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

Russ Altman, Principal Advisor
(Section on Medical Informatics)

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

Jeffrey Ullman
(Department of Computer Science)

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

Mark Musen
(Section on Medical Informatics)

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

Arthur Keller
(Department of Computer Science)

Approved for the University Committee on Graduate Studies

Dean of Graduate Studies and Research

Abstract

Decision-support applications in medicine require patient-specific data to provide advice in the context of clinical cases. The emergence and expansion of on-line clinical databases provides the potential for decision-support applications to access patient-specific data automatically. Because the means by which data are represented and retrieved vary widely among clinical databases, decision-support applications that automatically access patient data currently cannot be shared among health-care institutions . The goal of my research has been to overcome the *representational heterogeneity* of clinical databases so as to provide medical decision-support applications with a *uniform interface* to patient-specific data. I developed and implemented a formal methodology, called TransFER, to bridge the representational heterogeneity of relational databases and to increase the portability of medical decision-support applications.

The TransFER methodology supports the definition of a global *reference schema* of clinical information against which applications may formulate requests for patient data. The reference schema is specified using a semantic data model, called FER, that is a hybrid of the functional data model ("F") and the entity-relationship data model ("ER"). The FER data model includes a declarative query language called ReFER. A formal *mapping language* based on the relational algebra allows database administrators to specify the correspondence between the global reference schema and the local relational database at their site. A *translating compiler* at each site uses the encoded mappings to translate automatically queries that are specified against the global model to semantically equivalent queries that can be executed by the local relational database .

I implemented a prototype of this model and conducted an experiment to test the validity and generality of the TransFER methodology. The experiment entailed that three groups design a relational database schema to represent a single clinical data set and that I specify a reference schema for the same data set using the FER data model. My hypothesis was that the TransFER methodology could map the FER reference schema effectively to the three independently designed

relational database schemas and that the TransFER methodology could translate automatically queries formulated against the reference schema to queries that could be executed by all three schemas relational databases. The results of the evaluation indicated that (1) the FER data model and ReFER query language are well-suited for representing clinical database queries, (2) the mapping language is effective at mapping FER schemas to heterogeneous relational database schemas, and (3) the translating compiler is capable of correctly translating ReFER queries to equivalent SQL queries. The primary shortcoming of the TransFER approach is that automatically generated query translations are currently less efficient than manually generated translations, but the evaluation results indicate that formal query-optimization techniques that I have developed hold promise for eliminating this difference.

Acknowledgments

The ideas and work represented in this document would not have come about without the support of many people. My principal advisor, Russ Altman, guided me through the most difficult parts of this undertaking with his tireless intellectual and moral support. Russ's many insights into my work and endless creative suggestions helped me turn a collection of raw ideas into a scholarly document. Russ's spirited encouragement and fine sense of humor made it fun to do so. I am also indebted to Jeff Ullman, who contributed significantly to the quality of the technical content of this dissertation. Professor Ullman patiently assimilated and rigorously examined my ideas to help me place my work on a formal foundation that would have been otherwise unattainable. Mark Musen's valuable perspectives on the fields of medical informatics and artificial intelligence encouraged me to generalize and hone my research, and Arthur Keller's knowledge of database theory and database literature helped me place my work in an appropriate context.

Gio Wiederhold has contributed to my research ideas and my academic development from the time I began graduate school. Gio taught me much of what I know about databases, provided me with generous support, and extended me his friendship. He and Thierry Barsalou stoked my interest in database technology, and their innovative work on object–relational models influenced my own research. I was also fortunate to work with Peter Ludemann, who taught me the “art of Prolog” and saved me countless programming hours.

I am grateful for the superior research environment that was available to me at the Medical Information Sciences Program at Stanford. Through his vision and administrative wizardry, Ted Shortliffe has developed and sustained an academic program that has exceptionally talented people and the highest intellectual standards. I will surely miss the superb state-of-the-art computing environment that Tom Rindfleisch and his staff provided. My fellow students have been a consistent source of creative input and technical assistance, especially those who volunteered their time to design and build the databases that I needed to evaluate my work: Keith Campbell, Bill Detmer, Doug Fridsma, Vatsala Sarathy, and Huma Waheed. I still owe you pizza. Blackford Middleton and Doug Owens were also instrumental in formulating the clinical queries for the evaluation. Darlene Vian has been a dear friend and consistent supporter who has probably helped me in ways that I don't even know. I also appreciate the editing talents of Lyn Dupre and Debbie Fife, who taught me to express my thoughts succinctly and to reunite my infinitives correctly.

Finally, I must thank the special people in my life. I am grateful to my father for teaching me the value of hard work and perseverance. The camaraderie of Matt Daley always made the good times great and the difficult times tolerable. I have also been very lucky to have the affection and companionship of Ulrike; her patience, warmth, and cheerfulness have sustained me much more than she knows. And the greatest debt is to my mother, Eva, who has supported me in many ways, unwaveringly and unconditionally, throughout my graduate education and always. It is to her, with great love, that I dedicate this dissertation.