal interpretation method. In particular, many temporal abstraction methods only considered quantitative data types for their input parameters, yet generated qualitative data type abstractions as output. These qualitative abstractions can then be further abstracted into other qualitative abstractions.

3.3: Dimensions of Domain Knowledge

All temporal interpretation tasks require domain knowledge to ground the interpretation. Temporal interpretation methods vary with respect to representation, acquisition and the potential for re-use of domain knowledge. In this context, knowledge representation refers to the data structures and information models used to store domain knowledge for temporal interpretation. Knowledge acquisition refers to the tools used by domain experts and knowledge engineers to encode domain knowledge for a particular domain temporal interpretation task. Knowledge acquisition methods and tools have been previously reviewed in (76,77). Most temporal interpretation methods use a combination of domain expert interviews and literature review to acquire domain knowledge (17). The domain knowledge is then coded into the temporal interpretation method by a knowledge engineer. The task of domain knowledge acquisition can also include reuse of subsets of the domain knowledge that have already been acquired and encoded. This is particularly helpful
when two domain interpretation tasks vary only slightly from one another, such as when two domain response criteria only vary with respect to their defined thresholds for response categorization.

Temporal abstraction domain knowledge representation formalisms have been previously reviewed in (17). In most temporal interpretation methods, the domain knowledge required to perform the interpretations is integrated within the reasoning method itself. This includes methods developed for VM (53), by Kahn (49), and Bellazzi (52,59). This can facilitate a certain amount of computational efficiency, but with this approach neither the methods nor the knowledge can be reused across domains. Thus reuse of the framework requires recoding domain knowledge and methods from one application domain to the next. Most of these systems utilize rule-based approaches to encode the temporal abstraction domain knowledge (17). Rules are a convenient way to represent declarative knowledge in a way that is easily understood and offers an easy inference mechanism.

A key innovation by Shahar (58) utilized a knowledge-based approach where all of the domain knowledge required for the temporal abstraction task was modeled as input to a generic temporal abstraction method. Shahar’s Knowledge-Based Temporal Abstraction (KBTA) Framework (67) is a general, reusable method that takes as input time oriented clinical findings and events, and the domain knowledge required for the temporal abstraction task, and generates as output qualitative abstractions at the patient level (Figure 3.2). The domain knowledge required to perform the temporal interpretation tasks is modeled in an ontology that includes the event ontology, context ontology, parameter ontology, and abstraction-goal propositions. Combi in (78) utilized an object-oriented approach to modeling the temporal data inputs and outputs, as well as the knowledge required for temporal reasoning. These approaches enable reuse of the information models required to encode temporal events and domain knowledge.
The temporal interpretation reasoning methods themselves can also be described along several dimensions including the particular class of interpretation method, the degree of context-based reasoning performed, the complexity of the reasoning methods, their knowledge representation and reuse. We discuss the literature of temporal interpretation tasks along these dimensions.

3.4.1: Class of Temporal Interpretation Method

In addition to several classes of domain interpretation tasks, temporal interpretation methods can be classified as data driven or goal driven reasoning methods. Goal-driven reasoning is an efficient way to solve problems that can be modeled as structured selection problems. That is, the aim of the system is to pick the best choice from many enumerated possibilities. The goal of formal treatment response criteria is to quantify and classify response. When all possible combinations
of answers can be enumerated in advance, a goal driven reasoning method may be the best approach. For example, both the RECIST oncology criteria (Figure 2.9) and the EURAL rheumatoid arthritis criteria (Figure 2.12) have a discrete set of possible combinations of response parameters that yield a classification of an overall response category.

Table 3.2: Summary of data driven vs. goal driven temporal interpretation methods

<table>
<thead>
<tr>
<th>Methods</th>
<th>Data driven</th>
<th>Query tool to access data driven temporal interpretations</th>
<th>Goal Driven</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kahn - ATN (1986) (56)</td>
<td>Enumerates all possible deviations from expected response</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kahn – TNET/ETNET/T-Query 1991 (79)</td>
<td>Enumerates all possible deviations from expected response</td>
<td>T-Query tool to ask questions of abstractions</td>
<td></td>
</tr>
<tr>
<td>Kahn - TOPAZ (1991) (49)</td>
<td>Enumerates all possible deviations from expected response</td>
<td>Rhetorical schema asks a specific series of queries over the data to reach conclusions that are summarized to the user.</td>
<td></td>
</tr>
<tr>
<td>TrendDx (1995) (80, 81)</td>
<td></td>
<td></td>
<td>Matching temporal templates for diagnosis task</td>
</tr>
<tr>
<td>VIE-VENT (1996) (82,83)</td>
<td>Enumerates trends and oscillations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resume using KBTA - Shahar (1997) (67)</td>
<td>Enumerates all possible temporal abstractions</td>
<td>Query language to ask questions of temporal abstractions</td>
<td></td>
</tr>
<tr>
<td>Combi (1999) (78)</td>
<td>Enumerate all possible temporal abstractions</td>
<td>Query driven: Retrieve the MVT's of stationary ejection fraction, with a threshold of 15% for patient ID 2254.</td>
<td></td>
</tr>
<tr>
<td>CAPSUL - Chakravarty and Shahar (2001) (60)</td>
<td>Extensions to KBTA for enumerating patterns from data</td>
<td>Query language to identify complex patterns</td>
<td></td>
</tr>
<tr>
<td>IDAN/KNAVE11-Boaz and Shahar (2005) (84)</td>
<td>KBTA plus CAPSUL</td>
<td>Query driven: Did complex temporal pattern occur for patient ID 2254 and if so when?</td>
<td></td>
</tr>
<tr>
<td>Probabilistic Temporal Abstraction (PTA) – Ramati and Shahar (2005) (85,86)</td>
<td>Data driven, probabilistic temporal interpolation method</td>
<td>Plans to integrate into IDAN (76)</td>
<td></td>
</tr>
<tr>
<td>RASTA – O’Connor (2006) (61) (62)</td>
<td>Same as KBTA</td>
<td>Query driven temporal mediator</td>
<td></td>
</tr>
</tbody>
</table>
However, many forms of temporal interpretation and response assessment tasks are less formal and simply describe trends and patterns in the data. The reasoning method does not have a specific query that it is trying to answer, but instead enumerates multiple answers. When the user or decision support system has a question, it can then query the interpreted data for the answer. In such situations, a data driven reasoning method is more appropriate. The data driven approach uses keeps track of the current state of problem solution and looks for rules which will move that state closer to a final solution.

Table 3.2 summarizes the temporal interpretation methods according to a primary goal driven reasoning method or a data driven approach. For data driven approaches, it also describes the query tools that are employed to ask specific questions of the temporal interpretations. Goal driven temporal interpretation methods include TrenDx\(^{80,81}\), and the quality assessment classification procedure in Bellazzi’s work\(^{52}\). In each of these methods, the temporal interpretation task is goal directed in that it is answering a specific diagnostic or response assessment question using a pre-defined set of enumerated solutions. In TrenDx\(^{80,81}\) the matching method compares the input data set to a set of trend templates that define the possible diagnostic categories for growth curves. Bellazzi’s procedure for classifying the quality of a dialysis session on the other hand is more procedural in nature and uses a predefined set of parameter restriction values to classify the session as a success or failure\(^{52}\).

The remaining methods listed in Table 3.2 including Shahar’s Knowledge Based Temporal Reasoning Method (KBTA)\(^{1}\) and its extensions in CAPSUL\(^{60}\), IDAN\(^{84}\) and the Probabilistic Temporal Abstraction (PTA)\(^{85,86}\) method utilize a form of data driven temporal abstraction method to enumerate all possible temporal interpretations from the data and implement a query language to access the abstractions. Figure 3.3 shows an example of the interpretations a data driven temporal interpretation task such as the KBTA method might generate taking as input the long axis measurement of two cancer lesions and the respective start of two treatment events. A data driven approach would generate all possible combinations of
temporal interpretation from the data including state abstractions, gradient abstractions that have been temporally interpolated, and patterns. A query tool is then required to enable users or decision support systems to ask ad hoc questions of the temporal interpretations. These query languages enable users to ask questions of the form, “Did the complex temporal pattern A occur for patient ID 2554 and if so, when did it occur?” In this way, users may ask any number of questions of the temporal interpretations, increasing the range of possible questions that may be answered and solved. However, these temporal interpretation methods do not in and of themselves generate an answer to a specific question and thus require a query interface to access and make use of the interpretations.

Figure 3.3: Example of a Data Driven Temporal Interpretation Task where all possible combinations of temporal interpretations are generated from the data. This example shows the long axis measurements of two cancer lesions over time and start treatment events. This example shows how context can be generated from the start treatment events, and examples of state, gradient, temporal interpolation and pattern abstractions. The blue downward arrow indicates a decreasing gradient abstraction while the red upward arrow indicates an increasing gradient abstraction.
Kahn’s temporal interpretations methods were similarly data driven. In all of the three approached reviewed (49,56,79) data driven temporal interpretations were created and then repeating patterns of interpretations were summarized back to the user. In particular, Kahn’s temporal interpretation methods identified deviations from expected responses to chemotherapy, and summarized the count of these “unexpected” events and temporal intervals over which they occurred. The unexpected nature of these events would make a goal driven approach to this problem very difficult from a knowledge acquisition perspective.

3.4.2: Context Based Reasoning Methods

As described above, the response assessment task requires a treatment context for temporal interpretation. A temporal context defines the interval of time over which reasoning should occur. In reviewing the literature, the treatment context may be implied or explicit. Temporal interpretation methods vary in their use of context-based reasoning. Many temporal interpretation methods such TrenDx (80,81) and those by Combi (78), do not perform context based interpretation at all. In others, such as Bellazzi (52), the treatment context is assumed or implied given the nature of the data set being acquired while monitoring a therapeutic event such as dialysis. In others, such as Kahn (56), context based reasoning is widely used but the context is provided by the data set. Only in Shahar’s KBTA method (67) is the context actually generated from the treatment event data.

However, there are certain requirements for context-based reasoning for the response assessment task that are not met by the KBTA context generation method (Figure 3.4). First, the KBTA method defines the specification for a reasoning context such as the follow-up assessment context based on a single temporal event such as the primary treatment event (Figure 3.4A). However, some response assessment contexts require two events for full specification, such as a start of a follow-up assessment context defined by the start of the primary treatment, and the finish of the context defined by the start of the next treatment (Figure 3.4B). Second, context based reasoning within KBTA includes all of the data within a given context (Figure 3.4C). On the other hand, context based reasoning for response assessment requires further
temporal restriction of the data included in reasoning (Figure 3.4D). For example, multiple CT scans of the chest might have occurred within the 4 weeks prior to the start of treatment that defines the baseline response assessment period for RECIST. However, the criteria specify that the study closest to the start of treatment (closest to the end of the baseline assessment period) should be used as the baseline disease assessment and all others ignored for purposes of response assessment.

Figure 3.4: Limitations of the KBTA method for context based reasoning with respect to the requirements of the response assessment task. A. KBTA generates reasoning contexts with a single event. B. Some response assessment contexts require specification with two events. C. Context based reasoning in KBTA includes all data within the context. D. Context based reasoning for response assessment requires additional temporal restrictions on the data within a context that should be reasoned over. In this example only the highlighted data closest to the end of the baseline context should be used for reasoning.

The response assessment task also requires context based reasoning within a single context and across multiple contexts. The KBTA method (67) does enable single context reasoning called vertical temporal inference as well as multiple-context reasoning in the form of temporal interpolation and pattern detection. Similarly, in (79) Kahn performs single context reasoning using the ETNET to associate rules with data structures that assert new intervals or features. Multiple context reasoning is achieved in that system through temporal interpolation using a temporal hierarchy to
construct interpretations based on successively longer temporal intervals. The temporal interpretation methods in VM (54) performed only single-context based reasoning based on the present mode of mechanical ventilation. The context changed whenever the mode of mechanical ventilation changed.

![Diagram](image)

Figure 3.5: Inputs and output requirements of the response assessment method. Highlighted boxes represent aspects of the temporal abstraction task not handled in previous knowledge based frameworks.

### 3.4.3: Complexity of Reasoning Methods

The response assessment task as defined by formal response criteria requires a range of complex reasoning methods including classification and statistical interpretation methods that reason over individuals and cohorts of subjects (Figure 3.5). Classification methods are those that generate a qualitative temporal interpretation. Statistical methods on the other hand, generate a quantitative temporal interpretation.

#### 3.4.3.1: Classification Methods

Classification methods are those that generate a qualitative temporal interpretation (Figure 3.6). Classification methods can be further categorized based on the data type of their input parameters. Quantitative classification methods take as
input temporal data with a quantitative parameter data type. Qualitative classification methods on the other hand, take as input parameters with a qualitative data type.

Quantitative classification is the common sub-method of all temporal interpretation methods given that all temporal interpretation methods take as input data with a quantitative data type and generate qualitative abstractions (Table 3.3). There are several types of quantitative classification methods including quantitative relation restriction, quantitative change classification, and quantitative set restriction methods. The quantitative relation restriction method takes as input a quantitative parameter value and classifies the value based on thresholds and relation operators. An example of a quantitative relation restriction task is as follows: a cancer lesion with a long axis greater than or equal to 10 mm is classified as a measurable cancer lesion. Another example is a glucose value greater than 115 is classified as high. Many of the temporal interpretation methods reviewed in the literature perform this type of quantitative relation restriction method. These include KBTA’s Vertical Temporal Inference Method (67), Kahn’s augmented transition network structure (ATN) (56), Fagan’s VM (54), Bellazzi’s (59) (52) and Combi’s (78) temporal abstraction methods that generate state abstractions of the type high or low for parameter values.
<table>
<thead>
<tr>
<th>Temporal Interpretation Methods</th>
<th>Quantitative Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Quantitative Relation Rest</td>
</tr>
<tr>
<td>VM – Fagan (1980) (53) (54) (55)</td>
<td>Classification of physiologic parameters as &quot;acceptable&quot;</td>
</tr>
<tr>
<td>Kahn - ATN (1986) (56)</td>
<td>Augmented transition network structure (ATN)</td>
</tr>
<tr>
<td>Kahn – TOPAZ/ETNET (1991) (49)</td>
<td>ETNET implements context-specific reasoning by associating rules with data structures called ETNODEs</td>
</tr>
<tr>
<td>TrendDx (1995) (80,81)</td>
<td>None</td>
</tr>
<tr>
<td>Combi (1999) (78)</td>
<td>Quantitative State Abstraction</td>
</tr>
<tr>
<td>CAPSUL - Chakravarty and Shahar (2001) (60)</td>
<td>same as KBTA</td>
</tr>
<tr>
<td>IDAN/KNAVE11- Boaz and Shahar (2005) (84)</td>
<td>same as KBTA</td>
</tr>
<tr>
<td>Probabilistic Temporal Abstraction (PTA) – Ramati and Shahar (2005) (85,86)</td>
<td>None</td>
</tr>
<tr>
<td>RASTA – O’Connor (2006) (61) (62)</td>
<td>same as KBTA</td>
</tr>
</tbody>
</table>

Many temporal interpretation methods also have sub-methods that classify changes in parameter values over time and generate abstractions of the type gradient (e.g. increasing, decreasing, stable) or rate (e.g. slow, fast). See for example the gradient abstractions generated in Figure 3.3 that show intervals of increasing and decreasing quantitative parameter values between two time points. KBTA’s Horizontal Temporal Inference Method (67) generated temporal abstractions of the form gradient and rate, while Bellazzi (52,59), Combi (78), and Kahn (49) only generated abstraction of the form gradient. In TOPAZ (49), Kahn also classified deviations of
observed to predicted response parameters as “surprising predictions”. Thus we see examples of change classification over concurrent time points and across time.

Finally, quantitative set restriction methods classify a set of quantitative parameters based on a set restriction. A method that classifies the five largest measurable cancer lesions as target lesions is an example of a quantitative set restriction method. None of the temporal interpretation methods reviewed performed this type of quantitative classification method.

Qualitative classification is also a common sub-method of all temporal interpretation methods. However, many temporal interpretation methods do NOT take as input parameters with a qualitative data type, but instead utilized the qualitative interpretations from the output of their quantitative classification methods. There are several types of qualitative classification methods including qualitative restriction, temporal interpolation, and pattern matching methods (Figure 3.6) (Table 3.4). Qualitative restriction methods take as input a parameter with a qualitative data value and a restriction upon the possible set of values for that parameter that define how that parameter is to be further classified. An example of a simple qualitative restriction is a non-measurable cancer lesion classified as a non-target cancer lesion. In our review, only the KBTA Vertical Temporal Inference Method (67) was capable of this type of qualitative classification. Qualitative restrictions may also occur over sets of values including membership or non-membership to a set. For example, the anatomic location of a cancer lesion is compared to a set of lymph node anatomic locations and if classified by inclusion as a Lymph Node Lesion or by exclusion as a Non-Lymph Node Lesion. None of the temporal interpretation methods reviewed performed this type of qualitative restriction method.

Temporal interpolation on the other hand, is a common type of qualitative classification task among the reviewed methods. Temporal interpolation takes as input qualitative state, rate or gradient interval abstractions and combines them into sequentially longer intervals of the same category. Figure 3.3 shows an example where two gradient abstractions with a decreasing value are interpolated together to make one longer interval of a gradient abstraction with the same value. Many
temporal abstraction methods including KBTA (67), Kahn (49,56), VM (53), Combi (78), and Bellazzi (59) (52) perform some type of temporal interpolation.

Table 3.4: Summary of Temporal Interpretation Methods that perform Qualitative Classification Methods

<table>
<thead>
<tr>
<th>Temporal Interpretation Methods</th>
<th>Qualitative Classification</th>
<th>Temporal Interpolation</th>
<th>Pattern Matching</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kahn - ATN (1986) (56)</td>
<td>none</td>
<td>Use temporal hierarchy to construct successively longer temporal intervals</td>
<td>Identify recurring events</td>
</tr>
<tr>
<td>Kahn – TOPAZ/ETNET (1991) (49)</td>
<td>none</td>
<td>Temporal searching methods used in ETNET are designed to distinguish between persistent and intermittent events</td>
<td>Identify pattern in the model states or parameter values that can “explain” the observed discrepancy</td>
</tr>
<tr>
<td>TrendDx (1995) (80,81)</td>
<td>none</td>
<td>none</td>
<td>Complex data driven pattern matching method</td>
</tr>
<tr>
<td>Combi (1999) (78)</td>
<td>none</td>
<td>Temporal interpolation through extension of maximal validity interval</td>
<td>Multiple context pattern matching</td>
</tr>
<tr>
<td>CAPSUL - Chakravarty and Shahar (2001) (60)</td>
<td>same as KBTA</td>
<td>same as KBTA</td>
<td>Enhanced pattern matching of KBTA and repeating pattern detection</td>
</tr>
<tr>
<td>IDAN/KNAVE11- Boaz and Shahar (2005) (84)</td>
<td>same as KBTA</td>
<td>same as KBTA</td>
<td>Enhanced pattern matching via query tool</td>
</tr>
<tr>
<td>RASTA – O’Connor (2006) (61) (62)</td>
<td>same as KBTA</td>
<td>same as KBTA</td>
<td>same as KBTA</td>
</tr>
</tbody>
</table>
Temporal pattern detection is also a common form of qualitative temporal interpretation task. Pattern detection takes as input parameters with qualitative data types and generates as output a qualitative pattern. Pattern detection is of two basic types, pattern matching where a set of parameter values match a pre-defined pattern, and repeating pattern detection where cyclical events are identified. Pattern matching can occur over a single or multiple contexts. Single context pattern matching typically involves pattern matching across multiple parameters. In VM (53), complex rules are used with multiple antecedent clauses to define a particular pattern across multiple physiologic parameters within a context of a particular mode of ventilation. The antecedent clauses evaluate for the presence of a particular quantitative classification state and for quantitative change thresholds. In this way, VM combines multiple qualitative and quantitative classification tasks into a single pattern-matching step. Similarly, Bellazzi in (52) used a type of single context pattern matching in the pre-processing step to classify the success or failure of the hemodialysis service from the assessment of six treatment efficacy parameters.

Multiple context pattern matching can include single or multiple parameters. Methods have included the use of Allen’s temporal calculus to describe patterns where one qualitative temporal abstraction interval follows or precedes the other such as “Decreasing followed by Increasing Pattern” classification (Figure 3.3) used in Shahar’s KBTA pattern matching sub-method (67) and Bellazzi’s (59) (52) temporal abstraction methods. CAPSUL (60) further extended KBTA to handle more complex linear patterns that included constraints on events as well as parameter values. As opposed to temporal interpolation methods that extend a temporal interval for the same state of a parameter, these types of pattern matching methods handle parameter state transitions.

Repeating pattern detection is another common type of pattern detection method among temporal interpretation methods. Kahn in (49) identified and summarized the occurrence of repeating events of bone marrow toxicity across multiple contexts. CAPSUL (60) in addition to describing linear patterns defined
semantics for representing complex repeating patterns and a query service for accessing them.

3.4.3.2: Statistical Methods

Statistical interpretation methods generate as output quantitative abstractions over time oriented data. Input data types are typically quantitative but may be qualitative in the context of tasks that require counting the number of instances of a parameter. Statistical methods may be further categorized according to their use of set based reasoning or mathematical formulas (Figure 3.6).

Set based reasoning may either be for aggregation such as taking the sum, mean, median, or count of a set of parameters, or for set restriction in the case of identifying the minimum or maximum of a set of values. These types of reasoning methods are required for response assessment tasks such as RECIST which take the sum of longest diameters of the target lesion as a quantitative abstraction of tumor burden and which identify the minimum sum of diameters over time as the best response. A review of the temporal interpretation methods shows that very few include statistical methods (Table 3.5). The KBTA method (67) did not perform any statistical interpretation tasks, however the CAPSUL (60) extension to KBTA did include in their query tool set aggregation methods to calculate the mean or variance of a parameter and the set restriction task of identifying the minimum of a parameter. Kahn in (56) did provide a count of the number of repeating events identified in the pattern matching method. Bellazzi’s pre-processing methods in (52) facilitated data reduction by taking the median value of a parameter over the entire hemodialysis session as a quantitative abstraction of the value of that parameter for assessment of treatment efficacy. Similarly, Bellazzi’s post-processing methods in (59) generated a count of temporal abstraction events and average daily insulin utilization over an assessment period. Finally, in Combi (78) the output of the query tool included an average calculation of the queried parameter over queried time interval.
Table 3.5: Summary of temporal interpretation methods that perform statistical methods. Method are categorized as set based reasoning or mathematical formulas.

<table>
<thead>
<tr>
<th>Temporal Interpretation Method</th>
<th>Statistical Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Set Based Reasoning</td>
</tr>
<tr>
<td></td>
<td>Quantitative Set Restriction</td>
</tr>
<tr>
<td>Kahn - ATN (1986) (56)</td>
<td>None</td>
</tr>
<tr>
<td>Kahn – TOPAZ (1991) (49)</td>
<td>Min or Max of parameter value within temporal interval</td>
</tr>
<tr>
<td>TrendDx (1995) (80,81)</td>
<td>None</td>
</tr>
<tr>
<td>Resume using KBTA - Shahar (1997) (67)</td>
<td>None</td>
</tr>
<tr>
<td>Combi (1999) (78)</td>
<td>None</td>
</tr>
<tr>
<td>Bellazzi (2000) (59)</td>
<td>None</td>
</tr>
<tr>
<td>CAPSUL - Chakravarty and Shahar (2001) (60)</td>
<td>Minimum of a parameter</td>
</tr>
<tr>
<td>Bellazzi (2005) (52)</td>
<td>none</td>
</tr>
<tr>
<td>RASTA – O’Connor (2006) (61) (62)</td>
<td>none</td>
</tr>
</tbody>
</table>
Quantitative interpretations may also be generated by mathematical formula. Mathematical formulae include non-set based aggregation equations, difference functions, and temporal functions. The rheumatoid arthritis Disease Activity Score (DAS) (41) (Figure 2.11) is an example of a non-set based aggregation function. None of the temporal interpretation tasks reviewed included complex mathematical formula for aggregation of a temporal interpretation.

Complex mathematical formula can also be used to extrapolate or smooth data for analysis. In Bellazzi’s data pre-processing method for the diabetes application (2000), a complex Bayesian signal reconstruction method was used to extract smoothed time series that were then analyzed by the temporal abstraction methods. Similarly, in TOPAZ (49), Kahn developed a physiologic model of bone marrow response to chemotherapy using differential calculus. These complex mathematical formulas were used for the response prediction task within the TOPAZ system by generating a predicted set of time stamped data. Response assessment was then determined through point-wise comparison of predicted versus observed host response. Mathematic formulae that calculate change such as the example in Kahn, are common in temporal interpretation methods and especially in the response assessment task (Figure 2.15). Bellazzi’s pre-processing method in (52) also calculates the absolute difference and percent difference of planned versus observed outcomes that are then used to classify the response for that particular hemodialysis session.

Other temporal interpretation methods calculate temporal formulas that include rates of change or duration calculations. CAPSUL (60) for instance, calculates a duration function for determining the platelet half-life. VM (53) calculates and reports the duration of temporal interpolation intervals. Bellazzi in (59) also calculates the duration repeating events. However, none of the temporal interpretation methods provide a knowledge-based framework or reusable methods for addressing temporal interpretation tasks.

Finally, the Probabilistic Temporal Abstraction (PTA) (85,86) method utilizes stochastic probability functions to handle the problems of missing data and imprecise data to create probabilities on the truth of the generated temporal abstractions. Three
sub-methods were developed including temporal interpolation, temporal coarsening and temporal transformation applied in tandem to generate probability distributions on the reliability of the generated temporal interpolation and pattern abstractions.

**3.4.3.3: Single Subject vs. Cohort Reasoning**

Most of the temporal interpretation methods described reason over a single subject. Even most of query tools that ask questions of temporal interpretations, reason at the subject level only (60,84). Two methods in particular have tried to address at least in part, the need to query or create temporal interpretations across a patient cohort. Combi’s CARDIOTABS (78) query tool enabled query of temporal abstractions of both a single subject and a cohort of subjects. At the cohort level, the query tool summarized a count of the total number of cases queried and the count of the subset of cases that satisfied the query. The Probabilistic Temporal Abstraction (PTA) (85,86) method, describes two sub-methods specific for cohort evaluation including: 1) the temporal aggregation subtask that generates minimum, maximum, and average values over a cohort, and 2) the temporal correlation subtask that compares two patient populations. Plans were described for incorporating the PTA into the IDAN (84) infrastructure. Other than these two examples, cohort based temporal interpretation methods are limited.

**3.4.4: Method Knowledge Representation**

As described previously, many of the temporal interpretation methods incorporated their domain knowledge directly into their reasoning methods. Many of these systems used machine code (59) (52) or rule-based methods (53) (49) (17) to implement the temporal interpretation method. This can be a very computationally efficient approach at the expense of having the ability to reuse the reasoning method. One of Shahar’s major contributions in KBTA is the ability to reuse the temporal abstraction reasoning method across multiple domains by creating a domain independent reasoning method and encoding new domain knowledge in an input ontology. The reusable KBTA method itself was modeled in the CLIPS rule language but unlike the domain knowledge ontology, could not be inspected. The KBTA method was re-implemented in the RASTA system (61) (62) to improve the scalability.
of the method to larger data sets. RASTA implemented KBTA using ontologies for the knowledge base and Java code to implement the interpretation methods. The KBTA domain independent methods were later formalized in the Temporal Abstraction Rule (TAR) language (87) encoded in Datalog extended with temporal arguments. The TAR was integrated with IDAN (84) where the ALMA reasoning module translates the KBTA knowledge required to execute a query into TAR rules which process the query. In this way, rules are generated on the fly containing all of the domain and patient specific parameters required to execute the reasoning method. These implementations of the reasoning methods are likewise black box implementations that are not inspectible.

Most of the rule-based systems described above perform classification tasks. Statistical methods on the hand may require complex mathematical formulae that may be difficult to represent in a formal rule-based language. Most of the temporal interpretation tasks that perform statistical subtasks did so using machine code (49) (52).

3.5: Summary of Limitations of Temporal Interpretation Methods for Response Assessment

The temporal interpretation methods reviewed above have some significant limitations with respect to goal-driven, procedure oriented response assessment tasks. First, the majority of the methods are data-driven (Table 3.2), and despite the addition of query tools to access the data-driven interpretations, they could not be used to encode procedural response criteria such as RECIST. Second, while several methods perform context-based reasoning, only the KBTA method performs a context generation task needed to interpret response and treatment data in the context of a clinical trial protocol. While KBTA does perform context generation, the method is not sufficiently robust to account for the need to define context from multiple events, and to restrict interpretation to sub-sets of the data within a given context.

Furthermore, the majority of the methods only perform classification tasks (Table 3.3 and Table 3.4). The most commonly implemented classification tasks are quantitative relation restriction, change classification, temporal interpolation, and temporal pattern matching. Almost none of the methods perform qualitative
restriction, and none perform quantitative set restriction. The few methods that do perform statistical interpretation tasks (Table 3.5) have a limited range of the types of mathematical formulae, temporal calculations, or set based reasoning tasks. These statistical methods are also typically encoded in machine code for use in specific cases and are not general reusable methods. Finally, the methods reviewed have limitations with respect to their knowledge representation formalism for both their domain knowledge encoding and the interpretation methods themselves. Only KBTA and its predecessors provide a formal ontology to represent the input and output parameters, events, goals, and propositions for the input to their KBTA method. While the KBTA method itself is domain independent, the method itself is a black box set of rules that are not human inspectable. This limits the ability to extend KBTA to support a more expressive set of classification and statistical interpretation task necessary for response assessment.

Table 3.6: Summary of limitation of temporal interpretation methods for response assessment task

<table>
<thead>
<tr>
<th>Summary of Limitations of Temporal Interpretation Methods for Response Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Lack of a goal-driven interpretation method capable of encoding procedural response criteria</td>
</tr>
<tr>
<td>• Lack of knowledge based approaches for encoding statistical interpretation methods</td>
</tr>
<tr>
<td>• Lack of integration of statistical and classification methods into a single procedural decision logic</td>
</tr>
<tr>
<td>• Prior context based reasoning methods lack sufficient expressivity for context generation based on multiple events</td>
</tr>
<tr>
<td>• Prior context based reasoning methods lack context restriction to limit interpretation of data according to a response assessment protocol</td>
</tr>
</tbody>
</table>

Given these limitations, it is not possible to encode the procedure and goal-driven response criteria reviewed in chapter 2 using existing methods. In the next chapter, we present the Rule-Based Response Assessment Framework, a goal-driven, context-driven, and knowledge-based temporal interpretation framework that implements procedural response assessment tasks.
CHAPTER 4
RULE-BASED RESPONSE ASSESSMENT FRAMEWORK

4.1: Rule-Based Response Assessment Framework

The Rule-Based Response Assessment Framework is a goal-driven, context-driven, and knowledge-based temporal interpretation framework that is sufficiently expressive to implement procedural response assessment subtasks. The Framework consists of several knowledge models shown as triangles and several methods shown as rectangles in Figure 4.1. The methods take as input knowledge and data encoded in the models and generate as output new knowledge that is asserted back into the knowledge models. The knowledge models include the Protocol, Clinical Data and Response Criteria knowledge models and the temporal data and temporal restrictions models that they share. The Response Assessment Method consists of the Context Restriction and Interpretation sub-methods.

The Rule-Based Response Assessment Framework consists of several knowledge models shown as triangles and several methods shown as rectangles in Figure 4.1. The methods take as input knowledge and data encoded in the models and generate as output new knowledge that is asserted back into the knowledge models. The knowledge models include the Protocol, Clinical Data and Response Criteria knowledge models and the temporal data and temporal restrictions models that they share. The Response Assessment Method consists of the Context...
Restriction and Interpretation sub-methods.

The Response Assessment Method is context-driven in that it performs both context generation from treatment events and clinical trial protocol knowledge, and context restriction in that it limits the data to be interpreted with a given context to those define by the response criteria. The method is goal-driven in that it implements procedural response criteria through chaining of classification and statistical interpretation rules. The method is knowledge-based in that it provides a formal representation of both the parameters and interpretation rules to fully represent response criteria. It also provides a formal knowledge representation of the entities and relationships between entities for response assessment. This includes a representation of the response data and the clinical trial response assessment protocol and response criteria needed to interpret response data. Finally, the Rule-Based Response Assessment Framework is sufficiently expressive to encode classification and statistical interpretation tasks, single and multiple context interpretation tasks, and tasks that take as input single or multiple parameters that may have qualitative or quantitative data types.

For purposes of describing the Rule-Based Response Assessment Framework, it is useful to separate the knowledge and methods required for context restriction (Section 4.2) and those of context-based interpretation (Section 4.3). The following conventions are used to denote the concepts and relationships in these models. Concepts are italicized with initial capital letters such as Entity, while small upper case letters denote concept properties such as PROPERTY. Where appropriate, a property’s domain and range are specified as PROPERTY (Domain, Range). Those ranges that are a class themselves are denoted by a capital Range while a lower case range denotes those that are a data property. Graphical representations of concept relationships are presented in Figures where boxes represent concepts, solid lines represent IS-A relationships between concepts, and dotted lines represent property relationships.
4.2: Knowledge and Methods for Context Restriction

Context restriction is the procedure where by the planned response assessment contexts in the clinical trial protocol are mapped onto a specific patient given their treatment events and time stamped response data. Figure 4.2 shows a portion of the context restriction procedure where a clinical study’s response assessment plan is summarized as a table of assessment modalities and assessment periods. For example, this particular protocol has several assessment modalities including physical exam, laboratory and imaging modalities that take place at variable assessment periods. An assessment period is defined in relation to some event, typically a treatment event. For instance, the Baseline Assessment Period (B) in Figure 4.2 is defined by the start of the protocol treatment. The beginning of the Baseline Assessment Period is 28 days prior to the start of treatment and finishes just prior to
the start of treatment. The Context Specification that defines the assessment period is a knowledge model in the temporal restriction model that stores information about the events and temporal offsets that define the assessment and thus interpretation periods. These will be described in greater detail in section 4.2.2. The Context Restriction Method then takes this protocol specification and the patient’s treatment events and generates the respective Context Periods with valid date-time instances that are then mapped onto the respective time stamped response data. The following sections describe the knowledge models and methods in the Framework that describe the context restriction procedure for response assessment. The knowledge models include the protocol model and patient data model, as well as the temporal data model and temporal restriction model. Finally, the Context Restriction methods and its sub-methods will be described.

4.2.1: Protocol Specification Model

Cohort response assessment requires a specification of the study cohort to be analyzed, the planned study intervention, and the planned response assessment protocol. This enables identification of the correct patients, treatment events, and findings relevant to the response assessment task for a given study. Figure 4.3 shows concepts and relationships in the Protocol Model. A Research Study has a Study Protocol that has Planned Study Arms. A Planned Study Arm has Planned Interventions and Planned Response Assessment Protocols. A Planned Response Assessment Protocol has Context Specifications that describe the planned temporal periods in which response assessment will occur. The protocol model also specifies the Assessment Modalities that will be used to evaluate patient responses such as imaging studies or laboratory studies, and the Response Criteria to be used to classify response. The protocol model distinguishes between a Planned Study Arm and an Actual Study Arm where the former contains a specification of planned activities and the later specifies instances of patients enrolled on a particular study arm.
4.2.2: Temporal Restriction Model

The planned study protocol utilizes the temporal restriction model to specify the planned assessment periods and interpretation context for reasoning. The basic temporal restriction is the Context Specification briefly described above. A Context Specification defines a single interpretation context. However, several interpretation contexts can be combined into a composite interpretation context that facilitates reasoning over multiple contexts. A clinical trial protocol thus specifies a set of contexts and composite contexts that will be used for interpretation of response data. Thus, several types of knowledge are needed to specify the temporal restrictions for context based interpretation for a particular response criteria and clinical trial protocol (Figure 4.4): 1) Anchor Specification, 2) Anchored Duration Specification, 3) Context Specification, 4) Composite Context Specification, and 5) Context Data Constraints. These are described below.
Context-based interpretation for response assessment requires the ability to generate temporal contexts that are defined by multiple types of events. As such, a START-ANCHORED-DURATION SPECIFICATION and a FINISH-ANCHORED-DURATION SPECIFICATION define the essential properties of a Context Specification (Figure 4.5).

An Anchored Duration Specification is defined by the properties HAS-ANCHOR SPECIFICATION (Anchored Duration Specification, Anchor Specification), HAS-OFFSET (Anchored Duration Specification, integer), and HAS-GRANULARITY (Anchored Duration Specification, Temporal Granularity). Examples of instances of Temporal Granularity include Day, Week, Month, and Year. Anchored durations can specify a time point prior to the anchor time point via a negative offset and a time point after the anchor time point with a positive offset. An Anchor Specification defines the Clinical Event and the temporal property of the event that should be used to generate a time point anchor. For example, a Start Primary Treatment Anchor Specification defines an anchor that is the start of some the primary treatment event for a clinical research protocol. Context Specifications are further classified hierarchically as Baseline Context Specification which define the baseline assessment period, Follow-up Context Specification which define the individual follow-up assessment periods for the clinical trial.
Figure 4.5: Context Specification. A Start Anchored Duration Defines the start of the Context with a time point, offset and temporal granularity. Anchored durations can specify a time point prior to the anchor time point via a negative offset.

A Composite Context Specification enables reasoning over multiple contexts. It defines the relevant Contexts for reasoning over multiple contexts and the valid temporal period from those contexts. A START-CONTEXT SPECIFICATION and FINISH-CONTEXT-SPECIFICATION define the start and finish of the Composite Context Period (Figure 4.6). The remaining sub-contexts in between the start and finish are also specified for inclusion in the reasoning task.

Figure 4.6: Composite Context Specification. The start time point of the Start Context Specification defines the start of the Composite Context Specification, while the finish time point of the Finish Context Specification defines the finish of the Composite Context Specification. Contexts in between the start and finish contexts are also specified for inclusion in multi-context reasoning methods.
4.2.3: Clinical Data Model

The clinical data model contains information regarding Patients and their Clinical Events including Treatment events, Assessment Events, Findings, and Response Events (Figure 4.3). Findings include Data Parameters that are derived from direct observation and have an associated Assessment Event, and Abstract Parameters that are derived from an Interpretation Task within a Response Criteria. Patients also have Context Periods that are derived from a Context Specification in the Planned Response Assessment Protocol. Context Periods are Temporal Propositions within the temporal data model and have a Valid Period. Context Periods are only relevant for a particular Planned Study Arm and Context Specification. Response Events are events that have been observed as part of the response assessment procedure and include Disease Progression Event, Toxicity Events, and Death events. The temporal representation of the Clinical Data Model is described below and shown in Figure 4.8.

![Clinical Data Model Diagram](image-url)

Figure 4.7: The Clinical Data Model mapped to the Protocol Model
4.2.4: Temporal Data Model

The temporal data model includes a representation of both instant and interval based temporal primitives (Figure 4.2). Temporal *Instants* are represented as a single time point `HAS-TIME (Instant, time)`. Temporal *Periods* are represented by a start and finish time point, `HAS-START (Interval, time)` and `HAS-FINISH (Interval, time)`. The model also contains a concept of *Valid Time*, including the sub-classes *Valid Instant* and *Valid Period*. A *Temporal Proposition* is a concept with the property `HAS-VALID-TIME (Temporal Proposition, Valid Time)`. The Clinical Data Model is mapped to the Temporal Data Model as is shown in Figure 4.8. *Clinical Events* are a sub-class of *Temporal Propositions* and thus inherit the `HAS-VALID-TIME` property. For example, certain types of *Intervention* events might have a *Valid Period* temporal representation, while most *Finding* events have a *Valid Instant* representation. To facilitate context based reasoning, the *Anchor Instant* and *Context Period* are included.
as sub-classes of *Temporal Proposition*. Data that will be used in the reasoning task are assigned the property HAS-CONTEXT (*Finding, Context Period*).

**4.2.5: Context Restriction Method**

The Context Restriction Method limits the scope of reasoning by using the clinical trial protocol specification to assign interpretation contexts to the patient response data. The Context Restriction Method consists of several sub-methods including the Anchor Generation Method, the Context Generation Method, the Context Classification Method, and the Composite Context Generation Method.

**4.2.5.1: Anchor Generation Method**

The Anchor Generation Method takes as input a patient’s *Clinical Events* and the clinical protocol’s *Anchor Specifications* and generates as output patient *Anchor Instants*. The *Anchor Specification* identified which patient event should be used to generate the anchor and which time point should be used within that event. For example, the start, finish, or middle of an event interval could be used to define an anchor. Examples of patient *Events* that might generate an *Anchor Instant* include *Interventions* and *Response Events* such as the *Disease Progression Event*.

**4.2.5.2: Context Generation Method**

The Context Generation Method takes as input patient *Anchor Instants* and the clinical protocol’s *Context Specifications* and generates as output a set of patient *Context Periods*. As described previously, a *Context Specification* has a start and finish *Anchored Duration Specification* that identifies the *Anchor Specification*, temporal *offset* and temporal *granularity* used to generate the start and finish times of the *Context Period* from the *Anchor Instant* time point.

**4.2.5.3: Context Classification Method**

The Context Classification Method takes as input patient *Context Periods*, time stamped patient *Data Parameters* and *Context Data Constraints* and generates as output context labels on patient *Data Parameters*. The Context Classification Method utilizes Allen’s Temporal Intervals (70) to identify the *Data Parameters* with time points that intersect the *Context Period*. The method further utilizes the *Context Data
Constraints to identify the subset of Data Parameters that meet the restriction. For example, a Context Data Constraint may specify that the data closest to the finish of the Context Period should be selected. Identification of this subset of the data within the context requires calculation of the temporal distance between the data time point and the finish of the Context Period, and subsequent identification of the data point with the shortest temporal distance.

4.2.5.4: Composite Context Generation Method

The Composite Context Generation Method takes as input patient Context Periods, and the clinical protocol’s Composite Context Specifications, and generates as output patient Composite Context Periods. The goal of this method is to generate a temporal period that defines the start and finish of the Composite Context Period in order to later assign this valid time period to an Abstract Parameter that was generated from the parameters that match the sub-contexts in the Composite Context Specification. As such, the start and finish of the START- and FINISH-CONTEXT-SPECIFICATIONS define respectively the start and finish of the Composite Context Period.

4.3: Knowledge and Methods for Context-Based Interpretation

Several knowledge models and methods are required to perform the context-based interpretations of patient data for treatment response assessment. Response criteria specify a set of interpretation tasks as context based rules and formulas that define finding level, clinical state and response state interpretations as well as cohort response summary statistics. For example, Figure 4.9 shows the context restricted patient response data that will be used as input to the interpretation methods. The response criteria specifies a finding level rule that classifies a baseline lesion with a long axis greater than or equal to 1.0 cm as a measurable lesion. The two image findings that satisfy this rule are highlighted in green. Other classification rules are used to classify the remaining cells in the table for finding level interpretations. A mathematical formula is also shown that describes the procedure for estimating the tumor burden. These are the types of knowledge that are used to generate
interpretations at the finding level, the clinical state level, the response state level, and the summary statistic level.

The rules and formula that comprise response criteria were analyzed for their common features and syntactical patterns. Figure 4.10 shows two rules for the RECIST criteria used for finding level and response state level classification. Both rules have input and output restrictions. The input restrictions also have temporal, class and property restrictions. The output restrictions have class and property restrictions (for the second rule only) as well. Several rule patterns emerge depending upon whether the rule performs single context or multiple context interpretations and whether it performs classification or statistical interpretations. These rule patterns comprise a set of rule templates that make up the context-based interpretation methods for response assessment. Rule templates are domain independent rules that take as input a domain specific interpretation task specification and patient response data, and
generate interpretations that are asserted back into the knowledge base (Figure 4.11). The response criteria knowledge model specifies the knowledge needed to encode domain specific knowledge for response criteria interpretation tasks that can be used as input by domain independent rule templates that comprise the context-based interpretation methods.

Figure 4.10: Common interpretation rule patterns. These two rules both perform single context classification using a quantitative property restriction. Temporal restrictions are show as boxes, class restrictions are highlighted in blue, and property restrictions are highlighted in grey.

Figure 4.11: A rule template and its respective interpretation task specification
4.3.1: Response Criteria Knowledge Model

![Response Criteria Knowledge Model Diagram]

Figure 4.12: Response Criteria Knowledge Model: Response Criteria are defined by a set of Interpretation Task Specifications which themselves have temporal restrictions and input and output parameter-property restrictions. Context Specifications are generally used for single context interpretation task specifications which composite context specifications are used for multiple context task specifications. When relevant, Parameter-property restrictions are defined by a Parameter Restriction and Property Restriction with its associated property value and operators.

The response criteria knowledge model defines the concepts and tasks that are used as input to the context-based interpretation rule templates. **Response Criteria** are defined by a set of **Interpretation Task Specifications** which themselves have **Temporal Restrictions** and input and output **Parameter-Property Restrictions** (Figure 4.12). **Temporal Restrictions** include **Context Specifications** that are generally used for **Single Context Interpretation Task Specifications** while **Composite Context Specifications** are used for **Multiple Context Interpretation Task Specifications**. **A Response Parameter** defines the class restriction for **Parameter-Property Restrictions**. When relevant, the associated **Property Restriction** is further defined by its property value and operators. The number of class restrictions, and the data type of the property restrictions further classify **Parameter Property Restrictions**. **Qualitative**
**Property Restrictions** have a property with a class value, Boolean value, or categorical data type. **Quantitative Property Restrictions** on the other hand have a property with a quantitative data type. The following section describes the Response Parameter Model and the various types of Interpretation Task Specifications.

### 4.3.1.1: Response Parameters

The parameters for the response criteria are the inputs and outputs to the interpretation methods. They include data parameters and abstract parameters, their properties and relationships (Figure 4.13). The Parameters in the Response Criteria Knowledge Model are distinct from those in the Clinical Data Model in that the former are a specification of the parameters, while the latter are an instantiation of the parameters for a particular patient, property values, valid time and context.

![Figure 4.13: Response Criteria Parameter Model. Data Parameters and Abstract Parameters are defined for the Model. Abstract Parameters include parameters related to Findings, the Clinical State, Response State, and Summary Statistic.](image)

**Data Parameters** further specify the domain specific knowledge needed in the clinical data model to enable response assessment. This includes the data properties, data type, and range of property values for clinical observations. **Data Parameters** also have a relationship to the assessment modality and assessment measurement technique from which they are derived, and as such are often classified hierarchically by
assessments, modality.

Abstract Parameters are the outputs of the response assessment interpretation tasks that may later be used as input to another interpretation task. In order to facilitate truth maintenance, an Abstract Parameter has the properties HAS-CONTEXT, HAS-INTERPRETATION-TASK, HAS-RESPONSE-CRITERIA, and HAS TRANSACTION TIME that define how and when it was derived. Abstract Parameters are further classified as Finding Abstract Parameter, Clinical State Parameter, and Response State Parameter, and Summary Statistic Parameter (Figure 4.8). A Finding Abstract Parameter is an abstraction of a patient finding and not of the patient himself. Domain independent sub-classes of the Finding Abstract Parameter include New Finding, Resolved Finding, and Persistent Finding. These types of finding abstractions are used to generate Clinical State Parameters. Clinical State Parameters are further classified as quantitative or qualitative parameters where a Quantitative Clinical State has a continuous data type, and a Qualitative Clinical State has a categorical data type that is typically Boolean or ordinal. The Response State Parameters describe the patient’s response to the intervention and include Quantitative Response States that have a continuous data type and Qualitative Response States that have a categorical data type. Examples of domain independent Qualitative Response State parameters include the Follow-up Period Response Category and the Overall Response Category. Finally, the Summary Statistic Parameters define the data types for the cohort statistics that are typically continuous variables. The Time to Event statistic is further described by a temporal duration and a temporal granularity.

4.3.1.2: Interpretation Task Specification

An Interpretation Task Specification includes the knowledge needed to transform data in one parameter into another parameter. A collection of Interpretation Task Specifications define the knowledge needed to perform the higher level response assessment tasks of determining the clinical state, response state and summary statistic for a clinical trial patient cohort. Interpretation Task Specifications are defined by Temporal Restrictions and Parameter Property Restrictions (Figure 4.12). Input and
output Parameter Restrictions are specified by Response Parameter Restrictions, Property Restrictions, Property Value Restrictions and where relevant Operator Restrictions. Task Specifications are categorized along two major axes as 1) Classification or Statistical Reasoning Tasks, and as 2) Single Context or Multiple Context Reasoning Tasks (Table 4.1). Task Specifications can further be classified according to the number and type of Parameter Property Specifications required for full specification of a particular task.

Table 4.1: Classification of Interpretation Tasks by Single Context, Multiple Context, Classification or Statistical Subtasks. Grey boxes indicate classes of Interpretation Tasks without a respective domain rule example for the implemented response criteria.

<table>
<thead>
<tr>
<th>CLASSIFICATION OF INTERPRETATION TASK SPECIFICATIONS</th>
<th>SINGLE CONTEXT</th>
<th>MULTIPLE CONTEXT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Single Parameter</td>
<td>Two Parameters</td>
</tr>
<tr>
<td></td>
<td>Single Parameter Restriction</td>
<td>Class Restriction</td>
</tr>
<tr>
<td></td>
<td>Class Set Restriction</td>
<td>Class Set Restriction</td>
</tr>
<tr>
<td></td>
<td>Class Property Restriction</td>
<td>Class Property Restriction</td>
</tr>
<tr>
<td></td>
<td>Two Qualitative Parameter Restriction</td>
<td>Class Restriction</td>
</tr>
<tr>
<td></td>
<td>Qualitative Classification</td>
<td>GREATER THAN</td>
</tr>
<tr>
<td></td>
<td>Single Quantitative Relation Restriction</td>
<td>GREATER THAN</td>
</tr>
<tr>
<td></td>
<td>Single Quantitative Set Restriction</td>
<td>LEAST</td>
</tr>
<tr>
<td></td>
<td>GREATEST N</td>
<td>GREATEST N</td>
</tr>
<tr>
<td></td>
<td>Set Aggregation</td>
<td>Single Quantitative Property Aggregation</td>
</tr>
<tr>
<td></td>
<td>TEMPORAL FORMULA</td>
<td>Single Temporal Granularity</td>
</tr>
<tr>
<td></td>
<td>MATHEMATICAL FORMULA</td>
<td>Single Quantitative Property Formula</td>
</tr>
<tr>
<td></td>
<td>Two Quantitative Properties Formula</td>
<td>Multiplication</td>
</tr>
</tbody>
</table>

4.3.1.2.1: Specification of Single Context Interpretation Tasks

A Single Context Interpretation Task reasons only over a single Context Specification. However, while the interpretation task is performed as related only to a single context, the task itself may be repeated in multiple contexts. For example, the
statistical task that generates the quantitative clinical state estimate performs reasoning within a single context. However, this task is performed during the baseline context and all of the follow-up contexts. The reasoning context for Single Context Interpretation Tasks can thus be is specified by multiple Context Specification Classes.

4.3.1.2.2: Specification of Multiple Context Interpretation Tasks

In contrast, a Multiple Context Interpretation Task reasons over multiple contexts. Similarly, the interpretation task may be performed several times with different combinations of contexts. Multiple Context Interpretation Tasks utilize Composite Context Specifications to reason simultaneously over multiple types of Contexts. The START-CONTEXT-SPECIFICATION and FINISH-CONTEXT-SPECIFICATION that define the start and finish of the Composite Context Period can also be used as anchors for reasoning over multiple contexts.

4.3.1.2.3: Specification of Classification Tasks

Classification Tasks are those that reclassify an individual or generate a new qualitative property. Classification Tasks may take as input single or multiple parameters, with single or multiple properties that may have qualitative or quantitative data types (Table 4.1). The specification of a Classification Task requires knowledge of the parameters, properties and property restrictions that define the inputs, and the respective output parameter, property and property value that meet these restrictions. Qualitative Classification uses a qualitative property restriction. Qualitative property restrictions may be classes, categorical individuals, or Boolean. The Task Specification defines the qualitative values that make up this property restriction. Quantitative Classification uses a quantitative property restriction. An Operator and a threshold value define a quantitative property restriction. For example, a Relation Operator (i.e. less than, greater than) has an associated quantitative threshold value that defines a type of quantitative property restriction. Classification tasks may have a single qualitative or single quantitative constraint, or multiple or mixed qualitative and quantitative constraints. Classification tasks with more than one parameter in a single context or a single parameter in multiple contexts are a form of pattern matching task.
Furthermore, some classification tasks also require set reasoning. For example, among the set of measurable findings, the five with the longest diameter are classified as Target Lesions. Specification of this task requires knowledge of the \textit{Set Operator} used (i.e. GreatestN) and the value of the count restriction (e.g. five).

\textbf{4.3.1.2.4: Specification of Statistical Tasks}

\textit{Statistical Tasks} are those that generate a parameter with a quantitative property. \textit{Statistical Tasks} may take as input single or multiple parameters, with single or multiple properties that may have qualitative or quantitative data types. Single context, single parameter statistical tasks are those that have multiple values for a parameter property. These statistical tasks require a specification of the \textit{Set Operator} such as average, median, sum, or count. Multiple parameter statistical tasks require specification of the mathematical or temporal \textit{Formula} that is used to generate the quantitative output parameter. Each parameter within the formula requires a specification of its associated context and quantitative property.

\textbf{4.3.2: Context Based Interpretation Methods}

The Context-Based Interpretation Methods are categorized along the same axes as the Interpretation Task Specification as performing: 1) Single or Multiple Context Reasoning, and 2) Classification or Statistical Reasoning. A single Interpretation Method can execute multiple Interpretation Tasks across Response Criteria. The Interpretation Methods are described as rule templates, domain independent rule patterns that take as input domain specific Interpretation Task Specifications.

\textbf{4.3.2.1: Single Context Classification Methods}

There are several types of Single Context Classification Methods depending upon the type of property restrictions and operators used. These methods range in complexity with an increasing number of Parameters and Property Restrictions. The simplest logical statement includes a single Class and a single Context Restriction. For example, a finding classified as a Non-Measurable Lesion in the Baseline Context is reclassified as a Non-Target Lesion in the same context.
Rule Template:
If a Finding F is Class A and has Context X
Then Finding F is Class B

Example:
If a Finding F is an Image Finding and has a Baseline Context
Then Finding F is a Non-Measurable Lesion

The rule template becomes more complicated when property restrictions are included. A quantitative property restriction includes the relation operator and the threshold value for the relation. For example, a finding that is classified as a Lymph Node Lesion in the Baseline Context that has a long axis measurement greater than 15 mm is reclassified as an Enlarged Lymph Node in the same context.

Rule Template:
If a Finding F is Class A and has Context X
and has Quantitative Property p1 with Value V
and Value V meets a relation restriction
for relation operator O and threshold value T
Then Finding F is Class B

Example:
If a Finding F is a Lymph Node Lesion and has Baseline Context
and has a long axis with Value V
and Value V is Greater Than 15
Then Finding F is an Enlarged Lymph Node

Similarly, a qualitative property restriction can be applied to a classification task. Qualitative property restrictions can include restrictions on classes, categorical individuals or Boolean values. For example, an Image Finding in the Baseline Context or Follow-up Context with the property of anatomic location that matches one of a set of Lymph Node classes is reclassified as a Lymph Node Lesion.
Rule Template:

If a Finding F is Class A and has Context X
and has Qualitative Property p1 with Value V
and Value V is Class C
Then Finding F is Class B

Example:

If a Finding F is an Image Finding and has Baseline Context
and has an Anatomic Location with Value V
and Value V is a member of a class of Lymph Nodes
Then Finding F is a Lymph Node Lesion

Combinations of qualitative and quantitative property restrictions can also be used for single or multiple parameters. Multiple parameters with qualitative or quantitative property restrictions in a single context define single context pattern matching interpretation tasks. For example, classification of the Target Lesion Response Category requires knowledge of the Quantitative Response Category and the Target Lesion Resolution Status. Various combinations of qualitative property values for these two parameters result in the new parameter Target Lesion Response Category and its respective qualitative Response Category property value.

Rule Template:

If Finding F1 is Class A and has Context X
and property p1 with Value = R1
and Finding F2 is Class B and has Context X
and property p2 with Value = R2
Then Finding F3 is Class C with Context X
and property p3 and Value = R3

Example:

If Finding F1 is Target Lesion Resolution Status and has Follow-up Context
and Response Category = Some Lesions Persist
and Finding F2 is Quantitative Response Category and has Follow-up Context
and Response Category = Partial Response
Then Finding F3 is Target Lesion Response Category
and has Follow-up Context
and Response Category = Partial Response
Classification tasks also can utilize set operators. Quantitative properties can be used to classify members of a set of findings in a single context. For example, Measurable Cancer Lesions are classified as Target Lesions or Non-Target Lesions based upon their size in relation to the set of measurable findings. RECIST restricts the number of Target Lesions to the five with the longest diameters.

Rule Template:
If Set₁ contains all Findings F that are Class A and have Context X and have quantitative property p₁ with Value V and Finding F is contained in the Sub-Set₁ with the greatest N Value V Then Finding F is Class B

Example:
If Set₁ contains all Findings F that are Measurable Cancer Lesions and have Baseline Context and have long axis with Value V and Finding F is contained in the Sub-Set₁ with the greatest 5 Value V Then Finding F is a Target Lesion

Similarly, a sets of individuals can be defined and set operators used to identify if a second individual is a member of that set or not. For example, Image Findings that are not Lymph Nodes can be identified with the Not Contains set operator.

Rule Template:
If Set₁ contains all Findings F that are Class A and have Context X and Finding F₂ is Class B and has Context X and Set₁ does not contain Finding F₂ Then Finding F₂ is Class C

Example:
If Set₁ contains all Findings F that are Lymph Node Lesions and have Baseline Context and Finding F₂ is an Image Finding and has Baseline Context and Set₁ does not contain Finding F₂ Then Finding F₂ is a Non-Lymph Node Finding
4.3.2.2: Multiple Context Classification Methods

Classification methods also occur over multiple temporal contexts. For example, an individual clinical finding such as a Cancer Lesion is observed over time in multiple temporal contexts. Classification of the finding in one context can influence the classification of the same finding in another context. For example, classification of an Image Finding as a Target Lesion in the Baseline Context will result in classification of the same finding in the Follow-up Context as a Target Lesion. Such classification tasks require creation of a finding unique identifier so that it can be classified in multiple contexts.

Rule Template:

If Finding F1 is Class A and has Context X
   and has property identifier with Value V
and Finding F2 has Class B and has Context Y
   and has property identifier with Value V
Then Finding F2 is Class C

Example:

If Finding F1 is Target Lesion and has Baseline Context
   and has property identifier with Value V
and Finding F2 has Image Finding and has Follow-up Context
   and has property identifier with Value V
Then Finding F2 is Target Lesion

Similarly, the property of a finding may be assessed in multiple contexts, where the temporal pattern of the property value defines a response state. In HIV for example, a clinical state of (High, Low, Low) in the (Baseline, 8 week, 24 week) contexts respectively, defines the response state of Complete Response.

Rule Template:

If Finding F1 is Class A and has Context X
   and property p1 with Value = R1
and Finding F2 is Class A and has Context Y
   and property p1 with Value = R2
and Finding F3 is Class A and has Context Z
   and property p1 with Value = R3
Then Finding F4 is Class B and has Composite Context (X, Y, Z)
   and has property p3 with Value = R3
Example:

If Finding F1 is Qualitative Clinical State and has Baseline Context and state = High
and Finding F2 is Qualitative Clinical State and has 8 week Context and state = Low
and Finding F3 is Qualitative Clinical State and has 24 week Context and state = Low
Then Finding F4 is Qualitative Response Category and has Composite Context (X, Y, Z) and has Response Category with Value = Complete Response

Finally, set operators can be used to classify individuals based on their properties in relation to a set of individuals with the same class observed over multiple contexts. For example, the quantitative clinical state is typically assessed in multiple contexts. RECIST defines two quantitative response states, one based on the percent change in the clinical state from baseline to the follow-up context, and the other based on the percent change in the clinical state from the minimum clinical state to the follow-up context. Classification of the Minimum Clinical State requires set operators where the minimum value is identified from the set of clinical states in multiple contexts.

Rule Template:

If Set₁ contains all Findings F that are Class A and have Composite Context (X, Y, Z) and has Quantitative Property p1 with Value V and Finding F2 is the minimum value of Set₁ Then Finding F2 is Class B

Example:

If Set₁ contains all Findings F that are Quantitative Clinical State and have Composite Context (X, Y, Z) and has continuous value property with Value V and Finding F2 is the nadir value of Set₁ Then Finding F2 is Minimum Quantitative Clinical State
4.3.2.3: Single Context Statistical Methods

Statistical methods are highly dependent upon their mathematical formulas or set operators. Single Context Statistical Methods reason over a single context, but again may be performed independently in multiple contexts. For example, with the WHO criteria, the product of the short and long axis is calculated for inclusion in the sum of products quantitative clinical state estimate. This calculation is performed for all image findings that are within the relevant interpretation contexts.

Rule Template:

If a Finding F is Class A and has Context X
and has Quantitative Property p1 with Value V1
and has Quantitative Property p2 with Value V2
and mathematical formula(V1, V2) = V3
Then Finding F has Quantitative Property p3 with Value V3

Example:

If a Finding F is an Image Finding and has Baseline Context
and has long axis with Value V1
and has short axis with Value V2
and multiplication (V1, V2) = V3
Then Finding F has product of short and long axis with Value V3

Set operators are also commonly used to aggregate quantitative and qualitative properties of findings in a single context to estimate qualitative and quantitative clinical states. For example, the sum of diameters of the Target Lesions is used to estimate the quantitative clinical state for RECIST.

Rule Template:

If Set₁ contains all Findings F that are Class A and have Context X
and have Quantitative Property p1 with Value V
set-operator(Set₁) = V3
Then Finding F2 is Class B and has Context X
and has Quantitative Property p2 with Value V3

Example:

If Set₁ contains all Findings F that are Target Lesion
and have Baseline Context and have long axis with Value V
sum(Set₁) = V3
Then Finding F2 is Quantitative Clinical State and has Context X
and has continuous value property with Value V3
4.3.2.4: Multiple Context Statistical Methods

Lastly, Statistical Methods can also be applied over multiple contexts and are specified with mathematical formulas and set operators. For example, the percent change in the quantitative clinical state from baseline to follow-up period can be specified with the following logical statement:

Rule Template:

If a Finding F1 is Class A and has Context X and has Quantitative Property p1 with Value V1 and Finding F2 is Class B and has Context Y and has Quantitative Property p2 with Value V2 and mathematical formula(V1, V2) = V3
Then Finding F3 is Class C and has Context Y and has Quantitative Property p3 with Value V3

Example:

If a Finding F1 is a Quantitative Clinical State and has Baseline Context and has continuous value property with Value V1 and Finding F2 is a Quantitative Clinical State and has Follow-up Context and has continuous value property with Value V2 and percent change(V1, V2) = V3
Then Finding F3 is Percent Change from Baseline and has Follow-up Context and has continuous value property with Value V3

Set operators can be used to aggregate results taken over multiple contexts, such as aggregations to estimate the summary statistic. For example, the mean of a quantitative response state is taken over multiple contexts.

Rule Template:

If Set₁ contains all Findings F that are Class A and have Composite Context (X, Y, Z) and have Quantitative Property p1 with Value V set-operator(Set₁) = V3
Then Finding F2 is Class B and has Composite Context (X, Y, Z) and has Quantitative Property p2 with Value V3

102
Example:

If $\text{Set}_1$ contains all Findings $F$ that are Best Quantitative Response State
and have Composite Context
and have continuous value property with Value $V$
$\text{mean}(\text{Set}_1) = V_3$
Then Finding $F_2$ is Mean Best Response Rate and has Composite Context
and has continuous value property with Value $V_3$

This set of rule templates provides a framework for encoding domain
independent rules in a given rule based modeling formalism. The information model
for the domain specific interpretations tasks can be used as input to these domain
independent interpretation rule templates. The context restriction method and these
context based interpretations methods together with the response criteria knowledge
model, data model, protocol model and temporal models define the Rule-Based
Response Assessment Framework. The next chapter describes the implementation of
the Framework.
5.1: Knowledge Representation Implementation Choice

Several knowledge based technologies were considered for implementation of the Rule-Based Response Assessment Framework. These technologies are considered from the perspective of the desiderata for expressivity, reusability, and inspectbility. Consideration was also made for the existence of ontologies that could be used and extended for the implementation to further support interoperability and reuse of existing knowledge. A knowledge representation must be sufficiently expressive and sharable to aid the primary goal of an extensible, reusable method for response assessment. Ontologies and rules are a common formalism for encoding knowledge-based systems. Within this domain are several dominant implementations for ontologies and rules respectively: 1) Frames and JESS, and 2) OWL and SWRL. The Unified Modeling Language (UML) and the Object Constraint Language (OCL), a declarative language for describing rules that apply to UML, provide an object-oriented parallel to ontologies and rules.

Frames were one of the original formalisms for encoding ontologies. Reasoning with frames-based ontologies requires a custom mapping to a rule engine such as CLIPS and JESS, a java implementation of CLIPS. However, the rules themselves did not have any knowledge about the ontology structure itself. Reasoning methods implemented using JESS and frames-based ontologies are capable of performing closed world reasoning required for set based reasoning tasks.

The Semantic Web Ontology Language (OWL) (88) has replaced Frames as the emerging standard for the Semantic Web (89). The Semantic Web Rule Language (SWRL) (90) (91) is the defacto standard rule language for the semantic web and enables reasoning with OWL ontologies. SWRL allows users to write rules that can be expressed in terms of OWL concepts to provide more powerful deductive reasoning capabilities than OWL alone. Semantically, SWRL is built on the same description
logic foundation as OWL and provides similar strong formal guarantees when performing inference. OWL uses description logic as its logical formalism, which is based on the open world assumption. This makes it difficult to perform set reasoning, requiring an extension the SWRL rule language and systems that implement its logic. UML is a modeling formalism but is not based on a formal logic. As such, all reasoning about UML models must occur within the rules. OCL (92) is a declarative language for describing rules about UML models and enables closed world reasoning.

All of these formalisms however, are by default atemporal and thus require extensions to enable temporal representation and reasoning. A temporal ontology (93) exists for OWL to represent both instant and period temporal primitives and also provides representations for valid time and temporal granularity. The SWRL Temporal Built-in Library (94) provides a library of temporal operators that enable reasoning with all of Allen’s temporal intervals (70). These temporal extensions to SWRL utilize the temporal ontology.

The SWRL library includes the Semantic Query-Enhanced Web Rule Language (SQWRL) (95), a SWRL-based query language that can be used to query OWL ontologies. SQWRL provides SQL-like operations to format knowledge retrieved from an OWL ontology. SWRL rules have the basic features of if-then statements, where all of the predicates in the antecedent clause must be true in order to assert statements in the precedent clause. This structure is ideal for classification tasks, where a set of constraints in the antecedent clause defines the conditions for assertion of new knowledge. In general, SWRL rules reason about individuals within ontologies and their roles (properties). This works well for domain specific rules sets however does not work for domain independent rule templates. The Abox reasoning extension to SWRL (96) can be used for two types of assertions: (1) class assertions – an individual is an instance of an entity and (2) role assertions – an individual participates in a role (relation). Class and roles assertions are used extensively in the rule template implementation to enable identification of individuals in the ontology that meet the temporal, parameter and property restrictions.

The core SWRL built-ins library also contains mathematical operators (i.e.
add, subtract, multiply) and relation operators (i.e. lessThan, greaterThan) that maintain the WC3 specification of SWRL. The mathematical operators can be used to both specify and execute the mathematical formulae within SWRL rules. For instance, these mathematical operators can be used to encode the Pythagorean theorem needed to calculate the length of a line from two spatial coordinates, and to calculate the percent change in a parameter over multiple contexts.

The primary limitation in using OWL and SWRL for the response assessment task is the open world assumption of the description logic formalism. The open world assumption prevents set based reasoning and negation because it cannot make the assumption that the data set is complete. A recently developed extension to the SWRL Tab enabled the closed world reasoning required for set reasoning (97). For example, set reasoning is required to calculate the sum of an indeterminate number of values such as is required to calculate the sum of diameters for RECIST. The SWRL implementation of the set specification enabled the creation of sets and the use of logical set operators such as least, greatest, sum and average. The set operators also enable union and intersection statements as well as negation, all of which were required in part to implement several of the rule templates for the response assessment method.

Several ontologies are available in OWL that could be reused for purposes of implementation. This includes the temporal ontology (93) previously described available in OWL to model the temporal data (93), and ontologies to model aspects of the clinical trial protocol specification including the Biomedical Research Integrated Domain Group (BRIDG) model in UML and OWL, and the Ontology of Clinical Research (OCRe) model in OWL. The BRIDG project (98) is a collaborative initiative between the National Cancer Institute (NCI), the Clinical Data Interchange Standards Consortium (CDISC), the Regulated Clinical Research Information Management Technical Committee (RCRIM TC) of Health Level 7 (HL7), and the Food and Drug Administration (FDA) to develop a model of the shared understanding of the semantics of clinical research. The BRIDG model uses the Unified Modeling Language (UML) class, activity and state diagrams to represent declarative and
procedural knowledge. BRIDG has also been recently implemented in OWL. Aspects of the BRIDG model can be used to model clinical trial protocol specifications of the response assessment plan.

The Ontology of Clinical Research (OCRe) (99) is a formal ontology consisting of the entities concerning the design of human experiments, and the logical relationships among them. OCRe is a modular ontology of clinical investigation that includes (1) a representation of the structure of human studies and associated entities such as persons and organizations that play significant roles in studies, (2) informational entities such as study protocols and outcome variables, that are produced in the life cycle of studies, (3) terms for describing study characteristics, and (4) bindings to standard terminologies such as SNOMED CT and NCI Thesaurus. OCRe also incorporates relevant aspects of the BRIDG ontology and is available in OWL.

Considering the availability of several existing ontologies, the fact that OWL is the emerging standard for ontology development with the Semantic Web, and the fact that it is an inspectable language, I chose OWL and SWRL to define the ontologies and rules needed for the response assessment framework. The Protégé-OWL (100) ontology authoring environment and its associated SWRLTab plugin (101) was used for developing the ontologies and rules.

5.2: Ontologies and Rules for Context Restriction Procedure

Context restriction utilizes the Protocol, Temporal Restriction, Clinical Data, and Temporal Data Knowledge Models and the Context Restriction Method (Figure 5.1). Development began with the Protocol Knowledge Model. I collaborated with the developers of the Ontology of Clinical Research (OCRe) (99) to extend their clinical study protocol representation to include treatment response assessment plans. This includes the planned clinical trial protocol specification, study arms and study interventions, and the planned assessment modalities and the timing of those assessment modalities. OCRe also required extensions to model actual trial data opposed to the planned trial protocol. The protocol ontology is further extended by inclusion of the temporal restriction ontology that enables specification of response
assessment contexts and composite contexts for a particular trial (Figure 5.2A).

A specific clinical trial protocol is then instantiated in the protocol ontology. Several elements of the clinical study protocol are required for response assessment. Specifically, instantiation of a planned study protocol, with planned arms, planned intervention for each arm, and the treatment regimen name associated with the planned intervention. For this case, the GITrial has a PlannedGIProtocol with a single experimental GIPlannedArm with an associated GITreatmentSpecification containing the GITreatmentRegimen. In addition, the planned study protocol must have baseline and follow-up context specifications with their associated anchor specifications, composite context specifications, and context reasoning group specifications. The PlannedGIProtocol has a single StartTreatmentAnchorSpecification, a single GIBaselineContext, and four Follow-up Context Specifications with their four respective Composite Context Specifications. Lastly, the planned study protocol must specify a response criterion that will be used for response assessment, in this case the RECIST criteria is assigned. An actual study arm is then created that correlates with the planned study arm. The actual study arm specifies a list of patients enrolled on that arm. In this case the ActualGIStudyArm was linked to the GIPlannedArm, and the 10 patients on the study arm are assigned using the minimum required unique patient
The Clinical Data Model is instantiated by the Clinical Data Ontology. This is a very simple ontology that enables representation of patients and their respective events and findings. The Clinical Data Ontology is populated by a set of patients identified by a unique identifier and their respective treatment events and response assessment data. Raw patient response data are then imported as Data Parameters and valid time stamps. The temporal data model is instantiated by the temporal ontology and used in full without modification (Figure 5.2B). This enables temporal representation of the patient treatment events and response data.

The four sub-methods of the Context Generation Method represent the most generic methods in the implementation. The four sub-methods are themselves implemented by four rules that can be reused with any application of the Response Assessment Method. For example, the start treatment anchor generation rule is shown implemented in SWRL in Figure 5.3. The rule begins with a query for the study’s anchor specification information and matches the specification to an actual study arm subject. Given the possibility that a patient may have multiple treatment events in the clinical data model, the rule limits the generation of start treatment anchors to those that are relevant to a particular actual study arm, with its associated planned protocol and planned intervention. Anchors are only generated for patients enrolled on that particular study arm, with a treatment event that matches study protocol planned intervention. Each anchor instant created is annotated with the actual study arm,
planned study protocol and treatment event from which it was derived.

The SWRL Temporal Built-in Library (94) provides a library of temporal operators that are used to implement the temporal reasoning required by the Context Restriction Methods. In contrast, the interpretation tasks are context-based and thus for the most part do not require the use of temporal operators. The SWRL Temporal Built-in Library is based on the temporal ontology and implements among other operators all of standard Allen temporal operators (70). For instance, Allen’s temporal:intersects operator is used to implement the Context Classification Method by identifying data parameters with valid time instants that intersect a particular context period. Likewise, the temporal:add operator is used to implement the Context Generation Method, where an integer offset and temporal granularity are added to the anchor time point to generate the start and finish of the context period.

5.3: Ontologies and Rules for the Context Based Interpretation Procedure

The Context Based Interpretation Procedure includes the Clinical Data and Response Criteria Knowledge Models and the Context Based Interpretation Methods (Figure 5.4). The Clinical Data Model implemented by the Clinical Data Ontology is
described above. The Response Criteria Knowledge Model is implemented by the Response Criteria Ontology that is composed of a domain independent Response Parameter Ontology (Figure 5.5A) and a domain independent Interpretation Task Specification Ontology (Figures 5.5B, 5.6, 5.7). Seventeen and nine domain independent SWRL rule templates implement the classification and statistical interpretation methods respectively.

Figure 5.4: Ontology Development for the Context Based Interpretation Procedure. A: The Knowledge Models and Methods for the Context Based Interpretation Procedure. B: The respective ontologies and rules that encode the Knowledge Models and Methods for Context Based Interpretation Procedure. The Response Criteria ontology is composed a domain independent Parameter Ontology and Interpretation Task Specification Ontology. Seventeen classification rule templates and ten statistical rule templates compose the Interpretation Methods.

5.3.1: Response Criteria Parameter Ontology

Figure 5.5: Domain independent Ontologies for the Context-Based Interpretation Task Procedure. A: Response Parameter Ontology. B: Interpretation Task Specification Ontology
The Response Parameter Ontology (Figure 5.5A) includes domain independent response parameters that represent Finding, Clinical State, Response State and Summary Statistic Parameters. A domain specific implementation requires extension of the Parameter Ontology with lower level domain specific parameter entities that will be described in Chapter 6.

5.3.2: Interpretation Task Specification Ontology

Interpretation Task Specifications were organized along the axis of single context or multiple context interpretation task, and classification or statistical task (Table 4.1 and Figure 5.5B). For purposes of implementation in the OWL and SWRL formalism, interpretation tasks were further categorized according to the number of input parameters and properties they required. The interpretation task specification hierarchy is shown in Figure 5.5B. Table 5.1 shows a more detailed classification of the interpretation task specifications for Single Context Single Parameter Interpretation Task. The respective ontology implementation is shown in Figure 5.6. Highlighted leaf nodes represent the lowest level of task categorization relating to a respective SWRL rule template. The lower task ontology was driven in large part by the expressive features of the SWRL rule language. The categories of interpretation tasks also contain relevant property relationships to enable encoding of the interpretation task knowledge to be used by a respective interpretation rule template.

Table 5.1: Classification of Interpretation Task Specifications for Single Context Classification with a Single Input Parameter. Tasks have either a single or two parameter restrictions and perform re-classification of existing parameters or create a new parameter as a new classification.

<table>
<thead>
<tr>
<th>Single Context Classification Task Specification</th>
<th>Single Input Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qualitative Classification</td>
<td></td>
</tr>
<tr>
<td>Single Parameter Restriction</td>
<td>Re-Classification</td>
</tr>
<tr>
<td>• Class Restriction</td>
<td>• Class Restriction</td>
</tr>
<tr>
<td>• Class Property Restriction</td>
<td>• Class Property Restriction</td>
</tr>
<tr>
<td>Two Parameter Restrictions</td>
<td>New Classification</td>
</tr>
<tr>
<td>• Class Restriction</td>
<td></td>
</tr>
</tbody>
</table>
A Temporal Restriction (Figure 5.2A) and a set of input and output Parameter Restrictions define the specification for an Interpretation Task. The Parameter Restriction Ontology is shown in Figure 5.7 with the most commonly used restrictions highlighted in grey. At the top of the hierarchy is the simplest Parameter Restriction that only specifies the class of a Parameter to be used for reasoning. Further down the hierarchy are more complex Parameter Restrictions that specify a particular property value or relation restriction. For instance, the Single Quantitative Property Value Relation Restriction (PPVR_SingleRelationRestriction) specifies a parameter, property, quantitative value, and relation operator that will be used for reasoning in a Single Context Single Parameter Quantitative Classification Task. Qualitative Parameter Property Value Restrictions similarly specify the parameter, qualitative property, and qualitative values that will be used for reasoning in either classification or statistical tasks.
5.3.3: Context-Based Interpretation Rule Templates

The goal in implementing the interpretation rule templates was to create domain independent methods that could be re-used by multiple instances of domain specific interpretation task specifications. To this end, the rule templates were architected in a one-to-one relationship with their respective class of Interpretation Task Specification. This results in a one-to-many relationship between rule templates and instances of interpretation task specifications.

The knowledge needed to execute a particular interpretation task is encoded in the ontology as an instance of an Interpretation Task Specification with its respective Temporal and Parameter Restrictions. A rule template does not contain any domain specific knowledge and thus must access the ontology for this knowledge. The architecture of a rule template thus consists of three parts: 1) a query of instances of the task specification knowledge in the ontology, 2) a query of instances of response data that meet the task specification input constraints, and 3) an assertion of new instances of response data that meet the output constraints. The rule templates contain the architecture for manipulating different combinations of input and output restrictions as well as domain independent mathematical formulas.

Twenty-seven domain independent interpretation rule templates were implemented in order to execute the Interpretation Task Specifications for the RECIST solid tumor response criteria. Table 5.2 shows the classes of interpretation rule templates that were created for this implementation. Seventeen rule templates were
classification rules while ten were created to implement the statistical rules. Nineteen were single context rules while eight were multiple context rules.

Table 5.2: Interpretation rule templates created to implement the RECIST response criteria Interpretation Task Specifications.

<table>
<thead>
<tr>
<th>INTERPRETATION RULE TEMPLATES</th>
<th>Single Context</th>
<th>Multiple Context</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification Rule</td>
<td>13</td>
<td>4</td>
<td>17</td>
</tr>
<tr>
<td>Statistical Rule</td>
<td>6</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>Total</td>
<td>19</td>
<td>8</td>
<td>27</td>
</tr>
</tbody>
</table>

Figures 5.8 and 5.9 show examples of two implemented interpretation rule templates for classification and statistical task reasoning respectively. Both rule templates follow the rule architecture described above. For example, the rule template in Figure 5.8 performs reasoning for a single parameter, in a single context, with a quantitative property relation restriction. First the rule queries the ontology for an interpretation task specification of that class. It queries for the context restrictions and parameter restrictions for that interpretation task. In this particular rule template, the input parameter restriction includes specification of the parameter, quantitative property, relation operator, and quantitative value threshold. Next, the rule template queries the ontology for subject level response data that meet the input parameter restrictions. The rule templates utilized the abox reasoning SWRL built-ins to identify the subject parameters that meet these restrictions, there-by enabling domain independent instantiation of the rule templates. Subject level findings that meet these restrictions are then re-classified with new assertions according to the output parameter restrictions, again with the use of abox reasoning.
Figure 5.8: Rule template for Single Context Single Parameter Quantitative Property Relation Restriction Classification Task. The rule template consists of three parts that 1) query instances of the task specification knowledge in the ontology, 2) query instances of response data that meet the task specification input constraints, and 3) assert new instances of response data that meet the output constraints.

Figure 5.9 shows the implementation of a rule template that executes the single parameter single context set aggregation statistical task. This statistical rule maintains the same overall rule architecture as the classification rule in Figure 5.8 by querying the ontology for the knowledge related to an instance of a specific interpretation task specification class, querying for subject level response parameters, and asserting new subject findings into the ontology. This rule however, makes use of the SWRL set operators to take the sum of an undefined number of parameters with quantitative properties. This aggregated value is then asserted back into the ontology as a new entity and classified according to the output parameter restrictions.

The twenty-seven interpretation rule templates all follow this same architecture with variations for the number and type of input parameters, single or multiple context reasoning, and classification or statistical tasks. The rule template format also makes it easy for knowledge engineers to extend the rules to handle new types of reasoning tasks. This also has the advantage of enabling integration of statistical and
classification tasks within a single decision logic. The Rule-Based Response Assessment Framework is thus completely implemented in the ontologies and rules described above. The next chapter describes how the implemented framework is instantiated for two oncology response criterion over two cancer clinical trial data sets, and for a rheumatoid arthritis criteria.

Figure 5.9: Rule template for Single Context Single Parameter Set Aggregation Statistical Task.
CHAPTER 6
DOMAIN APPLICATIONS

The Rule-Based Response Assessment Framework is applied to response criteria in the clinical domains of oncology and rheumatology. Two oncology response criteria are applied to two oncology clinical trial data sets. The framework is then applied to instantiate a Rheumatoid Arthritis response criteria. These instantiations are described in detail below.

6.1: Application of the Framework to Solid Tumor Response Criteria

6.1.1: Solid Tumor Clinical Trial Data

Data was acquired from an ongoing Phase II metastatic colon cancer therapeutic clinical trial being conducted at Stanford Comprehensive Cancer Center to evaluate the application of the Framework to the RECIST 1.1 solid tumor response criteria. The clinical trial is a single arm experimental combination chemotherapy and biologic therapy trial with the primary endpoint of objective response rate. Patients received a maximum of 6 cycles of the investigational regimen every 3 weeks. Response was assessed using the RECIST method with CT as the primary assessment modality. CT scans of the chest, abdomen and pelvis were taken at baseline no more than 4 weeks prior to the start of treatment in accordance with the RECIST guidelines, and then at follow-up after every 2-3 cycles (6-9 weeks). Ten patients were enrolled in the study from 2005 to 2008. Each patient had one baseline study and at least one follow-up study for a total of 30 CT scans of the chest, abdomen and pelvis. The longest diameter of the target lesions and the presence of non-target lesions were recorded on RECIST flow sheets with response rates manually calculated by the primary clinical trial investigator. The primary investigator’s RECIST flow sheets were used to establish the human interpretation standard to assess accuracy of the automated implementation of the Response Assessment Method.
6.1.1: Image Annotations

The imaging data from CT scans in the form of image annotations was the primary response biomarker used for the oncology solid tumor use case. Response data for the clinical trial had been recorded on paper RECIST flow sheets. In order to make this computable and leverage the semantic and quantitative content of the data for automated reasoning, the images were re-evaluated to identify and measure the selected target and non-target lesions using an image annotation tool called the image Physician Annotation Device (iPad) (102). iPad is a plugin to the OsiriX (103) image viewing application and implements the Annotation and Image Markup (AIM) standard. AIM, a project of the National Cancer Institute cancer Informatics Grid (caBIG) (104), provides the data structure for storing the key semantic lesion information (102) such as lesion identification, location, size measurements, method of measurement, and other quantitative features. AIM utilizes the RadLex (105) controlled terminology for describing the contents of medical images, and provides a standard information model for semantic annotations. Information about image annotations is recorded in AIM as XML compliant with the AIM schema, enabling the consistent representation, storage, and transfer of the semantic meaning of imaging features.

6.1.2: Annotation Image Markup Ontology and Data Rules

In order to import the image annotation data into the framework, the Annotation Image Markup Ontology was developed along with a set of data rules that modified the annotation data to make it accessible to the response assessment reasoning method. Since the image annotation data from iPad is stored in XML, it was necessary to transform the data into a format that was more suitable for knowledge-based reasoning. To enable ontology-based reasoning, we transformed the AIM information model, which is described by XML Schema, into an ontological representation that defines a semantically equivalent information model. This model can both represent all the concepts in the AIM XML, and it can be used to store OWL instances of AIM annotations. Classes and properties were created in the OWL information model that corresponded to their respective components in AIM. We
utilized a tool called XMLMaster (106) to define this transformation. XMLMaster is as a plugin to the Protégé-OWL ontology development environment (100) and provides a graphical user interface that allows users to interactively define mappings between entities in an XML document and concepts in an OWL ontology. It can be used to define mappings between an XML model and an existing OWL ontology or can generate a new OWL ontology as the target of these mappings. We used this latter mode to create an OWL AIM information model.

![Diagram of data flow](image)

Figure 6.1: Image Annotation Data for Response Assessment. XMLMaster is used to define the AIM.owl model from AIM.xml model. XMLMapper is used to populate instances of AIM data from xml to owl. Finally, data rules are used to transform the AIM owl instances into the clinical data model for reasoning by the response assessment method.

The second step is to define a mechanism to transform existing AIM XML documents to their equivalent annotations encoded using the OWL information model. We utilized a tool called XMLMapper (107) to perform this task. XMLMapper uses the mappings defined by users of XMLMaster when they are specifying an XML–to-OWL transformation. XMLMaster stores these transformations in a mapping ontology that specifies of how entities can be mapped from an XML document to instances in an OWL ontology. XMLMapper uses this mapping ontology to automatically transform XML documents to OWL ontologies. It can process streams of XML
documents and populate an OWL knowledge base with the resulting transformed content. Using this approach we transformed 115 AIM XML instances into a single AIM data ontology.

While the AIM data ontology improved the capacity to reason over the AIM instances, additional data preprocessing was required prior to performing the response assessment method. This included asserting the valid time of the annotation from the date of the imaging study, calculation of length measurements from a set of pixel coordinates, identification of the longest diameter from a pair of length annotations for a given lesion, and assertion that a finding was identified based on the geometric shape of the markup used for annotation. These assertions were made with a set of SWRL data rules that are outside of the framework model. Finally, the properties of the image annotation instances required for response assessment were further mapped to the data parameter information model using additional data rules. These properties include the lesion ID, valid time, location, qualitative descriptions, long axis, short axis, and a property that indicates whether an image finding was identified or not.

6.1.3: Instantiation of the Response Evaluation Criteria in Solid Tumors

Several steps were required to utilize the framework implementation for automated assessment of treatment response in the colorectal clinical trial using the RECIST criteria (Figure 6.2): 1) the details of the study protocol required for response assessment were modeled in the protocol ontology, 2) patient image annotation and
treatment data was aligned with the clinical data ontology, and 3) the response framework ontology was extended with RECIST specific parameters & interpretation tasks. These steps are discussed in greater detail in this section.

6.1.3.1: Instantiation of Study Protocol Specification

Instantiation of the protocol specification for the GI protocol were previously described in Section 5.2. In summary, the GITrial has a PlannedGIProtocol with a single experimental GIPlannedArm with an associated GITreatmentSpecification containing the GITreatmentRegimen. In addition, the planned study protocol must have baseline and follow-up context specifications with their associated anchor specifications, composite context specifications, and context reasoning group specifications. The PlannedGIProtocol has a single StartTreatmentAnchorSpecification, a single GIBaselineContext, and four Follow-up Context Specifications with their four respective Composite Context Specifications. Lastly, the planned study protocol must specify a response criterion that will be used for response assessment, in this case the RECIST criteria is assigned. An actual study arm is then created that correlates with the planned study arm. The actual study arm specifies a list of patients enrolled on that arm. In this case the ActualGIStudyArm was linked to the GIPlannedArm, and the 10 patients on the study arm were assigned using the minimum required unique patient identifier.

6.1.3.2: Importing Image Annotation Data to Clinical Data Ontology

The next step was to align the patient image annotation and treatment data with the clinical data ontology and the research data ontology. As described previously in section 6.1.2, data rules were used to align the patient image annotation data with the response criteria parameter data model to enable efficient reasoning. The specific dates of treatment administration were not available for each cycle of therapy but rather for the start and end of the entire treatment regimen. These were derived from patient case report forms and manually entered into the treatment event section of the clinical data ontology. Finally, a data rule was used to align the patient identifiers in the AIM data ontology with those of the patients on the clinical study to generate a single unified response data ontology over which reasoning would take place.
6.1.3.3: Domain Extension of Response Parameter Ontology

The Response Parameter Ontology was extended with domain specific data parameters, finding abstract parameters, clinical state parameters, response state parameters, and summary statistic parameters (Figure 6.3). The imaging data parameter was added to the ontology complete with all of the imaging features required for RECIST response assessment. This parameter was aligned with the aim data model using data rules described previously. The finding abstract parameters were extended to include oncology specific findings such as Enlarged Lymph Node.
and Non-Target Lesion (Figure 6.3A). In total, 27 new finding abstract parameters where added. Similarly, 4 oncology and RECIST specific clinical state parameters were added to the parameter ontology (Figure 6.3B). Fourteen response state parameters were added to the ontology including Non-Target Lesion Response Category and Target Lesion Response Category to name a few (Figure 6.3D). Lastly, five new parameters were added to the response summary statistic parameter ontology including Count of Complete Response and Time to Disease Progression to name a few (Figure 6.3).

6.1.3.4: Instantiation of Domain Specific Interpretation Tasks

Table 6.1: Categories of Interpretation Task Specifications for RECIST implementation.

<table>
<thead>
<tr>
<th>RECAST</th>
<th>INTERPRETATION TASK SPECIFICATION</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Finding Level</td>
</tr>
<tr>
<td>Single Context Classification Task</td>
<td>28</td>
</tr>
<tr>
<td>Single Context Statistical Task</td>
<td>2</td>
</tr>
<tr>
<td>Multiple Context Classification Task</td>
<td>4</td>
</tr>
<tr>
<td>Multiple Context Statistical Task</td>
<td>0</td>
</tr>
<tr>
<td>TOTAL 105</td>
<td>34</td>
</tr>
</tbody>
</table>

The Response Framework was further extended through the addition of RECAST specific interpretation task specifications. In total 105 instances of interpretation tasks were specified in the interpretation task specification ontology (Table 6.1). The majority of the interpretation task specification instances specified knowledge for single context classification tasks at the Finding Level (28) and the Response State Level (45). Of the 27 interpretation rule templates, 13 had only one associated interpretation task instance (Table 6.2). The majority of the rule templates with this low ratio of rule to interpretation task instance were not surprisingly statistical tasks executing specific mathematical formulas or aggregation tasks. On the other hand, five of the classification rule templates had 5 or more associated interpretation task instances making up 65 percent of all the interpretation task
instances over only five rule templates. For instance, the rule template for Single Context Single Parameter Quantitative Property Relation Restriction Classification Task depicted in Figure 5.9 had 18 interpretation task specification instances for the RECIST implementation. Overall, the RECIST implementation had an average of 3.8 interpretation task specifications per interpretation rule template.

Table 6.2: Ratio of rule templates to interpretation task specification instances for RECIST implementation. Average of 3.8 interpretation task specifications per interpretation rule template.

<table>
<thead>
<tr>
<th>Ratio of Rule Templates to Interpretation Task Specifications</th>
<th>Number of Classification Rule Templates</th>
<th>Number of Statistical Rule Templates</th>
<th>Total Number Rule Templates</th>
<th>Number of Interpretation Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 to 1</td>
<td>6</td>
<td>7</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>1 to 2</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>1 to 3</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>1 to 4</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>1 to 5 or more</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>68</td>
</tr>
<tr>
<td>Total</td>
<td>17</td>
<td>10</td>
<td>27</td>
<td>105</td>
</tr>
</tbody>
</table>

Many of the RECIST clinical rules were broken down into several interpretation task specification instances in order to fit the rule template framework. For instance, RECIST has a clinical rule where a lymph node with a short axis between 10 and 15 millimeters is classified as enlarged but not measurable in the baseline context. This clinical rule was broken down into four interpretation task specifications using three rule templates. First, the Single Context Single Parameter Quantitative Property Relation Restriction Classification Task described in Figure 5.9 was used to classify the lymph node according to its short axis measurement as a *Lymph Node with a short axis greater than or equal to 10 millimeters* and a *Lymph Node with a short axis less than 15 millimeters* using two different interpretation task specification instances. Another rule template was used to combine the two parameters into a single parameter *Lymph node with short axis between 10 and 15 millimeters*. Finally, a third rule template was used to classify a baseline *Lymph node with short axis between 10 and 15 millimeters* as a *Non-Measurable Lesion*. Several other clinical rules in RECIST were decomposed in a similar manner to make up the set of
interpretation task specifications for the RECIST implementation.

Similarly, RECIST performs single context pattern matching across multiple parameters to determine the response category for a specific follow-up period. In the published RECIST document, a table is provided that specifies the pattern matching details for three parameters that define the overall response at a particular time point (Figure 2.8). Instead of creating a new three-parameter single context interpretation rule template for this case, the three-parameter pattern-matching task was broken down into a series of two-parameter pattern-matching tasks. In this way, the number of rule templates was minimized through decomposition of clinical tasks into a set of simpler tasks. As a result of this approach, an increased number of interpretation task specifications were created.

Once the clinical and research data had been processed, and the RECIST knowledge implemented, all the Context Generation Rules and Context-Based Interpretation Rules were run in the Protégé SWRL tab environment where the output of was asserted back into the ontology as new knowledge. The SWRL tab rules execution engine provides native technology for chaining SWRL rules according to the domain semantics. However, our implementation of the Rule-Based Response Assessment Framework using rule templates created technical limitations for use of this native rule-chaining infrastructure. In particular, the combined use of set operators and abox reasoning for creating assertions in the knowledge base created for set based rules resulted in an asynchrony of rule template firing. As a result, we had to manually rule groups of rule templates separately in order to maintain the appropriate chaining of the rules for this implementation.

6.2: Application of the Framework to Lymphoma Response Criteria

6.2.1: Lymphoma Clinical Trial Data

Data from an ongoing Phase II Follicular Lymphoma vaccine therapeutic clinical trial was used to evaluate the application of the Framework for the Lymphoma response criteria. Fourteen patients were enrolled on the trial from 2004-2007. Thirteen of the fourteen patients had measurable disease by imaging criteria, with one
patient that had only cutaneous disease to follow. Patients had a baseline study no earlier than 28 days prior to the start of therapy utilizing either CT or MRI of the Neck, Chest, Abdomen and Pelvis. Patients then received 2 days of radiation therapy to a single affected lymph node site followed by 10 weekly injections of the vaccine therapy to the same site. At the completion of treatment, the first follow-up scan was performed, followed by scans every 3 months until the patient’s disease progressed to the point where they were started on the next treatment. Several patients on study have yet to progress and continue to be evaluated every three months.

The imaging data was analyzed by secondary central review. The radiologist utilized the iPad OsiriX plug-in to annotate the lesion information. He simultaneously recorded the data in an excel spreadsheet. Comparing the user’s data entry to the intended data entry utilizing the spreadsheet, six patients had complete data that is used for the Lymphoma use case. The radiologist’s excel spreadsheet is used as the human interpretation standard to assess accuracy of the automated implementation.

6.2.2: Instantiation of the International Harmonization Criteria

![Diagram](image)

Figure 6.4: Instantiation of the International Harmonization criteria for interpretation of a lymphoma clinical trial using the Rule-Based Response Assessment Framework. Blue areas indicate implementation with domain specific knowledge.

A second oncology response criterion was then implemented to evaluate the extensibility of the Rule Based Response Assessment Framework. The International Harmonization Criteria (IHC) (1) is a response criterion for lymphoma very similar to the RECIST criteria but utilizing the sum of the products of bi-dimensions tumor
lesion measurements as an estimate of tumor burden. For the instantiation of the International Harmonization Criteria in the Response Assessment Framework, we evaluated the requirements for extending the original RECIST implementation to implement a second criterion in the same clinical domain. We followed the same three steps in order to utilize the framework implementation for automated assessment of treatment response in the lymphoma clinical trial using the International Harmonization Criteria (Figure 6.4). First, the aspects of the study protocol required for response assessment were modeled in the research data ontology exactly as was previously described for the colorectal clinical trial. Next, patient image annotation and treatment data was aligned with the clinical data ontology as was also previously described. And finally, the RECIST implementation was extended with IHC specific response parameters and interpretation task specifications.

Table 6.3: Shared Rules and Interpretation Task Specifications between RECIST and IHC.

<table>
<thead>
<tr>
<th>Interpretation Task Specification</th>
<th>RECIST</th>
<th>IHC</th>
<th>Shared</th>
<th>Percent of IHC Shared with RECIST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finding Level</td>
<td>34</td>
<td>35</td>
<td>22</td>
<td>63%</td>
</tr>
<tr>
<td>Clinical State</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>100%</td>
</tr>
<tr>
<td>Response State</td>
<td>50</td>
<td>50</td>
<td>44</td>
<td>88%</td>
</tr>
<tr>
<td>Summary Statistic</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>100%</td>
</tr>
<tr>
<td>Total</td>
<td>105</td>
<td>106</td>
<td>68</td>
<td>77%</td>
</tr>
</tbody>
</table>

Another primary distinction between IHC and RECIST is their definition of abnormal lymph nodes and the use of the product as opposed to the length for estimating the quantitative disease burden. As such, much of the encoded knowledge for the RECIST implementation was reused for the IHC implementation with the following exceptions. First, twelve new IHC specific domain parameters were added to the Response Ontology. These were largely used for intermediate classification steps such as those described above for RECIST where complex clinical rules were broken down into several interpretation task specifications and thus required the creation of intermediate parameters. Many of these additional parameters were for
IHC specific quantitative thresholds that are distinctly different from the RECIST implementation such as the use in IHC of the length measurement of both the long and the short axis of lymph nodes to classify target lesions.

Like the RECIST implementation, the IHC implementation used 106 interpretation task specification instances to encode the response criteria task knowledge (Table 6.3). Seventy-seven percent of the RECIST interpretation task instances were re-used in whole for the IHC implementation demonstrating the significant overlap in these two oncology criteria. In particular, the two response criteria had 100% overlap in their clinical state and summary statistic interpretation task specification instances. The majority of the difference between the two criteria was accounted for in their finding level and response state interpretation task specifications, yet they maintained 63% and 88% overlap in those respective types of tasks.

6.3: Application of the Framework to Rheumatoid Arthritis Response Criteria

![Figure 6.5: Instantiation of the European League of Associations for Rheumatology (EURAL) Rheumatoid Arthritis response criteria using the Rule-Based Response Assessment Framework. RA=Rheumatoid Arthritis.](figure)

Lastly, we used the Rule-Based Response Assessment Framework to instantiate the European League of Associations for Rheumatology (EURAL) Rheumatoid Arthritis response criterion using the Rule-Based Response Assessment Framework (Figure 6.5). The EURAL response criterion is used to classify treatment
response in therapeutic rheumatoid arthritis clinical trials (40). The EURAL response criterion uses the Disease Activity Score (DAS) as a quantitative estimate of rheumatoid arthritis disease activity at each assessment time point. The DAS is a quantitative clinical state estimate derived from the complex mathematical formula shown in Figure 2.11. The Disease Activity Score parameters include: 1) the Ritchie Articular Index (RAI), a count of the number of tender joints found on physical examination, 2) the SJC44, a swollen joint count for 44 joints, 3) the Erythrocyte Sedimentation Rate, which is a blood test, and 4) the Global Health score, a clinical history finding describing the patient’s overall impression of their health. The DAS is calculated at baseline and at each follow-up assessment period. The quantitative response is defined by the absolute change in the DAS from the baseline assessment period to the follow-up assessment period and by the reached DAS value at the follow-up assessment period. The qualitative response category of “Good”, “Moderate” and “None” are defined by combining thresholds of changes in the DAS and reached DAS. Summary statistics include a count of patients in each of the three response categories.

![Data Parameter Ontology](image)

Figure 6.6: Extensions to the Data Parameter Ontology for the Rheumatoid Arthritis Criteria Instantiation. Five new data parameters were added.

As before, domain specific response parameters and interpretation task specifications were added to the ontology to encode the new response criteria. Five new data parameters were added to the data parameter ontology to represent data derived from the new assessment modalities that include physical exam, laboratory and clinical history findings (Figure 6.6). Twenty-nine new abstract parameters were added to the parameter ontology (Figure 6.7). Of those, 12 were new Finding level abstract parameters such as the Ritchie Articular Index, which is a calculated sum of
all examined tender joints (Figure 6.7A). Seven new clinical state parameters were created including the Disease Activity Score (DAS) (Figure 6.7B). Seven new response state parameters were created including the Absolute Change in the Quantitative Clinical State to account for the absolute change in the DAS (Figure 6.7C). Finally, three new summary statistic parameters were created to account for differences in classification of response such as Count of Good Response (Figure 6.7D). In total, 34 new domain parameters were added to the Parameter Ontology.

Figure 6.7: Extensions to the Abstract Parameter Ontology for the Rheumatoid Arthritis Criteria Instantiation. Total of 29 new Abstract Parameters added to the Response Assessment Ontology. A: Finding Abstract Parameters for Rheumatoid Arthritis (12 new) B: Clinical State Parameters for Rheumatoid Arthritis (7 new), C: Response State Parameters for Rheumatoid Arthritis (7 new), D: Summary Statistic Parameters for Rheumatoid Arthritis (3 new).
Table 6.4: Number of Interpretation Tasks for Each Type of Response Parameter.

<table>
<thead>
<tr>
<th>Interpretation Task Specification (EURAL)</th>
<th>Classification</th>
<th>Statistical</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finding Level</td>
<td>0</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Clinical State</td>
<td>5</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Response State</td>
<td>15</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>Summary Statistic</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td>15</td>
<td>35</td>
</tr>
</tbody>
</table>

Table 6.5: Comparison of the number of classification and statistical rule templates between RECIST and EURAL response criteria.

<table>
<thead>
<tr>
<th>Rule Templates</th>
<th>Number of Classification Rule Templates</th>
<th>Number of Statistical Rule Templates</th>
<th>Total Number Rule Templates</th>
<th>Ratio Classification: Statistical Rule Templates</th>
</tr>
</thead>
<tbody>
<tr>
<td>RECIST</td>
<td>17</td>
<td>10</td>
<td>27</td>
<td>1.7:1</td>
</tr>
<tr>
<td>EURAL</td>
<td>4</td>
<td>7</td>
<td>11</td>
<td>1:1.75</td>
</tr>
<tr>
<td>% EURAL rule templates shared with RECIST</td>
<td>100%</td>
<td>29%</td>
<td>55%</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.6: Ratio of rule templates to interpretation task specification instances for EURAL implementation with an average of 3.2 interpretation task specifications per interpretation rule template.

<table>
<thead>
<tr>
<th>Ratio of Rule Templates to Interpretation Task Specifications</th>
<th>Number of Classification Rule Templates</th>
<th>Number of Statistical Rule Templates</th>
<th>Total Number Rule Templates</th>
<th>Number of Interpretation Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 to 1</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>1 to 2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>1 to 3</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>1 to 4</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>1 to 5 or more</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>17</td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
<td>7</td>
<td>11</td>
<td>35</td>
</tr>
</tbody>
</table>
Due to the complexity of the mathematical formula for the Disease Activity Score (DAS) (Figure 6.6), five new statistical rule templates were created. These rule templates included mathematical formulas such as taking the square root and natural logarithm of quantitative parameter properties. Two statistical rule templates were also re-used for a total of seven statistical rule templates to instantiate the EURAL criterion. Fifteen statistical interpretation task specifications utilized these seven statistical rule templates (Table 6.4). Four classification rule templates were completely reused from the oncology implementation along with twenty new domain specific interpretation task specifications to encode the classification logic for the EURAL criterion. Only one of these 35 total interpretation task specifications was reused from the oncology implementation.

Compared to the oncology response criteria described above, the EURAL criterion is relatively simple requiring approximately one third (35 compared to 105) the number of interpretation task specifications to encode the response criterion domain logic. Likewise, the EURAL criteria only utilized 11 compared to 27 rule templates. The ratio of statistical rule templates to classification rule templates was also inverted for the EURAL response criterion (Table 6.5). This is in part due to the fact that the RECIST criteria had a large number of finding level classification interpretation task specifications where the EURAL response criteria had none (Table 6.4) and the fact that the formula for the EURAL Disease Activity Score (DAS) is complex. Overall, the EURAL implementation was able to reuse all of the classification rule templates from the RECIST implementation but only 29% of the statistical rule templates with approximately 55% of the rule templates shared (Table 6.5). However, like the oncology implementations, the ratio of interpretation task specifications to interpretation rule templates was approximately 3.2 for the EURAL implementation (Table 6.6) compared to a ratio of 3.8 for the RECIST implementation (Table 6.2).

In summary, we have implemented the Rule-Based Response Assessment Framework for three clinical domain response criteria across oncology and rheumatoid arthritis domains. The next chapter describes the evaluation of this approach.
CHAPTER 7
EVALUATION

Evaluation of the Rule-Based Response Assessment Framework includes assessment of the implementations expressivity, clinical accuracy, the ability to reuse the code for multiple data sets, demonstration of extensibility across clinical domains, and the ability to use the framework to compare knowledge and response outcomes between response criteria.

7.1: Evaluation of Expressivity of Framework

In order to evaluate the expressivity of the implemented framework, we tested the hypothesis that a complex response criterion could be completely encoded by the implemented framework as to perform automated response assessment for a clinical trial cohort. A sufficiently expressive implementation would encode the knowledge and methods needed for the context restriction procedure and context-based interpretation procedure. This includes all of the specialized temporal reasoning tasks for the former, and the classification and statistical tasks for the later.

Chapters 5 and 6 detail the successful implementation of Rule-Based Response Assessment Framework in OWL and SWRL for two oncology response criteria. The two criteria are applied to two clinical trial data sets for treatment response assessment at the individual patient and cohort level. Full expressivity of the interpretation rules was only achieved by the use of newly developed set extensions to the SWRL formalism. Instantiation of these two criteria required the addition of domain specific response parameters to the Parameter Ontology and the development of domain specific interpretation task specifications. These tasks were executed by a set of domain independent interpretation rule templates.
7.2: Evaluation of Automated Outcomes Compared to Human Interpretation

In order to evaluate the clinical accuracy of the implementation, we tested the hypothesis that the outputs of the interpretation rules would be clinically similar to an independent human interpretation of the data using the response criteria. Human interpretations were derived from the response data flow sheets for the lymphoma clinical trial data set that were generated by an independent radiologist performing a central review of the response data. We compared the interpretation of the findings, clinical states, response states, and summary statistics between the human and the automated method. Six patients from the lymphoma clinical trial were included in the comparison analysis. These six patients had 22 imaging studies representing 22 response assessment periods, thirty-six baseline lesions, 98 follow-up lesions, 6 best response rate and category assessments, and 1 overall quantitative cohort response rate (Table 7.1). These represent three types of qualitative outcomes generated by classification methods and three types of quantitative outcomes generated by statistical methods.

Table 7.1: Results of Comparison of Human and Automated Interpretation.

<table>
<thead>
<tr>
<th>6 Patients 22 Assessment Periods</th>
<th>Percent Agreement Human-Automated Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification 36 Baseline Lesions</td>
<td>98%</td>
</tr>
<tr>
<td>Classification 98 Follow-up Lesions</td>
<td>99%</td>
</tr>
<tr>
<td>Classification 6 Best Response Categories</td>
<td>100%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Percent (Range) Difference in Quantitative Interpretations</th>
</tr>
</thead>
<tbody>
<tr>
<td>22 Quantitative Tumor Burdens</td>
</tr>
<tr>
<td>6 Best Tumor Response Rate</td>
</tr>
<tr>
<td>1 Mean Best Tumor Response Rate</td>
</tr>
</tbody>
</table>
Before moving forward with the results of the comparison, we should note that a transcription error was discovered in the flow sheet as part of this evaluation procedure. The transcription error on the flow sheet incorrectly recorded the dimensions of a baseline target lesion, and as such the error propagated thru the findings, clinical state, response state for that patient and even impacted the summary statistic. The lesion size was recorded as 2.5 x .13 cm on the flow sheet, when in fact it was 2.5 x 1.3 cm as determined by evaluation of the raw image annotation data. This gave a different product calculation for the long and short axis as required by IHC. Since it was a baseline lesion, the error changed the baseline sum of products, and as such the best percent change in sum of products from baseline was inaccurately assessed as -27% when it should have been -37%. This would not have changed the category of response for that patient but changed the summary statistic for the cohort’s overall response rate. This illustrates the potential human errors that can occur when performing an analysis that requires manual transcription of findings from one medium to another. The error in the human interpretation was corrected for the remainder of the analysis that proceeds since it was due to an error in transcription, not in interpretation.

Comparison of the finding level results includes the classification of the baseline image findings as target, non-target, or normal lesions, and classification of the follow-up image findings as persistent, resolved, new, or normal. Thirty-six image findings were recorded at baseline. Of those, thirty-five were classified exactly the same between the human and automated interpretations. The single difference in classification of the finding is accounted for by an increased precision of the automated method compared to the human who rounded decimal places when recording lesions on the flow sheet. In this specific case, the user very appropriately rounded the short axis of a lymph node from 10.04 to 10 mm on the flow sheet. The short axis of a lymph node must be greater than 10 mm to meet criteria for being enlarged, and as such the finding was considered a normal baseline lymph node. The automated tool however, used the precision of 10.04 mm that technically meets the criteria of greater than 10 mm and thus classified the image finding as an enlarged
lymph node. Human annotation of image findings can be variable within several millimeters when using manual image caliper tools. As such, rounding to the nearest decimal place is a common practice in dictated radiology reports and reflects the appropriate precision of the measurement tool. However, users might inconsistently classify findings under these borderline circumstances, where an automated method would not.

Similarly, the follow-up lesions were classified as persistent target or non-target lesions, resolved target or non-target lesions, new lesions, or normal findings. The 36 baseline lesions were observed in follow-up a total of 98 times. Of those, the human interpretation and automated interpretation agreed 97 of 98 times. As before, the disagreement in classification of this one image finding was due to a resolution of precision of the baseline finding being classified as normal and later the follow-up finding being classified as new.

The quantitative clinical state calculation of the sum of products at each baseline and follow-up assessment period was compared between the human and automated calculation. Over the 22 assessment periods, there was on average a 1.2% difference in the sum of products calculation with a range of 0.15% to 5%. The single value with the 5% difference was due to inclusion of a different set of target lesions between the two groups due to the difference in lesion classification described earlier. Similarly, the quantitative response state was compared between the two groups. The best overall percent change in sum of products from nadir to baseline was compared for the 6 patients. On average, the quantitative response state calculation differed by 1% with a range of 0.25%–4.4%. This difference in the quantitative response state however did not impact the qualitative response state classification between the two groups. As such the two groups had 100% agreement in classification of best overall response category. As such the groups also had 100% agreement in the summary statistics that presented a count of the response categories, but disagreed on the calculation of the mean best quantitative response state by 3.1%. Overall, there was a high degree of agreement between the human and automated response assessment interpretations.
7.3: Evaluation of Reuse of Implementation

In order to evaluate the reuse of the implementation, we testing two hypotheses: 1) that the code could be reused to apply a response criteria to multiple data sets and 2) that knowledge encoded in the framework could be reused between two closely related response criteria in the same domain. In order to test the first hypothesis, we applied the Solid Tumor Response Criteria (RECIST) to both the colorectal clinical trial data set (Figure 6.2) and the lymphoma clinical trial data set (Figure 7.1). We thus demonstrate that the encoded implementation can be reused with multiple clinical trial data sets.

In order to test the second hypothesis for knowledge reuse, we evaluated the percentage of the domain specific knowledge that could be reused between the two implementations of the oncology response criteria. To do this, we evaluated the number of domain specific parameters and interpretation task specifications that were shared between the two implementations. We found almost complete reuse of the oncology specific parameters, and 73% reuse of the interpretation task specifications (Table 6.3). This represents a significant amount of reuse of domain knowledge between two different oncology response criteria.

![Diagram](image)

Figure 7.1: Reuse of the implementation through application of a single response criteria to multiple trial data sets. Application of the Solid Tumor Response Criteria (RECIST) to the lymphoma clinical trial data set.
7.4: Evaluation of Extensibility of Framework

In order to evaluate the extensibility of the framework, we tested the hypothesis that a minimum amount of domain specific knowledge would need to be added to the ontologies and rules to instantiate multiple response criteria across multiple clinical domains using the Rule-Based Response Assessment Framework. There is no agreed upon threshold that defines a “minimum amount” of knowledge needed to extend a knowledge representation system for another instance. As such, we performed an evaluation of the amount of knowledge engineering that was required to extend the framework in two settings: 1) to instantiate a response criterion in the same clinical domain of oncology, and 2) to instantiate a response criteria in a different clinical domain of rheumatoid arthritis. The experiment for extensibility of the framework consisted of holding constant the RECIST implementation and extending the ontologies and rules as needed for the IHC and rheumatoid arthritis criteria.

As described previously in Chapter 6, the IHC implementation reused 100% of the context generation and interpretation rule templates from the RECIST implementation without requiring the creation of any additional rule templates. In order to fully encode the IHC response criteria, the parameter ontology was extended with 12 new IHC specific parameters (Figure 6.4). Sixty-eight interpretation task specifications were re-used from the RECIST implementation for the IHC implementation with the creation of 38 additional IHC specific interpretation task specifications (Table 6.3). This experiment demonstrated that the context generation and interpretation rule templates were completely reused with the domain knowledge of the second oncology response criteria completely encoded within the ontology.

For the second experiment of extensibility of the framework we encoded a response criteria from different clinical domain. Domain specific response parameters and interpretation task specifications were added to the ontology to encode the EURAL Rheumatoid Arthritis Response Criterion. In total, 34 new domain parameters were added to the Parameter Ontology. Thirty-five interpretation task specifications were used to encode the EURAL response criteria logic. Of those, only
one interpretation task could be reused from the RECIST implementation. The EURAL implementation only required the use of 11 rule templates. Of those, 4 were classification rule templates that were completely re-used from the RECIST implementation. The remaining 7 were statistical rule templates, of which only 2 could be reused from the RECIST implementation. Five new statistical rule templates were created to account for the unique mathematical formulas used in the EURAL response criteria.

This demonstrates how the implementation of the Rule-Based Response Assessment Framework is not complete but extensible. By design, the Framework allows for the expression of any complex mathematical statement. As such, it is expected that the rule templates would need to be extended for cases that include new types of mathematical formulas. The Rheumatoid Arthritis criteria utilized mathematical formulas such as the square root and the natural logarithm of quantitative parameter properties. Since these mathematical formulas were not used in the Oncology response criteria, rule templates needed to be created to express these specific formulas for the Rheumatoid Arthritis criteria. This extensibility is one of the major strengths of this implementation of the Rule-Based Response Assessment Framework.

At the same time, the classification rule templates were completely reused for the EURAL implementation. This demonstrates the differences in the completeness of the framework for classification versus statistical tasks. While we cannot guarantee that the classification rule templates are absolutely complete in this implementation for cases not tested, we do believe that they represent a wide range of simple and complex classification tasks used in the clinical response criteria domain.

**7.5: Comparison of Response Criteria**

In order to evaluate the ability of the framework to enable comparison of criteria knowledge and outcomes we tested the two following hypotheses: 1) that comparison of domain parameters and interpretation task specifications can enable knowledge level comparison of two response criteria and 2) that the impact of these knowledge level differences could be assessed by comparing the outcomes of applying
two response criteria to the same data set. Intuitively, the two oncology response criteria that were implemented have many similarities since they are both evolutionary descendents of the WHO oncology criteria, and both predominantly utilize imaging response biomarkers to perform response assessment. In addition, the two criteria have similar therapeutic goals of evaluating for tumor shrinkage and monitoring for tumor growth or progression. Our framework enables an evaluation of the knowledge level similarities and differences between these two different criteria.

Specifically, at the parameter level, the two criteria are practically identical with the exception of the introduction of intermediate response parameter concepts. Some of the differences between these response parameters were extremely subtle and perhaps arguably clinical insignificant. For example, there was a subtle distinction in the cut offs for lymph node classification with lymph node short axis GREATER-THAN-OR-EQUAL-TO 10 in RECIST and GREATER-THEAN 10 for IHC. These are trivial differences between the two criteria, yet we wanted to maintain the exact classifications according to the response criteria documents in our implementation. More significant parameter level differences are for the cut offs for classification of response where RECIST uses the sum of the diameters and IHC uses the sum of the products. Thus very appropriately, the two criteria have different thresholds for defining response with between a single dimension and two dimension quantitative estimate of disease burden. These differences were implemented as different response parameter and interpretation task specifications.

Differences between the two criteria are better evaluated by demonstrating the ability to reuse interpretation task specifications between the two implementations. Table 6.3 shows the overlap in the interpretation task specifications and interpretation rules between the RECIST and International Harmonization Criteria. Overall, the two criteria share approximately 77% of their task specification knowledge. Interestingly, 100% of the rules are shared at the clinical state level and for the summary statistics suggesting a strong concordance in their overall approach to response assessment. Differences emerged primarily at the finding level classifications where they shared only 63% of their interpretation task specification instance. Largely, the differences
between the two criteria stem from variation in definition of thresholds for classification tasks. Specifically, the two criteria have different interpretation task specifications for defining enlarged lymph nodes. Since they also use different quantitative estimates of tumor burden (sum of diameters for RECIST and sum of products for Lymphoma criteria), they very appropriately have different thresholds for defining response categories thus accounting for the only 88% overlap in response level interpretation task specification instances. Again, many of the differences between the two criteria were subtle, where RECIST uses the 5 largest measurable lesions and IHC uses the 6 largest. These don’t represent major breaks in the reasoning methodology, but simply subtle differences in the definitions for the interpretation task specifications.

Table 7.2: Comparison of Lymphoma and RECIST criteria outcomes

<table>
<thead>
<tr>
<th>6 Patients 22 Assessment Periods</th>
<th>Percent Agreement Lymphoma/RECIST Response Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification 36 Baseline Lesions</td>
<td>81%</td>
</tr>
<tr>
<td>Classification 98 Follow-up Lesions</td>
<td>82%</td>
</tr>
<tr>
<td>Classification 6 Best Response Categories</td>
<td>83%</td>
</tr>
</tbody>
</table>

In order to test the second hypothesis, that the framework could enable comparison of the output of two different response criteria to the same patient data set, we performed a small experiment as a proof of concept. We applied both the RECIST and Lymphoma Criteria to the 6 patient lymphoma data set (Figure 7.1) and compared their outputs (Table 7.2). Given that the quantitative clinical state calculations differ considerably between the two criteria, comparison of the quantitative outcomes was not considered relevant. However, the impact of these quantitative outcomes can be assessed through their ultimate impact on the qualitative response state values. With respect to finding level classifications, there was 81% agreement between IHC and RECIST in classification of baseline image findings as normal, target or non-target
lesions. Disagreements occurred due to differences in threshold definitions of measurable disease and inclusion of five vs. six largest lesions as target lesions for RECIST and IHC respectively. Similarly, there was an 82% agreement in classification of follow-up image findings as resolved, persistent, new or normal. Again, this was due to differences in threshold definitions for enlarged and normal lymph node lesions between the two criteria.

These differences in finding classification resulted in only an 83% agreement of the best overall qualitative response category between the two criteria for the six patients in the data set. Arguably, our data set is small, and the significance of this difference would need to be assessed with a much larger data set. Furthermore, in an ideal setting, differences in outcomes of response criteria should also be correlated with survival data to identify which method most closely correlates with this gold standard of efficacy. This small experiment however, demonstrates another use of the Rule-Based Response Assessment Framework to evaluate the response criteria themselves. In conclusion, these evaluations demonstrate that the rule-based framework for response assessment enables an extensible, reusable method for representing, applying and comparing response criteria.
CHAPTER 8
DISCUSSION

This dissertation describes the Rule-Based Response Assessment Framework, a goal-driven, context-driven, and knowledge-based temporal interpretation method implemented with ontologies and rule templates to perform procedural response assessment tasks. We have demonstrated that the implementation of this framework is sufficiently expressive to encode response criteria in the clinical domains of oncology and rheumatoid arthritis and clinically accurate as applied to clinical trial cohort data sets. This chapter summarizes the informatics and clinical contributions and limitations of the Rule-Based Response Assessment Framework

8.1: Informatics Contributions

8.1.1: Definition of the response assessment task

The work presented in this dissertation makes several informatics contributions in the area of temporal interpretation methods and rule-based systems. First, this work explicitly defines response assessment as a distinct type of temporal interpretation task. Historically, response assessment has been classified as a type of diagnosis task (50). A careful analysis of clinical tasks in this work however, shows that the response assessment task is a distinct type of temporal interpretation task that requires a treatment context (Figure 1.4). Furthermore, we have distinguished between the closely related tasks of response prediction and response assessment (Figure 1.5). While these two sub-tasks may be closely integrated in some temporal reasoning methods, they are distinctly different in that response prediction is a type of projection task that reasons about future states, while response assessment is a type of temporal interpretation task that reasons about the past up to the present.

Several temporal interpretation methods have solved the response assessment task for various clinical toxicity and efficacy assessments (Table 3.1). However, most
methods solve the response assessment task using a data driven approach where trends of increasing and decreasing values or quantitative thresholds define the response state. These methods are not sufficient to solve the goal-driven, procedural approach required by formal response criteria. The Rule-Based Response Assessment Framework describes both the knowledge models and methods required to solve a goal-driven response assessment task. We have demonstrated that the implementation of this framework is sufficiently expressive to encode three procedural domain response criteria.

8.1.2: Contributions to temporal interpretation methods

The Rule-Based Response Assessment Method is a goal-driven and context-driven temporal interpretation method. The method is goal-driven in that it models procedural response criteria where all of the states can be anticipated in advance in order to classify patient response. As opposed to data-driven approaches that generate all possible states and trends, the goal-driven approach solves a specific complex problem and generates a final set of usable answers. This has significant advantages where the problem that is trying to be solved is known in advance and is solved for multiple patients over and over again as in the case of response criteria applied to clinical trial data sets. A data-driven approach is more flexible in situations where ad-hoc querying is desired and the combination of states cannot be anticipated in advance.

The Rule-Based Response Assessment Method also makes contributions with its domain independent Context Restriction Methods and the Context-Based Interpretation Methods. The Context Restriction Methods not only perform the context generation task, but also a context restriction task that limits the data to be interpreted with a given context to those defined by the response criteria. The context generation task in particular enables specification of context from more than one clinical event, a minor enhancement over the KBTA method (67). None of the other temporal interpretation tasks reviewed performed context restriction as required by domain response criteria.
Finally, the Rule-Based Response Assessment Method dramatically increases the range of classification and statistical interpretation sub-tasks that can be used for temporal interpretation. The Context-Based Interpretation Methods are domain independent temporal interpretation methods that can be chained together to perform the response assessment task for a given response criteria. The domain knowledge for the response criteria is encoded in the interpretation task specifications in the knowledge model, completely separate from the rule template methods. The implemented interpretation task specifications and interpretation rule templates perform a variety of subtasks including single context, multiple context, classification and statistical tasks (Table 4.1). These tasks also include single or multiple input parameters with qualitative or quantitative data types. Very few of the temporal interpretation methods reviewed performed any type of statistical reasoning task (Table 3.5). Furthermore, the rule template framework allows classification and statistical rules to be seamless integrated with one another as opposed to calling outside quantitative reasoning services. The statistical interpretation methods implemented for the three domain response criteria demonstrate the expressiveness of the implemented framework. However, by their nature, statistical interpretation methods can be highly variable with almost limitless combinations of parameters making up mathematical formulae. Therefore, we claim that the statistical interpretation methods in the Rule-Based Response Assessment Method are extensible but not complete. Classification tasks on the other hand are much more restrictive. While the framework is sufficiently extensible such that new classification methods can be implemented if needed, we did not find that new classification rule templates were required for implementation of our second and third domain response criteria. This argues for but does not guarantee, completeness of the classification subtasks for the Rule-Based Response Assessment Method.

8.1.3: Knowledge-based implementation

One of the key advantages of the Rule-Based Response Assessment Framework is the knowledge-based approach to implementation of this complex temporal interpretation task to enable reuse of the domain independent components
from one domain implementation to the next. The first method to take this approach was Shahar’s KBTA (61) method that used ontologies to store domain knowledge and reusable domain independent interpretation methods to implement the reasoning task. Like KBTA, the Rule-Based Response Assessment Framework is implemented as a domain independent upper response assessment ontology that can be extended with domain specific parameters and interpretation task specifications, and a set of domain independent reasoning methods. The domain independent reasoning methods in KBTA were implemented in the CLIPS rule language and represent a form of black box reasoning in that the knowledge engineer cannot inspect or extend the rules but their behavior is predictable. On the other hand, the rule templates for the Rule-Based Response Assessment Method are implemented in SWRL as part of the formal knowledge representation formalism of OWL. Like OWL, SWRL is based on the description logic formalism. SWRL rules have a formal serialization with OWL that enables a standard way of storing rules with OWL so that they can be saved and transported with the ontology. This enables sharing not only of the ontology, but of the deductive methods as well. Furthermore, the SWRL rule templates can be inspected and extended by the knowledge engineer and directly reference interpretation task specification concepts in the Response Assessment Ontology.

8.1.4: Rule Templates for domain knowledge acquisition

The domain independent rule template implementation also enables fewer rules to be developed for each response criteria implementation. The RECIST implementation for example, required 105 interpretation task specifications for only 27 interpretation rule templates. A domain specific implementation where each clinical rule is encoded as its own SWRL rule would have resulted in at least 105 domain specific SWRL rules. There would likely have been even more rules given that a key contribution of the rule template implementation is the ability to specify multiple contexts over which the rule is valid. For example, the sum of longest diameters is a statistical interpretation rule that is valid for both the baseline and follow-up context. Without the rule template implementation, this rule would have had to be written twice, once for the baseline context and once for the follow-up
context. Thus, the rule template implementation provides developers some efficiency in modeling the knowledge within the ontology in not having to re-code what can otherwise be abstracted as generic rules. The rule templates can also be largely re-used as in the case of the classification interpretation rule templates for the second and third response criteria implementations, or extended as in the case of the statistical interpretation rule templates for the rheumatoid arthritis response criteria implementation. Extending the rule templates is very simple given that the existing rule base has a very consistent structure that enables copying of rules and making minor modifications as needed to extend the rule-base.

8.1.5: Evaluation of SWRL extensions

The rule template implementation of the Rule-Based Response Assessment Method would not have been possible without the recent extensions to the Protégé implementation of the SQWRL rule language to enable abox, temporal, and set based reasoning (97). The Abox reasoning extension to SWRL (96) can be used for two types of assertions: (1) class assertions – an individual is an instance of an entity and (2) role assertions – an individual participates in a role (relation). Class and roles assertions are used extensively in the rule template implementation to enable identification of individuals in the ontology that meet the temporal, parameter and property restrictions. The temporal extensions to the SQWRL language are extensively used to model the Context Restriction Methods. They provide a concise set of operators for efficiently creating temporal queries. The temporal interpretation rules on the other hand are largely atemporal in that they did not directly utilize temporal primitives to reason, but instead utilize derived temporal contexts. The temporal interpretation rules however make extensive use of the newly developed set operators for SQWRL. These set operators effectively close the world in OWL which otherwise maintains an open world assumption. This was a key innovation without which this implementation would not have been possible. In a co-development effort with the SQWRL developer, the response assessment use case tested and pushed the limits of the set operators with a myriad of set based reasoning requirements.
8.1.6: **Reuse and integration of existing ontologies**

Finally, to facilitate the possibility of sharing of knowledge and data between systems, we have reused and extended existing clinical research ontologies where possible. Utilization of ontologies for clinical research such as BRIDG and AIM that are emerging as standards has the advantage of enabling a direct mapping with the response assessment ontologies instead of having to manually encode the trial protocol or data each time. We have demonstrated in part what is required to do this for the AIM data by transforming xml instances from the iPad image annotation tool into OWL instances. From there a mapping into the response assessment ontology clinical data model was performed using SWRL rules. However, this was not an automated process and additional development would be required to automatically import the AIM xml instances into the ontology.

8.2: **Informatics Limitations**

8.2.1: **Trade-offs of rule template implementation**

The Rule-Based Response Assessment Framework however, has several informatics limitations. One of the key tradeoffs in choosing a rule template implementation is the need to formally chain the rule templates for a particular response criteria implementation. This would not have been necessary if we had not utilized a rule template approach and instead implemented domain specific rules. In particular, the addition of set-based reasoning limited the ability to utilize the native SWRL rule-chainer given the extensive use of abox reasoning. Thus chaining of rules and running them in a particular order became a requirement. This requirement however can be handled programmatically using meta-rules. Tools are being developed in the Protégé SWRL tab to handle these dependencies.

Another tradeoff in utilizing domain independent rule templates is the possibility that the domain specific rules are less comprehensible to a reviewer. The rule templates are very abstract in that they reason about generic contexts, parameters, and properties. Domain specific rules on the other hand can be quite easy to read and comprehend. There will always be a trade-off for domain independent knowledge-
based reasoning methods to provide a comprehensible framework for knowledge engineers while maximizing reusability.

Finally, embedding mathematical procedures into a logical framework has several advantages and disadvantages. One could argue that such statistical tasks could be implemented outside of logical frameworks in machine code more efficiently and perhaps with more expressivity. However, the statistical tasks in response criteria and closely integrated into the complex decision logic of classification and statistical tasks. Full integration of the statistical reasoning tasks along with the classification tasks enables a complete and logical approach.

8.2.2: Missing data

In addition, the current implementation of the Response Assessment Method assumes that there is no missing data. Specifically, it assumes that if a finding is evaluated in the baseline assessment period, then it has been evaluated at each subsequent follow-up assessment period. However, even in strictly protocol driven care, missing data is not uncommon. For instance, during the process of central review, one of the contributing factors to differences in response assessment between local providers and central reviewers is that up to 10% of data is often unavailable for the central reviewer at the time of their secondary analysis (8). The RECIST response criterion does provide guidance on how to handle the situation of missing data and this was implemented for the current version of the criteria. However, OWL and SWRL only reason about actual entities and individuals that can be queried, they do not reason about the absence of individuals. This implementation of the RECIST response criteria for example, requires input data that accounts for and classifies response data as missing.

8.2.3: Lack for temporal truth maintenance

Furthermore, in prospective data collection, data for analysis can arrive out of order. In such cases, a truth maintenance system is required to update interpretations upon the arrival of new data. OWL does not allow knowledge to be re-asserted and thus cannot be used for truth maintenance. Truth maintenance for this implementation
would need to be provided by a system outside the current implementation such as a database that could be updated.

8.3: Clinical Contributions

8.3.1: Formal model of dimensions of response criteria

This dissertation makes several clinical contributions to the domain of treatment response assessment in medicine. First, an analysis was performed of domain response criteria across oncology, HIV, and rheumatoid arthritis identifying the common features of response criteria. This analysis motivated the development of a formal model for treatment response assessment in medicine. This model was encoded in the Rule-Based Response Assessment Method for automated interpretation of response data for clinical trial data sets. The model of response assessment provides benefits beyond automation, it clearly identifies the dimensions required to specify response criteria (Figure 2.14). These dimensions can be used to evaluate the completeness of the specification of published response criteria. In our review of the response criteria, we demonstrated that both the HIV and Rheumatoid Arthritis criteria were underspecified in defining the timing of the baseline assessment with respect to the start of treatment (Table 2.1). The goal of response criteria is to provide a consistent approach for interpreting response so that novel treatments can be compared within and across cohorts of patients. When response criteria are underspecified, clinical trial response assessment protocols tend to have more variance. This has resulted in the inability to reasonably compare the response outcomes of two clinical trials in the same domain. The developers of response criteria could use the defined model for response criteria as a means of evaluating the completeness of their specification to avoid deviations that result from variations in interpretation.

8.3.2: Complete decision-logic for three domain response criteria

For each of the three domain response criteria, we apply the model of response criteria to encode a specification of the response parameters and chaining of classification and statistical tasks. This provides a significant advantage for application of response criteria over the prose-based approaches, with the goal of
reducing ambiguity in interpretation. Examples of the decision-logic are shown for the RECIST 1.1 oncology response criteria in Figures 2.6, 2.7, 2.8, and 2.9. These decision trees and decision tables demonstrate the complexity of response criteria and the need for explicit specification of their underlying logic. This effort to formally specify three domain response criteria represents a clinical contribution irrespective of the technical implementation.

This effort also exposes areas of the response criteria decision logic that are underspecified. In particular, Figure 8.1 shows the decision logic for classification of a follow-up lymph node lesion in RECIST 1.1. The rule for classifying a follow-up lymph nodes lesion with a lesion response status of “unequivocal progression” is underspecified in the prose document of the RECIST 1.1 response criteria. In contrast, a specific rule exists for the IHC lymphoma response criteria for defining the response state of unequivocal progression for a follow-up lymph node lesion. Similarly, the IHC response criteria are underspecified in classifying response for non-
lymph nodes lesions, and extrapolations from RECIST had to be made to fill this gap. More surprisingly, while the IHC criteria make explicit definitions for classification of the follow-up quantitative target lesions response, it is completely underspecified with respect to classification of the follow-up qualitative response status. The follow-up qualitative response status takes into account the target lesion and non-target lesion follow-up response status as well as the new lesion status. Logically, the approach taken by RECIST for classifying follow-up shown in Figure 2.9 could be applied to other oncology response criteria and was used where the IHC criteria were underspecified for this implementation.

8.3.3: Automated response assessment for consistent application of response criteria

In this dissertation, I have developed an automated method for treatment response assessment utilizing response criteria domain knowledge and clinical trial response assessment protocols. Automation of interpretation for complex reasoning problems such as response criteria could provide a mechanism for consistent application of response criteria both within and across clinical trials. Within the Rule-Based Response Assessment Framework, the RECIST implementation required 105 interpretation task specifications to be fully encoded. With so many discrete logical steps, it is not surprising that there is inter-observer discordance in application of response criteria to patient data. Furthermore, we have shown that the implementation is reusable for multiple clinical trial data sets and can thus be applied broadly.

8.3.4: Comparison of response criteria knowledge and results

We have also demonstrated with a thought experiment, the use of this framework for evaluating the knowledge level differences and impact on response outcomes between two criteria in the same domain applied to the same patient data set. We compared the knowledge overlap between the RECIST and IHC oncology response criteria and found that they shared almost all of their parameter concepts and approximately 77% of their task specification knowledge (Table 6.3). We also applied the RECIST criteria to the Lymphoma clinical trial data set and found that the two
criteria had approximately 82% agreement in their classification of cancer lesions and follow-up response status (Table 7.2). These types of experiments allow direct comparison of response criteria within domains to identify differences in interpretation of patient data. This is especially important as response criteria evolve over time such as from RECIST 1.0 to RECIST 1.1 to understand the impact of these changes on comparison of results using the new versus the old criteria. As will be discussed in future research directions in Chapter 9, we anticipate that this framework will allow us to continue to develop and experiment with new response criteria and compare them to current standards.

8.4: Clinical Limitations

8.4.1: Lack of decision support for input data set

Our implementation of the three domain response criteria has some clinical limitations as well. First, while our methods for automated interpretation of response assessment should improve the consistency of interpretation of the same response data set, it does not take into account variations in inter-observer measurement of response data itself. For image based-approaches in particular, the current paradigm of human selection and measurement of cancer lesions for inclusion in the interpretation method still leaves room for inter-observer variation that can impact interpretation results. This automated method for response assessment simply creates a consistent interpretation of a given set of input data, but cannot control what input data are included. Automated image processing methods are needed to assist users in consistent identification and measurement of cancer lesion image features.

8.4.2: Limited modeling of assessment modalities

Second, our implementation of the oncology response criteria was limited to interpretation of Computed Tomography (CT) image data sets. This was in part because the clinical trial response assessment protocol for the two clinical trials utilized in our evaluation only required CT as their response assessment modality. These data sets were chosen in part for this simplifying assumption. However, even within oncology response criteria, multiple assessment modalities can be used to
assess treatment response both quantitatively and qualitatively and are incorporated into response criteria. For instance, breast cancer and prostate cancer clinical trials often assess patient response with both CT and bone scans given the distribution of cancer lesions for those diagnoses. Similarly, the lymphoma response criteria also incorporate evaluation of disease with PET imaging for certain types of lymphoma. PET however is not typically used to assess follicular lymphoma and was not used for our particular clinical trial protocol. The current implementation of the International Harmonization Criteria within the Framework does not include the rules governing analysis of PET data, although it could be extended with a set of interpretation tasks to do so. The decision was made not to include rules for which we did not have data to test. If we had done so, the number of interpretation task specifications would have been much greater than 105 for the RECIST implementation. This represents an area of future modeling work for the oncology response criteria. The rheumatoid arthritis response criterion on the other hand, did include multiple assessment modalities to generate a quantitative estimate of the clinical state.

8.4.3: Retrospective versus prospective data acquisition and analysis

This implementation performs a retrospective assessment of treatment response by taking as input a complete set of response data. This has several clinical limitations for real life interpretation of response data. First, as described previously, retrospective assessment of treatment response can give different outcomes compared to prospective response assessment (8). The Rule-Based Response Assessment Framework assumes a complete data set when assessing response and thus performs retrospective data analysis. However our implementation of the oncology response criteria uses the baseline assessment period as the basis for selecting the measurable disease for quantitative response assessment. This implementation does not use knowledge of how the tumor lesions change over time in making its selection of baseline target lesions and as such performs response assessment in prospective manner. However, patient data collected retrospectively often gives attention to minimal disease in the baseline assessment period that has progressed in subsequent
follow-up periods. Data collected prospectively might not give attention to such minimal disease. While the Response Assessment Method will give a consistent answer for response with the same data set, it cannot control for changes made to the input data set when data is collected prospectively vs. retrospectively.

Along the same lines, our method for interpretation of response assessment does not provide real-time interpretation of patient data as it is being collected, nor does it provide updates to interpretations as new information is collected. The OWL and SWRL implementation of our method takes as input a complete data set and give interpretations from that data set alone. If given a partial data set, the method will provide an interpretation assuming the data set is complete. If the method is run again with a more complete data set, it will provide a new interpretation but does not provide a truth maintenance infrastructure. Enhancements to the method would need to be made to enable more realistic application in real clinical settings where data is being generated in an asynchronous manner and were interpretations are required prospectively, not just retrospectively.

8.4.4: Limited to protocol driven care

Finally, the Rule-Based Response Assessment Framework requires knowledge of the planned assessment periods for response assessment in the clinical trial protocol as well as knowledge of the treatment events. This type of knowledge is available at the time of data analysis for clinical trial protocols, but limits the ability to generalize this method to non-protocol based care situations. Typically, response criteria are only applied to clinical trial contexts and not to everyday patient care. This is in part due to the fact that it is time consuming and thus expensive to apply response criteria in the context of everyday clinical practice. In addition, many providers are not educated in complex procedures for applying response criteria to patients in their practice. However, it is desirable to be able to introduce quantitative response assessment methods to everyday patient care. This approach however is limited by ability to merge response data sets with treatment histories and by the variability in timing of assessment periods in non-protocol based setting. Without knowledge of the treatment history, a new version of response criteria would thus need to be developed
to account for response assessment in the non-protocol based environment. Specifically, use of the quantitative aspect of criteria could be used, but without knowledge of the treatment start and stop time, it would be difficult to classify the nature of the response as opposed to describing it quantitatively.

In summary, the Rule-Based Response Assessment Framework provides several clinical and informatics contributes including a formal model and task specification for response assessment, and an automated method for interpretation of clinical trial response data. The framework provides a goal-driven approach to temporal interpretation tasks while extending the scope to include both classification and statistical reasoning tasks. The framework including the knowledge model and reasoning methods is also fully implemented with a knowledge representation formalism that pushes the limits of set-based and rule template reasoning within OWL and SWRL. The next chapter discusses future research directions in the domain of treatment response assessment.
The Rule-Based Response Assessment Framework is just one piece of a larger system that is required to support treatment response assessment. While we demonstrate that the Framework can be used in clinical domains outside oncology, the remainder of this section will focus on the functional requirements for systems to support the use and development of treatment response assessment criteria for oncology. While these approaches can also be generalized to other domains, we focus on the specific needs of image based response criteria for purposes of this discussion. First, we describe the functional requirements for systems needed to apply response criteria in everyday clinical and research practice, then describe additional features needed to support research and development of oncology response criteria.

9.1: Proposed system for application of response criteria

A comprehensive system to support prospective application of treatment response criteria requires integrated tools for data acquisition, data interpretation, data visualization, decision support, and results reporting, as well as standards for data storage and transfer. As described in Chapter 2, the workflow associated with the application of treatment response criteria is complex and involves multiple providers and systems. For tumor response assessment using imaging studies as the assessment modality, this procedure involves numerous providers including radiologists, oncologists, and clinical trial data managers; and numerous clinical systems including specialized imaging systems, data management systems, and clinical systems that currently do not communicate with one another. Thus, a comprehensive system to support prospective application of treatment response criteria requires integrated tools for data acquisition, data analysis, data visualization, decision support, and results reporting, as well as standards for data storage and transfer (Figure 9.1). This section describes the functional requirements for such a system, some existing tools that address these requirements at least in part, and places the Rule-Based Response Assessment Framework within this greater context.
9.1.1: Data Acquisition Tools

Two of the primary limitations of the current workflow relate to lack of coordinated communication between radiologists and oncologists regarding the measurable disease to be tracked, and lack of unambiguous identifiers for tracking the same lesions over time. Two approaches to this problem are currently being pursued, a semantic annotation approach and an image processing approach.

We previously described the image Physician Annotation Device (iPad) (102), a plugin to the OsiriX (103) image viewing application that implements the Annotation and Image Markup (AIM) standard. The image annotation tool enables semantic annotation of image findings, however requires several extensions to support longitudinal lesion tracking described in (44). A critical step in tackling these challenges is recording and communicating the key "semantic" image information—the identity, location, and features of the measurable disease. Our hypothesis is that consistency and accuracy in the reporting of objective criteria such as RECIST can be
improved by adopting semantic image annotation tools for recording and communicating the key information necessary to guide radiologists in applying objective image-based criteria of neoplastic disease. Our objective is to develop informatics tools to support collecting and disseminating semantic image annotations in oncology to improve communication and coordination of oncologists and radiologists.

Towards this end, we have modified the image annotation tool to enable labeling of image annotations with unique identifiers. Ongoing work includes extension of the image annotation tool to enable annotation of image findings with concepts specific to oncology response criteria such as Target Lesion or Resolved Lesion. In addition, we are actively developing an image annotation flow sheet within the image annotation environment to facilitate communication between radiologists and ontologist as to the measurable disease to be tracked for response assessment. We have already tested our adaptations with the radiologist who annotated the cases for the Lymphoma study. Preliminary results show that image annotation alone is insufficient to support the lesion-tracking task. While users found the annotation task to be straightforward, they lacked tools to help track the lesions over time, navigate to the lesions previously annotated, and audit conditions of incomplete annotation. We hope that the addition of the flow sheet will enable more consistent image annotation.

In addition to seamless integration of these data acquisition tools into the radiologist’s workflow, knowledge drive decision support tools could further support accurate and consistent application of response criteria. While the Rule-Based Response Assessment Framework was designed for interpretation of response data, it is anticipated that aspects of the criteria knowledge representation and clinical trial protocol specification could be re-appropriated for use in knowledge driven decision support tools for response data acquisition.

An alternative approach to human annotation of image data is machine interpretation for lesion tracking. A complete discussion of image processing algorithms is beyond the scope of this section except to note that the goal of these methods is to improve consistency of lesion measurement and tracking (108), and
enable multi-dimensional lesion assessment (109). Ideally, the output of these image-processing methods could be annotated with AIM such that the Rule-Based Response Assessment Framework could be used for interpretation.

9.1.2: Data Analysis Tools

The implementation of the Rule-Based Response Assessment Framework is an example of a data interpretation tool that could be integrated into the larger response assessment system architecture for data analysis and decision support in application of response criteria at the point of care. These data interpretation tools would work at both the individual patient level and the cohort level. At the cohort level however, they would require additional statistical tools to analyze outcomes between cohort arms.

9.1.3: Data Visualization Tools

Data abstraction and data visualization are powerful methods for summarizing large amounts of data. We have piloted several visualization techniques for summarizing the results of the response interpretation method both at the individual patient level and cohort level. These visualization techniques were incorporated into a pilot system called LesionViewer that allowed the user to evaluate multiple dimensions of the data (110). We created anatomic visualizations of the temporal changes in lesion features (Figure 9.2) as well as graphical summaries of these changes (Figure 9.3).

A system usability study was performed consisting of direct observation of participants using LesionViewer to complete a multiple choice and fill-in-the blank questionnaire of 17 clinical tasks across 4 of the clinical cases used to build the prototype system. Comprehension tasks included determining the change in the number of lesions, SLD and disease state from one time point to another. Data review tasks were designed to test the user’s ability to navigate to the primary imaging data using the interface features. Of the 9 users, 4 had perfect scores and the remaining 5 users performed 16 out of 17 tasks correctly. Four users incorrectly calculated the difference between two dates and one user incorrectly assessed disease behavior due to lack of feature awareness. There was no difference in performance between the
Internists and Oncologists. Users took approximately 10 minutes to complete the 4 cases.

Figure 9.2: Anatomic Visualization of Temporal Changes in Lesion Features. Anatomical Summary View with three CT scans of the Chest, Abdomen and Pelvis. A lesion located in the second study is moused over, triggering the appearance of a pop-up box describing the lesion’s basic characteristics. The same lesion is later clicked. Since that lesion is present in all three studies, its axial slices from all the studies appear at the bottom of the screen.

After completing the clinical tasks associated with the clinical cases, users were given a structured questionnaire detailing their satisfaction with the system. Users rated the ease of performing tasks using LesionViewer on a 5-point scale (1=Very Easy, 2=Easy, 3=Somewhat Easy, 4=Somewhat Difficult, 5=Difficult). The nine users gave an average usability score of 2. All nine users stated that they would use the system in clinical practice if it were integrated into their current radiology
workstation and felt it would save them time. Work is ongoing to create an integrated lesion image annotation tool that can also facilitate visualization of trends in lesion features. The open source OsiriX DICOM viewer enables integration of such visualizations into the clinical workflow.

Figure 9.3: Graphical Summary of Temporal Changes in Cancer Lesion Features. Disease Summary View with eleven studies. The top graph charts the number of lesions over time. The bottom graph charts the composite length score over time. (See text for explanation of composite length score.) The timeline at the bottom displays the disease state (automatically generated using the RECIST guidelines) during each interval between studies.

9.1.4: Treatment Decision Support tools

Through the evaluation of our prototype LesionViewer system, we have shown that visualizations of response data and response interpretations can facilitate treatment decision-making. This is a passive form of treatment decision support. However, active treatment decision support could also be facilitated through integration of response outcomes with protocol-based decision support algorithms.

9.1.5: Tools for reporting response assessment outcomes

In addition to providing the input to decision support at the individual patient level, the output of the response assessment method can be used to report cohort level outcomes. This is important with respect to generating data for use in clinical trial
publications and for reporting to regulatory agencies. Specifically, the clinical trial outcomes may be modeled using the Ontology of Clinical Research (OCRe) such that outcome methods and results can be queried and compared across trials. Reporting of response results is important for clinical trial publications, data repositories and FDA regulatory reporting.

9.1.6: Standards for data storage & transfer

Essential to the interoperability of the response assessment system are standards for data storage and transfer. In chapter 5 we described the Annotation Image Markup (AIM) information model that has been developed by the caBIG imaging workgroup. In addition, a minimum amount of treatment related information is needed to support context restriction for the response assessment method. Imaging systems and treatment systems are typically separate in most clinical systems. However, image annotations provide an opportunity to dramatically reduce the data volume of image data and facilitate incorporation into clinical systems.

9.2: Proposed system for development of response criteria

Similarly, a system to support research and development of response criteria requires a large data repository of response biomarkers, treatment and outcomes data, with an open architecture to enable evaluation of new assessment modalities and measurement techniques, new data acquisition approaches, and new response criteria. Chapter 2 described some of the current limitations with respect to research and development of response criteria, the greatest of which is the lack of a publicly available database containing baseline and follow-up imaging studies, as well as diagnosis, treatment and survival outcomes information. Such a system would enable experimentation on new imaging modalities and measurement techniques, new data acquisition approaches, and new response criteria (Figure 9.4).

9.2.1: Data repository

A core requirement for supporting further development of response criteria is the development of a large data repository containing baseline and follow-up imaging studies, as well as diagnosis, treatment and clinical outcomes data. In order to validate
new response criteria, such a database should contain data from patients at multiple institutions with multiple diagnoses if relevant. One model would be a centralized database like the ACRIN project (111); another could be a federated grid model such as proposed by the caBIG initiative (112). Ideally, portions of the data would be publicly available for development of new data acquisition techniques and response criteria by the larger research community. Furthermore, data could be prospectively collected as part of clinical trial data management or everyday patient care as described above. Similarly, as above, standard information models for data storage and transfer would be required.

9.2.2: Data acquisition tools

Ongoing research on data acquisition methods would benefit greatly from the ability to access a well-structured gold standard dataset. This would facilitate comparison of methodologies with respect to reproducibility, accuracy and correlation with clinical outcomes. Multiple data sets would be required covering multiple
imaging modalities and types of malignancies. Data acquisition tools are needed to automate the process of 1) lesion identification, 2) lesion tracking, 3) semantic annotation of location and 4) extraction of quantitative and qualitative lesion imaging features such as lesion dimensions, metabolic activity and density. Ideally, automated methods would replace much of the somewhat tedious work done by human reviewers, and could be validated with multiple training and testing data sets.

9.2.3: Knowledge acquisition tools

Knowledge acquisition tools are also needed to dynamically modify the parameters of response criteria for purposes of experimentation. Such tools could be used in conjunction with the Rule Based Response Assessment Framework to modify the values of existing parameters and create new parameters for inclusion into experimental response criteria. This would also enable comparison of the knowledge level differences between proposed criteria, as well as enable application of new methods to data sets for to evaluate the differences in outcomes. Furthermore, the knowledge representation systems could be used to publish new response criteria in machine accessible formats.

9.2.4: Tools to enable experiments on response criteria

In the application of response criteria, typically one specific criterion is applied to analyze the clinical trial cohort data. Development of response criteria requires the ability to apply multiple criteria to the same data set, and store and compare the results using various statistical tools. In addition, large feature sets could also enable the use of machine learning techniques to discover clinically relevant imaging features that correlate with survival endpoints.

The Rule-Based Response Assessment Framework provides a key link in the complex workflow for application and development of oncology response criteria. Further integration of the framework within the larger infrastructure of data acquisition, results reporting and visualization, and decision support will further the applicability of the Rule-Based Response Assessment Framework beyond the scope of this dissertation work.
9.3: Conclusions

This dissertation describes the Rule-Based Response Assessment Framework, a goal-driven, context-driven, and knowledge-based temporal interpretation framework that performs procedural response assessment tasks. The Rule-Based Response Assessment Framework defines the dimensions of response assessment including the knowledge and tasks required to apply response criteria to patient data. The Framework consists of several knowledge models and several sub-methods that make up the Response Assessment Method. The methods take as input knowledge and data encoded in the models and generate as output new knowledge that is asserted back into the knowledge models. The Response Assessment Method consists of the Context Restriction Method and the Context-Based Interpretation Methods.

The Rule-Based Response Assessment Framework is implemented using ontologies and rules with the semantic web formalisms of OWL and SWRL respectively. The framework is applied to two response criteria in the oncology domain and one response criterion for rheumatoid arthritis. The implementation of the framework in OWL and SWRL was sufficiently expressive to generate clinically accurate response interpretations compared to human interpretation. In addition, we demonstrate that approximately 87% of the criteria knowledge is shared between the two oncology criteria, and that the framework enables a comparison of the outputs of the two criteria applied to the same patient data set. We have demonstrated that the implementation of the Rule-Based Response Assessment Framework is extensible, reusable and enables consistent application of response criteria to patient data and comparison of response criteria within a clinical domain.

This dissertation makes several clinical and informatics contributions. Clinical contributions include the development of a formal model of the dimensions of response criteria and complete decision-logics for three domain response criteria. In addition, automated application of response criteria could enable more consistent application within and across clinical trials. Finally, the framework enables
comparison of response criteria knowledge and results, essential for the ongoing development of response criteria.

Informatics contributions include a formal definition of the response assessment task as a type of temporal interpretation method. In addition, we have developed the Response Assessment Method as a goal-driven, context-driven, temporal interpretation method that extends the range of interpretation tasks for temporal interpretation methods to include classification and statistical interpretation tasks across single and multiple-contexts, for single and multiple parameters and subjects. We have also introduced the rule template approach to domain knowledge acquisition and implemented the framework using the knowledge representation formalism of OWL and SWRL that can be inspected, extended, and transported. Future research directions will incorporate the Rule-Based Response Assessment Framework into a larger informatics infrastructure for application and development of oncology response criteria.
REFERENCES


7. Thornton. The next front in the war on cancer. The wall street journal 27 February 2009


9. United States Food and Drug Administration Oncology Drugs Advisory Committee Briefing Document for NDA 21-649 (Oblimersen Sodium).


44. Levy MA, Rubin DL. Tool support to enable evaluation of the clinical response to treatment. AMIA Annu Symp Proc 2008:399-403.


55. A symbolic processing approach to measurement interpretation in the intensive care unit; Proceedings of the annual symposium on computer application in medical care. 1979.


64. Representation and use of temporal information in ONCOCIN; Proceedings of the annual symposium on computer application in medical care. 1985.


83. Miksch S, Seyfang A, Horn W, Popow C. Abstracting steady qualitative


111. www.Acrin.Org;