CHEMOTHERAPY PLAN ABSTRACTION METHOD

By

Haresh Lokumal Bhatia

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Biomedical Informatics

August, 2012

Nashville, Tennessee

Approved:

Professor Mia Levy

Professor Joshua C. Denny

Professor Thomas A. Lasko

ACKNOWLEDGEMENTS

I would like to thank my thesis committee for their guidance throughout this work. I thank my advisor Mia Levy for all her help, cooperation and patience while working with me on this project and the thesis document. I thank Cindy Gadd for their help in advancing my education in Biomedical Informatics and guiding me through the formalities of graduate school.

I am grateful for those who helped at various junctures of this project. I would like to thank Ioana Danciu for arranging the initial data set of chemotherapy medications for lung and breast cancer patients from Pharmacy. I would also like to thank Matt Marshall who provided valuable information concerning the charges recorded in pharmacy system, and how to match and account for the negative charges. The information provided by him was crucial to working with the pharmacy information system data. I extend my thanks to Melissa Basford, Alexander Saip and James Cowan who were instrumental in porting the pharmacy data from EDW into the SD and de-identifying it, so I could use it for my project.

I would like to express my gratitude to the Department of Biomedical informatics. I appreciate the support of fellow students, faculty and staff. My special thanks to the department secretary Rischelle Jenkins for all her support in arranging the meetings and rooms whenever necessary. I am grateful to the National Library of Medicine for their support (5 T15 LM 7450-10).

Finally, I am very grateful for the support of my family, my lovely wife Parneeta and adorable daughter Karishma for their unconditional love, patience and encouragement.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

This thesis describes the motivation for, development of, and evaluation of the chemotherapy plan abstraction method (CPAM), and its use for cohort plan analysis. This introduction defines the concepts related to the generic and clinical tasks of planning and plan abstraction, and their importance in clinical practice and research.

1.1. Plan abstraction

Plan abstraction is an important generic and clinical reasoning task. Table 1.1 defines the concepts related to planning and plan abstraction. An *intended plan* is a sequence of future actions specified by an actor to achieve a goal. *Planning* is the task of specifying a plan. When a plan is executed, the sequence of pre-specified actions is carried out. Some of those actions will have associated artifacts (e.g., execution timestamp and serial ordinal number) that can be reviewed by an outside observer. The *executed plan*, therefore, can include modifications due to execution time changes in the state of the plan components. *Plan abstraction* is the task of inferring the existence of a plan from the pattern of observed events(1; 2). *Plan recognition*, on the other hand, is the task of establishing a correspondence between a sequence of observed events and a known set of established plans(1). *Cohort plan analysis* is the task of conducting a collective analysis of plans executed by a cohort of actors.

Table 1.1. Definition of concepts related to plan abstraction.
Definitions include the generic (domain non-specific) context, general clinical context, and chemotherapy context.

Term / Concept	Generic Task	Clinical Task	Chemotherapy Task
Intended plan	A sequence of actions specified by an actor to achieve a goal.	A sequence of diagnostic or therapeutic events specified by a clinician with the goal of improving a patient's duration or quality of life.	A sequence of chemotherapy medication events specified by an oncologist with the goal of treating a specific cancer and improving the patient's duration or quality of life.
Executed plan	A completed plan with observable artifacts in terms of a sequence of events corresponding to the plan implementation.	A completed diagnostic or therapeutic plan with a sequence of recorded clinical events corresponding to the plan.	A completed chemotherapy plan with a sequence of recorded chemotherapy medication events corresponding to the plan.
Planning	Task of specifying a plan.	Clinical task of specifying a clinical plan.	Clinical task of specifying a chemotherapy plan.
Plan abstraction	Task of inferring executed plans from an observed sequence of past events.	Clinical task of inferring a clinical plan from a sequence of past diagnostic or therapeutic events.	Clinical task of inferring a chemotherapy plan from an observed sequence of the past chemotherapy medication events.
Plan recognition	Task of establishing a correspondence between an observed sequence of past events and a known set of established plans.	Clinical task of matching an observed set of clinical events to a known set of clinical plans.	Clinical task of matching an observed set of chemotherapy medication events to a known set of chemotherapy plans.
Plan selection	Task of selecting from among a set of plans that can all achieve the same goal.	Clinical task of selecting from among an established set of diagnostic or treatment plans for a given disease.	Task of selecting from among an established set of chemotherapy protocols to treat a given type and stage of cancer.
Cohort plan analysis	An analysis of the executed plans of a cohort of actors	A clinical analysis of the executed diagnostic and therapeutic plans of a cohort of patients.	A clinical analysis of the executed chemotherapy plans of a cohort of cancer patients.

1.2. Plan abstraction in medicine

An *intended clinical plan* is a sequence of future diagnostic or therapeutic events specified by a provider with the goal of improving the patient's duration or quality of life. *Clinical planning* is the task of specifying a clinical plan. Many clinical plans are highly specific for a particular disease, and are established as a standard of care through clinical research. A clinician prescribes an appropriate clinical plan for a given disease condition through a process of *plan selection*, while balancing between likely relative efficacy, side-effects and cost. An *executed clinical plan* consists of the completed sequence of clinical events corresponding to the intended plan. The actual events, though, can deviate from the intended course of the plan (e.g., due to inability of the patient to tolerate the side-effects). The task of *clinical plan abstraction* involves inferring the clinical plan from actual clinical events. *Clinical plan recognition* is the task of establishing a correspondence between the observed sequence of clinical events and the set of known clinical plans established as standard of care.

Clinical care follows a cyclical process of *diagnosis*, treatment plan selection, treatment *plan management*, and *response assessment* tasks (Figure 1.1). *Diagnosis* is the task of ascertaining the disease condition while *response assessment* is the task of evaluating how the disease condition changes in the context of treatment. Treatment *plan management* is the task of scheduling, customizing, and iteratively refining the intended plan events based on the outcome of the response assessment task. A *treatment summary* is a concise statement summarizing the clinical events associated with an executed treatment plan. *Plan abstraction* is the essential task of deriving the treatment summary

from observed clinical events. The treatment summary provides important feedback for both the treatment selection and the treatment management tasks (figure 1.1).

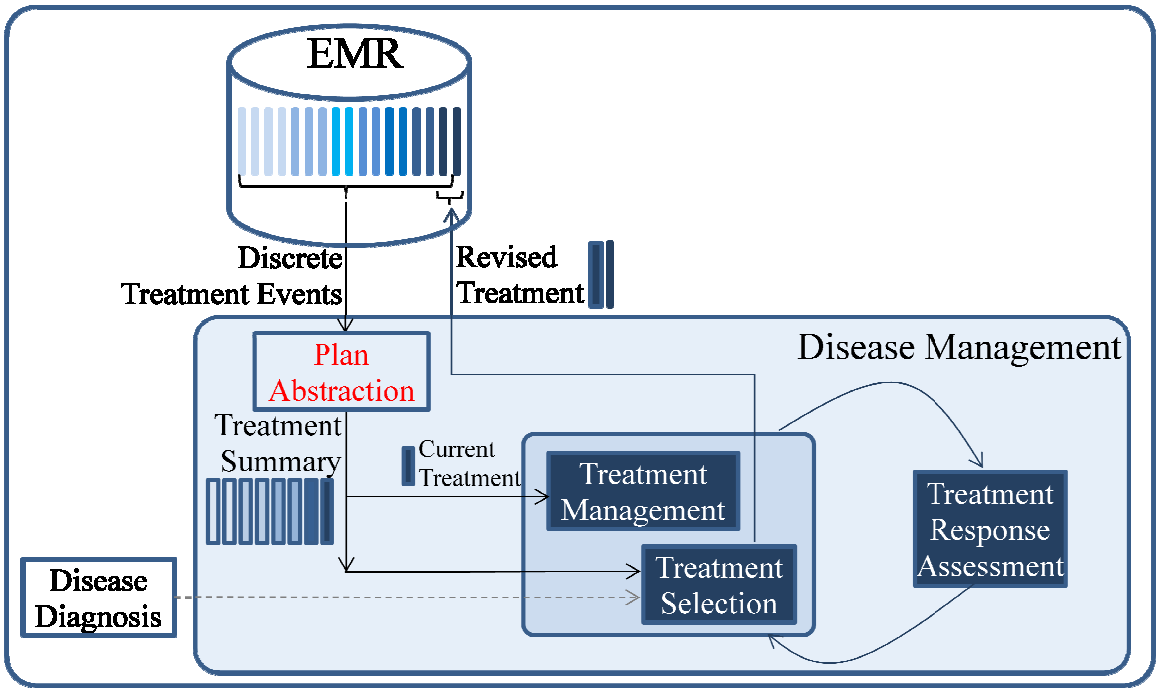


Figure 1.1. Disease management and plan abstraction tasks.
Disease management is a cyclical process, consisting of diagnosis, treatment selection, treatment plan management and treatment response assessment tasks. The treatment history is reviewed in terms of treatment plans that correspond to the respective treatment protocols, and plan abstraction task produces treatment plans from the treatment events stored in EMR. The treatment history is reviewed during the treatment selection task and treatment plan management task. The current treatment is reviewed and managed with respect to the disease response and is either continued or revised.

Clinical cohort plan analysis is the task of evaluating a set of executed clinical plans for a cohort of patients. Such cohort level analysis can address questions related to patient adherence to standard of care plans, physician practice patterns, the comparative effectiveness of multiple treatment plans, and the comparative cost of multiple plans.

1.3. Chemotherapy plan abstraction

The chemotherapy version of the concepts listed in table 1.1 can be defined in a similar fashion as those for the clinical tasks, except that in case of chemotherapy all events represent the chemotherapy drug events.

Chemotherapy plans are a class of protocol-based treatments consisting of specialized drugs to kill cancer cells while minimizing toxicity to the patient. Individual patients show varying degrees of tolerance to chemotherapy drug treatments. As such, it is common for the executed chemotherapy plan to have minor or major deviations from the intended plan to account for patient toxicity. Chemotherapy protocols are highly complex treatment plans specific to a particular cancer type, and are established as standard of care through rigorous clinical research. Some chemotherapy plans span many months or even years, and can consist of hundreds of distinct drug events. A given plan can consist of multiple drugs and a given drug may be part of multiple plans.

Figure 1.2 shows several examples of simple and complex intended (panel A) and executed (panel B) chemotherapy plans and their respective treatment summaries (panel C). Intended chemotherapy plans (Figure 1.2-A) are specified as a set of medication events that repeat at a given frequency and number of cycles. For example, plan 1 is a simple chemotherapy plan consisting of a single drug paclitaxel (‘P’) that is repeated every twenty-one days for four cycles. Plan 3 on the other hand is a more complex plan that consists of two drugs repeated in multiple nested cycles. In plan 3, cisplatin (‘C’) is given on day 1 only while etoposide (‘E’) is given daily for three days starting on day1. The whole set is repeated every twenty-one days for six cycles.

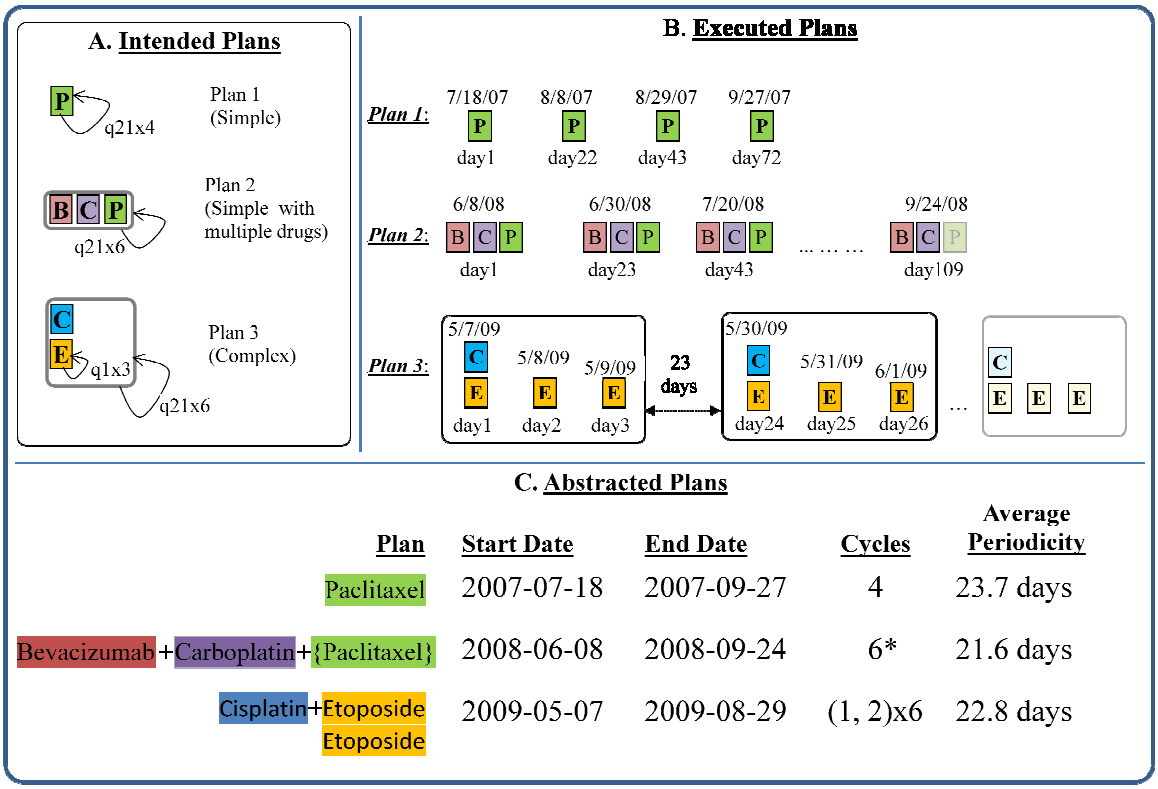


Figure 1.2. Plan representations.
Panel A shows intended plans. The looping arrow indicates 'repetition' of the drug marked in the box, *q* gives the frequency of repetition, and x gives the number of repeats (e.g., q21x4 indicates repeat every 21 days 4 times). Plan 1 is a simple plan consisting of only one drug (paclitaxel – 'P') repeating every three weeks. Plan 2 is another simple plan that repeats every three weeks, but it consists of multiple drugs (bevacizumab – 'B', carboplatin – 'C' and paclitaxel – 'P'). Plan 3 is complex and consists of drugs cisplatin – 'C' given on day 1 and etoposide – 'E' repeated 3 times daily starting day 1; the whole set is then repeated ; that are given together on the first day, Etoposide is then repeated six times every twenty-one days. **Panel B** shows executed plans corresponding to the intended plans shown in panel A. Actual dates and day#s are stated for individual medication events. **Panel C** gives a representation of abstracted plans as inferred from the events in panel B. The list of abstracted plans shows the start-date, end-date, number of cycles and average periodicity discerned from the observed events.

Panel B of figure 1.2 shows the respective examples of the actual sequence of events that can result from execution of these plans. Panel C shows the corresponding abstracted versions of these plans, which forms a *treatment summary* for a given patient. In addition to inferring the constituent drugs, the abstracted version also shows the start date, end date, number of cycles, and average periodicity (rate of repetition) for the respective

plans. The actual rate of repeat (averaged over all the cycles) for each of these plans is longer than that suggested by the respective plans, which demonstrates that the intended plan is not always executed as designed in order to account for patient response to the treatment.

Cancer care follows the same cyclical process of diagnosis, chemotherapy plan selection, chemotherapy plan management and response assessment (in terms of tumor size and patient tolerance) tasks, as shown in figure 1.1. During the plan selection task, the oncologist uses the treatment summary as the record of therapies the patient has previously completed to inform selection of the next treatment. During the chemotherapy plan management task, the oncologist uses the summary of the current executed plan to recall any deviations from the intended plan due to toxicity.

Organizations like ASCO recommend using cancer treatment summaries(3) to record the details of an executed cancer treatment. ASCO's chemotherapy treatment plan and summary templates were developed to help improve documentation and coordination of cancer treatment and survivorship care. They are intended to facilitate provider-to-provider and provider-to-patient communication. The completed treatment summaries are recorded in the patient chart and can be distributed to the patient and to their providers. Importantly, the treatment plan and summary are not intended to replace detailed chart documentation, including complete patient histories or chemotherapy flow sheets.

However, manual generation of a chemotherapy treatment summary can be a very time consuming task for clinicians in practice. In addition to cognitive stress of discerning patterns pertaining to complex plans there are challenges associated with the extraction of medication events from the medical records and technical intricacies

accompanying clinical data sources. Chapter 2 further elaborates on the challenges of extracting medication events and intricacies of accessing clinical data sources.

Table 1.2. Questions that can be addressed for cohort analysis.
Operational, quality, and research questions addressed when plan history information is available for cohort analysis. Questions may be across or within plans, across or within disease and with disease feature restrictions. Listed are several questions (Q) across these dimensions and their clinical utility (U).

Available Data Sources	Cohort Analysis Across Plans	Cohort Analysis Within a Plan
Treatment History alone	Q: What are the most frequently used plans across all cancer diagnosis? U: Resource utilization U: Cost analysis U: Prioritize CPOE order set implementation U: Pharmacy supply management	Q: What is the variance in sequencing, total number of cycles, and cycle frequency for a given plan across cancer diagnosis? U: Estimate of variance in plan utilization across cancer diagnoses
Treatment History plus Cancer Diagnosis	Q: What are the most frequent plans for a given cancer diagnosis? Q: Which plans are most often used first, second, or third in treatment sequencing for a given cancer diagnosis? U: Analysis of variance in provider practice patterns within and across institutions	Q: What is the variance in sequencing, total number of cycles, and cycle frequency for a given plan for a single cancer diagnosis? U: Estimate of patient toxicity to plan U: Estimate of average disease progression on plan
Treatment History plus Cancer Diagnosis, Cancer Stage, Tumor Biomarkers, and Patient Survival	Q: What are the most frequent plans for a given cancer diagnosis, stage, and set of biomarkers? U: Analysis of variance in provider practice patterns within and across institutions U: Analysis of provider compliance with standard guidelines Q: What is the comparative efficacy of plans for a given cancer diagnosis, cancer stage and set of tumor biomarkers? U: Comparative effectiveness research on large populations	Q: What is the variance in sequencing, total number of cycles, and cycle frequency for a given plan for a single cancer diagnosis, stage and biomarker? U: Estimate of patient toxicity to plan U: Estimate of average disease progression on plan Q: What is the comparative efficacy of a given plans for a given cancer diagnosis, cancer stage and set of tumor biomarkers? U: Predictive and prognostic biomarker discovery for a given plan.

When detailed chemotherapy plan histories are available for a large cohort of patients, many questions can be addressed related to quality, efficacy, and cost of care (Table 1.2). For example, when chemotherapy plan history alone is available, a cohort

level analysis across all plans can reveal the frequency of treatment plans at a given institution. A computerized provider order entry (CPOE) implementation team could use this information to help prioritize creating order set templates at their facility. Several more questions can be addressed when the cancer diagnosis information is added. For example, what are some of the most frequently administered chemotherapy protocols for a given cancer diagnosis? This can give information on provider practice patterns within and across institutions, including insight into provider awareness of existing knowledge and resource utilization(4). Likewise, a within plan analysis of a given disease can include an evaluation of the variance in sequencing, total number of cycles, and cycle frequency for a given plan. In the case of a plan for a metastatic cancer for instance, the median duration of treatment could correlate with the median time to disease progression in that patient population.

Such cohort level analyses, however, are currently very time-consuming to perform since researchers must manually recreate the treatment history from clinical documents stored in the EMR, as exemplified by the study conducted by Zafar et al.(5)

1.4. Chemotherapy plan abstraction as a temporal abstraction task

Chemotherapy plans involve one or more drugs repeated over a number of cycles, for a specified periodicity. Considering the discrete drug events as instances along the temporal dimension it is possible to apply temporal reasoning methods to solve the problem of chemotherapy plan abstraction. Identifying individual instances by corresponding drug-name and time-stamp can help establish temporal patterns among the sequence of events. Application of temporal logic enables extraction of such attributes as the periodicity and cycle length of plans inferred from the temporal patterns. Whereas

plan recognition needs to use an external knowledge base of standard plans as a reference to *recognize* the plan from the executed sequence of events, a plan abstraction method could be created that relies solely on the content of data to infer plans and derive the related attributes.

I therefore hypothesize that it is possible to create a chemotherapy plan abstraction method that takes as input distinct chemotherapy drug events and accurately generate as output a temporal sequence of chemotherapy treatment summaries in terms of abstracted plans.

1.5. Overview of the thesis document

The following six chapters describe the development of the Chemotherapy Plan Abstraction Method (CPAM) and its application to cohort plan analysis. Chapter 2 describes the challenges of performing the chemotherapy plan abstraction and cohort plan analysis tasks in the clinical and research settings. Chapter 3 reviews the temporal reasoning literature related to the computational task of plan abstraction. This chapter includes a discussion of the dimensions of the plan recognition and plan abstraction tasks, prior work related to these dimensions, and their limitations.

Chapter 4 describes the CPAM, a data-driven temporal reasoning method that takes as input chemotherapy medication events and generates as output a sequence of abstract chemotherapy plans for patients in multiple cancer domains. This chapter describes the details of the data extraction method, pre-processing method, and plan abstraction method. Chapter 5 describes the evaluation methodology and results, including an evaluation of the performance of the CPAM at a patient level, and an evaluation of the clinical utility of the abstracted plans for cohort plan analysis.

Chapter 6 discusses the contributions and limitations of the CPAM in the domains of informatics and medicine. Chapter 7 discusses possible future work including iterative improvements to the CPAM and its potential applications for patient care and clinical research.

CHAPTER 2

CHALLENGES OF CLINICAL PLAN ABSTRACTION

As recommended by ASCO(3), oncologists will often create a treatment summary for each course of treatment as an unstructured text document. These treatment summaries have several limitations. First, each individual summary provides an overview of a specific set of treatment events, but lacks an overall view of the patient's entire treatment history. Second, they lack structure to enable use by downstream systems for clinical decision support or cohort analysis. Third, they are time consuming to produce since the only way to obtain the executed plan summary is by abstracting the plans from the chronology of past medication events. Finally, there are limitations with respect to the accuracy and truthfulness of past medication events that can be derived from the EMR. This chapter discusses the challenges associated with accessing the medication events from the EMR system for the purpose of chemotherapy plan abstraction.

2.1. Challenges related to extraction of medication events from the EMR

An executed clinical plan consists of a sequence of recorded clinical events that correspond to an intended clinical plan (table 1.1, figure 1.2). The EMR records these clinical events in many different and complementary ways. Some event records correspond more closely to the intended event (e.g., clinical order record) than to the executed event (e.g., nursing medication administration record). Some events are recorded in highly structured ways while others are recorded in free text, resulting in a variable accuracy in extracting these events. The following subsections describe the challenges related to medication event data extraction from the EMR systems.

2.1.1. Clinical information artifacts for medication events

Treatment information is recorded in several types of clinical records and in various formats that have varying degrees of accessibility to automated systems. Providers often use multiple clinical data sources to extract the treatment summary. Each data source has advantages and disadvantages for manual or automated plan abstraction.

2.1.1.1. Truthfulness of event and data accuracy.

Plan abstraction requires accurate medication event data. Two levels of accuracy are apparent, event truthfulness and data accuracy. Event truthfulness refers to the confidence associated with the event having actually occurred. *Data accuracy* refers to the completeness and faithfulness with which event data is reproduced from a given data source. For a given EMR implementation, and the corresponding clinical data sources, there is an implicit decision process of optimization between event truthfulness and data accuracy.

2.1.1.2. Clinical data sources

The data sources considered for the purpose of chemotherapy plan abstraction method are clinical notes, provider orders, provider order sets, pharmacy dispensing records, and nursing medication administration records. Table 2.1 presents a categorical summary of each clinical data source across the following dimensions: temporal context, degree of structure, storage format, data completeness, and medication information. These dimensions are analyzed with respect to truthfulness and data accuracy.

The *temporal Context* refers to the temporal context of medication events recorded in each type of document including reference to the past, current or intended medication

events or plans. The information pertaining to the future events only conveys an intention, limiting the truthfulness of the event having actually occurred.

The *degree of structure* refers to whether the data elements of medication events are in structured or free-text format. The degree of structure affects the data accuracy of the events.

The *storage format* refers to the medium of data storage. Data stored on paper can be accessed manually, but only by a single user at a time. Data stored in digital media can be accessed by multiple simultaneous users as well as by automated or programmable systems. Text files stored as image files are more difficult for automated systems to process than text stored in an ASCII format. The storage format affects the data accuracy of the events.

Data completeness refers to the extent of availability of data pertaining to different types of medication events, e.g., events pertaining to the medications taken at home, or taken outside a given institution. This data informs the degree of truthfulness of events.

Medication information refers to the data elements of the individual medication events, e.g., the drug name, drug code, dosage, route, and schedule (time-stamp). This dimension along with degree of structure influences the data accuracy of the events.

Table 2.1. Categorical summary of clinical data sources.
Summary of clinical data sources by various dimensions affecting the *truthfulness and data accuracy*. [V ≡ Implementation level at Vanderbilt University Medical Center; * ≡ Naming / Coding convention may be local to the institution.]

Dimension	Attribute	Clinic Notes	Provider Orders	Provider Order Sets	Pharmacy Dispensing Records	Nursing Medication Administration Record
Temporal Context	Current Medication Events	+	+	+	+	+
	Past Medication Events	+	-	-	-	-
	Future Medication events	+	-	-	-	-
	Current plan	+	-	+	-	-
	Past plan	+	-	-	-	-
	Future plan	+	-	-	-	-
Degree of structure	Free text	+	+	+	-	+
	Semi-structured	+/-	+ (V 95%)	+ (V 95%)	-	+ (V 95%)
	Highly structured	-	+ (V 5%)	+ (V 5%)	+ (V 100%)	+ (V 5%)
Storage formats	Paper	+	+	-	-	+
	Scanned Image Files	+ (V)	+ (V 95%)	+ (V 95%)	-	+
	Digital	+ (V)	+ (V 5%)	+ (V 5%)	+ (V)	+ (V 100%)
Data Completeness	Administered Medications	+	+ (V 5%)	+ (V 5%)	+ (V 100%)	+ (V 100%)
	Medications taken at home	+	+ (V 45%)	-	-	N/A
	Within Institution record	+	+	+	+	+
	Outside Institution record	+	-	-	-	-
Medication Information	Plan Name	+*	-	+*	-	+/-
	Drug Name	+*	+*	+*	+	+
	Drug Code	-	+*	+*	+	+*
	Dosage (e.g., mg/m^2)	-	+	+	+	+
	Dose Amount (e.g., mg)	+	+	+	+	+
	Route (e.g., iv)	+	+	+	+	+
	Schedule	+	+	+	+	+

The following paragraphs describe the data sources in greater detail with respect to these dimensions. Table 2.2 lists the relative advantages and disadvantages of the data sources.

Clinical notes: For every patient encounter, the provider creates a clinical note to record the details about the patient history, physical exam, test results, treatment plan and response assessment. The information can include a summary of the past, current, and intended medication events and treatment plans. When cancer treatments consist of multiple cycles that span many months, the providers often record information regarding the current treatment plan in their clinic notes and typically refer to the most recent clinical note to remind them of the state of the current plan. Even with these reminders, the providers often lose track of the current cycle number and must use other sources to reconstruct the most recent history.

Clinical notes are typically recorded in free text format. In the absence of an EMR implementation, the notes are handwritten on paper, and any subsequent information extraction is only feasible through manual review. EMR implementation allows direct entry of notes into the system that can then be accessed by programmable processes. The notes prior to the EMR implementation can be scanned and brought into the EMR system as image files with limited ability for data extraction. The information contained in notes, however, is largely unstructured. To obtain any meaningful information, moderately sophisticated natural language processing (NLP) methods are required. The accuracy of information extracted from such methods however is limited and highly variable(6), (7), (8).

Table 2.2. Advantages and disadvantages of various clinical data sources. [BCMA= Bar code medication administration.]

Data Source	Advantages	Disadvantages
Clinical Notes	• Text documents ubiquitously available. • Can provide data about other events temporally correlated to medication events	• Requires NLP to extract medication events and plans with variable accuracy, and temporal ambiguity(6)(7)(8) • Mention of medication indicates intention that may not result in corresponding administration event.
Provider Orders	• Computerized provider order entry (CPOE) systems contain well structured medication event data • CPOE systems provide electronically readable data that can be used by automated and programmable systems • Can provide data about other treatment events temporally correlated to medication events	• Medications administered may differ from the order in dose or schedule, or may not actually be administered • The level of uncertainty of corresponding administration event is considerably smaller than that of clinical notes
Provider Order Sets	• Contain an order set name that may correspond to an abstract plan name • When unchanged, the order set name is consistently used across providers within an institution • Group medication events that are part of a plan	• Order set name is institution specific and not standardized • Order sets may be used as a starting point template for a completely different plan • The level of uncertainty of corresponding to the plan is smaller than that of clinical notes
Pharmacy Dispensing Records	• Most hospitals have electronic pharmacy systems • Contain well structured medication event data • Provide electronically readable data that can be used by automated and programmable systems	• Medication administered may differ from the pharmacy order in dose or schedule or may not actually be administered. • The level of uncertainty of corresponding administration event is smaller than that of provider orders or order-sets
Nurse Medication Administration Record	• Most accurate document of what patient actually received (drug, dose, schedule, including time-stamp) • BCMA contains structured data about the administered drugs including time-stamp of administration, coded drug name, dose, units, route, schedule and frequency	• Free text requires NLP to extract medication events • BCMA systems have limited use in most outpatient chemotherapy infusion centers

Provider orders: Provider orders contain information about the provider's intended plan of treatment and include orders for medications, procedures and lab tests. In many

systems orders are documented on paper, which are stored as such or scanned into a digital storage as image files. In either case, to obtain medication information, the accessibility to these documents is limited to manual processing.

CPOE systems allow providers to enter structured orders that are recorded electronically. If CPOE is implemented the order information can easily be retrieved using standard tools. Unlike clinical notes no NLP is required and the data available is highly accurate, as to the ordering event of the medication. The medication event data obtained from provider orders is only the intention to administer the medication treatment. Some proportion of orders may be cancelled before being fulfilled by the pharmacy. Cancellations occur for many reasons including, lack of available drug, or a change in the patient state or treatment plan. These cancellations can affect the accuracy of medication event data obtained from provider orders and special processing is thus required. Clinical orders, however, have a higher degree of certainty with respect to current medication event than clinical notes, but not as certain as medication administration event records. Finally, provider orders are limited to a single institution and do not contain information about the orders outside a given medical facility.

Provider order-sets: Many implementations of CPOE allow for the creation of order sets, a collection of orders that are defined by a specific treatment protocol. Order sets facilitate ordering a complete set of drugs in the protocol all at once, rather than creating multiple individual orders from memory. Medications ordered using order sets retain all of the properties of regular orders described above. However, for the purposes of the plan abstraction task, order sets make two main contributions: 1) grouping medications together that belong to a plan creates a partial knowledge base of plans, and 2) order sets

are often assigned an institution specific identifiers including unique names and numbers, and their utilization can thus be tracked. Many institutions with CPOE order sets for chemotherapy solve the chemotherapy plan abstraction problem by looking at the order set name (or number). This approach has some advantages and disadvantages.

When order sets are used, the order set name provides a consistent naming convention within the institution for that plan. For example, the breast cancer plan "Dose Dense Adriamycin and Cytoxan" is a common adjuvant breast cancer plan that is also called "Dose Dense AC" or "dd AC" when written in short hand in clinical notes. As such, providers will have variable representations of plan names in clinical notes, but order sets provide a consistent representation within the institution. The order set names and identifiers, however, do not follow any national standard such that they can be compared across institutions. Yet, order sets are a convenient way to acknowledge plans that have a high likelihood of being the same *intended* plan across providers in a single institution, and are a more reliable representation of plan names than clinical notes.

While this is useful for commonly prescribed plans, most institutions do not have the resources to create order sets for every variation of every possible chemotherapy plan. One institution's recent implementation of a chemotherapy CPOE system for impatient and outpatient treatment required creation of over one thousand chemotherapy sets(9). Some plans are only used once every few years for rare diseases. Other plans are simply slight modifications of a more commonly used plan, and as such providers simply use the more common plan as a starting point template and modify it to transform to an alternate plan. Modifications can include changes in medication dose, deletion or addition of a medication or a modification to the plan schedule. It is, therefore, difficult to rely on the

order set name as the definitive evidence that the plan elements were administered to the patient.

Furthermore, order sets do not capture the temporal frequency of plans. While each cycle may be represented by a separate order set, the provider has to remember how frequently each cycle should be given. This is not represented in an order set view of chemotherapy plans. Finally, while inpatient CPOE order sets are relatively common, outpatient chemotherapy order management systems are less common. Availability of this type of data is lacking in many institutions.

Pharmacy dispensing records: The hospital pharmacy information system (PIS) stores information for every medication dispensed that is intended for administration on the medical facility premises. This includes inpatient medications and outpatient medications to be administered in infusion centers. The PIS documents the medication events that are the closest representation of the corresponding administration event, short of administration itself. Thus the event data obtained from pharmacy dispensing records is a good representation of the executed medication events. Furthermore, most hospital facilities have implemented PIS, however they only store information pertaining to the dispense events at a single institution.

Medication events in PIS are highly structured and can be readily retrieved. The pharmacy dispensing records store information related to the drug name, drug-code, dispense time-stamp, frequency, quantity, and billing charge. Many systems use National Drug Codes (NDC) as well as other drug coding systems to represent the names of drugs.

Like CPOE systems, PIS are transaction systems that facilitate pharmacy workflow and billing. Orders are cancelled through negation of charges. Cancellation of a

chemotherapy medication is recorded as a negation of the exact amount of the billing charge, while those for other medications can be a partial charge negation. Conventions for such cancellations vary with implementations and business rules at different facilities. The accuracy of reconstructing medication events obtained from pharmacy dispensing records is very high as long as the cancellations are addressed appropriately.

Nurse medication administration records: Nurses create a record of each medication administered during an inpatient hospitalization or an outpatient infusion center visit. The nurse medication administration record is the highest level of truth that the patient received a particular medication. Each medication administration event records the drug name, dosage, administration time-stamp, and quantity given. These records reflect only the medication administration events within the premises of a given facility.

Nurse medication administration records, at many facilities, are a paper-based or are documented in free text digital formats. Obtaining medication event information from these records involves manual or, if stored electronically, NLP processing. Some institutions have implemented bar-coded medication administration (BCMA) systems or structured nursing documentation systems. The medication administration event information in such cases is available in structured and complete format. This includes a structured coded drug identifier that uses a terminology similar to pharmacy systems. However, BCMA is rarely implemented at outpatient infusion centers where most of the chemotherapy drugs are administered.

2.1.1.3. Advantages and disadvantages of sources for chemotherapy events

There is an implicit decision process for optimization between truthfulness of chemotherapy medication event and corresponding data accuracy for the task of chemotherapy plan abstraction. A comparison between the provider's clinical notes and nurse administration records can provide a good example for event accuracy. A chemotherapy drug event recorded in clinical notes is only an intention of such an event to occur, whereas the same event recorded in nurse administration records provides the conclusive proof of occurrence of the event. If both the records are in free text format, data accuracy for either of these sources is limited. If the same chemotherapy event, however, is recorded in the CPOE or in the pharmacy dispensing records the event data accuracy would be very high.

The BCMA records, if implemented and available, would provide the most accurate account of medication administration event in terms of event truthfulness and data accuracy. Given a choice between the CPOE and the pharmacy dispensing records, the later would provide a more accurate picture of chemotherapy drug events by virtue of being closer in time to the corresponding administration event. Also, relative to a given administration event, the corresponding provider orders in the CPOE system have a wider time-precedence as compared to the corresponding drug dispensing event. Moreover, electronic implementation of the pharmacy dispensing records has wider penetrance compared to CPOE implementations or provider order-set implementations. In an ideal scenario all of these sources could be used together and triangulated to provide a statistical certainty as to the degree of confidence in the occurrence of a medication event.

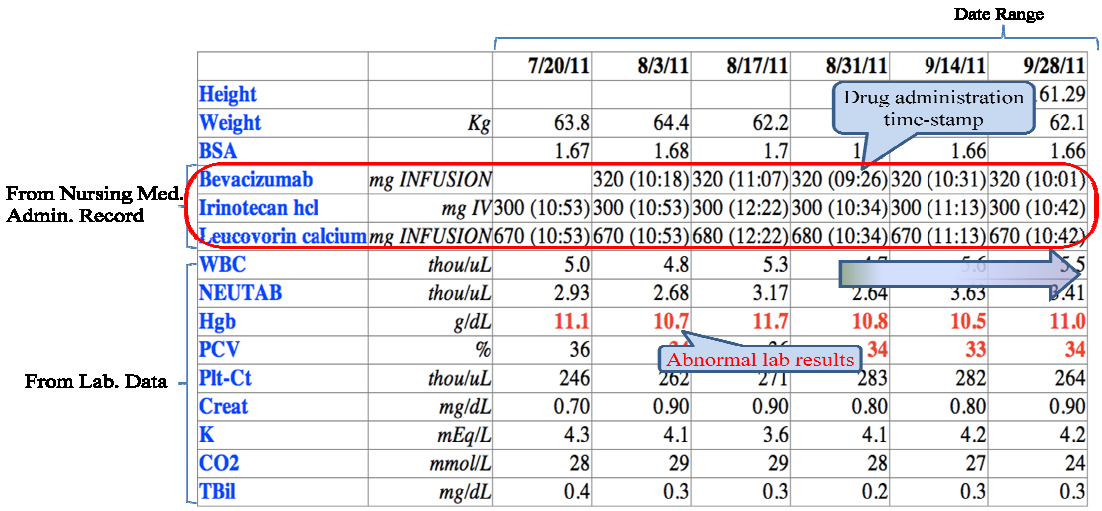
2.2. Challenges related to extracting chemotherapy plan history

The executed chemotherapy plans are not recorded in structured abstracted version (figure 1.2C). The abstract versions of the plans therefore must be derived from the distinct medication events. Any task or process that derives the chemotherapy plans from the distinct medication events must be able to deduce the plan attributes for individual plans.

An abstract representation of chemotherapy plan provides such attributes as the constituent drugs, number of elapsed cycles, average periodicity, the start date, and the end date. A chemotherapy plan history is a listing of executed plans (figure 1.2C) that are abstracted from the chronology of executed medication events (figure 1.2 B). A chemotherapy plan history *is* the *treatment summary* that an oncologist would refer to during the plan selection task (figure 1.1). The treatment summary not only contains the details of the actual plans administered, it also implicitly conveys the knowledge of the intended plans. It informs the oncologist regarding the sequence of the plans and the number of elapsed cycles of the past and current plan.

Oncologists usually record such information as the current plan and the cycle number in clinical notes, to refer back to it during the subsequent encounter. The sequence of treatments is also recorded in summary sections of documents. Such documentation practices can be prone to transcription error and are often difficult to find in the sea of clinical records. Another approach is to reconstruct the treatment history by reviewing the history of distinct chemotherapy administration events, and creating an abstract conceptualization of the chemotherapy plans from the details and chronology of those events. Current interfaces, however, provide only simple formatting with basic

segregation of the event data. An example of one such interface is a chemotherapy flow sheet shown in figure 2.1.



		7/20/11	8/3/11	8/17/11	8/31/11	9/14/11	9/28/11
Height							61.29
Weight	*Kg*	63.8	64.4	62.2			62.1
BSA		1.67	1.68	1.7	1	1.66	1.66
Bevacizumab	*mg INFUSION*		320 (10:18)	320 (11:07)	320 (09:26)	320 (10:31)	320 (10:01)
Irinotecan hcl	*mg IV*	300 (10:53)	300 (10:53)	300 (12:22)	300 (10:34)	300 (11:13)	300 (10:42)
Leucovorin calcium	*mg INFUSION*	670 (10:53)	670 (10:53)	680 (12:22)	680 (10:34)	670 (11:13)	670 (10:42)
WBC	*thou/uL*	5.0	4.8	5.3			5.5
NEUTAB	*thou/uL*	2.93	2.68	3.17	2.64	3.63	3.41
Hgb	*g/dL*	11.1	10.7	11.7	10.8	10.5	11.0
PCV	*%*	36			34	33	34
Plt-Ct	*thou/uL*	246	262	271	283	282	264
Creat	*mg/dL*	0.70	0.90	0.90	0.80	0.80	0.90
K	*mEq/L*	4.3	4.1	3.6	4.1	4.2	4.2
CO2	*mmol/L*	28	29	29	28	27	24
TBil	*mg/dL*	0.4	0.3	0.3	0.2	0.3	0.3

Figure 2.1. An example of chemotherapy flows-sheet.
A visual aid interface provided for oncologists. For a given patient and date-range, it provides patient specifics, some lab-test data and lists the chemotherapy drugs given to the patient along with the chronology of the administration events.

Manually abstracting chemotherapy plans from distinct chemotherapy drug events is, at best, a sub-optimized process. Keeping mental record of the event order, inferring the plan attributes, and calculating the average periodicity of the plan cycles is not feasible during a patient encounter.

2.3. Challenges related to cohort plan analysis

Many prospective and retrospective cancer cohort studies take into account the patient's treatment history as part of their analysis. Researchers commonly reconstruct and record the patients' treatment history through manual data collection or retrospective chart reviews. At the individual patient level, they use similar techniques and data sources as the clinicians to perform the plan abstraction task. Such efforts are very costly,

time consuming, and sub-optimal at best. Due to the high cost of manual data extraction, these efforts tend to be restricted to small patient samples.

2.3.1. Data sources beyond the EMR for cohort plan analysis

The comparative effectiveness research (CER) between two or more chemotherapy plans is feasible only when detailed treatment history is available for a cohort of patients. CER is designed to inform health-care decisions by providing evidence on the effectiveness, benefits, and harms of different treatment options(10). There has recently been a big push by the US government to encourage CER(11). However, CER usually requires large data sets that historically have only been available in large state and national cancer registries. One such cancer registry is the SEER(12) database that contains patient demographic, date of diagnosis, cancer type, cancer stage, first line of treatment and vital status. It is used to create national cancer incidence statistics(13). However, the treatment history in SEER and other cancer registries is limited to a listing of the drugs used in the first line of treatment and their start date. It does not contain any information regarding the doses of medications, their frequency, number of cycles, duration of treatment, or treatment response. Nor does it contain any information on subsequent therapies, an important piece of information for CER beyond first line therapy. Furthermore these registries are incredibly expensive to maintain since they require manual data abstraction and data entry.

Other researchers outside oncology have performed cohort plan analysis using administrative health data such as claims management system (CMS) databases(14). The CMS databases contain coded information on medication events that are billed to insurance as part of patient care. The medication billing records include a date of service

and a structured medication code. Yet others have performed analysis of large databases of pooled outpatient pharmacy records.

2.3.2. Learning cancer systems for cohort analysis

A learning cancer system (LCS) has recently been described to facilitate secondary use of EMR data for continuous CER with feedback to clinical decision support systems(15). Given the heterogeneity of clinical information artifacts for treatment history and medication events, automated methods will be needed to facilitate chemotherapy plan abstraction to realize the promises of a LCS. Both clinicians and researchers thus have a need for a method that can create a rich set of treatment history attributes similar to that of ASCO's treatment summary guideline(3).

Cancer research is currently limited to learning from the clinical outcomes of only the 3% of the cancer patients who participate the clinical trials. The recent progress in molecular testing has increased the number of sub-types of cancer patients, thus shrinking the individual pools of the study participants. Some molecular variants could be too small to necessitate data aggregation over multiple institutions. Given this trend, it would be compelling to learn from the experience of all cancer patients (with their data anonymized), rather than a small proportion of them.

2.4. Hypothesis

To satisfy these requirements, I hypothesize that an automated plan abstraction method can accurately abstract medication plans from the temporal sequence of medication event records across multiple cancer domains. The following approach was used to test this hypothesis:

- Create a data-driven method for automated chemotherapy plan abstraction.
- Test the performance of the method against a manually annotated gold standard set of chemotherapy plans.
- Train and test the performance of the method on a data set limited to two cancer diagnoses.
- Test the generalizability of the method performance on a separate data set that includes all cancer diagnoses except those in the previous step.
- Demonstrate the utility of the method for cohort plan analysis using a large data set of medication events from a single cancer diagnoses at a single institution.
- Perform an across plan analysis by identifying the most frequent plans for a given cancer diagnosis.
- Perform a within plan analysis by exploring the variance in sequencing, total number of cycles, and cycle frequency for a given plan for a single cancer diagnosis.

CHAPTER 3

TEMPORAL REASONING FOR PLAN ABSTRACTION

Temporal reasoning is essential to successfully addressing problems of a time-sensitive and dynamic world. Plan abstraction is a type of temporal reasoning task that takes as input time stamped events and produces as output a sequence of abstracted plans. Temporal representation and reasoning as an area of informatics research has been extensively reviewed(16–21). This chapter focuses on the dimensions of temporal abstraction methods as they relate to plan abstraction, prior work on temporal abstraction, and limitations of the prior approaches to perform the task of temporal abstraction.

3.1. Dimensions of temporal abstraction methods

Before discussing the abstraction methods and techniques, it is important to understand the dimensions involved in temporal abstraction (TA) methodologies. The dimensions of a typical TA method include input data, input knowledge, reasoning methods used to perform the temporal abstraction task, the abstracted output data, and the clinical domains of application. The following sections describe details of each of these dimensions. Table 3.1 lists three of the several methods discussed in the following sections, along with their dimensions.

3.1.1. Input data

TA creates higher level concepts from input of distinct event data represented as instances or intervals in time. In the medical domain higher level concepts are abstracted from clinical event data. Some examples of clinical event data are medications, clinical procedures, lab tests, and vital sign (temperature, blood pressure, pulse rate)

measurements. Clinical data event may itself be structured and contain data components. For example, in chemotherapy plan abstraction, the medication event has a drug-name and dose amount.

Table 3.1: Summary of the temporal abstraction methods.

Method	Input	Type	Clinical Domain	Output
RESUME (based on KBTA)	Time-stamped observational or treatment events	Knowledge-based	AIDS, CGVHD	Set of interval-based, context specific parameters at the same or higher level of abstraction, along with their respective values
RASTA (distributed algorithm)	Time-stamped event data and case-identifiers (patient-ID)	Knowledge-based	Hypertension	Structured datasets passed as XML to invoking application or stored in relational database
CAPSUL	Time-stamped event data pertaining to the procedures, treatments, and lab results	Knowledge-based	Bone Marrow Transplant (BMT)	Interval-based abstractions directly used by applications or displayed using visualization tool (e.g., KNAVE)

The temporal granularity of input data defines the conceptual representation of the timestamp associated with the input event (e.g., a second, an hour, or a day). The RESUME system(22), for example, uses timestamps at specific predefined level of granularity. Complex temporal abstractions can be inferred from the input event data, but the set of granularity levels (and thus the implied temporal uncertainty) is limited to the finest granularity of the input data.

The complexity of TA task increases if multiple granularities are used for the input timestamps(19; 20); it is therefore desirable that the timestamp used for input events be uniform. Given the disparate and varying nature of data sources that provide the input event data (see chapter 2), it is often required to conform the input data to use consistent temporal granularity. Such procedures are covered by pre-processing steps. Conformation of the temporal granularity is but one of the pre-processing steps. Other pre-processing steps include filtering of input event data with-respect to desired type of events and fixing it with respect to the data content.

3.1.2. Input knowledge

The input knowledge used by TA methodology is typically known as a '*knowledge base*'. Some TA methods that use a 'knowledge base' as input can be applied across multiple domains. A knowledge base is useful as input in the creation of generalizable TA methods that are not domain- or problem-specific. In such implementations the methodology remains the same while the knowledge base used is domain-specific and therefore changes depending on the applied domain. Depending on the method, the input knowledge base can be a combination of rules, ontologies, or semantics.

One of the most well-known TA methods is Y Shahar's KBTA framework(23), developed in 1980s. In this framework the external knowledge base provides domain specific structural and semantic knowledge to perform TA tasks. It uses four domain specific knowledge types: structural, classification, temporal semantic and temporal dynamic knowledge. As another example, ChronoMiner developed by R. Raj(24) is an ontology-driven method, which uses a mining ontology as an input knowledge base to

mine patterns of HIV mutations associated with the drug-resistance from the time-oriented research data.

The abstractions produced by methods using a knowledge base are dictated and constrained by the input knowledge. Even though the TA methods that use a knowledge base can be applied across multiple domains using domain-specific knowledge, they require the knowledge to be maintained and kept up-to-date to produce accurate abstractions. Often a separate effort is established to acquire and curate the knowledge base for each domain of application; "Knowledge acquisition" as the step is termed in KBTA, involves collecting and curating the knowledge and building onto it moving forward; this step also involves amending or improving the existing pool of knowledge(25).

3.1.3. Reasoning methods

The reasoning methodology itself can be classified broadly into knowledge-driven and data-driven methods. A knowledge-driven method uses an external knowledge base as a guide to recognize abstractions from the input data and consists domain independent reasoning subtasks. A data-driven method, on the other hand, uses the input data itself to perform the abstraction task. The following two subsections describe the knowledge-driven and data-driven TA approaches and corresponding examples.

3.1.3.1. Knowledge-driven methods

The knowledge-driven TA methods use domain-specific structural or ontological knowledge to perform generic TA tasks.

Shahar's KBTA(23) method, for example, uses domain specific structural, classification, temporal semantic and temporal dynamic knowledge as its knowledge base. The KBTA method itself is decomposed into five sub-tasks: temporal context restriction, vertical temporal inference, horizontal temporal inference, temporal interpolation and temporal pattern matching. Each of these sub-tasks is solved by a set of corresponding domain independent rules. These sub-tasks produce abstractions of several types: state (e.g., high, low), gradient (e.g., increasing, decreasing), rate (e.g., slow, fast), and pattern (e.g., crescendo). The RESUME system(22) is an implementation of KBTA framework. The knowledge base for RESUME is called "TA ontology" and it defines ontologies of events (e.g., drug administration), of parameters (e.g., blood-glucose values), and of interpretation contexts. The RESUME system takes as input the time-stamped patient data and clinical events, and produces abstractions that can be stored for additional analysis or for subsequent use by other applications. The TA mechanisms iterate alternately, activated by the input data and by the previously derived abstractions. This setup, as Shahar and Musen describe, makes KBTA versatile enough to be used over a variety of clinical domains(26).

RASTA is another knowledge based approach developed by O'Connor et al(27). RASTA incorporates many ideas and concepts used by RESUME (which uses KBTA framework), and acts as a basis of a scalable architecture for performing temporal reasoning with clinical data. RASTA uses a distributed algorithm to allow independent evaluation of abstractions in abstraction hierarchies. The algorithm allows the methodology to work in parallel on very large datasets and supports varying configuration options to deal with different application requirements. RASTA uses an

"*abstraction knowledge base*" for *input knowledge*, which is a detailed description of all the temporal abstractions that it can perform in a particular domain with the time-stamped *input data* (termed as *primitive data* in RASTA terminology). The input data is assumed to be sourced from a relational database specified in the '*mapping knowledge base*', which also specifies the database table and column name for each data component. Each piece of the input (primitive) data is time-stamped. The abstractions are associated with a particular *context* (another part of knowledge base) – a proposition that intuitively represents a state of affairs (e.g., an abstraction may be relevant only during the administration of a certain type of drug). The TA algorithm itself, like the KBTA framework, contains four sub-tasks: context restriction, vertical temporal inference, horizontal temporal inference and temporal interpolation.

Chkravarty et al. proposed CAPSUL(28) as a 'pattern specification language' to acquire and evaluate the knowledge for the knowledge base, and to **perform TA by analyzing patterns** among the time-oriented clinical data. CAPSUL allows the specification of components (what repeats), pattern constraints (how it repeats) and the corresponding context to define the ontology of patterns. CAPSUL allows 3 levels of constraints, local, global, and repeating, which are defined as ranges to enhance flexibility. Based on the given ontology of patterns (the knowledge base), CAPSUL relies on the RESUME system as its computational tool to perform the temporal abstractions. Abstractions are associated with respective rules that govern how they are derived from the input time-stamped data points for a given set of constraints. As is the case for RESUME, the newly created abstractions are added to a general pool of instances from which further abstractions can be derived.

3.1.3.2. Data-driven methods

Data-driven methods rely on the content of the '*input data*' to abstract and infer temporal information. These methods adopt statistical, machine-learning, or heuristic approaches for TA over the data.

The data-driven methods adopting statistical approach use tools of regression analysis or association rules to perform TA. For example Lin et al. used logistic regression, association rule analysis and classification trees (a data mining technique) to impute associations between antiretroviral drugs administered (as a predictor) to the HIV patient and corresponding mutation of the HIV(29). The temporal analysis used was the time-window of drug administration and the extent to which the HIV mutated during that time. The temporal abstraction produced at individual patient level was the length of the time window and corresponding number of mutations in the HIV.

Bramsen et al. used a supervised machine-learning approach to identify pair-wise temporal relations using temporal anchors(30). They used manually annotated samples for supervised training and used segment boundaries (events) and anchors (e.g., *yesterday*) to discern the relationship between events. This method would produce a set of event pairs and their temporal relationship identified by the method in terms of *before*, *after* and *incomparable*, along with a corresponding score (a higher score indicates higher confidence). The method may produce, for example, an event pair (insulin injection, blood glucose measurement) with a relationship of 'before' – indicating that the event 'insulin injection' occurred 'before' the event of 'blood glucose measurement'.

A heuristic approach seeks to gather temporal information by exploration of possibilities, rather than following pre-set rules. Cousins et al. have used the temporal

granularity heuristics(17) to associate the level of importance of a medical event to the temporal granularity being considered. For example, even in an acute setting like an ICU, previously recorded information is manipulated at a different level of temporal granularity than the current events; also, once discharged from the hospital, the entire ICU course can be combined into a single abstract fact.

Data-driven TA methods, unlike the knowledge-based methods, do not use any external reference to perform the task of abstraction. The abstraction task is performed either by using standard algorithms (e.g., statistical methods) or by using / following the features contained in the data itself.

3.1.4. Output data

Temporal abstraction methods can produce outputs of various types typically with a time interval temporal representation. The KBTA method is capable of producing temporal abstractions of interval, state, gradient, rate and pattern, depending on the application. The RESUME system(22), for example, produces temporal abstraction for chronic graft-versus-host disease (CGVHD). The abstraction (shown in Figure 5 of the paper) shows the respective grades for mylo-, platelet or granulocyte toxicity, along with corresponding trends of decreasing or increasing platelet and granulocyte counts.

The RASTA system produces abstractions as structured data sets of temporal intervals. These abstractions are passed to the invoking application as a custom XML data structure or stored in a relational database. The output of CAPSUL is a set of interval-based abstractions (including pattern abstractions) that can be directly used by applications, or can be displayed and explored using visualization tool, such as knowledge-based abstraction visualization and exploration (KNAVE)(31).

Other customized TA applications, such as the temporal granularity heuristics created by Cousins et al. produces an output on a graphical user interface (GUI) to communicate the time-line of the events that can be interactively manipulated to varied degree of granularity and observe the events(17). The time-line of diabetes data for a DM patient shows events of illness, stress, hospitalization, along with varying levels of blood glucose. The ChronoMiner(24) by R. Raj also produces the GUI output showing the subject-wise longitudinal view of clinical data showing the viral load, mutations, the chronology of the drug treatment intervals for each drug. Such visual displays are informative and can communicate valuable information to clinicians.

Bramsen et al. produce pair-wise temporal associations between events(30) and respective events as identified by the segmentations that can be useful for processing clinical narratives. The output abstractions produced by the KBTA based methods used by Shahar and Musen can be of different types, state, gradient, rate or pattern; and are determined by using goal-oriented task specific controls.

3.1.5. Clinical domains

Given the fact that patient health and medical data is time-sensitive, temporal representation and abstraction touches almost every domain of medical practice. There is a long history of using temporal abstraction methods in the domain of cancer dating back to the 1980's. Kahn et al. developed TOPAZ(18) to interpret time-varying patient data for applications in cancer chemotherapy treatments and generate narrative summary of the temporal events found in the EMR.

The KBTA method and CAPSULE have both been used for assessment of graft versus host disease in bone marrow transplant. The KBTA method has also been used in

the domain of therapy for insulin-dependent diabetes(26), and AIDS patients' therapy(22; 32). Levy's Rule-based Response Assessment Method was used to classify tumor response to treatment in solid tumors and lymphoma(33). In the domain of medical research Lin et al.(29) and Raj et al.(24) have used customized TA methods to study HIV mutations associated with drug-resistance. Shahar and Musen reason(34) that clinical guidelines are a common format in medical domains (be it oncology, AIDS, or diabetes) for prescribing a set of rules and policies that a provider should follow. They have demonstrated that automated support for clinical guidelines could be enhanced considerably by formal representation of *therapy-planning-operators' efforts*, *plan-revision strategies*, and *the underlying goals and policies of the guideline* in the form of temporal abstraction patterns to be maintained, achieved or avoided.

TA is a very useful mechanism available to analyze medical information. With the advent of EMR systems, it is imperative that various applications of TA can be designed and devised. Adlassnig et al.(20) have provided a detailed account of promising directions of research in the field of temporal representations and reasoning in medicine, and Augusto(19) suggests that more research is needed to make time-based systems for widespread use in medicine.

3.2. Limitations of prior work on temporal abstraction

The existing methods such as those based on KBTA(23), (26), use external knowledge base as a reference to perform TA. To accomplish the TA task accurately, these methods require a carefully compiled and curated knowledge base. Additionally they need to continually maintain this knowledge base to keep it up to date. The tasks of knowledge acquisition and evaluation require the creation of the additional elaborate

tools. In the cancer domain itself, as was discussed in chapter 2, there are over one thousand protocols of chemotherapy treatment(9) with several thousand more under investigation in therapeutic clinical trials(35). If a knowledge-based TA method were to be used for chemotherapy plan abstraction, it would require maintaining a knowledge base of all these chemotherapy protocols and regular updates that result from the developments of the new protocols culminating at the end of the clinical trials in progress. This regular maintenance of a knowledge base requires significant effort from experts in various cancer domains.

Other TA methods have restricted application for detecting patterns from the input events and are otherwise not generalizable or reusable methods. The data-driven method for prediction of HIV mutations by R. Raj(24), for example, indicates only temporal association between the HIV mutations and corresponding treatment. It is used as a preliminary step to provide predictors for domain experts to perform confirmatory analysis. The elegant interface produced by Cousins(17) for display and manipulation of temporal information does a very good job of displaying temporal events in appropriate sequence, but achieves little in creating patterns at the output with associated attributes to aptly describe features of the distinct treatment plans abstracted from the input events. The data-driven method for finding temporal order in discharge summaries by Bramsen(30) is also restricted to indicating temporal association between two events in terms of their relative temporal order.

3.3. Informatics opportunity

Considering the limitations of the knowledge-based TA methods, there is an opportunity to create a data-driven plan abstraction method that does not rely on external

knowledge base to abstract medication plans. In an ICU setting knowledge-based and data-driven methods were used for TA to derive features that would be used to predict whether postsurgical patients needed mechanical ventilation (MV)(36). The knowledge-based method uses knowledge from practitioners to derive qualitative patterns of state changes. The data-driven method, on the other hand, searches through a large number of data summaries to discover those that have predictive value for the need for MV. An assessment of the two methods by Verduijn showed that the knowledge-based method had better sensitivity, with a lower misclassification rate. Moreover, the data-driven method provided additional statistical summaries.

In addition to the need for an integrated solution for a plan abstraction method at clinical level, it is expected that such a tool would be useful in performing cohort analysis. As described in chapter 2, the task of extracting treatment history for individual patients in a cohort study is laborious, lengthy and costly. A tool that can abstract treatment history from distinct medication events can prove to be a valuable resource.

CHAPTER 4

CHEMOTHERAPY PLAN ABSTRACTION METHOD

The chemotherapy plan abstraction method takes as input chemotherapy drug events and produces as output a sequence of patient level chemotherapy treatment plans. These plans can be used for patient care and research, or can be used as input to a cohort plan analysis method.

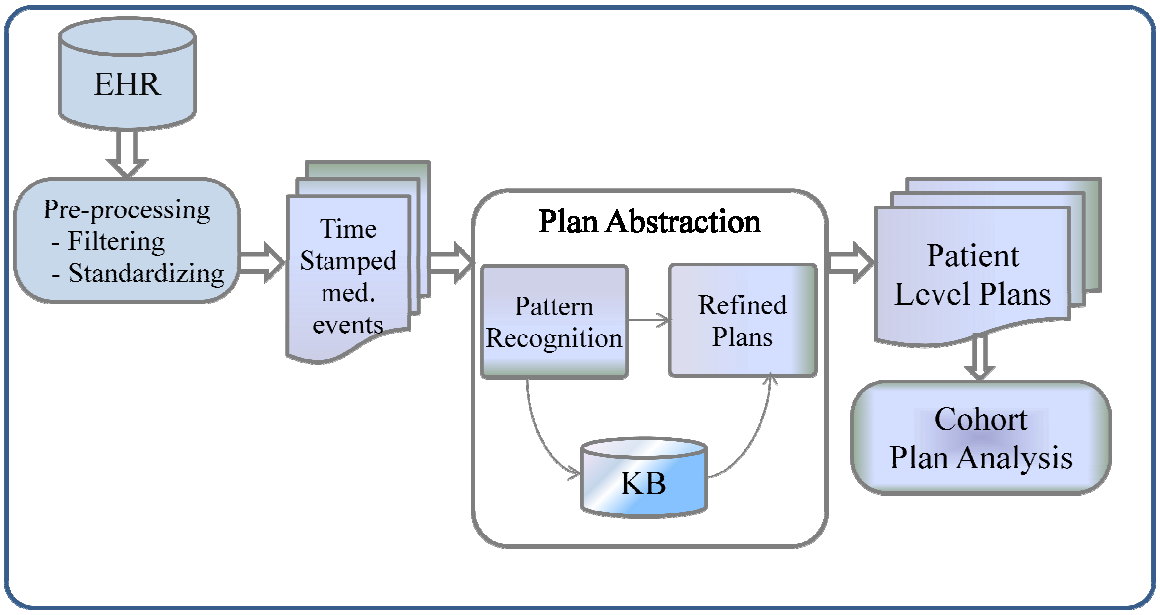


Figure 4.1. Schematic of chemotherapy plan abstraction method. The pre-processing module takes its input from a data source residing in the EHR (Electronic Health Record) and produces standardized, time-stamped medication event data. The plan abstraction method reads the time-stamped medication event data as input and produces patient level plans as output. These plans can be analyzed at the cohort level to provide informative aggregate data.

4.1. Overview of chemotherapy plan abstraction method

The development approach was to create a method that would process a standardized set of time-stamp medication events, detect temporal patterns among them, and produce

medication plans at the output. The method uses a 'data-driven' knowledge base (KB) to refine the plans before delivering the plans as output. Figure 4.1 shows the schematic of this approach. Depending on the particular requirements of the data source for medication events, a corresponding pre-processing module creates standardized medication event data that can be used as input to the method. The chemotherapy plans and corresponding information attributes that the method produces provide valuable information to the practicing oncologists. This information, when considered collectively over patient groups, can serve as input for cohort level analysis.

4.2. Pre-processing of data

The method expects the medication events to have a minimum of three important attributes, Patient ID, Drug Name, and the date-time stamp of drug administration. Very few data repositories in EHR subsystems can provide medication event data with these attributes without some amount of pre-processing. A pre-processing module imparts flexibility to the method by enabling it to read input data from disparate sources. Pre-processing incorporates filtering, cleansing and transforming of the input data to accommodate variances inherent in the sources of data.

4.2.1. Data Sources

Table 2.1 summarizes the various clinical data sources available at the Vanderbilt University Medical Center (VUMC) (denoted by a "V") that could be used to extract medication events.

At the time of this study, these data sources were available with variable completeness (denoted by the "%"). Specifically, bar-coded nursing administration records were not available for the patients treated in the outpatient cancer infusion center,

where the vast majority of cancer drug therapies are administered. Pharmacy dispensing records, however, were available electronically for both the outpatient and the inpatient population, providing a complete record of the chemotherapy medications administered to the cancer patients at our institution. The pre-processing method discussed here creates well-formatted medication events from the pharmacy transaction database for use as input to the CPAM.

4.2.1.1. Pharmacy database and the Synthetic Derivative

The Enterprise Data Warehouse (EDW) at the VUMC contains the schema where pharmacy transaction data are stored. Four DB tables in this schema contain the medication dispensing information. The EDW, however, contains patient identified data. To develop and test the plan abstraction method while maintaining patient privacy, we chose to de-identify the pharmacy datasets used for training and testing the method. At the same time, we took the opportunity to incorporate the comprehensive version of the de-identified pharmacy data into our institution's de-identified synthetic derivative of the EHR.

In 2006, the VUMC began the effort of creating a comprehensive de-identified relational research database called the Synthetic Derivative(37) (SD). The SD contains clinical data (physician notes, orders, diagnoses, lab tests, etc.) in de-identified form(38), *derived* from the VUMC's EHR; all the personal information from the EHR is stripped, and dates shifted (*synthesized*) before it is inserted into the SD.

The SD makes use of DE-ID, a commercially available licensed de-identification tool, to scrub EHR records of the 18 HIPAA safe harbor provisions, along with the other significant pre- and post-processing techniques. The de-identification process ensures that

the corresponding relevance of attributes to individuals is maintained, but re-identification (reverse of de-identification) is not possible. The medical record numbers (MRNs) identifying individual patients are replaced by respective Research User IDs (RUIDs), and dates (of birth, admission, discharge, diagnoses, etc.) are randomly shifted. To ensure temporal relativity of the dates – age at the time of admission, for example – the date shift is consistent for a given patient. **Use of the SD requires appropriate IRB approval, which was obtained for this study**.

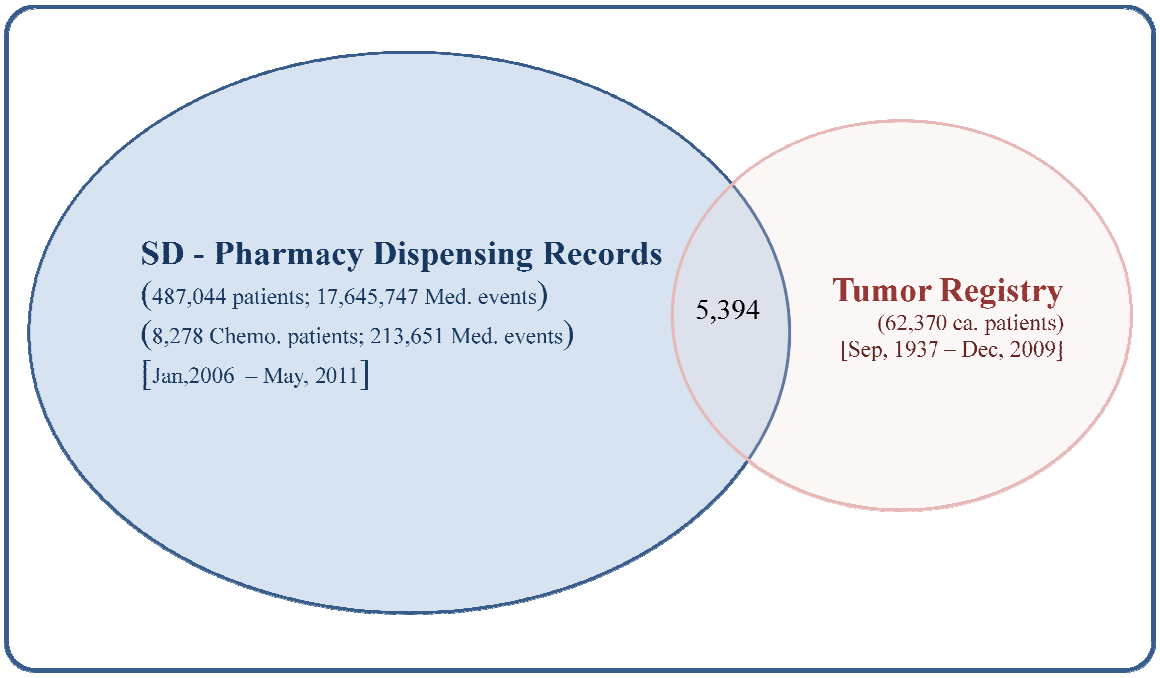


Figure 4.2. SD and Tumor Registry.
Figure showing pharmacy dispensing records set in SD and corresponding overlap with the Tumor Registry data in terms of number of patients that have respective data in either sets. SD = Synthetic Derivative.

Currently over 1.5 TB in size, the SD contains information for over 1.5 million patients from the majority of the clinical activities at the VUMC, including laboratory results, vital signs, imaging, pathology reports, billing codes, clinical narratives etc. The

SD has been used as a stand-alone resource for clinical research. The Vanderbilt Ingram cancer center (VICC) tumor registry (TR) data for over 62,000 patients has recently been added to the SD, to enhance cancer research. As part of research effort for this thesis, pharmacy data was added to the SD consisting of 17,645,747 charged item records between January 2, 2006 and May 22nd, 2011. Figure 4.2 shows the pharmacy dispensing records set and the TR data set, and the number of patients with data in both sets.

4.2.2. Data extraction and processing methods

For the purpose of testing the chemotherapy plan abstraction method, only medication dispensing transactions pertaining to chemotherapy drugs were extracted. Vanderbilt's pharmacy system uses National Drug Codes (NDC) as its controlled terminology for representing drug concepts. A sub-set of these codes is manually classified as "cancer drugs" for operational purposes. This classification was used as a reference to filter the pharmacy dispensing records for chemotherapy drug instances. A single de-normalized dataset of chemotherapy drug events was extracted from the pharmacy data in the SD. Each record in the dataset had the following attributes: 1) a surrogate of patient identifier, 2) name of the drug, 3) NDC identifier, 4) drug dose, 5) frequency at which the drug is to be administered, 6) route of administration, 7) the quantity of the drug dispensed, 8) the dollar amount charged to the patient's account, 9) the date-time-stamp of the charge, and 10) the date of charge. The date-time-stamp of charge corresponds to both the pharmacy dispense date and the nursing drug administration date.

This marked the starting point of the data with reference to the pre-processing module. Figures 4.3 and 4.4 show the schematic of the pre-processing steps for extracting pharmacy dispensing data from the EDW and SD. Pre-processing consists of 1) eliminating canceled orders, 2) removing records with oral drugs, 3) normalizing drug names, and 4) drug dose consolidation.

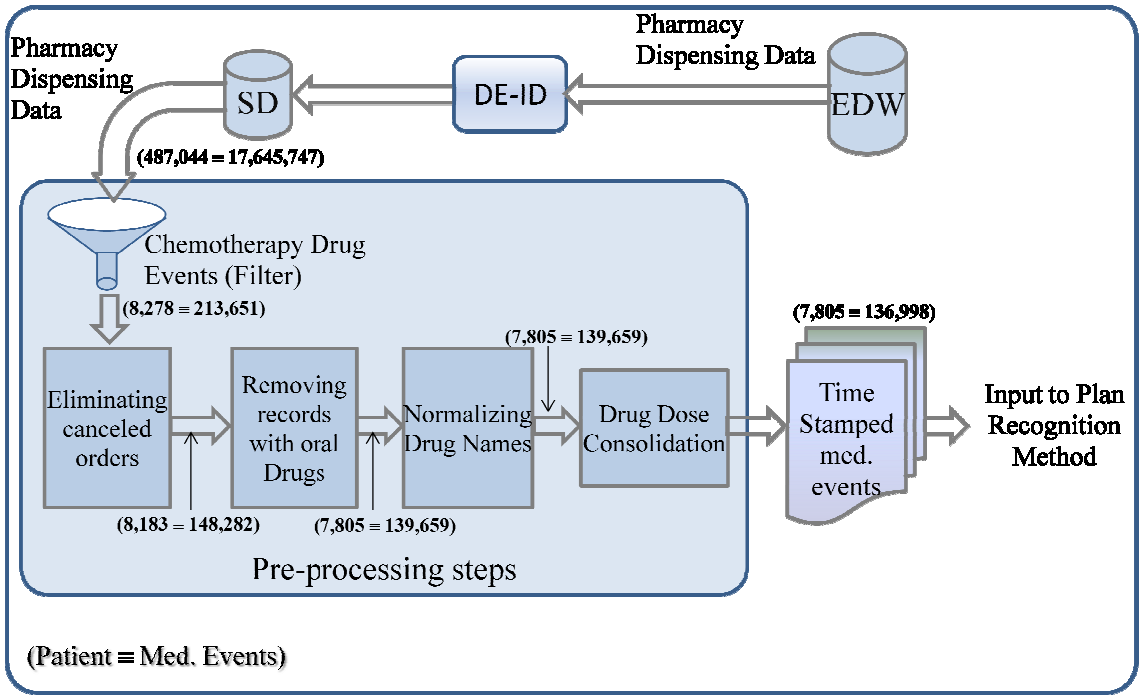


Figure 4.3. Schematic of preprocessing steps.
Pharmacy dispensing records extracted from EDW (Enterprise Data Warehouse) are de-identified using commercial software DE-ID and brought into SD (Synthetic Derivative). The pre-processing module extracts the chemotherapy-dispensing events from pharmacy dispensing records in SD by applying chemotherapy drug-filters. The chemotherapy drug events are then normalized through several pre-processing steps including 1) eliminating cancelled drug orders, 2) removing oral drugs events, 3) normalizing drug names, and 4) consolidating drug doses. At the end, the pre-processing module produces time-stamped chemotherapy events that are used as input to the plan abstraction method. [The numbers at respective steps indicate the patients and corresponding medication events.]

Eliminating canceled orders: Pharmacy systems are clinical transaction systems that allow the pharmacy department to record and store orders and bill for the dispensed medications. If the order is valid when the patient presents to the infusion center to get their medication, the nurse administers the medication and the corresponding charge is processed by the billing system. If due to some reason the dose needs to be changed or the medication canceled, the pharmacy system records that cancellation as a reversal of the charge for that dispensing event. A small proportion of the transactions have positive as well as corresponding negative charge records. To eliminate these canceled orders, the pre-processing method groups the pharmacy records by patient, drug name and credit date of charge. The corresponding charge amounts are aggregated to obtain the net amount charged for a given drug-dispensing event. Dispensing events with a net zero charge were eliminated, since they represent cancelled orders, as illustrated in figure 4.4 drug B is eliminated because B_+ and B_- charges for drug B result in 0 net charge.

Removing records with oral drugs: Anti-cancer therapies that are given orally are typically taken at home daily or for several consecutive days. Most of these therapies are not dispensed by the hospital chemotherapy pharmacy but rather by the patient's preferred outpatient pharmacy. Rarely, these drugs are dispensed by the hospital chemotherapy pharmacy when the patient is admitted to the hospital or on a research protocol. Since these events were rare and did not represent a complete history of oral anti-cancer therapies prescribed to our patient population, they were eliminated for the purpose of this analysis. All the dispensing records with oral administration route (attribute ROUTE with either of these values – 'ORAL', 'PER TUBE', 'PO', 'PO/PT' –

which are all equivalent of oral route) were therefore removed. As illustrated in figure 4.4, drug F with oral route is removed.

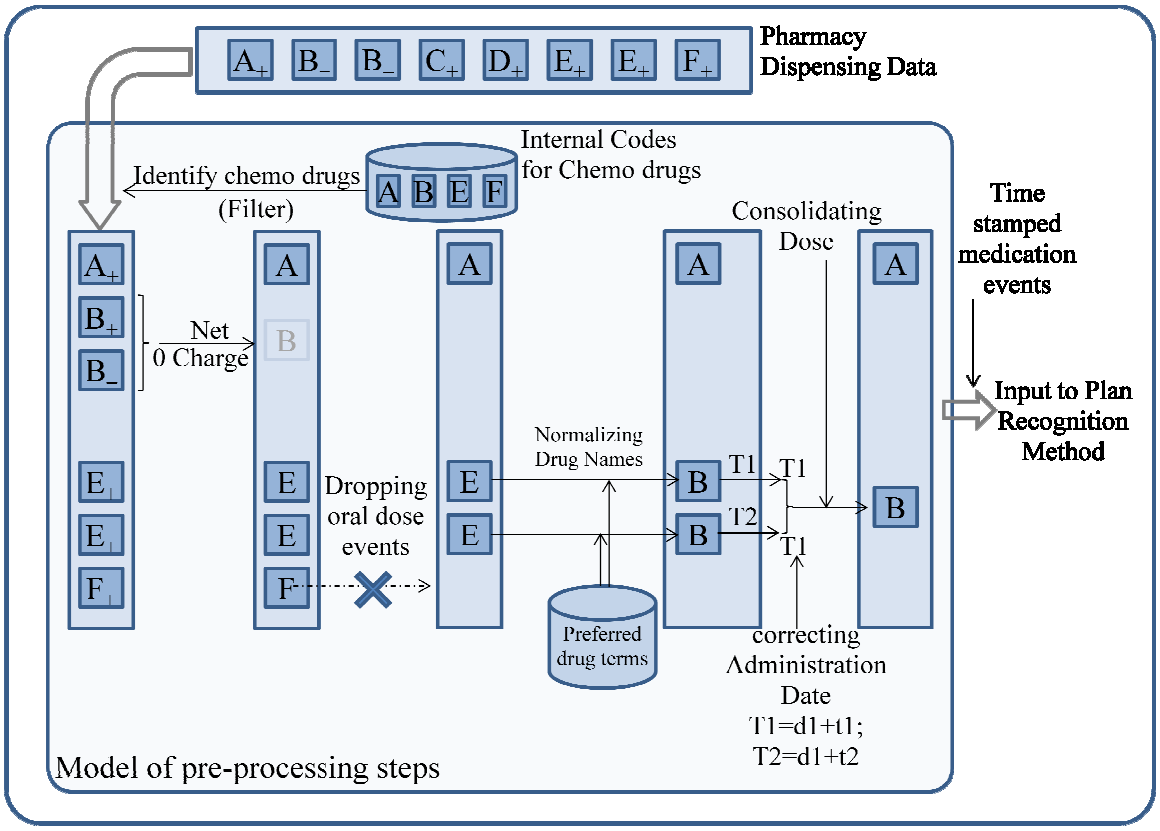


Figure 4.4. An illustration of pre-processing steps.
A reference of internal codes for chemotherapy drugs is used to select chemotherapy dispensing events from the pharmacy dispensing data in SD. After that the events go through 1) charge consolidation to get the net charge to eliminate canceled orders, 2) oral drugs are dropped 3) drug name variants are modified to used common names by referencing the preferred drug-names (provided by practicing oncologist) and 4) administration dates are corrected to use common time-stamps for the drug events of the same drug occurring on the same day (d1) for a given patient and then the drug-doses are consolidated for drugs administered at the same time for a given patient.

Normalizing drug-names: Pharmacy transaction systems often contain multiple names for the same drug. For the purposes of this methodology, synonymous drugs needed to be normalized so as to use a common terminology. To translate the drug name

variants to respective common terms, distinct drug names were extracted from the initial dataset. A practicing oncologist manually mapped the synonyms to a common preferred drug term. The original dataset was then augmented with a new attribute – DRUG_NAME_USED; for each record with drug name variant, the corresponding preferred term was assigned to this new attribute. As illustrated in figure 4.4, Drug E is a variant of drug B and therefore name B is assigned to this drug. For example, drug names Docetaxel and Docetaxel (Taxotere) are both equivalent to Docetaxel; therefore the name Docetaxel assigned to the new attribute DRUG_NAME_USED for both cases.

Drug dose consolidation and controlling temporal granularity: Some medication orders have multiple dispensing records in the pharmacy system because the total dose of a given medication order is dispensed in multiple vials. Depending on the quantity of the drug ordered, the corresponding transactions are appropriately split if it exceeds the constrained size. Also, the finest granularity of administration event frequency considered for this method is one day. Any given drug with multiple doses dispensed on the same day, for a given patient, is counted as a single drug event in the temporal sequence of drugs administered to the patient. This was done by assigning the earliest date-time stamp to the group of records that matched for the patient, drug name and dispense date. For example, drug B in figure 4.4 (which was originally named drug E) is dispensed as two separate doses on the same day. It is consolidated into a single dose, with the earliest time-stamp of the two. As an example, the drug Doxorubicin is typically dispensed in two vials, corresponding to a single dose administered to a patient. These two dispensing events were consolidated into a single dispense event.

At the end of the pre-processing steps, the dataset consisted of 136,998 chemotherapy drug events for 7,805 distinct cancer patients, spanning 5 years and 5 months, with the following attributes: 1) a surrogate of patient identifier, 2) name of the drug (with common terminology) and 3) dispense date.

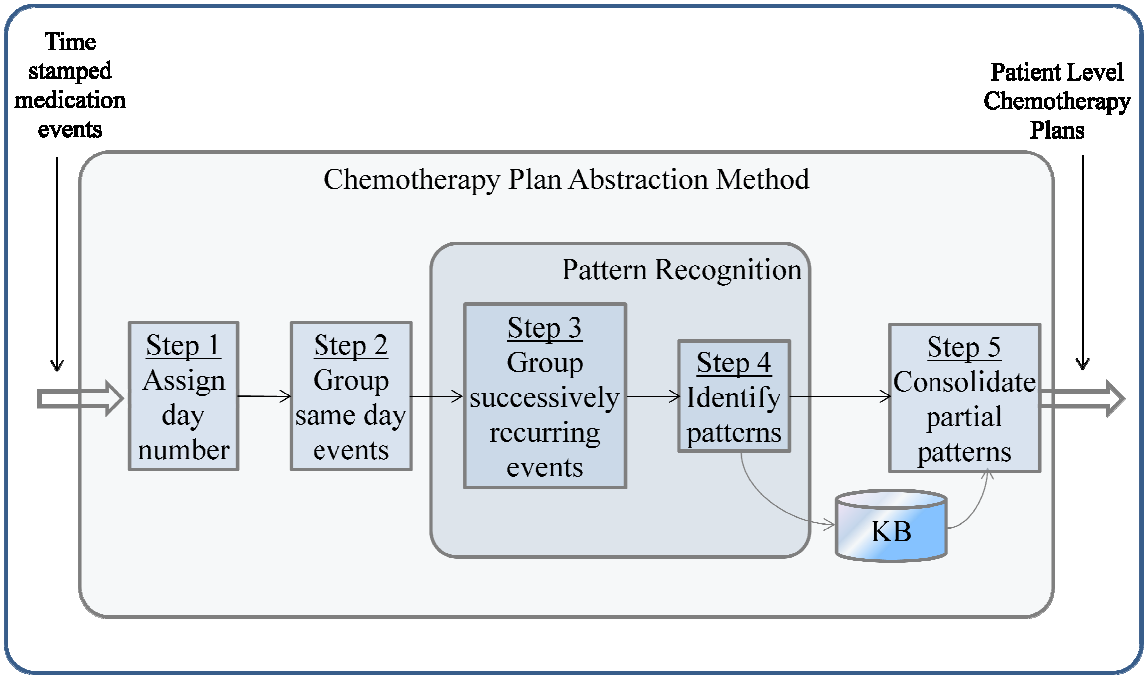


Figure 4.5. Plan abstraction Method steps.
1) Assign day number, 2) group same day events, 3) group successively recurring events, 4) identify patterns and 5) consolidate partial patterns by referencing knowledge base (KB).

4.3. Description of the chemotherapy plan abstraction method

The chemotherapy plan abstraction method takes as input time-stamped chemotherapy medication events including the patient ID, drug name, and administration date, and generates as output a sequence of patient level chemotherapy plans. The method achieves this abstraction in a 5 step process shown in figure 4.5: 1) assign day number, 2) group the same day events, 3) group the successively repeating events, 4) identify

patterns, and 5) consolidate partial plans. The drug events are processed one patient at a time to arrive at the patient level chemotherapy plans. The first two steps, being relatively simple are achieved by using queries written in structured query language (SQL). The later three steps involve relatively complex processing and use Perl scripts.

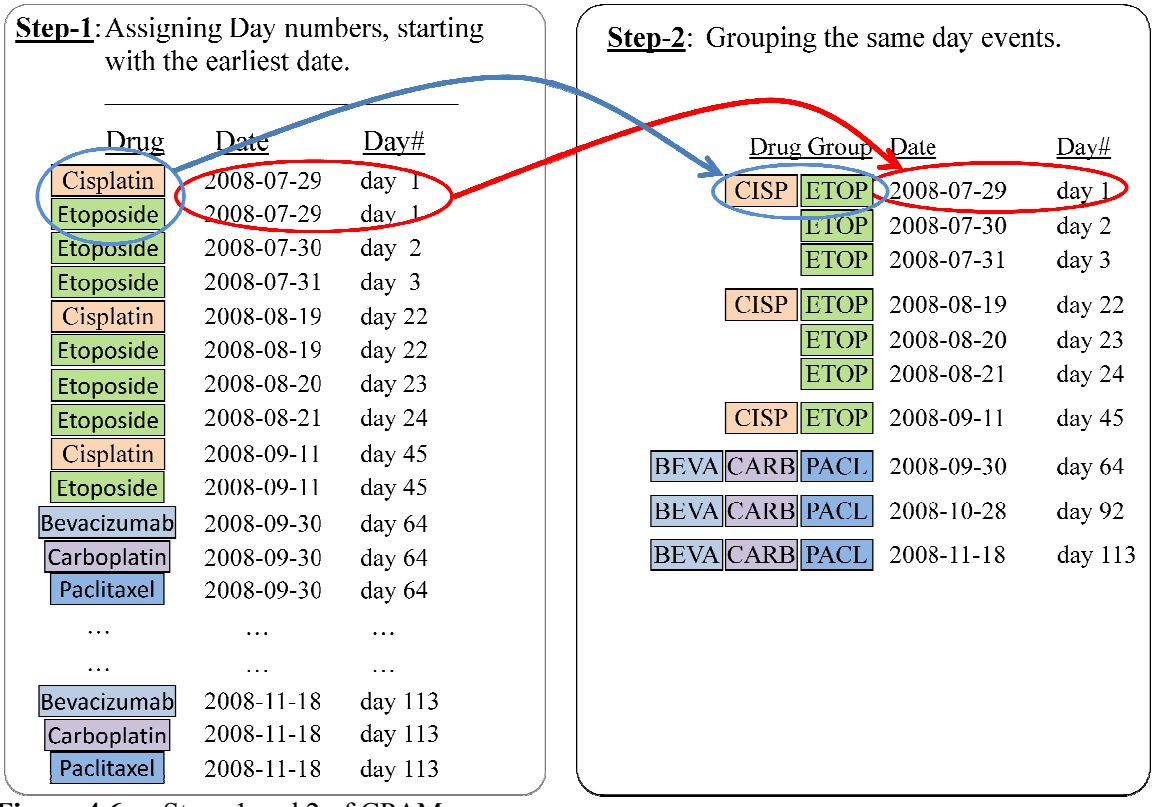


Figure 4.6. Steps 1 and 2 of CPAM.
Step1 – Assigning day number: Medication events for a given patient are ordered chronologically and then, assigned a relative day number starting with 1. Subsequent events are assigned day numbers relative to this first event. Day# = relative day number.
Step2 – Grouping same day events: Events occurring on the same day are grouped together with drug names abbreviated and concatenated.

Step 1: Assign day number

The method starts with assigning a day number for each event, for a given patient. After arranging events in chronological order, the earliest event is assigned the day

number of 1 with each successive event numbered based on its date relative to the earliest event. This step thus converts absolute time instances into their corresponding relative time instances. For example, as shown in figure 4.6, the earliest drug event for Cisplatin occurring on 2008-07-29 is assigned the day number of 1. A subsequent event occurring on 2008-08-19 is assigned the day number of 22, that date being the 22nd day from 2008-07-29. Events occurring on the same day are assigned the same day number; for example, events for Bevacizumab, Carboplatin and Paclitaxel occurring on 2008-09-30 are all assigned the day number of 64.

Step 2: Group the same day events

In this step, the events occurring on the same day are grouped together to form a common event and the corresponding drug names are concatenated. Before concatenating, the names are capitalized and abbreviated to the first four characters, and a '+' sign is used as a separator between consecutive drug names. Figure 4.6 highlights these actions. For example events for the drugs Cisplatin and Etoposide occur on the same day; these events are merged together as a single event. The drug names are abbreviated to CISP and ETOP respectively, concatenated and separated by a '+' sign.

Abbreviating the drug names and concatenation thereof with a '+' separator creates a new vocabulary for the CPAM.

Step 3: Group the successively repeating events

By this step, for a given patient on any given day, only a single event comprised of one or more drugs remains. With the events arranged in chronological order, the method looks for successively repeating events. Each group of repeating events is merged

together into a single record instance. For example, as shown in figure 4.7, events for Etoposide (ETOP) repeat twice and are merged into a single instance with a relative date of '2, 3' as an attribute. As another example, events comprising of the drugs Bevacizumab (BEVA), Carboplatin (CARB) and Paclitaxel (PACL) are merged together into another single record instance of drug combination BEVA+CARB+PACL, a chemotherapy plan commonly prescribed to lung cancer patients. Several new attributes are created at this step including *StartDate*, *DayString* and (number of) *Cycles*, which are carried forward to the next step (and defined in table 4.1). Simple plans start emerging at the end of this step, but more complex plans require additional processing.

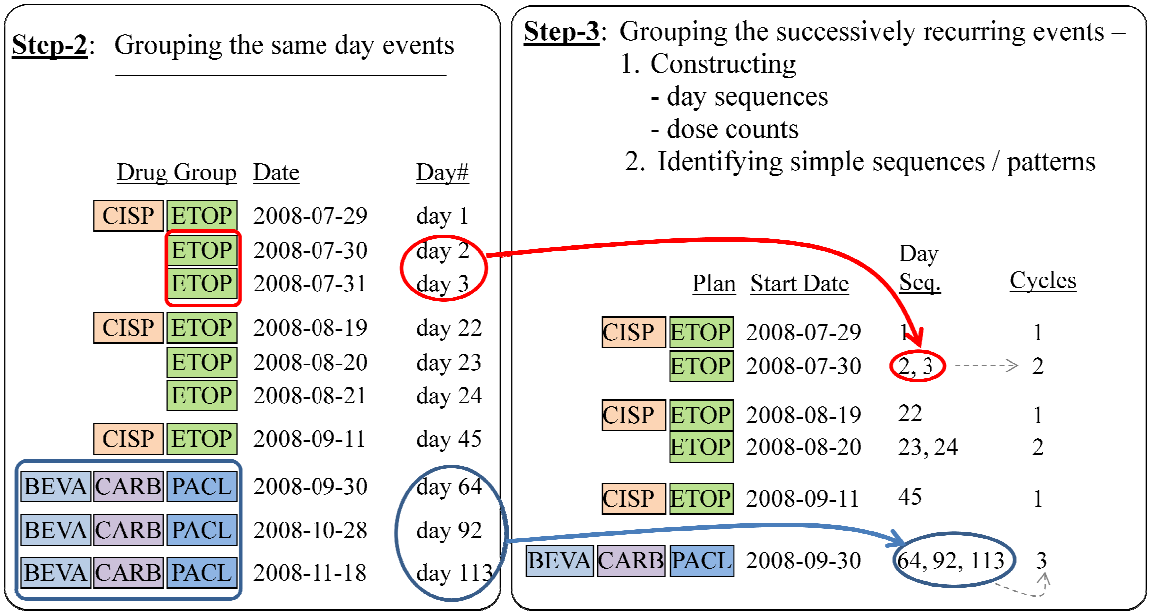


Figure 4.7. Step 3 of CPAM.
Group *the successively repeating events*. Events that repeat successively are merged into a single instance with corresponding start-date, day-sequence and number of cycles.

Step 4: Identify patterns

In this step, the method attempts to detect the repeating patterns of drugs among the sequence of record instances produced in the previous step. As shown in figure 4.8, the

pattern consisting of an instance of CISP+ETOP followed by an instance of ETOP is repeating twice. Record instances exhibiting such patterns are merged together with corresponding pattern constructed. For example, the pattern of CISP+ETOP followed by ETOP is constructed as "CISP+ETOP, ETOP", where a comma (",") separates each drug component in the pattern. This is another example of the emerging vocabulary for this method. Whenever such patterns are identified corresponding record instances are merged and several attributes are re-computed as described in table 4.1. Any records not participating in pattern sequences are carried forward unaffected. For example, as shown in figure 4.8, the record instances consisting of the drug group CISP+ETOP and another one consisting of the drug group BEVA+CARB+PACL.

Step 5: Consolidate partial patterns

Close examination of step 4 output reveals some plans that appear similar to other plans, but are somehow incomplete. For example, the record corresponding to plan 'CISP+ETOP' (the second in the list at step 4 of figure 4.8) appears to be an incomplete version of the previous plan 'CISP+ETOP, ETOP'. This could have happened due to the patient intolerance to the toxicity of the drugs and thus likely represents an incomplete cycle of the same plan. To be able to infer such incomplete or non-repeating complex patterns as plans, a list of distinct plans is compiled at the end of step 4. Every record produced at the end of step 4 that repeats more than once (attribute value for *Cycles* > 1) is classified as a plan at this stage and included in a data-driven knowledge base (KB). Non-repeating plans may be incomplete plans and are thus not included in this list. For example, in figure 4.8 plans 'CISP+ETOP, ETOP' and 'BEVA+CARB+PACL' are added to this knowledge base.

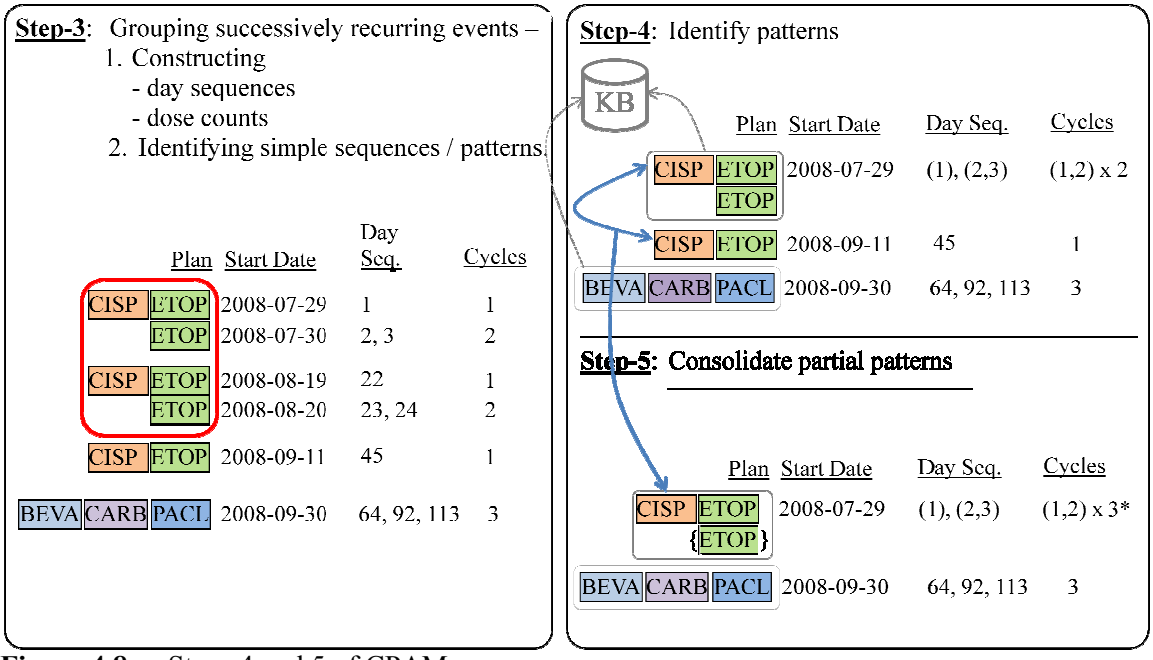


Figure 4.8. Steps 4 and 5 of CPAM.
Step 4: Identify patterns. The method attempts to identify repeating patterns among the record instances produced by the previous step. For example, the pattern CISP+ETOP, ETOP is seen repeating twice. Record instances for these patterns are grouped together and corresponding *StartDate*, *DaySequence*, *Cycles* and other attributes are computed as described in table 4.1. [KB = Knowledge Base]
Step 5: Consolidate partial plans. If a partial sequence similar to the plan immediately prior to it exists, the partial sequence is merged with the prior plan. The drug-string in the plan is modified to indicate the drug-component missing from the last instance of the plan. The *Cycles* string is modified to increment the number of plan cycles by 1, marking it with an asterisk to indicate that the last cycle of the plan was incomplete.

The plans produced by step 4 are parsed one more time using knowledge base as a reference to identify partial plans among the drug-sequence. If any drug-sequence matches partially with a plan in the KB, and the matching plan happens to be the one immediately prior plan for a given patient, the prior plan record and the partial sequence are merged together. As shown in figure 4.8, the plan instance CISP+ETOP occurs only once, and has a similar drug sequence to that of the prior plan, 'CISP+ETOP, ETOP', and thus two plan instances are merged. The number of cycles is incremented by 1 and is

marked with an asterisk, indicating that the last cycle of the plan was incomplete. The drug-sequence in the plan is modified, enclosing the component that was missing in the last cycle with a pair of braces. The marking of '*cycles*' string with an asterisk and enclosing the missing drug-component with braces is yet another example of new vocabulary used by the method to communicate specific meanings.

The output of the chemotherapy plan abstraction method is a set of structured data, with each record representing an *abstract* form of chemotherapy plan inferred from the input data of distinct chemotherapy drug events. A new 'vocabulary' is devised to convey specific and helpful details for each of the plans. The abstract versions of the plans produced by the method are accompanied by a rich set of attributes that are consistent with those suggested by ASCO in their treatment summary guidelines(3). A complete list of attributes that is produced by the method, for each plan, is given in table 4.1 along with the corresponding description.

Table 4.1. Plan attributes produced by chemotherapy plan abstraction method.

Attribute	Description
Patient Id	Surrogate identifier of the patient for whom the chemotherapy plans are produced from the corresponding drug events.
Serial Number	For a given patient this is a running number assigned to each plan, from the earliest to latest, starting with 1. SerialNumber facilitates a simple way of assigning cardinality and order to the set of plans for a given patient.
Drugs	The string listing the chemotherapy drugs that constitute the plan. Individual drug components are separated by comma (‘,’). Each of the drug components themselves may consist of multiple drugs – when multiple drugs are administered on the same day – in which case these drugs are separated by “+” sign.
Cycles	This is the number of cycles by which the chemotherapy plan repeats. In case of the compound plans (plans consisting of multiple drug components separated by comma) this attribute is a set of cycles delimited by comma and enclosed in a pair of parentheses – with each component within the parentheses having one-to-one correspondence to the drug-component in the plan. This string represents the dose count of the first dose set of the drugs. The total number of cycles for the compound plan is indicated by the right-most numeric, outside the parentheses and separated by ‘x’ from the parenthesized set, and indicates the number of times for which the whole plan is repeated. In cases where the last cycle of a compound plan is incomplete (as shown in figure 4.8) an asterisk appears at the end of this attribute.
StartDate	This is the date when the first drug event of the chemotherapy plan (as listed in the attribute *Drugs*) was started.
StartDay	This is the day number corresponding to the *StartDate* of the plan
DayString	The string of day numbers when the drug-events corresponding to this plan occurred. For a compound plan, this attribute is a set of day number values enclosed in a pair of parentheses and delimited by comma – with each component having one-to-one correspondence to the drug-component in the *Drugs* string.
DaysToChange	For a given patient, this is the number of days between the *StartDate* of the current plan and that of the one immediately prior. [By definition, this attribute will have meaning only for the plan records with *SerialNumber* > 1.]
DaysBetween	For a given patient, this is the number of days between *StartDate* of the current plan and the last drug event of the one immediately prior.
Periodicity	This is the periodicity with which the chemotherapy plan is repeated.

CHAPTER 5

EVALUATION

The evaluation of the CPAM consists of an assessment of its accuracy and utility for cohort plan analysis. The former evaluates the extent to which the method correctly infers patient level treatment plans. The later analyzes the concordance of the method's cohort level output with expected standard of care chemotherapy protocols.

5.1. Evaluation of patient level performance

The chemotherapy plan abstraction method was iteratively trained using a manually curated gold-standard training set of chemotherapy drug events for breast and lung cancer patients. The method was then tested on two manually curated datasets of chemotherapy drug events for 1) breast and lung cancer patients and 2) non-breast, non-lung cancer patients.

5.1.1. Training and testing data sets

The training and testing data sets were derived from the pharmacy-dispensing database in the SD. Specific subsets were grouped by cancer diagnosis as determined by the tumor's site and histology from the tumor registry data. Table 5.1 shows the counts of patients and corresponding drug-events used in the training and testing data sets for the CPAM.

To ensure that the method could satisfactorily process the drug events for mixed cancer domains, the initial training and testing data sets consisted of patients with either lung cancer or breast cancer. The initial training set consisted of 163 patients with 2,402 chemotherapy drug events before pre-processing and 2,298 events after pre-processing.

Similarly, Test Set 1 consisted of 341 breast and lung cancer patients with 5,713 chemotherapy drug events after pre-processing.

Table 5.1. Training and testing datasets.
Training and testing data sets including a description of the cancer domain covered, the patient count, and the drug events counts before and after the pre-processing.

	Training set	**Non-trained test sets**		
		Test set 1	**Test set 2**	**Test set 2 (Sample for evaluation)**
Cancer Domain	breast & lung	breast & lung	non-breast/non-lung solid tumor	
Patient count	163	341	7,805	168
Pre-filtered drug events	2,402	6,050	139,659	3,366
Pre-processed drug events	2,298	5,713	136,998	3,214

To test the generalizability of the CPAM, a testing set was created of patients with non-breast, non-lung cancer solid tumors. 7,805 patients with 136,998 chemotherapy drug events were identified in the SD. A random sample was draw from this data set to create Test set 2. The size of the statistically relevant random sample was estimated by taking into account the expected performance of the method. Based on earlier tests, the recall and precision values of 0.8 each were considered. Using these values for either of the characteristics, the sample size figure of 300 was obtained [using proportionality test function `prop.test` of statistics package R - version 2.12.2 (2011-02-25)], with a 95% confidence interval of (0.7493, 0.8428). This suggests that if the true recall and precision of the method were to be 0.8 each, for any random sample of 300 plans obtained at the output of the method, these parameters would evaluate to be between (0.7493, 0.8428) 95% of the time. Though the calculated sample size was 300 plans, drawing a set of

exactly 300 plans from the output was not feasible, since for a given patient, all or no plans needed to be included. The resulting testing set consisted of 306 plans for 168 cancer patients.

After applying the CPAM to both the training and the testing sets, a trained medical oncologist manually classified the output to create the gold standard. The oncologist compared the individual plans and the corresponding input medication events for each patient, marking each abstracted plan as a true positive (TP), false positive (FP), or false negative (FN) plan. For example, if the method produced multiple output plans for what should have been a single plan, each abstracted plan would be classified as a false positive plan. Likewise, the collection of FP plans would be classified as single false negative plan corresponding to the actual plan that was not detected.

Table 5.2. Performance results of the CPAM.
Performance results of the CPAM showing the recall, precision, F1-score and accuracy for the training test set and the two non-trained test sets.

	Training test set (95% C.I.)	Non-trained test sets (95% C.I.)	
		Test-1	Test-2 (Sample)
Recall	0.888 (0.8432, 0.9223)	0.913 (0.8870, 0.9329)	0.899 (0.8537, 0.9316)
Precision	0.752 (0.6996, 0.7973)	0.829 (0.7989, 0.8563)	0.755 (0.7020, 0.8012)
F1-Score	0.814	0.869	0.821
Accuracy	0.687 (0.6348, 0.7346)	0.768 (0.7361, 0.7979)	0.696 (0.6427, 0.7442)

5.1.2. Performance results of the plan abstraction method

Table 5.2 shows the performance of the CPAM for the training and testing data sets along with their respective confidence intervals. The original training set had a recall rate

of 88.8%, a precision of 75.2% and an accuracy of 68.7%. For the testing round, test set 1 had a recall rate of 91.3%, a precision of 82.9% and an accuracy of 76.8%. Test set 2, had a recall rate of 89.9%, a precision of 75.5% and an accuracy of 69.6%. The results for test set 2 represent generalized performance of the chemotherapy plan abstraction method for solid tumors with a confidence level of 95%.

5.2. Cohort level plan analysis

One of the primary goals of the CPAM is to extract sequences of patient level chemotherapy protocols from discrete medication events for cohort level analysis. The second part of the evaluation process demonstrates a simple use of cohort level analysis to evaluate the practice patterns and the variance in plan adherence for several chemotherapy plans as compared to the standard of care. For a single disease breast cancer, we demonstrate an across plan analysis comparing the frequency of the plans administered, and a within plan analysis to understand variances in the administration of a single plan.

5.2.1. Across plan analysis

The first analysis assessed the most frequently administered chemotherapy protocols for a cohort of breast cancer patients at the VUMC. This gives insight into the practice patterns for this disease at our institution. A cohort of 554 breast cancer patients treated with chemotherapy was isolated in the SD using a combination of the tumor registry and the pharmacy data sets. For this set, a list of abstracted chemotherapy plans was created along with the count of patients receiving each plan, and ordered by descending patient count. For this cohort, 107 unique breast cancer plans were identified. Table 5.3 shows the 5 most frequently prescribed plans. A full list of breast cancer plans abstracted by the

CPAM is listed in Appendix A. The most commonly administered plan in this cohort was an adjuvant therapy protocol Cyclophosphamide + Doxorubicin, received by about 36%. of patients. In comparison, the 5th most commonly prescribed plan, an adjuvant protocol consisting of the drugs Cyclophosphamide + Docetaxel, was administered to only about 8% of this population. This demonstrates a clear dominance of the Cyclophosphamide + Doxorubicin adjuvant therapy protocol among this group of breast cancer providers.

Table 5.3. List of five most frequently administered chemotherapy plans for breast cancer. Five most frequently administered chemotherapy plans showing corresponding patient count along with average and standard of care figures for number of cycles and periodicity. The line of therapy column indicates whether the corresponding plan is used for adjuvant or metastatic treatment (or both). The values in columns with grey-background are provided by a practicing oncologist.

Name Chemotherapy Protocol	Patient Count	Line of therapy	Periodicity (days)		Number of Cycles	
			Standard of Care	Average (Min, Max, SD)	Standard of Care	Average (Min, Max, SD)
Cyclophosphamide, Doxorubicin	198	Adjuvant	14	15.8 (7, 28, 3.1)	4	3.6 (1, 6, 0.9)
Paclitaxel	138	Adjuvant, Metastatic	7	10.6 (6.4, 29.4, 4.7)	12, Unlimited	7.1 (1, 16, 4.3)
Trastuzumab	61	Adjuvant, Metastatic	7, 14, 21	22.1 (7.8, 63, 8.6)	1 year, Unlimited	9.0 (1, 64, 11.1)
Fulvestrant	55	Metastatic	28	27.3 (13, 107, 16.6)	Unlimited	6.4 (1, 53, 8.5)
Cyclophosphamide, Docetaxel	46	Adjuvant	21	21.7 (20, 29.7, 1.8)	4	3.7 (1, 6, 1.0)

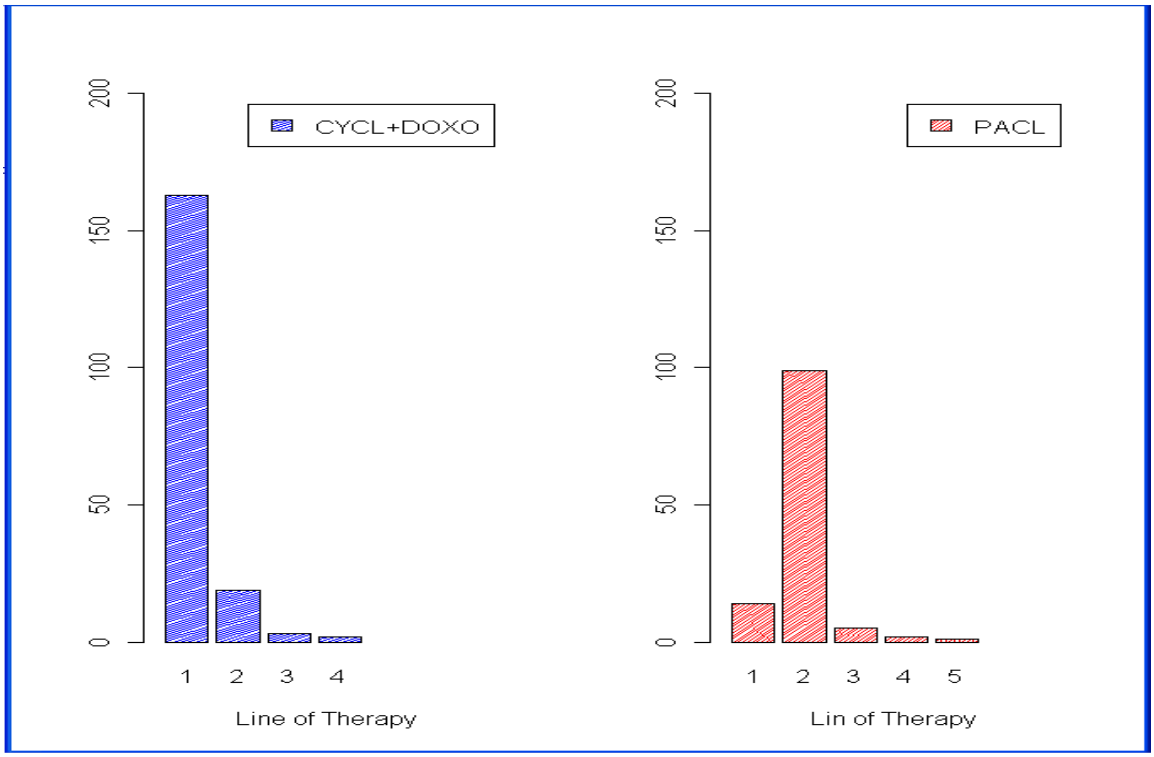


Figure 5.1. Line of therapy bar-plot for the two most frequently administered breast cancer plans.
The line of therapy bar-plot shows that most of the patients receiving the "CYCL+DOXO" plan received it as the first line of therapy, and most of those receiving "PACL" plan received it as the second line of therapy.

The sequence of plans also gives insight into provider adherence to standard of care protocols. Figure 5.1 shows the bar chart for the sequence of therapy for the two most frequently administered breast cancer plans. The sequence of therapy, represented by the Serial Number attribute produced by the CPAM indicates whether the plan was given first, second or third in sequence for a given patient. As is consistent with the clinical practice, the Cyclophosphamide + Doxorubicin plan is typically the first plan to be administered in the adjuvant setting(39). Depending upon the clinical context, the Paclitaxel plan is typically administered second following the Cyclophosphamide + Doxorubicin plan in the adjuvant setting(39), (40) or as a first line of treatment in the metastatic setting(41). This analysis demonstrates both that the CPAM produces results

concordant with what is expected for standard of care, and that the VUMC practice patterns are likewise concordant.

5.2.2. Within plan analysis

A within plan analysis also provides insights regarding plan adherence. As discussed in earlier chapters, patient toxicity may require deviation of the actual chemotherapy drug administration events from the planned periodicity and cycles defined by a chemotherapy protocol. Table 5.3 shows the observed average periodicity and number of cycles compared with the corresponding standard of care values for the five most frequent breast cancer plans.

The statistical parameters, including the standard deviation were calculated for each plan over the respective number of patients. A trained medical oncologist defined the corresponding standard of care values. Figure 5.2 shows violin plots of the observed values for periodicity and cycles for the two most frequently prescribed plans, “Cyclophosphamide + Doxorubicin” and “Paclitaxel” [statistics package R, version 2.12.2 (2011-02-25).] Over each of these violin plots, a reference line is drawn in red corresponding to the standard of care value and another line in green showing the average for the dataset. The distribution plots for “Cyclophosphamide + Doxorubicin” exhibit a pronounced modality at or near the standard of care values, indicating a close concordance to the respective standard of care values for this treatment plan.

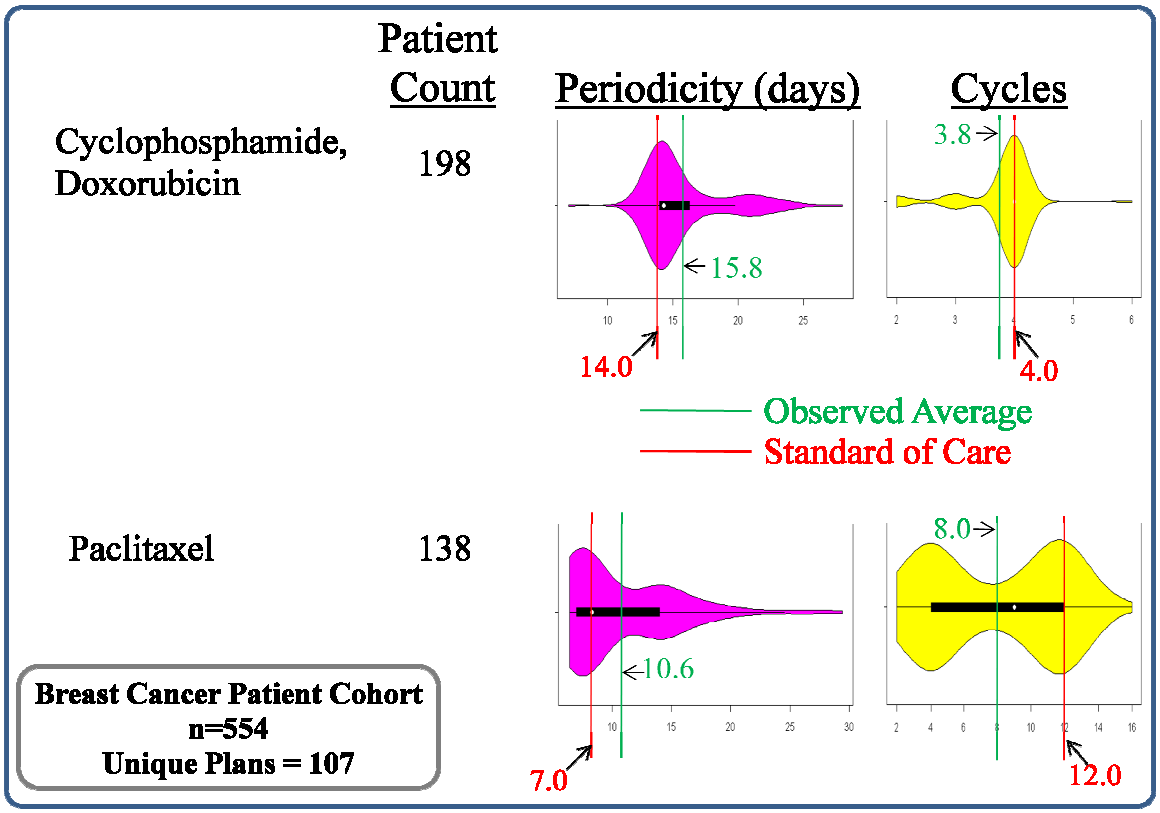


Figure 5.2. Periodicity and # of Cycles distribution for the two most frequently administered breast cancer plans.
Violin plots showing the distribution of the figures for periodicity and number of cycles for respective patient cohorts receiving one of the two most frequently administered chemotherapy plans for breast cancer.

For the second most common plan "Paclitaxel", however, the periodicity and number of cycles both show a bimodal distribution. The *cycles* plot exhibits a bulge at 4 cycles and another at 12 cycles; and the *periodicity* plot exhibits a bulge near 7 days and another near 14 days. These plots suggest the possibility of two different plans, both containing the single agent Paclitaxel. This information is consistent with two known standard of care protocols for Paclitaxel, one where Paclitaxel is given every two weeks for 4 cycles(39) and the other where it is given weekly for 12 cycles(42).

CHAPTER 6

DISCUSSION

The CPAM takes as input time stamped medication events and produces as output abstract level medication plans using a data-driven approach. This method has clinical as well as research applications. This chapter discusses the contributions of the CPAM to informatics and medicine, and its limitations.

6.1. Informatics contributions

The need for abstraction of temporal data is often encountered in medicine. Even though a wide literature concerning temporal abstraction is cited in various medical domains, processing the temporal data still remains a challenge(43), (19) due to the varied and interacting clinical dimensions involved (e.g., disorders, treatments, disease states). As referenced in this text earlier, there have been several proposals towards the framework of the knowledge-based approach for temporal abstraction and its applications. Though there are fewer references concerning data-driven approach to temporal abstraction, the efforts(36) in that direction are rapidly evolving(43).

6.1.1. Data-driven approach

Data-driven methods have the advantage of not requiring extensive knowledge acquisition and continued maintenance. Data-driven methods rely on and derive their knowledge from the features and content of the data provided as input. The CPAM performs reasonably well without the need for a manually derived knowledge base. Because the method does not rely on any external knowledge base, it is not limited by

any rigid rules pertaining to the specific parameters of the plans it abstracts as to the ranges of variations in periodicity or number of cycles it can detect.

6.1.2. Simplicity of the method

The CPAM is a simple heuristic approach that derives its knowledge base from the input data. The method does not use any complex mathematical algorithms or sophisticated probabilistic analysis; it uses basic grouping, pattern identification and matching processes. Due to the simplicity of the implementation, the method does not have any specific software or hardware constraints. It can be implemented in any environment and expected to perform equally well in terms its outcome. The performance speed of the method is also scalable with the bulk of input data. During this project, the final data bulks given as input to the method were more than 50 times larger than the initial input data and the method performed well without degradation in processing time.

The versatility of the method also lies in its flexibility to accept data from a variety of input sources. Given appropriately formulated pre-processing steps, it can take chemotherapy event data input from any of the clinical sources discussed in chapter 2.

6.1.3. Secondary use of pharmacy data for research

In addition to the development of the CPAM, this work has enabled secondary use of Vanderbilt's Pharmacy transaction database for research. Implementation of computerized pharmacy dispensing record systems is widespread, and the data stored in such systems is in a structured format. Prior to the development of the CPAM the pharmacy dispensing records data were only available to researchers via the enterprise data warehouse (EDW) with all the patient health information (PHI) intact. As part of the CPAM development effort, and to be able to provide sufficient data bulk, the SD

database was expanded to include the pharmacy dispensing records data in a de-identified form. This facilitated the use of pharmacy data for wider use by the Vanderbilt research community.

6.2. Informatics limitations

The CPAM has several limitations including issues related to controlled terminologies, pattern detection, and the lack of a query tool.

6.2.1. Drug terminology limitations

Medications often have several clinically relevant synonyms including the generic name (e.g., Vemurafenib), brand name (e.g., Zelboraf), and names used during drug development (e.g., PLX4032). In addition to variations in drug name, there are several widely used drug terminologies that create unique identifiers including RxNorm and NDC to name a few. Clinical information systems use a variety of drug name and drug coding terminologies along with institution specific variations. For instance, where possible, the VUMC pharmacy information system uses the First Databank (FDB) drug knowledge base as its reference terminology including the generic and brand drug names as well as the NDC codes. However, knowledge bases such as FDB often lag in their representation of investigational agents that have not yet been assigned an NDC. In order to process investigational drugs using the PIS, a custom term is created often with a non-specific NDC and a custom generated name. In the VUMC system for instance, the word 'INVEST' (for investigational) is appended to the end of the drug name. This allows the pharmacy to bill appropriate drug supply for a clinical trial.

The current implementation of the CPAM has several limitations with respect to the pre-processing approach to drug name normalization. First, it does not use a controlled

drug terminology to identify anti-cancer drug names by class. The current implementation leverages an existing manually annotated list of anti-cancer drugs specific to the VUMC pharmacy system. While this manually annotated list is continuously updated for operational purposes, this approach limits the generalizability of the method to other data sources, including pharmacy data sources from other institutions. Second, the method does not use a controlled terminology to identify drug name synonyms. Instead, a manual process was required to map drug name synonyms to a common preferred term. Given that several of the drugs were investigational agents, a combination of the RxNorm and National Cancer Institute Thesaurus (NCIT) terminologies could be leveraged to generalize the approach.

6.2.2. Limitations in pattern detection

The CPAM is based on the assumption that groups of medications are given as repeating events. This is an appropriate assumption for many oncology treatments that often repeat but may be less relevant for other diseases. However, there are some complex cancer treatments that span several consecutive days with multiple drugs that do not repeat, especially in hematologic malignancies. Due to the assumption that all plans must repeat for a patient, the CPAM does not recognize the non-repeating plans correctly. The method could be extended to look for non-repeating patterns across patients.

The CPAM had good performance for both simple and complex plans where all components of the plan recur for each cycle. However, the CPAM had more difficulty for those cases where some component of the plan is missing in one or more cycles. For example, one of the valid chemotherapy plans involves a combination of the drugs “Irinotecan” and “Leucovorin" given every two weeks. In some cases however, the

patient received only “Irinotecan” for some events, thus breaking the pattern of combination drugs. For those patients the method produced a plan sequence consisting of two distinct plans, one each corresponding to “IRIN+LEUC” and “IRIN” (represented by respective abbreviated drug names), instead of the actual plan “IRIN{+LEUC}” (the braces around the LEUC being indicative of the fact that the last component of the plan was missed during the trailing cycles). This happens because, 1) the event corresponding to IRIN by itself repeats, and 2) IRIN by itself is a valid chemotherapy plan.

Finally, the CPAM is limited in its ability to classify distinct plans with the same drugs but different cycle frequency. A good example involves the drug PACLITAXEL where there are two distinct plans that use the single drug, one that is given at the frequency of 7 days and the other at a frequency of 14 days. This is evident in figure 5.2, which shows a bimodal distribution of cycle frequency (periodicity). These two distinct plans could be easily resolved by considering the cycle frequency as part of the plan abstraction method.

6.2.3. Lack of query tool and incremental data analysis

The CPAM does not offer any GUI based query mechanism to efficiently conduct either patient level or cohort level plan analysis. Data abstraction and analysis is done using the SQL through standard query tools and requires adequate technical skills. Development of a query tool would enhance the utility of the method for researchers.

Finally, this initial implementation of the CPAM performs a one-time retrospective data analysis. The method does not allow for continuous updates of the plans as new data emerges in time. The technical implementation of the method would need to be extended to accommodate this requirement.

6.3. Clinical contributions

In addition to being a simple data-driven approach to temporal abstraction the CPAM has already demonstrated applications for clinical research including identification of cohorts by treatment plan and identification of practice patterns at our institution.

The CPAM was used for cohort identification to isolate lung cancer patients who had received two distinct chemotherapy plans as their first treatment. Using the tumor registry database to identify cases of lung cancer, the CPAM was able to accurately segregate two sub-populations, one following the chemotherapy plan "Bevacizumab + Carboplatin + Paclitaxel" and the other following the plan "Carboplatin + Paclitaxel". The two cohorts are being evaluated for the presence of genetic variations that may predict response to Bevacizumab. Prior to the development of the CPAM, it would have taken a significant amount of time, effort and cost to isolate the two patient cohorts through manual chart review of hundreds of lung cancer patients.

The CPAM has been used to evaluate the practice patterns for breast cancer patients at the VICC. The across plan analysis provided listing of the most frequently administered plans for this population (Table 5.3 and Appendix A). This demonstrates the trends in chemotherapy prescribing patterns at our institution over the last 5 years. The list has been used for operational purposes to help prioritize the creation of the order sets for the new chemotherapy CPOE system at the VICC. It could also be used to analyze resource utilization by the pharmacy department to help prioritize purchasing requirements and other resource planning tasks.

The within plan analyses facilitated important information regarding the provider practice patterns, patient tolerance to the treatment, and tumor response to the treatment. Figures 5.1 and 5.2 show the distribution of the periodicity, number of cycles and sequence of therapy for the two most frequently administered breast cancer plans. The analysis shows the degree of concordance with the established standard of care for these protocols. In particular, the most commonly prescribed plan "Cyclophosphamide + Doxorubicin", is typically administered every 14 days for 4 cycles. The data for 198 patients shows an average cycle frequency of 15.8 days with an average of 3.8 cycles. This means that most patients were able to tolerate the treatment on the standard schedule. This has important implications for both, assessing the quality of care and identifying therapies where toxicities often require deviation from the standard schedule.

A within plan analysis can also be used to estimate the time to disease progression (TTP) for a given therapy. TTP is typically measured from the date of the first treatment until the date the tumor is documented to progress, typically measured by imaging studies. The CPAM documents the duration of the treatment, which could be a surrogate for TTP for certain clinical settings. Figure 6.1 shows the distribution plot of the number of cycles for the "Fulvestrant" plan, a medication given only in the metastatic setting for breast cancer. Fulvestrant is typically given every 4 weeks, such that the number of cycles corresponds to the number of months the treatment was given. The distribution plot for this plan shows a median number of cycles to be about 4 and the average number of cycles is 6.4. These numbers have close correspondence to the median TTP of 5.5 months observed by Robertson et al. in a prospective analysis of two multicenter trial of Fulvestrant versus Anastrozole(44). While the patient sample size of this multicenter

study was much larger (n=428), and that of the VICC is much smaller (n=55), the proximity of the TTP value shows the importance of the cohort analysis that can be done with the output produced by the CPAM.

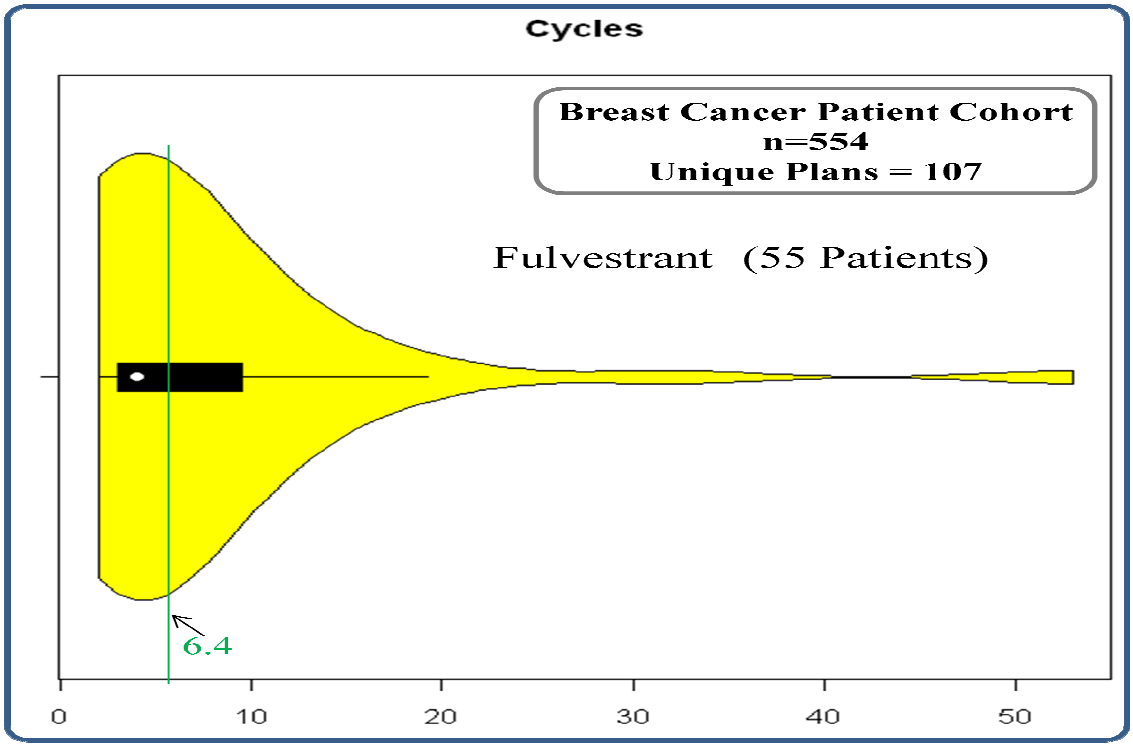


Figure 6.1. # of Cycles distribution for the treatment plan of Fulvestrant.
Violin-plot of the distribution for number of cycles for a cohort of 55 patients on chemotherapy treatment plan of Fulvestrant

Given that the chemotherapy plans can be readily abstracted from a set of medication events, the CPAM can significantly expedite retrospective studies of chemotherapy treatments. Using the CPAM to produce chemotherapy treatment plans can save laborious effort, significant amount of time to collect and analyze data for discrete drug events, and considerable amount of cost to the investigators of such studies.

6.4. Clinical limitations

We have demonstrated some utility of the CPAM for clinical research however, it does have some limitations. First, the method is currently restricted to medications that were administered at the institution's infusion center. In order to provide a complete treatment history, the data sources and the CPAM would need to be extended to include oral and intravenous medications not administered at the VUMC. Second, the current implementation of CPAM does not take into account any dose variations, and it also does not indicate at the output if there were any dose variations. The method only recognizes a complete absence of an administration event. Such enhancements would be valuable both clinically and for research.

CHAPTER 7

CONCLUSION AND FUTURE DIRECTIONS

7.1. Conclusions

The CPAM is a temporal reasoning method that accurately abstracts a sequence of chemotherapy plans at the patient level. The major advantage of the CPAM is its simple data-driven approach to plan abstraction that does not require maintenance of an external knowledge base of plans. The utility of the CPAM is further demonstrated through several cohort plan analyses that provide information on provider practice patterns, plan adherence to standard of care, patient toxicity, and tumor response to treatment.

7.2. Future directions

The CPAM could be further extended in several dimensions for both patient care and clinical research.

7.2.1. Implementation in clinical care systems for patient care

The CPAM could prove to be a useful tool in a clinical setting if incorporated into EHR systems or in the chemotherapy flow sheets. The tool can produce the abstract form of chemotherapy plans instead of the simple chronology of discrete drug events. In a clinical system, the CPAM could decrease the time spent by a clinician to perform chemotherapy plan abstraction. A system could also be developed that utilizes the CPAM to help create treatment summaries in the form suggested by ASCO(3).

7.2.2. Extensions for clinical research

Within oncology, the CPAM could be extended in a number of dimensions to further facilitate clinical research. First, the method could be extended to take medication event input from additional data sources including inpatient and outpatient CPOE, nursing administration records, and clinical notes. It could also be extended to continuously update plans as new drug events occur over time. The addition of a GUI for plan query would also facilitate research utilization of the important abstracted data.

In order to facilitate comparative effectiveness research, additional patient data would need to be integrated with the abstracted treatment plans produced by the CPAM. Preliminary work towards this end has been already demonstrated by the use of diagnosis and histology data from the tumor registry database. Additional tumor registry and clinical data features such as demographics, cancer stage, tumor biomarkers and vital status would also need to be integrated to facilitate CER. In addition, the cost of treatments could be included to perform comparative cost analysis.

7.2.3. Application to other clinical domains

Many chronic diseases such as diabetes, hypertension, cardiovascular disease, and HIV, require ongoing management and assessment by the special providers. The treatments of such diseases are continuous and prolonged, and involve multiple groups of treatments over time. On these premises, it should be possible to extend use of the CPAM to infer treatment plans for such diseases too. Given the fact that the medication administration for most of these diseases is on an outpatient basis, and the medication administration events are relatively unreliable, it would require incorporation of some

stochastic techniques to establish the authenticity of the medication events to accurately abstract corresponding plans.

Healthcare providers routinely perform the task of reviewing a patient's medication history, which can be quite laborious. Given that all the raw data required for such summaries are already available in many EHR systems, incorporation of abstraction tools like CPAM could significantly improve provider workflow. Similarly, secondary use of EHR data does not have to involve manual reviews and creation of ad-hoc processes. Researchers' resources are better spent at thinking in terms of abstract ideas rather than mundane task of sifting through distinct low-level data items. CPAM-like tools could provide such abstract level output for analysis of EHR data. The CPAM has great potential to assist both the clinical care and clinical research.

REFERENCES

1. Kautz H, Allen JF. Generalized plan recognition. 1986;32–37.

2. Kautz HA. A formal theory of plan recognition and its implementation. Reasoning about plans 1991;125:1–67.

3. ASCO Cancer Treatment Summaries | Cancer.Net [Internet]. 2011;[cited 2011 Oct 31] Available from: http://www.cancer.net/patient/Survivorship/ASCO+Cancer+Treatment+Summaries

4. Eng C, Kopetz S. Commentary: Practice Patterns and Potential Impact on Quality Measures for a Practicing Physician. Journal of Oncology Practice 2009;5(5):233 – 235.[cited 2011 Nov 11]

5. Zafar SY, Marcello JE, Wheeler JL, Rowe KL, Morse MA, Herndon JE, Abernethy AP. Longitudinal Patterns of Chemotherapy Use in Metastatic Colorectal Cancer. J Oncol Pract 2009 Sep;5(5):228–233.

6. Warner JL, Anick P, Hong P, Xue N. Natural Language Processing and the Oncologic History: Is There a Match? Journal of Oncology Practice 2011 Jul;7(4):e15 –e19.[cited 2011 Oct 9]

7. Gold S, Elhadad N, Zhu X, Cimino JJ, Hripcsak G. Extracting Structured Medication Event Information from Discharge Summaries. AMIA Annu Symp Proc 2008;2008:237–241.

8. Mork JG, Bodenreider O, Demner-Fushman D, Doğan RI, Lang F-M, Lu Z, Névéol A, Peters L, Shooshan SE, Aronson AR. Extracting Rx information from clinical narrative. J Am Med Inform Assoc 2010 Sep;17(5):536–539.

9. Sklarin NT, Granovsky S, O'Reilly EM, Zelenetz AD. Electronic Chemotherapy Order Entry: A Major Cancer Center's Implementation. Journal of Oncology Practice 2011 Jul;7(4):213 –218.[cited 2011 Nov 17]

10. What Is Comparative Effectiveness Research | AHRQ Effective Health Care Program [Internet]. 2011;[cited 2011 Nov 19] Available from: http://www.effectivehealthcare.ahrq.gov/index.cfm/what-is-comparative-effectiveness-research1/

11. Sox HC. Comparative Effectiveness Research: A Progress Report [Internet]. Annals of Internal Medicine 2010;[cited 2011 Nov 11] Available from: http://www.annals.org/content/early/2010/08/02/0003-4819-153-7-201010050-00269.short

12. SEER Web Site [Internet]. 2011;[cited 2011 Nov 19] Available from: http://seer.cancer.gov/

13. SEER Cancer Statistics [Internet]. 2011;[cited 2011 Nov 19] Available from: http://seer.cancer.gov/statistics/

14. SEER-Medicare Linked Database [Internet]. 2012;[cited 2012 Apr 25] Available from: http://healthservices.cancer.gov/seermedicare/

15. Abernethy AP, Etheredge LM, Ganz PA, Wallace P, German RR, Neti C, Bach PB, Murphy SB. Rapid-Learning System for Cancer Care [Internet]. Journal of Clinical Oncology 2010 Jun;[cited 2011 Oct 14] Available from: http://jco.ascopubs.org/content/early/2010/06/28/JCO.2010.28.5478.abstract

16. Combi C, Shahar Y. Temporal reasoning and temporal data maintenance in medicine: issues and challenges. Comput. Biol. Med. 1997 Sep;27(5):353–368.[cited 2012 Mar 27]

17. Cousins SB, Kahn MG, Frisse ME. The Display and Manipulation of Temporal Information. Proc Annu Symp Comput Appl Med Care 1989 Nov;76–80.

18. Kahn MG, Fagan LM, Sheiner LB. Combining physiologic models and symbolic methods to interpret time-varying patient data. Methods Inf Med 1991 Aug;30(3):167–178.[cited 2012 Mar 27]

19. Augusto JC. Temporal reasoning for decision support in medicine. Artif Intell Med 2005 Jan;33(1):1–24.[cited 2011 Oct 15]

20. Adlassnig K-P, Combi C, Das AK, Keravnou ET, Pozzi G. Temporal representation and reasoning in medicine: Research directions and challenges. Artif Intell Med 2006 Oct;38(2):101–113.[cited 2012 Mar 27]

21. Stacey M, McGregor C. Temporal abstraction in intelligent clinical data analysis: A survey. Artificial Intelligence in Medicine 2007 Jan;39(1):1–24.[cited 2011 Oct 15]

22. Shahar Y, Musen MA. RESUME: a temporal-abstraction system for patient monitoring. Comput. Biomed. Res. 1993 Jun;26(3):255–273.[cited 2012 Apr 10]

23. Shahar Y. A framework for knowledge-based temporal abstraction. Artificial Intelligence 1997 Feb;90(1-2):79–133.[cited 2011 Nov 20]

24. Raj R, O'Connor MJ, Das AK. An ontology-driven method for hierarchical mining of temporal patterns: application to HIV drug resistance research. AMIA Annu Symp Proc 2007;614–619.[cited 2011 Oct 10]

25. Shahar Y, Miksch S, Johnson P. The Asgaard project: a task-specific framework for the application and critiquing of time-oriented clinical guidelines. Artif Intell Med 1998 Oct;14(1-2):29–51.[cited 2011 Oct 15]

26. Shahar Y, Musen MA. Knowledge-based temporal abstraction in clinical domains. Artif Intell Med 1996 Jul;8(3):267–298.[cited 2011 Oct 5]

27. O'Connor MJ, O'connor MJ, Sc M, Grosso WE, Tu SW, Musen MA. RASTA: A Distributed Temporal Abstraction System to facilitate Knowledge-Driven Monitoring of Clinical Databases. In: In Proceedings of MedInfo 2001. IOS Press; 2001 p. submitted.

28. Chakravarty S, Shahar Y. Acquisition and analysis of repeating patterns in time-oriented clinical data. Methods Inf Med 2001;40(5):410–420.[cited 2011 Nov 13]

29. Lin RS, Rhee S-Y, Shafer RW, Das AK. Prediction of HIV Mutation Changes based on Treatment History. AMIA Annu Symp Proc 2006;2006:1011.

30. Bramsen P, Deshpande P, Lee YK, Barzilay R. Finding Temporal Order in Discharge Summaries. AMIA Annu Symp Proc 2006;2006:81–85.

31. Shahar Y, Cheng C. Model-based visualization of temporal abstractions. In: Temporal Representation and Reasoning, 1998. Proceedings. Fifth International Workshop on. 1998 p. 11 –20.

32. Shahar Y, Musen MA. A temporal-abstraction system for patient monitoring. Proc Annu Symp Comput Appl Med Care 1992;121–127.

33. Levy MA, O'Connor MJ, Rubin DL. Semantic reasoning with image annotations for tumor assessment. AMIA Annu Symp Proc 2009;2009:359–363.[cited 2012 Apr 25]

34. Shahar Y, Musen MA. Plan Recognition and Revision in Support of Guideline-Based Care. IN WORKING NOTES OF THE AAAI SPRING SYMPOSIUM ON REPRESENTING MENTAL STATES AND MECHANISMS 1995;118–126.

35. NCCN Clinical Trials at NCCN Member Institutions [Internet]. 2011;[cited 2011 Oct 16] Available from: http://www.nccn.org/clinical_trials/member_institutions.asp

36. Verduijn M, Dagliati A, Sacchi L, Peek N, Bellazzi R, de Jonge E, de Mol B. Comparison of two temporal abstraction procedures: a case study in prediction from monitoring data. AMIA Annu Symp Proc 2005;2005:749–753.

37. Roden D, Pulley J, Basford M, Bernard G, Clayton E, Balser J, Masys D. Development of a Large-Scale De-Identified DNA Biobank to Enable Personalized Medicine. Clin Pharmacol Ther 2008 May;84(3):362–369.[cited 2011 Nov 2]

38. Vanderbilt Institute for Clinical and Translational Research - VICTR [Internet]. 2011;[cited 2011 Aug 31] Available from: http://www.mc.vanderbilt.edu/victr/pub/message.html?message_id=182

39. Citron ML, Berry DA, Cirrincione C, Hudis C, Winer EP, Gradishar WJ, Davidson NE, Martino S, Livingston R, Ingle JN, Perez EA, Carpenter J, Hurd D, Holland JF, Smith BL, Sartor CI, Leung EH, Abrams J, Schilsky RL, Muss HB, Norton L.

Randomized Trial of Dose-Dense Versus Conventionally Scheduled and Sequential Versus Concurrent Combination Chemotherapy as Postoperative Adjuvant Treatment of Node-Positive Primary Breast Cancer: First Report of Intergroup Trial C9741/Cancer and Leukemia Group B Trial 9741. Journal of Clinical Oncology 2003 Apr;21(8):1431 –1439.[cited 2011 Nov 21]

40. Mamounas EP, Bryant J, Lembersky B, Fehrenbacher L, Sedlacek SM, Fisher B, Wickerham DL, Yothers G, Soran A, Wolmark N. Paclitaxel After Doxorubicin Plus Cyclophosphamide As Adjuvant Chemotherapy for Node-Positive Breast Cancer: Results From NSABP B-28. Journal of Clinical Oncology 2005 Jun;23(16):3686 –3696.[cited 2011 Nov 21]

41. Sledge GW, Neuberg D, Bernardo P, Ingle JN, Martino S, Rowinsky EK, Wood WC. Phase III Trial of Doxorubicin, Paclitaxel, and the Combination of Doxorubicin and Paclitaxel as Front-Line Chemotherapy for Metastatic Breast Cancer: An Intergroup Trial (E1193). Journal of Clinical Oncology 2003 Feb;21(4):588 –592.[cited 2011 Nov 21]

42. Sparano JA, Wang M, Martino S, Jones V, Perez EA, Saphner T, Wolff AC, Sledge GW Jr, Wood WC, Davidson NE. Weekly paclitaxel in the adjuvant treatment of breast cancer. N. Engl. J. Med. 2008 Apr;358(16):1663–1671.[cited 2011 Nov 21]

43. Guyet T, Garbay C, Dojat M. Knowledge construction from time series data using a collaborative exploration system. J. of Biomedical Informatics 2007 Dec;40(6):672–687.[cited 2011 Nov 14]

44. Robertson JFR, Osborne CK, Howell A, Jones SE, Mauriac L, Ellis M, Kleeberg UR, Come SE, Vergote I, Gertler S, Buzdar A, Webster A, Morris C. Fulvestrant versus anastrozole for the treatment of advanced breast carcinoma in postmenopausal women. Cancer 2003 Jul;98(2):229–238.[cited 2011 Nov 21]

APPENDIX – A

BREAST CANCER CHEMOTHERAPY PLANS

Table A–1. List of chemotherapy plans for breast cancer patient cohort, in descending order of frequency of administration.

	Counts		Periodicity				Dose Count			
PLAN	Plan Cnt.	Distinct Pt. Cnt.	Avg.	std. dev.	Min.	Max.	Avg.	Std. dev.	Min.	Max.
CYCL+DOXO	201	198	15.8	3.1	7.0	28.0	3.6	0.9	1.0	6.0
PACL	138	138	10.6	4.7	6.4	29.4	7.1	4.3	1.0	16.0
TRAS	72	61	22.1	8.6	7.8	63.0	9.0	11.1	1.0	64.0
FULV	55	55	27.3	16.6	13.0	107.5	6.4	8.5	1.0	53.0
CYCL+DOCE	48	46	21.7	1.8	20.0	29.7	3.7	1.0	1.0	6.0
VINO	47	47	14.6	12.6	6.0	67.0	6.1	8.1	1.0	47.0
GOSE	44	42	42.6	23.2	24.0	119.0	7.7	6.6	1.0	31.0
PACL+TRAS, TRAS	36	36	68.2	65.9	21.0	178.0	1.2	0.5	1.0	3.0
BEVA	33	32	17.2	3.5	13.8	26.3	10.7	16.5	1.0	92.0
DOXO	32	32	14.0	9.5	7.0	43.0	4.4	8.7	1.0	51.0
DOCE	31	31	14.6	4.5	7.0	25.2	3.9	3.7	1.0	21.0
GEMC	30	30	12.9	5.3	7.0	29.8	4.3	5.1	1.0	28.0
BEVA+PACL, PACL	27	25	18.8	5.4	14.0	34.0	5.0	6.4	1.0	33.0
CARB+GEMC	23	23	14.2	3.4	7.0	21.0	5.4	3.7	1.0	15.0
CISP+PACL	16	16	10.9	2.0	7.0	14.5	8.6	7.8	1.0	35.0
TRAS+VINO, VINO	16	14	30.3	14.2	21.0	63.0	3.9	2.8	1.0	10.0
BEVA+CYCL+DOXO	14	14	14.1	0.7	13.0	15.7	3.4	1.2	1.0	4.0
CARB+DOCE+TRAS, TRAS	13	13	28.0	0.0	28.0	28.0	1.2	0.6	1.0	3.0
CISP+PACL, PACL	12	11	85.3	68.2	14.0	150.0	1.3	0.5	1.0	2.0
PACL, BEVA+PACL	10	10	30.0	11.4	14.0	49.0	5.6	6.7	1.0	24.0
CISP, CISP+PACL	9	9					1.0	0.0	1.0	1.0
PACL+TRAS	9	9	19.3	26.1	7.0	77.0	5.1	4.5	1.0	12.0
NAB-	9	9	12.9	7.5	8.8	27.7	4.4	3.4	1.0	9.0
CYCL+FLUO+METH	8	8	22.2	1.8	21.0	25.5	3.6	2.2	1.0	6.0
BEVA+PACL, BEVA	7	7	218.0	0.0	218.0	218.0	1.3	0.8	1.0	3.0
PACL+TRAS, PACL	7	7	17.5	4.9	14.0	21.0	1.9	1.1	1.0	4.0
TRAS, PACL+TRAS	7	7	119.0	120.4	49.0	258.0	3.6	3.1	1.0	8.0
CISP+PACL, CISP	6	6	37.0	0.0	37.0	37.0	1.3	0.8	1.0	3.0
TRAS+VINO	6	6	14.0	0.0	14.0	14.0	1.2	0.4	1.0	2.0
CARB	5	5	8.4	1.3	7.0	9.3	2.4	1.5	1.0	4.0

CARB+DOCE+TRAS	5	5	21.1	0.1	21.0	21.2	2.2	2.2	1.0	6.0
CARB+GEMC, GEMC	5	5	46.3	25.6	21.0	82.0	3.6	4.2	1.0	11.0
TRAS, CARB+DOCE+TRAS	5	5					1.8	0.4	1.0	2.0
IXAB	5	5	7.9	2.4	6.3	11.4	12.6	12.5	1.0	33.0
BEVA+DOCE	4	4	20.8	5.7	14.0	28.0	3.8	2.1	2.0	6.0
CARB+DOCE	4	4	8.4	0.0	8.4	8.4	2.3	2.5	1.0	6.0
BSI-+CARB+GEMC, BSI-	4	4	10.7	9.0	5.0	21.0	6.3	5.1	1.0	12.0
CARB+PACL	4	4	10.2	6.7	6.0	20.3	5.5	5.7	2.0	14.0
DOCE+TRAS	4	4	7.9	1.3	7.0	8.8	2.3	1.9	1.0	5.0
RITU	4	4	7.5	1.0	7.0	9.0	4.0	0.0	4.0	4.0
FULV+GOSE	4	4	30.5	3.5	28.0	33.0	2.0	1.4	1.0	4.0
CYCL+DOXO, TRAS	4	4					1.0	0.0	1.0	1.0
BEVA+PACL+TRAS	4	3	14.0	0.0	14.0	14.0	2.0	2.0	1.0	5.0
BEVA+PACL	4	4					1.0	0.0	1.0	1.0
BEVA+GEMC, GEMC	3	3	24.5	4.9	21.0	28.0	1.7	0.6	1.0	2.0
BEVA+PACL+TRAS, PACL+TRAS	3	2	28.0	0.0	28.0	28.0	1.3	0.6	1.0	2.0
VINO, TRAS+VINO	3	3	574.0	0.0	574.0	574.0	1.3	0.6	1.0	2.0
TRAS, FULV	3	1	23.3	4.0	21.0	28.0	3.3	0.6	3.0	4.0
GOSE+TRAS, TRAS	3	3	35.5	9.2	29.0	42.0	2.3	1.2	1.0	3.0
CYTA	3	3	13.0	0.0	13.0	13.0	1.3	0.6	1.0	2.0
CARB+PACL+TRAS	3	2	7.0	0.0	7.0	7.0	1.7	0.6	1.0	2.0
BEVA+CARB+GEMC	2	1	21.0	0.0	21.0	21.0	1.5	0.7	1.0	2.0
BEVA+GEMC	2	2	14.0	0.0	14.0	14.0	1.5	0.7	1.0	2.0
BSI-+CARB, BSI-	2	2	18.0	4.2	15.0	21.0	3.5	2.1	2.0	5.0
FLUO	2	2					1.0	0.0	1.0	1.0
FULV+GOSE, FULV	2	2	28.0	0.0	28.0	28.0	1.5	0.7	1.0	2.0
GEMC, GEMC+TRAS	2	2	20.0	0.0	20.0	20.0	2.0	1.4	1.0	3.0
VINO, TRAS	2	2	22.0	0.0	22.0	22.0	1.5	0.7	1.0	2.0
VINO, BEVA+VINO	2	2	24.0	5.7	20.0	28.0	3.5	2.1	2.0	5.0
TRAS, PACL	2	2	133.0	0.0	133.0	133.0	2.0	1.4	1.0	3.0
TRAS, CARB+PACL	2	2	20.0	0.0	20.0	20.0	1.5	0.7	1.0	2.0
T-DM	2	2	23.0	0.3	22.8	23.2	21.0	15.6	10.0	32.0
PACL, CARB+PACL	2	2	21.0	0.0	21.0	21.0	2.0	1.4	1.0	3.0
GEMC+TRAS	2	2	14.0	0.0	14.0	14.0	1.5	0.7	1.0	2.0
FLUO+METH	2	2	7.0	0.0	7.0	7.0	1.5	0.7	1.0	2.0
DOCE+TRAS, TRAS	2	2	75.0	0.0	75.0	75.0	4.5	4.9	1.0	8.0
CISP+DOXO	2	2	10.7	0.0	10.7	10.7	2.5	2.1	1.0	4.0
BEVA+VINO	2	2					1.0	0.0	1.0	1.0
BEVA+DOCE, DOCE	2	2	14.0	0.0	14.0	14.0	3.5	3.5	1.0	6.0
ALDE	1	1	1.0	0.0	1.0	1.0	4.0	0.0	4.0	4.0
CARB+PACL,	1	1	14.0	0.0	14.0	14.0	3.0	0.0	3.0	3.0

BEVA+CARB+PACL										
DOCE+NAB-	1	1					1.0	0.0	1.0	1.0
DOCE+GEMC	1	1					1.0	0.0	1.0	1.0
DOCE+DOXO	1	1					1.0	0.0	1.0	1.0
DENO	1	1	39.0	0.0	39.0	39.0	3.0	0.0	3.0	3.0
CYCL+FLUO	1	1	19.0	0.0	19.0	19.0	2.0	0.0	2.0	2.0
CYCL+EPIR	1	1					1.0	0.0	1.0	1.0
CYCL+DOXO+RITU+VINC	1	1	21.0	0.0	21.0	21.0	3.0	0.0	3.0	3.0
CYCL	1	1					1.0	0.0	1.0	1.0
CISP+PACL, PACL, CISP	1	1					1.0	0.0	1.0	1.0
CISP+ETOP, ETOP	1	1	28.0	0.0	28.0	28.0	4.0	0.0	4.0	4.0
CISP	1	1	12.3	0.0	12.3	12.3	8.0	0.0	8.0	8.0
CARB+PACL, CARB	1	1					1.0	0.0	1.0	1.0
VINB	1	1	9.0	0.0	9.0	9.0	2.0	0.0	2.0	2.0
TRAS, GEMC+TRAS	1	1	61.0	0.0	61.0	61.0	2.0	0.0	2.0	2.0
TOPO	1	1	6.7	0.0	6.7	6.7	16.0	0.0	16.0	16.0
NAB-+TRAS	1	1					1.0	0.0	1.0	1.0
MM-1	1	1	7.0	0.0	7.0	7.0	3.0	0.0	3.0	3.0
MITO	1	1	26.0	0.0	26.0	26.0	2.0	0.0	2.0	2.0
METH	1	1					1.0	0.0	1.0	1.0
IXAB+TRAS	1	1					1.0	0.0	1.0	1.0
IRIN	1	1	28.0	0.0	28.0	28.0	2.0	0.0	2.0	2.0
GEMC, DOCE	1	1					1.0	0.0	1.0	1.0
GEMC+TRAS, GEMC	1	1	22.0	0.0	22.0	22.0	5.0	0.0	5.0	5.0
FULV, TRAS	1	1	14.0	0.0	14.0	14.0	2.0	0.0	2.0	2.0
FULV, DENO	1	1	65.0	0.0	65.0	65.0	2.0	0.0	2.0	2.0
FULV+TRAS	1	1					1.0	0.0	1.0	1.0
ERIB	1	1	14.0	0.0	14.0	14.0	4.0	0.0	4.0	4.0
EPIR	1	1	7.0	0.0	7.0	7.0	2.0	0.0	2.0	2.0
DOXO+IFOS	1	1	8.0	0.0	8.0	8.0	14.0	0.0	14.0	14.0
CARB+DOCE, CARB	1	1	35.0	0.0	35.0	35.0	2.0	0.0	2.0	2.0
CARB+DOCE+TRAS, TRAS, CARB+DOCE	1	1	50.0	0.0	50.0	50.0	2.0	0.0	2.0	2.0
BORT	1	1	8.2	0.0	8.2	8.2	13.0	0.0	13.0	13.0
BEVA+TRAS	1	1	21.0	0.0	21.0	21.0	16.0	0.0	16.0	16.0
BEVA+NAB-	1	1					1.0	0.0	1.0	1.0
BEVA+PACL, PACL, BEVA	1	1	28.0	0.0	28.0	28.0	2.0	0.0	2.0	2.0
BEVA+CARB+GEMC, BEVA+GEMC	1	1	80.0	0.0	80.0	80.0	2.0	0.0	2.0	2.0