The RAND Corporation is a nonprofit institution that helps improve policy and decisionmaking through research and analysis.

This electronic document was made available from www.rand.org as a public service of the RAND Corporation.

Skip all front matter: Jump to Page 1

Support RAND

Browse Reports & Bookstore
Make a charitable contribution

For More Information

Visit RAND at www.rand.org
Explore the Pardee RAND Graduate School
View document details

Limited Electronic Distribution Rights
This document and trademark(s) contained herein are protected by law as indicated in a notice appearing later in this work. This electronic representation of RAND intellectual property is provided for non-commercial use only. Unauthorized posting of RAND electronic documents to a non-RAND website is prohibited. RAND electronic documents are protected under copyright law. Permission is required from RAND to reproduce, or reuse in another form, any of our research documents for commercial use. For information on reprint and linking permissions, please see RAND Permissions.
This product is part of the Pardee RAND Graduate School (PRGS) dissertation series. PRGS dissertations are produced by graduate fellows of the Pardee RAND Graduate School, the world’s leading producer of Ph.D.’s in policy analysis. The dissertation has been supervised, reviewed, and approved by the graduate fellow’s faculty committee.
Designing Efficient Systematic Reviews Using Economical Allocation, Creation and Synthesis of Medical Evidence

Mike Scarpati
Designing Efficient Systematic Reviews Using Economical Allocation, Creation and Synthesis of Medical Evidence

Mike Scarpati

This document was submitted as a dissertation in June 2014 in partial fulfillment of the requirements of the doctoral degree in public policy analysis at the Pardee RAND Graduate School. The faculty committee that supervised and approved the dissertation consisted of Siddhartha Dalal (Chair), Kanaka Shety, and Jeffrey Wasserman.
Contents

Abstract 1

Acknowledgments 3

1. Introduction 5
   1.1. Overview .................................... 5
   1.2. Policy Importance ............................... 6
   1.3. The Systematic Review ............................ 9
      1.3.1. How Reviews are Utilized ...................... 11
      1.3.2. Factors Influencing the Quality of Systematic Reviews ...... 13
   1.4. A Brief Economic Model of Systematic Reviews .............. 15
      1.4.1. Single Review Model .......................... 15
      1.4.2. Allocating Resources to Multiple Reviews ............. 17
      1.4.3. Implications .............................. 18
      1.4.4. Limitations .............................. 19
   1.5. Alternatives to Traditional Systematic Reviews .............. 20
   1.6. Initial Recommendations ........................... 21
   1.7. Conclusion ................................... 21

2. Estimating the Value of Systematic Reviews: Osteoporosis Case Study 23
   2.1. Introduction .................................. 23
      2.1.1. Brief Background on Cost Effectiveness Analysis .......... 25
      2.1.2. Other Benefits of Systematic Reviews ................. 27
   2.2. Flexible Osteoporosis Cost and Utility Simulation (FOCUS) ...... 28
      2.2.1. Initialization ................................ 28
      2.2.2. Changes During Model Execution .................. 31
      2.2.3. Transition Probabilities ........................ 31
      2.2.4. Treatment Strategies .......................... 35
      2.2.5. Utility and Costs ........................... 37
      2.2.6. Summary of Architecture and Parameters Used .......... 38
   2.3. Identifying Optimal Treatment Strategies .................. 40
   2.4. Costs and Utility Effects of Individual Fractures .......... 41
   2.5. Combining the Pieces to Calculate the Population Value of Information . 44
      2.5.1. Limitations .............................. 46
   2.6. US Benefits of Computer Aided Screening .................. 48
   2.7. Conclusion ................................... 49
3. **Screening Times in Systematic Reviews**  
   3.1. Introduction .................................. 51  
   3.2. Background .................................. 52  
   3.3. Data ...................................... 53  
   3.4. Descriptive Statistics ............................. 56  
   3.5. Factors Associated with Screening Times .............. 59  
   3.6. Application of Results .................................. 59  
   3.7. Conclusion ................................... 62  

4. **De Novo Reviews**  
   4.1. The Systematic Review Process .......................... 65  
   4.2. Domain-Specific Concerns .......................... 66  
   4.3. Proposed Approach .................................. 67  
      4.3.1. Notation and Background .......................... 67  
      4.3.2. Differential Error Costs .......................... 69  
      4.3.3. Classification as Bipartite Ranking ................. 69  
      4.3.4. Active Learning .................................. 71  
   4.4. Performance Evaluation .................................. 74  
   4.5. Data ...................................... 75  
   4.6. Results .................................... 76  
      4.6.1. OHSU Experiments .................................. 76  
      4.6.2. Low Bone Density Experiments .............. 79  
   4.7. Conclusions .................................. 82  

5. **Updating Existing Reviews**  
   5.1. Introduction .................................. 83  
      5.1.1. Prior Work .................................. 84  
      5.1.2. Different Types of Updates ...................... 84  
   5.2. Surveillance Updates .................................. 85  
      5.2.1. Approach .................................. 85  
      5.2.2. Experiments .................................. 86  
   5.3. Full Updates .................................. 88  
      5.3.1. Approach .................................. 88  
      5.3.2. Experiments .................................. 90  
   5.4. Discussion .................................. 93  

6. **Conclusion**  
   6.1. Implementation Considerations .......................... 95  
      6.1.1. Challenges .................................. 95  
      6.1.2. Opportunities .................................. 97  
   6.2. How This Work Could Affect Policy Development .............. 98  
      6.2.1. Reframing the Goals of Evidence Creation and Synthesis .... 99  
      6.2.2. Organizational Implications .....................101  
   6.3. How Stakeholders Benefit from this Work .......................... 103
6.4. Conclusion ........................................................................................................ 106

Appendix .................................................................................................................. 107

A. Acronym List ....................................................................................................... 109

B. Value of Information and Low Bone Density .................................................. 111
   B.1. Conversion of Maternal Hip Fracture Risk to Parental Hip Fracture Risk  111
   B.2. Initialization Regression Results ................................................................. 112
   B.3. Group Risks from Relative Risks ............................................................... 112
   B.4. Treatment Strategies Implemented ......................................................... 113
   B.5. Meta-Analysis Forest Plots ........................................................................ 113
   B.6. Recent AHRQ EPC Evidence Reports ....................................................... 113

C. Timing Details ..................................................................................................... 117
   C.1. Empirical Distributions of Screening Times ............................................ 117
Abstract

Medical literature and the actions of policymakers have emphasized the importance of evidence-based medicine in recent years, but basing clinical practice on an exploding base of evidence is challenging. Systematic reviews, which are very resource-intensive, are a crucial channel in the pathway from medical literature to clinical practice. This thesis begins by estimating the value of one systematic review, finding that synthesized evidence regarding treatments to prevent osteoporotic fractures generated a net benefit of approximately $450M. Next, the time taken to screen articles in systematic reviews is analyzed, showing that user interface changes can result in significant reductions in resource requirements. Presenting multiple articles on one screen while reviewing titles leads to a seven-fold reduction in time taken per article. Experience and mental state are also related to screening times, with abstracts reviewed at ideal session lengths requiring 33% less time than those at the beginning of a session.

To further increase the speed at which articles can be screened and decrease the cost of preparing systematic reviews, machine learning techniques allow avoidance of up to 80% of articles. When updating an existing review, savings are increased by utilizing the information present in original screening decisions to train the machine learning model. Finally, implementation issues are addressed, paying attention to technical, organizational, and institutional challenges and opportunities.
Acknowledgments

When beginning the process of writing this thesis, I did not realize the amount of work that would be required of those whom I asked for advisement. I truly appreciate the effort of the people who helped me reach this point. The program itself has given me flexibility to study challenging technical problems while remaining connected to practical considerations. Combining the valuable work and the incredible colleagues, PRGS was ideal.

I am deeply indebted to my dissertation chair, Sid Dalal, who has devoted substantial time to developing my research skills and my career—from technical details to a focus on the broader importance of research problems. It is my goal to bring the rigor, energy, and passion that Sid exudes to the challenges that lie ahead. The healthcare knowledge and experience in the policy sphere contributed by the other members of my committee, Kanaka Shetty and Jeffrey Wasserman, made this document far more organized and interesting. I am extremely grateful for the valuable insights and technical expertise of my external reviewer, Lauren Hannah, particularly for pointing out important details and connections to other work.

I further appreciate the support of the researchers at the Southern California Evidence-Based Practice Center at RAND. Without the support, data, and discussions provided, this work would not have been possible. Using real projects allowed me to get a deeper understanding of the requirements and implementation challenges relating to this work. Within the RAND community, I especially appreciate the mentorship of Matt Lewis. I could not have hoped for a better person with whom to start my research career.

Finally, I would like to thank my family. My parents’ encouragement of my curiosity played a big role in my decision to pursue graduate education. Along with my sister, Tracy, their love and support has been unwavering and unconditional. My wife, Lauren, helped in so many ways that listing them be be a disservice, and my dogs Penny and Leo helped keep me sane through the trials of these past few years.
1. Introduction

1.1. Overview

Evidence-based medicine (EBM), defined as “the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients”[SRG+96] has gained momentum rapidly in recent years. While there is some debate about the impact of the movement toward EBM[CSH04], its importance has been displayed in both the academic and clinical literature as well as in the actions of policymakers. As shown in Figure 1.1, which is derived from the National Library of Medicine’s MEDLINE service, between 2000 and 2005 the number of publications per year containing the term “evidence based” doubled, before doubling again over the next five years. Using bibliometrics in this way is a common approach to measure the scientific interest, and has been applied to many fields[TG13, RGABP13, RPMG08, KU05] Likewise, the American Recovery and Reinvestment Act authorized $1.1B to support comparative effectiveness research[Con09], a key component of EBM. Combining this with the observation that both clinicians and patients are becoming more interested in treatments supported by evidence, it is clear that EBM is valued by a wide variety of audiences.

For clinical decisions to be influenced by the best medical evidence, there must be a clear pathway of information from the medical literature to clinicians. While a more detailed description of this pathway is presented in Section 1.3.1, systematic reviews are

![Growth in Production of Evidence Based vs. All Medical Literature](image)

**Figure 1.1.** Explosive growth in evidence-based literature. Lines depict the number of articles in MEDLINE published each year, with the blue being all articles, and the red being articles matching the query “evidence based.” Both series are normalized to unity in the year 2000[Cor04].
Chapter 1

Introduction

a crucial piece\cite{MCD97}. In a systematic review, a team of medical experts undertakes a thorough analysis of the evidence available to answer a set of specific questions. In 2005, a systematic review such as this was estimated to cost approximately \$250K\cite{Fox05}, with increased literature bases and inflation making the current cost even higher. Given the rapid expansion of medical literature, the current practice is at risk of becoming unsustainable. Furthermore, these reviews are traditionally funded by the Agency for Healthcare Research and Quality (AHRQ), whose budget is severely threatened; in late 2012, a Congressional subcommittee recommended dissolution of the Agency and removing all discretionary spending for “patient-centered outcomes research”\cite{M12}. However, technological advances made in the private sector, academia, and organizations like RAND may be able to considerably reduce the cost of conducting high-quality systematic reviews. This dissertation will advance these tools, place them in the broader context of evidence-based medicine, examine issues surrounding their implementation, and discuss their importance to the broader health system. As such, hope for a sustainable information pathway lies with technological advances and updated institutions that provide a way for a high-quality research to reach clinicians quickly at a reasonable cost.

1.2. Policy Importance

The amount of medical information in existence far exceeds the memory limits of a single person, and new information is being created at an ever-increasing rate. Doctors specialize in large part because it is impossible to learn the best practices in all aspects of medicine, particularly when those best practices are frequently evolving.

This thesis discusses one product that aims to reduce the amount of raw information that an individual must digest in order to understand the state of medical evidence. With no knowledge of the health system, consider the case of a general practitioner, Dr. Smith, with a new patient. Our illustration is not meant to reflect a perfect model of realistic behavior, but to demonstrate the difficulty of utilizing current medical evidence in care decisions. This physician’s training is sufficient to realize that the patient may be suffering from Attention Deficit Hyperactivity Disorder (ADHD). Perhaps he has an established diagnosis pattern that has been successful in the past, and will use that pattern with his new patient. What if Dr. Smith realizes that he has not diagnosed or treated ADHD in a number of years, and desires to ensure that his old pattern is still a best practice?

One way to make this assessment is by searching MEDLINE, the largest database of medical literature. Unfortunately, this immediately presents problems. MEDLINE is designed for skilled users, and it is difficult to create searches that have capture all relevant articles without including many irrelevant articles. Even for expert researchers, it is difficult to capture the relevant articles without including 10 times as many irrelevant articles (see 4.2 for details). If Dr. Smith is able to create a search that performs well, he is faced with the task of reviewing the potentially relevant articles to return only those that provide concrete information about the diagnosis and treatment of ADHD. After
scanning perhaps hundreds of abstracts, he is left with dozens of articles that, more likely than not, will need to be purchased. Hundreds of dollars and many hours later, he has reviewed the articles, and found that each study used different subjects, different treatments, different durations, and different outcomes of interest. He makes a table to keep track of the relevant details and outcomes, eventually realizing that the evidence shows that multiple treatments are effective, but relative performance is unclear.

In the best-case scenario for this process, our physician has spent one day of time for the needs of a single patient—clearly this is unsustainable. A more likely case is that he either finds one article quickly and uses only its data, or perhaps he realizes the magnitude of this task and reverts to his old treatment pattern. With thousands of conditions and thousands of treatments, there is no way that a physician could undertake this sort of research regularly.

Hope for evidence-based medicine comes from the fact that many healthcare providers share the same challenges. Particularly in cases of common conditions like ADHD, a large number of providers must diagnose and treat the condition. If a small group of researchers undertake the process outlined for our hypothetical physician, the results could be shared with a broad audience. Because it shares properties with public goods (detailed in Section 1.4) and publications generate positive externalities, the private market under-provide these services.

Systematic reviews involve a small group of researchers synthesizing the important information related to a set of clinical questions. This essentially amounts to “sharing the load,” so that the same work is not repeated by hundreds of individual researchers or providers. However, the current policy, where the government funds nearly all systematic reviews, is not perfect. Because systematic reviews seek to answer a set of questions for a broad audience, they are often very long—recent reports from the AHRQ EPC program have exceeded 300 pages[TRF12, CND12]. Even if the systematic review were perfectly updated and answered the precise questions of the clinician, it would be difficult to read the entire review for one patient. Executive summaries and abstracts are helpful, but sometimes do not provide sufficient detail for the specific case at hand.

Even if all relevant information about the efficacy and adverse events of a given treatment are presented in a logical, concise manner, the provider’s task is not complete. In many cases, significant uncertainty exists, and costs of different treatments vary wildly. In the case of fracture prevention, which will be investigated in this thesis, multiple treatments are effective[CND12]. According to recent analysis, teriparatide has the highest probability of being the most effective treatment, but costs approximately 100 times the cost of generic alendronate[LMN06, NRG11], a drug that is also effective at reducing fractures. If it were possible to obtain error-free estimates of average fracture reductions for each potential treatment, the caregiver must still weigh the cost of treatment against the benefits, which are not readily in comparable units. How should the avoidance of one hip fracture be valued? The fracture will incur acute medical expenditures, reduce life expectancy, and be very painful for the patient. A provider may know these effects, but almost certainly will not possess reasonable, concrete valuations of each.
One policy option at this stage is to choose the most effective treatment, as some recommend against cost considerations in treatment decisions. When the Patient Protection and Affordable Care Act created the Patient-Centered Outcomes Research Institute, the new organization was forbidden from any measure that “discounts the value of a life because of an individual’s disability” for making treatment decisions or recommendations\[NW10\]. Unfortunately, the reality of resource constraints imply that this will lead to decisions that are suboptimal by most assessments. If Medicare reimburses providers for treatments with high costs but low incremental effectiveness, its budget will be depleted before funds can be spent on more cost-effective treatments\[WSG+96\].

A preferable option involves the use of cost-effectiveness models to compare possible treatment strategies. While it would be unwise to pursue one treatment approach for all patients because a model deemed it optimal, the standardization provided by these models allows for better informed decisionmaking. These models yield estimates of costs and health outcomes for the treatments studied (a more detailed overview is presented in Chapter 2). Tasks involving the comparison of various treatments are significantly simplified with these analyses. However, cost-effectiveness models are time-consuming to create, and benefit from the existence of systematic reviews. Because the assumptions about the effects (both beneficial and adverse) are primary drivers of the analysis, obtaining robust estimates in cost-effectiveness analysis involves research very similar to a systematic review.

Results of cost-effectiveness analysis depend critically on a wide range of assumptions, and reasonable values for these assumptions are likely to vary between providers. For example, nursing home stays are much more expensive in Alaska ($271K per year) than Texas ($47.5K per year)\[Gen13\], thus interventions that avoid time in nursing homes are more valuable in Alaska. Furthermore, the aim of these analyses is not always to directly lead to practice decisions. The most common goal is to answer the question “Is treatment X cost effective for condition Y?” Numerous analyses\[TMDH+08, SWP14, TBML08, PKY+13, PDW+11, MKR+07, LMN+06, SJD+05, HEB+09\] have answered that question for a number of drugs aimed at preventing fractures in patients with osteoporosis. Unfortunately, the question of which patients should be given treatment is addressed less frequently\[TMDH+08, NRG11\].

In order to assist clinicians, organizations ranging from providers like Kaiser Permanente \[DRKR+12\] to interest groups like the National Osteoporosis Foundation\[NOF13\] to hospitals\[GSD+07\] release “clinical practice guidelines,” which turn medical evidence into more concrete recommendations regarding appropriate treatment. These guidelines can be created for a national audience\[Age05\] or be developed for a single provider organization \[DRKR+12\]. While there has been mixed evidence regarding the efficacy of these guidelines \[MCNa99\], they are a reasonable approach to enabling physician behavior to reflect medical evidence.

The primary problem with reliance on clinical practice guidelines is the amount of effort required to reach the point of usability. They are most efficient when they utilize
systematic reviews and cost-effectiveness models, which in turn rely on publication of primary evidence. Systematic reviews can take multiple years (see Section B.6 in the Appendix), and require considerable resource expenditures[Fox05, DRKR012]. Cost-effectiveness models are also time-consuming to create and publish, and the primary research requires time to write and publish. In all, the process of turning original research into clinical practice guidelines can take approximately four years. Given that research is continually produced, this long cycle means that valuable research will have been published but not incorporated into the guidelines.

This thesis takes steps to reduce the long time lags between primary research and implementation, often referred to as “bench to bedside.” There are certainly cases of healthcare providers reading primary research and updating practice trends as a result, but this is rare enough that it is infeasible to develop policies around it. The aforementioned process can be effective, but its length reduces potential societal benefits. While the proposed tool takes a small step towards increasing throughput of the synthesis pipeline, they also outline important aspects of the problem that will make future research more practical.

1.3. The Systematic Review

With massive growth in the already broad medical literature, it is nearly impossible for a clinician to find the best available evidence in response to patient’s needs using standard information retrieval tools. Individual clinicians generally spend less than two minutes researching their clinical questions[SEY06], which is hardly enough time to process information even if the perfect article is presented immediately. Specialists who seek to provide guidelines that mesh with the best medical evidence would have a significant challenge time identifying all of the important primary research. Systematic reviews provide synthesis by aggregating and critically evaluating the evidence that answers a set of specific medical questions. Best practices have been established to increase the likelihood that these reviews are done properly and are clinically useful[eaF10, The11]. While many organizations fund systematic reviews[HSBM12], this document will focus primarily on AHRQ reviews conducted by the Evidence-Based Practice Centers(EPCs). Even within EPCs, there are variations in the exact steps, but the following procedure outlined by AHRQ is an example of a respected approach.

Beginning the process outlined in Figure 1.2, AHRQ asks medical experts to provide input into important questions that can be answered by a systematic review; these

![Figure 1.2: Overview of the systematic review process. Article screening is highlighted, because it is central to this thesis.](image)
suggestions are then prioritized according to the criteria outlined in [Age08]. When the group has decided to move forward with a specific review topic, that topic is refined. A “topic refinement team,” composed of experts with a variety of skills and knowledge related to systematic reviews and the topic in question, leads this section. At least one of the members of the topic refinement team will also belong to the group that will conduct the review [BAB+13]. In this phase, researchers ensure that conducting a review will be beneficial and that the goals of the review are specific enough for the research team to attain. This centers on identification and clarification of a set of “key questions” that the review seeks to answer. These key questions are initially presented in topic nomination, and refined by the topic refinement team and the EPC conducting the review [Age08]. The team may decide that the review is no longer warranted at any point during topic refinement. For example, when a team discovered that a very similar review was being conducted by another group, the team in question determined that the current topic no longer required a systematic review [BAB+13].

Once the topic is refined, the EPC conducts the review, which begins by ensuring that the key questions are actionable. The researchers proceed to develop the criteria that will be used to include or exclude individual studies. These are generally specific to each of the key questions, and specify the acceptable populations, interventions, outcomes, and any other restrictions imposed. A number of analytic frameworks provide consistent structure; one popular framework analyzes populations, interventions, comparators, outcomes, timings, and settings (PICOTS) [TRF+12]. By enumerating the inclusion criteria in terms of PICOTS, analysis is simplified and standardized across reviews, and the research is made more transparent. Next, researchers develop a search strategy, which is a formal specification of all data sources and queries used. Searches generally incorporate multiple scholarly databases, and involve the creation of complex queries [Age08]. It is not uncommon for these queries to span multiple pages in the final report. Because the time cost of reviewing an additional title and abstract is generally small compared to the cost of missing an important study, recall is emphasized over precision\(^1\); literature searches in recent AHRQ comparative effectiveness reports returned an average of 9,000 articles (see Appendix B.6).

After researchers define the inclusion criteria and obtained the search results, they manually screen those results to determine which articles may answer the key questions. The specific process varies at each organization, but often uses a tiered approach. When the list of potential studies is large, researchers quickly scan the titles. Articles containing titles that may be relevant pass to the next stage, in which screeners read the abstracts to answer slightly more detailed questions. If an article appears to meet the inclusion criteria given the information in its abstract, the full text is retrieved and reviewed. In order to ensure that important studies are not omitted, two reviewers independently assess each article at each stage; disagreements are discussed and resolved by consensus. With the relevant articles selected, researchers assessment the quality of each article;

\(^1\)Recall is the fraction of relevant articles returned, and precision is the fraction of returned articles that are relevant.
1.3 The Systematic Review

these results enable studies that are below the quality threshold to be included as part of a sensitivity analysis[BSS04]. Often concurrently, the critical data are abstracted from the articles and compiled into a standard format, so that methods and results can be rigorously analyzed. In the analysis phase, the estimates from each study are laid out in an evidence table and accompanied by an evidence synthesis that may be qualitative, quantitative, or both. A qualitative assessment presents relevant results along with assessments of the research. When quantitative synthesis are feasible, the estimates from relevant studies are combined in a statistically sound way—estimates with high precision carry more weight than those with high variance.

Quantitative synthesis is complicated by the fact only similar studies can be compared directly, and it is very difficult to obtain a rigorous definition of similarity [eaF10]. While qualitative synthesis is less concise, it allows readers to combine the data as they deem appropriate, and potentially re-weight estimates to match the conditions of a specific practice. In the final stage of drafting a report, the findings are interpreted and made clinically relevant.

After the report is drafted, it undergoes peer review. Experts with knowledge of the relevant literature and those with methodological expertise ensure that the analysis captured all relevant studies and results were analyzed properly. Though both internal and external experts have been involved throughout the process, the peer review provides feedback on the completed product. After passing review, results are ideally shared with clinicians and incorporated into practice guidelines[Age08].

1.3.1. How Reviews are Utilized

Systematic reviews distill the relevant literature to arrive at answers to a set of questions. Depending on the state of the literature, the answers may or may not be conclusive. In order to be useful, they must be accessible, accurate, and clinically relevant. Though the flow of information from laboratories to clinical practice is long and complex, we outline the information pathway and evidence regarding its effectiveness.

After a review (or any medical evidence) is published, it is available to practitioners—but it is unlikely that many will perform the research required to stay up-to-date on all relevant publications. While a good systematic review distills the best clinical evidence, it is structured as a research report. In order to make the results reflect a clinician’s decision process, practice guidelines are “developed to improve the process of health care and health outcomes, decrease practice variation, and optimize resource allocation” [CGEW97]. These guidelines are “systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances” [FK90]. Guideline creators utilize medical evidence provided by systematic reviews, practitioner experience, and expert opinion[LPD+04] to yield robust advice for practitioners.

A graphical summary of the flow of information from medical evidence to clinical outcomes is presented in Figure 1.3. In the first stage, evidence is distilled into systematic
reviews. Practice guidelines and performance measures are then created by combining the review with experience and expert opinion \cite{LPD04}; these hope to improve physician decision-making and lead to improved health outcomes.

It is important to note that while the diagram shows a linear flow, the real process involves many feedback loops. Articles cited in this thesis show that systematic reviews can be conducted not only on clinical trials, but also on practice guidelines and physician behavior. Systematic reviews frequently cite other systematic reviews, enabling reuse of the synthesized information. Furthermore, while the path shown is emphasized, every box to the left of physician behavior may directly influence practitioners.

At each step, there is a risk that the information present upstream is not appropriately transferred to the subsequent stages that require it. Some trials may be missed in the screening phase of a systematic review, practitioner experience may not generalize to a wider population, or physicians may fail to implement practice guidelines. The goal of evidence-based medicine is to facilitate this flow of information to enable medical evidence to improve outcomes.

While the flow of information may break down at any stage, this work is primarily interested in those following the systematic review. If the review is created but not used, the resources used to create it could certainly have been used better elsewhere. Practice guidelines are created to improve the flow of best practices to physicians, but clinicians do not always adhere to these recommendations. Indeed, guidelines have not considerably changed clinician behavior for a number of reasons. In a 1999 review investigating 46 studies on the implementation of clinical practice guidelines, 36 analyzed studies found that at least 10% of physicians surveyed were unaware that the guideline in question even existed. Other barriers to adoption included lack of familiarity, agreement, and self-efficacy. The habits of previous practice and external barriers like inconvenient guidelines and environmental factors also contributed to lack of adherence \cite{MCNa99}. Unfortunately, the plethora of explanations illustrates that there is no easy solution to this issue. One potential route towards increased adherence is the creation of evidence-based performance measures. This has been widely used, with sepsis
change bundles being one specific example [LPD+04]. Once these measures are created, it becomes possible to increase the pressure to adhere through public reporting [Jun10] or pay-for-performance. While the effect of performance payments have been mixed [EESS13], they remain a potential tool for modifying physician behavior.

Finally, it is valuable to understand the relationships between creation of these evidence-based practice guidelines and clinical outcomes. The Netherlands began a concentrated effort to bring medical evidence into practice guidelines beginning in the 1990s, and a 2009 systematic review by Lugtenberg, Burgers, and Westert analyzed the effects of this effort. They found that a vast majority of studies (17 of 19) demonstrated significant improvement to the “process and structure of care.” Nine studies measured the effects on patient outcomes, and six of these showed small improvements in this area [LBW09]. In a separate review, So and Wright found that the literature generally finds no or mixed effects when examining the hypothesis that practice guidelines improve quality of care [SW12]. Many studies showed no significant improvement in outcomes, and twelve studies found no positive effect of guidelines on outcomes, while another twelve found positive results for some outcomes and negative or zero effects for others [SW12]. Three systematic reviews, including the aforementioned [LBW09], found small effects with large variance. Explicit guidelines were more effective, but improvements were larger in process-related measures than outcomes [SW12].

1.3.2. Factors Influencing the Quality of Systematic Reviews

A traditional systematic review usually takes at least one year, with many taking longer; in some, the time between ending the literature search and report publication is nearly two years (see Appendix B.6). A 2010 study performed by researchers at McMaster University conducted a systematic review of medical literature exploring the range of methodological approaches from “rapid reviews” to traditional systematic reviews [GCT10]. A significant portion of this review investigates the effect of various speed improvements on bias of the resulting analysis. Table 1.1 briefly outlines the proposed procedural modifications and the effect each may have on the evidence synthesis. The majority of these strategies appear to have a relatively small impact on review quality, and biases are often in differing directions.

In addition to these items, which are explicit yes/no decisions, reviewers make a number of choices that are less easily measured, such as the creation of queries for literature databases. Limiting retrieval to documents containing only words that are obviously desirable allows for rapid creation of the search strategy and reduced screening requirements, but sacrifices recall. A broader search will return more documents and increase recall, but requires additional screening time. Even a task that may appear simple—searching MEDLINE for systematic reviews—is difficult to do at high precision. In 2005, the most sensitive strategy achieved sensitivity of 99% and specificity of 52% [MWMH05], indicating that a requirement to include systematic reviews could nearly double the number of articles retrieved and screened.
<table>
<thead>
<tr>
<th>Suggestion</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excluding Grey Literature</td>
<td>Increase publication bias; treatment effects more optimistic [HMCE07, MPTM00, HCLS07]</td>
</tr>
<tr>
<td>Excluding Small Studies</td>
<td>Decrease treatment effects; bias due to lower quality [SGE00]</td>
</tr>
<tr>
<td>Limit to English</td>
<td>More conservative treatment effect or no change [JHS+02, MPK+00, MPLK03]</td>
</tr>
<tr>
<td>Only Searching MEDLINE</td>
<td>Optimistic treatment effect, but bias likely small [SBM+03]</td>
</tr>
<tr>
<td>Only Single Screener</td>
<td>9% missed vs. double screening, on average [ECD+02]</td>
</tr>
</tbody>
</table>

Table 1.1.: Effect of various time saving strategies on review quality

1.3.2.1. Timeliness

In 2007, Shojania et al. analyzed a quasi-random sample of 100 systematic reviews published between 1995 and 2005. Defining important changes as those that change statistical significance or the estimated effect sizes by at least 50%, they found that the median time period between publication and being outdated to be 5.5 years. For 7% of the reviews studied, the conclusions were out of date by the time the report was published, and this increased to 23% within two years of publication [SSA+07].

Given that the rate at which trials and systematic reviews are published has approximately doubled since 2007, this is a troubling issue. The fact that a high-quality systematic review exists may make practitioners confident that the best evidence is shown in that report, when it may very well be outdated. Out of date reports are particularly problematic because they are not marked according to their timeliness. Dates can be easily understood, but what is a clinician to make of a report that is two years old? Its conclusions could be absolutely current, or they could be outdated. Chapter 5 discusses this problem, along with a potential solution, in increased detail.

1.3.2.2. Strength of Evidence

In addition to answering the key questions, systematic reviews assess the strength of evidence supporting various conclusions. The EPCs have a standardized grading system for medical evidence, with possible grades of high, medium, low, and insufficient. The first three refer to the researchers’ “confidence that the evidence reflects the true effect.” A grade of insufficient is assigned when no conclusion can be drawn from the available evidence. These grades are also tied to the stability of conclusions. If the evidence is of high strength, it is unlikely that future research will change the conclusions [OLA+10].

While strength of evidence assessment does not directly affect the methodological quality of a review, like timeliness it is related to the utility of the completed report. If
little evidence is available, or the evidence contradicts itself, a report will be unable to draw strong conclusions. Conversely, if the evidence is already very strong, conducting a systematic review may not provide much additional value over what is currently known. Assessing evidence strength can provide significant utility to organizations that fund and prioritize research.

1.4. A Brief Economic Model of Systematic Reviews

An economic model of the value created by a systematic review motivates the path taken in this thesis. While a complete model would be tremendously complex, useful decision aids can be obtained from a heavily simplified structure that captures the most salient features of the process. Borrowing the concept of the value of information from decision theory, a systematic review provides value when a practitioner either chooses a course of action more quickly or changes a clinical decision as a (possibly indirect) result of the review, and that choice changes either cost or health outcomes.

1.4.1. Single Review Model

Because the audience of a review is diverse, and the contents are non-rival and often non-excludable\(^2\), reviews can be viewed as public goods. While it is possible to prohibit a person from reading the contents by requiring payment for the article, a significant number, including all outputs of the AHRQ EPC program, are released publicly. Articles not freely available nearly always provide a public abstract that reports the most important findings.

Public goods have a well-known problem in economics, namely that society is better off if the good is provided, but it is in no individual’s best interest to purchase it individually[MCWG95]. In the case of systematic reviews, this problem is most commonly solved by government funding of review creation, though in some cases private donors or patient groups will support the synthesis[BMM\^07, LRM\^12]. This is important to the forthcoming explanation, as the mathematical framework used to analyze the provisioning of public goods differs from that of private goods. It also permits determination of conditions that must be satisfied if reviews are to be conducted in a socially optimal manner.

In order to simplify the model, consider the case where the goal of a review is to determine if treatment \(A\) or \(B\) should be utilized. Without loss of generality, the superior treatment is \(A\), but this may or may not be known to practitioners. By synthesizing information, the review should increase the probability that treatment \(A\) is chosen. However, not all treatments or systematic reviews are created equally. Let the value of choosing treatment

\(^2\)Non-rival goods are those where the consumption of the good by one individual does not impede consumption by another. Non-excludable goods are those from which individual consumption cannot be prevented.
Figure 1.4.: Relationship between quality $q(e)$ and effort $e$ on a single review

For each patient, there is a probability that the superior treatment will be chosen, which is denoted $P(A|q(e))$. The function $q(e)$ describes the quality of the review, and is itself a function of effort $e$. The functional form of $q(e)$ reflects methodological approach, including the factors specified in Table 1.1. While we do not attempt to specify the form of $q(e)$ explicitly, it will possess certain properties. Because a bad review will cause more harm than good, $P(A|q(0)) > P(A|q(e))$ for some small effort $e > 0$. Effort increases the quality of the review, thus $\frac{\partial q(e)}{\partial e} \geq 0$, but, diminishing marginal returns to effort imply $\frac{\partial^2 q(e)}{\partial e^2} \leq 0$. A notional function with these properties is shown in Figure 1.4. The cost of using the given level of effort and approach is $c(e)$, which we assume is increasing and convex $(\frac{\partial c(e)}{\partial e} \geq 0$ and $\frac{\partial^2 c(e)}{\partial e^2} \geq 0$).

Given this setup, we can describe the optimization problem undertaken by a welfare-maximizing social planner when deciding the appropriate level of effort to allocate to a single systematic review:

$$\max_{e} N \times [V(A) - V(B)] \times P(A|q(e)) - c(e)$$

In this problem, $N$ is the number of times physicians must decide on treatment for the condition in question. If effort is greater than zero ($e^* > 0$), a unique global optimum exists because $P(A|q(e))$ is concave in $e$ and we assume that $c(e)$ is convex. This solution
occurs when the following condition is satisfied:

\[ N \times [V(A) - V(B)] \times \frac{\partial P(A|q(e^*))}{\partial q(e^*)} \frac{\partial q(e^*)}{\partial e} = \frac{\partial c(e^*)}{\partial e} \]

Like many results in economics, this can be expressed as the point where the marginal benefit is equal to the marginal cost. On the left-hand side, the marginal benefit of effort is the number of cases multiplied by the incremental treatment effect of \( A \) over \( B \), multiplied by the rate at which increases in effort increase the probability of the correct treatment. On the simpler right-hand side, the marginal cost is the rate at which the cost of the review changes with effort.

### 1.4.2. Allocating Resources to Multiple Reviews

Agencies funding systematic reviews have a larger problem than deciding on the funding level of a single review—they must also determine which reviews should be conducted, and which reviews should be updated at what time. Our model can be extended in a straightforward manner to incorporate this reality. Letting \( T \) be the number of topics that could potentially be reviewed, and \( B \) the agency’s budget for systematic reviews, the new optimization problem is

\[
\max_{e_1, \ldots, e_T} T \sum_{i=1}^{T} N_i \times [V(A_i) - V(B_i)] \times P(A_i|q(e_i))
\]

s.t. \( \sum_{i=1}^{T} c(e_i) \leq B \)

Because we can no longer make the assumption that \( e_i > 0 \), this is no longer a convex optimization problem. While the interpretation is not as clean as in the prior case, useful observations can be made, particularly if the cost of individual reviews that are conducted is small compared to \( B \). In that case, each individual review will be budgeted at the level that solves the single-review optimization problem, and the problem will simplify to

\[
\max_{r_1, \ldots, r_T} T \sum_{i=1}^{T} N_i \times [V(A_i) - V(B_i)] \times P(A_i|q(e^*_i))
\]

s.t. \( \sum_{i=1}^{T} r_i \times c(e^*_i) \leq B \)

where \( r_i \) is an indicator function taking the value of one if review \( i \) is conducted, and zero otherwise. While this may not appear to be more simple, it is a version of the well-known knapsack problem. Each review has a value and cost when undertaken at the optimal level of quality, and the problem is to choose the set of reviews that maximize benefit without exceeding the total budget. While it is not optimal, there exists a simple
strategy that achieves a total surplus not worse than half of the optimal surplus. It involves first conducting the review with the largest surplus, followed by expending the remainder of the budget by selecting those reviews with the highest expected surplus per dollar\cite{Lag96}.

### 1.4.3. Implications

Though this representation omits a great deal of complexity in the medical decision-making process, we believe that it makes evident a number of useful observations. First, the same level of quality is unlikely to be optimal for all reviews. This implies that utilizing the same methodology, such as universally requiring duplicate screening, is not wise in all cases. While tailored approaches depart from a “best practice sharing” paradigm where all reviews follow the same process, it does not imply that standardization is undesirable. The aforementioned optimization problem is impossible to carry out exactly in practice, but structured processes can be used as optimization heuristics. The prioritization and topic refinement process attempts to intelligently allocate resources between reviews; independent work by Hoomans derived a similar model for this express purpose\cite{HSBM12}.

It follows that there is no value to a review that does not change the behavior of decision-makers. This claim alone may be controversial, as it would, at least retrospectively, place a very low value on inconclusive reviews. However, it encourages the evidence-based medicine community to focus on ensuring a rapid flow of information from “bench to bedside.” It may also aid decision-makers in balancing the allocation of resources between conducting trials and reviews with downstream efforts, such as refinement of practice guidelines.

This model also assists in the decision between updating existing reviews and conducting additional \textit{de novo} reviews. The cost required to update a review to a given level of quality is considerably lower than creating a new review of a same quality. This is due to the fact that the questions and search strategy have been decided, researchers have experience reviewing the articles, statistical analysis code is written, and the literature has already been analyzed through a certain point in time. Furthermore, conducting the original systematic review may provide researchers with a better understanding of the pace at which new literature is released–informing expert judgment of the the likelihood that the conclusions are out of date.

Finally, the model identifies areas where new reviews are most likely to be useful. The ideal topic investigates treatments for a common condition, where the treatment effect is large and new evidence synthesis is most likely to change practice. Clearly, common conditions tend to be heavily researched, and large treatment effects are easier to identify, so it is unlikely that any topic will be highly rated in all areas. Nevertheless, it provides a framework to assist in prioritization, and to guide resource allocations between topics.
1.4.4. Limitations

Resulting from its simplicity, this model has a number of limitations. First, it does not explicitly consider any concept of time, whereas in reality an excellent review 20 years ago is less useful than a recent, average-quality review. However, this can be incorporated by either specifying a decay in review quality over time or by recognizing that the state of the current literature must be included in \( P(A|q(e)) \). If a high-quality review was recently completed, \( P(A|q(e)) - P(A|q(0)) \) will be small even at high levels of effort.

Temporal dynamics are also at play in the adoption of review conclusions. If physicians only follow practice guidelines, the review yields no value until its conclusions are put into those guidelines.

A further temporal complication is that the cost of conducting a review of a given quality can change in undetermined ways if the review is postponed. Consider a decision at time \( t_0 \) to postpone a review for a given topic. At time \( t_1 > t_0 \), the funding agency decides to review this topic. If all of the evidence favors the same treatment, the cost may go down because a rapid review will capture sufficient relevant information. Conversely, if a study is released that contradicts the prevailing wisdom at \( t_0 \), a thorough review must be conducted in order to appropriately weight research findings, and this would require resources above those that would have been required at \( t_0 \). In spite of this limitation, the model still provides value; if a study contradicts previous beliefs, the appropriate systematic review would need to be updated, resulting in even larger costs. By combining our model with the temporal and other practical considerations found in [HSBM12], this limitation is minimized; we perform this analysis for one topic in Chapter 2.

Our model only explicitly considers comparative effectiveness reviews with two treatments. Extending this to multiple treatments is trivial, by considering each pair of treatments as a separate topic. The costs would be shared, but the implications would be very similar. Evaluating a single treatment is also straightforward, and can be accomplished by modeling one of the potential treatments as a placebo. The extension to reviews that do not investigate specific treatments, such as [TRF+12], is more difficult. These reviews are likely to influence behavior less directly, which makes estimation of their value more challenging.

Equity makes no explicit appearance in the framework. This could imply that it is best ignore conditions that primarily affect minority groups and rare diseases. Mathematically, equity concerns could be incorporated into the social planner’s optimization problem illustrated in section 1.4.2. However, when the functional form becomes more complex than the summation used herein, obtaining practical guidelines from the abstract formulation is challenging.

Finally, the functions referenced are almost entirely unobservable. While the cost of various approaches and levels of effort, and the number of patients with each condition could be estimated, other factors are more difficult. Indeed, a critical component is \( [V(A) - V(B)] \), the estimation of which is a goal of many reviews. Furthermore, the probability of correct treatment requires a great deal of information, such as the state
of the current practice, the ways that physicians make decisions, the likelihood of identifying the superior treatment, and patient-specific factors. While this is not ideal, the model still provides a variety of useful insights and provides a framework within which to discuss methodological approaches to systematic reviews in a resource constrained environment.

1.5. Alternatives to Traditional Systematic Reviews

If time and cost were no object for a systematic review, all literature with any potential for inclusion would be reviewed; this would be fantastically expensive. Even if fiscal resources were only a minor concern, the time spent screening for one review cannot be spent on other reviews. In 2003, it was estimated that more than 10,000 systematic reviews were required to synthesize the trials that were then present in the Cochrane Central Register of Controlled Trials (CENTRAL) [MC03]. The medical literature has expanded greatly since 2003, and not all medical evidence is in CENTRAL, making this a conservative lower bound. Furthermore, many of the reviews conducted by the EPCs are would not be considered as Cochrane reviews, indicating that most EPC reviews should be added to this estimate. For these systematic reviews to be useful, they must be current. Given the massive workload required, and the limited workforce of researchers able to conduct systematic reviews, completely exhaustive searching on each review is not optimal.

The fact that systematic reviews require significant time and resource expenditures has caused researchers to investigate alternatives. A logical approach would be to eliminate or reduce the intensity of the literature search and screening phase; variants of this have been proposed [GCT10, SDF+13, PCP+11, KKC+12, HK12]. The most important distinction between syntheses utilizing these approaches and a traditional systematic review is that they are explicitly setting a lower quality requirement in exchange for increased speed and decreased cost. This does not imply that the researchers are being lazy or sloppy; they are responding to realistic trade-offs between limited resources and prompt answers. In order to distinguish the methodological approaches that represent reasonable modifications from those that are ineffective, it is critical to know how various methodological choices affect the cost and quality of the review.

Procedures that seek to use a streamlined approach to obtain a systematic review-like synthesis have been called “rapid reviews,” “pragmatic reviews,” or “evidence summaries.” Their development is an important step towards operationalizing the framework presented in the Section 1.4. While the task of determining the optimal level of effort for a given review is quite abstract, it is simpler for decision-makers to choose from a menu of possible levels of detail. Thus far, the primary driver of these modifications has been timeliness, but they can be appropriate decisions for other reasons. If a particular review has a low probability of changing behavior, or may not have a large treatment effect, one of these simpler reviews may be a better choice.
1.6. Initial Recommendations

Many of the proposed actions in Table 1.1 of Section 1.3.2 have a significant effect on cost but only minor effects on outcomes of the review. For example, limiting searches to English seems to be a worthwhile tradeoff, particularly on topics outside of alternative medicine. The evidence is mixed, but if this restriction leads to biased estimates of treatment effects, it tends to create bias towards more conservative treatment effect [JHS+02, MPK+00, MPLK03]. The cost implications of requiring additional language expertise could be considerable if the language is not known to the available screeners. Single screening instead of double screening cuts the cost of a screening phase in half, but risks losing 9% of the relevant articles [ECD+02]. Because this figure appears to refer to the title screening phase of review preparation, it would double costs in only a small portion of the overall review. In the Chapter 3, we show that for one review, the cost of title screening is less than 1% of the total screening cost. Double screening, therefore, is wise—at least in the title phase.

These examples are meant to reiterate that all evidence synthesis face a tradeoff between comprehensiveness and efficiency. The fact that a review is systematic is not a guarantee that it will include every relevant study. Educated decisionmaking in systematic reviews requires three pieces of information. First, we must understand the estimated change to precision and bias of the proposed action. Second, we must estimate the effect on time and cost. Finally, and most difficultly, we must understand how the value of a systematic review varies with its bias and precision. Value increases with precision and decreases with bias, but other factors abound. The framework provided in Section 1.4 above provide a useful stylized model for making these decisions more explicit.

1.7. Conclusion

A systematic review is a rigorous synthesis of the evidence addressing a set of medical questions. While these reviews provide a very thorough answer to the specified questions, the increasing growth rate of medical literature and limited budgets make it difficult to conduct all of the needed reviews at the current high standards. In addition, reviews must be updated regularly to be useful; it is difficult to determine when an update be undertaken. In the following chapters, we first estimate the value created by systematic reviews and discuss factors related to the time taken to screen articles. Armed with a thorough understanding of the problem, we present a method that appreciably reduces abstract-level screening time with a small risk of bias, and is effective for conducting both de novo systematic reviews and updating existing reviews.
2. Estimating the Value of Systematic Reviews: Osteoporosis Case Study

2.1. Introduction

Millions of dollars are spent each year conducting systematic reviews [HSBM12] to make medical evidence more easily incorporated into clinical decisionmaking. While they are seen as important [MCD97] by researchers and the expenditures demonstrate governmental support, no research has quantified the value created by a systematic review. Doing so could enable better allocation of resource to individual reviews and inform budgeting decisions that must allocate scarce research funding. In general, estimating the value of systematic reviews is a daunting task. The body of literature related to each condition is unique—some are very mature, others are new. Even within mature domains, new knowledge can be generated, and young domains have varying proportions of hype versus medical evidence. A tool with the ability to prospectively assess the value of a review would be most useful, but even retrospective analysis is difficult.

Ideally, we would value the contribution of a single review by evaluating the effect of that review on practice decisions, outcomes, and research funding decisions. Because both of these types of decisions are the result of complex systems with many sources of influence, we use a simulation model rather than a purely statistical analysis. As an illustrative example, we quantify the value created by two systematic reviews on treatment for fractures associated with low bone density [CND+12, MAC+07]. We chose this condition both because it has a large societal burden and a large, rapidly growing literature base. The National Osteoporosis Foundation estimates that 8.9M Americans suffer from osteoporosis, and a total of 51.4M from low bone density [NOF13]. In 2005, there were more than 2M new osteoporotic fractures, which cost approximately $17B [BDHS+07]. In addition to these accounting costs, fractures can be traumatic, with self-judged quality of life on a 0-1 scale decreasing from an average of 0.76 before a fracture to 0.31 immediately following a hip fracture, and remaining at 0.66 for the remainder of the patients’ years [SWdGP14].

Coupled with large costs, the condition has been the subject of research for many years, with the first treatment approved by the FDA in 1984 [MAC+07] and a number of systematic reviews and meta-analyses [MAC+07, CND+12, KRL+00, HERR08, MDM+12, SWP14, SWdGP14, CWG+02, BMTB12, SJ+05, CWW+02, HMCE+10, WCP+11]. Two of these systematic reviews [MAC+07, CND+12] were conducted as part of the
Chapter 2  Estimating the Value of Systematic Reviews: Osteoporosis Case Study

Figure 2.1.: High level overview of the approach to calculating the value of the systematic review of fracture prevention

![Diagram](image1)

Figure 2.2.: Method for finding the optimal treatment strategy for a given set of beliefs

![Diagram](image2)

Comparative Effectiveness Review program of the Agency for Healthcare Research and Quality (AHRQ), and their data is used elsewhere in this manuscript. This wide research base provides sufficient data to populate the requisite models and allows formalization of the state of knowledge in this particular domain.

To enable estimation of quantities that cannot be obtained from existing data, we created the Flexible Osteoporosis Cost and Utility Simulation (FOCUS), a microsimulation model implemented in Java. It creates patients with realistic characteristics and fracture histories and supports a wide range of treatment strategies. In its standard configuration, simulated cost and health outcomes can be compared across various treatment strategies. Resulting from its open, object-oriented architecture, it is straightforward to change assumptions about patient characteristics, investigate new treatments, obtain additional outputs.

We propose a five-step approach to calculating the societal benefit of this systematic review; these steps are outlined in Figure 2.1. First, we find the optimal treatment from a set of possible strategies using a cost-effectiveness model, as shown in Figure 2.2. We run FOCUS once under beliefs that represent the pre-review state of knowledge, and once under post-review beliefs to obtain the corresponding optimal strategies. Secondly, we use a variant of the same model to calculate the long-term cost and quality-of-life effects of each type of fracture, using an approach detailed in Section 2.4. Figure 2.3 shows the following two steps, where the two optimal strategies are run through the original cost effectiveness model using the best current estimate of all parameters. Outcomes of interest are the costs and decreases in Quality-Adjusted Life-Years (QALYs) resulting from fractures, which we then convert to dollars using standard values. By taking the difference in monetized utility between the two strategies, we obtain an estimate of the annual benefit of better treatment decisions attributable to the systematic review. Finally, using an additional set of assumptions, we scale this estimate to reflect the population value of information.
2.1 Introduction

Figure 2.3.: Method for calculating the value of improved treatments due to information synthesis from a systematic review

<table>
<thead>
<tr>
<th>Age</th>
<th>Event</th>
<th>Utility</th>
<th>Cumulative QALYs</th>
</tr>
</thead>
<tbody>
<tr>
<td>70</td>
<td>None</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>71</td>
<td>Hip Fracture</td>
<td>0.8</td>
<td>1.8</td>
</tr>
<tr>
<td>72</td>
<td>None</td>
<td>0.9</td>
<td>2.7</td>
</tr>
<tr>
<td>73</td>
<td>Death</td>
<td>0.0</td>
<td>2.7</td>
</tr>
</tbody>
</table>

Table 2.1.: Simple worked example of QALY calculations

2.1.1. Brief Background on Cost Effectiveness Analysis

Both cost-benefit analysis and cost-effectiveness analysis involve modeling a usually complex process and presenting outcomes in terms of values that can be compared across options. In cost-benefit analysis, all outcomes are converted into economic terms, while cost-effectiveness presents the outcomes without forcing all outcomes to be converted into dollars[WS77]. Both formalisms share the underlying assumption that for a given decision being modeled, the societal goal is to maximize the sum-total welfare of society for a given resource level[WS77]. In this paper, we use both types of analysis: health outcomes and costs are reported, but to estimate the total value of a systematic review, the costs of health outcomes are converted to dollars.

While it is challenging to assign a value to life, longevity, and well-being, the most common way of doing so is Quality-Adjusted Life-Years(QALYs). To calculate this measure, each relevant health state is assigned a utility value between 0 and 1[VVM11]. A utility of 0 is equivalent to death, and 1 to perfect health. Between these states lies a continuum where relatively minor discomfort is close to one (the long-term utility of a forearm fracture is 0.81[SWdGP14]), and very painful events are much lower (utility immediately following a hip fracture averages 0.31 [SWdGP14])\(^1\). Each year, instead of counting “1” for each life-year, this utility value is added to a counter—a simple example is shown in Table 2.1. To obtain population estimates, this value is averaged over the relevant patients to yield average health outcomes. By standardizing this measure of well-being, it is possible to make resource allocation decisions across various conditions.

\(^1\)Note that these values differ from those used in the model as the former are population averages while the latter are multiplicative factors.
Chapter 2  Estimating the Value of Systematic Reviews: Osteoporosis Case Study

Figure 2.4.: Graphical depiction of possible cost-effectiveness results. A is dominated by B, and both C and D dominate E. Depending of the value of a QALY, B, C, or D could be rational decisions.

Without standardizing, it would be difficult to choose between averting some number of hip fractures and averting blindness from diabetic retinopathy, for example. While such a decision is still complex, the cost-effectiveness formalism provides structure that aids decisionmakers.

An additional benefit of this type of analysis is that it encourages users to make their values explicit. If superior treatments are available at a higher cost, there can be multiple rational treatments for different valuations of health states. An example of potential outcomes is shown in Figure 2.4. The desirable location in this figure is the bottom-right, where health outcomes (here, QALYs) are good and costs are low. Treatments A and E are examples of irrational treatment decisions. When comparing A vs. B, treatment B achieves better outcomes at a lower cost, so A is “dominated” by B. Similarly, E is dominated by both C and D. However, the decision between B, C, and D is more challenging. When comparing B vs. C, by spending $5,000 more, 1.5 QALYs can be gained, and an additional 1 QALY can be gained by spending $7,000 above that of C.

Given this formalism, it becomes possible to consider the cost per QALY that a society might use as a decision threshold. For some time, guidance in the United States has been to choose treatments if their cost per QALY did not exceed $50,000[BMK+08]. In spite of this established guidance, research allocation decisions have been shown to value QALYs at somewhere between $95,000 and $297,000. In order to remain conservative but realistic, this work uses a value of $100,000 per QALY. Using any of these plausible values, treatment D should be preferred in Figure 2.4.

The models used to determine the costs, lifespans, and quality of life that are required for cost effectiveness analysis can take many forms. One option is to perform the analysis analytically, with a system of equations regulating the critical outputs. This is not feasible for the current analysis for two reasons. First, history is very important in modeling osteoporosis; for example, women with a prior vertebral fracture are four times more likely than others to have a subsequent vertebral fracture[KRL+00], and excess mortality is large after a hip fracture but decreases with time[HMCE+10]. Secondly, the
primary decision in the prevention and treatment of osteoporosis is to determine which patients should receive pharmaceutical treatment—there is little demonstrated difference between approved treatments [MAC+07, CND+12, MDM+12], but these therapies are generally very effective [CND+12]. In spite of the similar efficacy, FOCUS demonstrates that alendronate can range from cost-saving to dominated by a lack of treatment simply by varying the rules governing which patients are treated. Correspondingly, this model must be accurate at the patient level rather than the level of national aggregate statistics. The model used to calculate the long-term effects of an individual fracture (in Section 2.4) could be seen as an analytic model, but involves calculating outcomes in more than $10^{38}$ possible states, which is intractable to calculate exhaustively. Because heuristics are required to carry out the estimation, and the model cannot be easily expressed in a small set of equations, it differs appreciably from the traditional concept of obtaining an analytic solution.

A commonly used alternative modeling approach would be to use a Markov model, which passes cohorts through a memoryless process, and has been widely used in previous work [SWP14]. Unfortunately, this type of model suffers from the same deficiencies for our problem—difficulty of incorporating history and targeting treatment decisions. The framework that does provide sufficient flexibility is that of microsimulation, which is defined by its working at the level of individual “units,” [Wil07], which are patients in this case. By simulating each patient, microsimulation provides the ability to make treatment decisions based on a patient’s characteristics. It enables the creation of treatment strategies that mimic the way a physician may operate; any characteristic observable by a physician (such as age, race and prior fractures) can be utilized in the formalization of a treatment strategy within the model. While some microsimulation models of bone density and fractures do exist [SWP14], none were publicly available and calibrated for the United States. Accordingly, we create a new model (FOCUS), with the structure heavily influenced by [HEB+09].

### 2.1.2. Other Benefits of Systematic Reviews

While cost-effectiveness analysis is a useful tool for structuring the quantification of improved treatment decisions resulting from the systematic review, there are numerous, more difficult-to-quantify benefits of systematic reviews. When researchers synthesize and critically evaluate the work in one domain, gaps in the current state of knowledge become evident, along with areas where the conclusions are unlikely to change. This can provide a great deal of value for research prioritization. Continuing the example of low bone density, it is unlikely that future trials would find that alendronate is not effective at preventing hip fractures. Conversely, future head-to-head trials of alendronate versus teriparatide could inform decisions about which agent is more effective. Looking only at individual trials, it would be challenging to make such conclusions.

---

2 Memory can be incorporated into Markov models by expanding the set of possible states, but this creates an exponentially increasing state space that quickly becomes difficult to manage.
Once medical knowledge is created, it is sometimes used as-is by providers, and sometimes used in additional research products, such as cost-effectiveness modeling and guideline creation. Models help providers and policymakers make informed decisions about therapeutic options, and the availability of systematic reviews makes it easier to populate these models. In this work, we used systematic reviews or meta-analyses for utility [HERR08, SWdGP14], costs [BMTB12], efficacy and adverse effects [CND+12, MAC+07, MDM+12], and hip fracture mortality [HMCE+10]. The reviews allow modelers to better understand the uncertainty surrounding parameters, use more robust estimates, and provide a curated list of relevant prior work. Finding such work can be tedious and error prone for large models with many parameters. When used to inform clinical practice guidelines like [NOF13], the evidence is made more readily usable by healthcare providers, and these guidelines are stronger when they are backed by a high-quality systematic review.

Finally, the act of preparing systematic reviews involves analyzing a vast quantity of medical literature, and careful thought about the strengths, weaknesses, and gaps of the prior work. This creates subject matter expertise in the preparers, which can be utilized to, among others, speed the preparation of future reviews, present to policymakers, and advise on future research prioritization. The value of the synthesis is not only in the finished document, but also in the minds of the preparers.

2.2. Flexible Osteoporosis Cost and Utility Simulation (FOCUS)

The cost-effectiveness model at the core of this work, FOCUS, is a microsimulation model implemented in Java. The overall structure is largely modeled after [HEB+09], which permits individuals to reach six possible “states”: hip, vertebral, forearm, or other fracture, death, or no fracture. Time is stepped in one-year increments, and all utility, costs, and fractures are recorded as simulation progresses. We obtained parameters from a wide array of published sources, and are described in the following sections, and present a simple illustration of the structure in Figure 2.5.

2.2.1. Initialization

In order for FRAX™ to be estimable, clinical risk factors must be assigned to patients, and to ensure validity, they must be assigned in a way that matches the distribution in the true population. To ensure that this requirement was met, parameters were obtained or estimated from a variety of published data sources, as shown in Table 2.5. The most heavily utilized dataset is the third iteration of the National Health and Nutrition Examination Survey (NHANES III), which was conducted between 1988 and 1994. This is the largest public dataset containing bone mineral density measurements, and also asks a large number of background, health, and nutrition related questions to participants.
2.2 Flexible Osteoporosis Cost and Utility Simulation (FOCUS)

(a) Overall FOCUS architecture. After initialization, the patient can have one of four fracture types, experience no fracture, or die. After each step, trackers and some characteristics are updated; the simulation stops when the patient dies. The blue node captures beliefs and decisions about treatments; the intervention is reflected there.

(b) Detailed illustration of transition probability estimation, which includes making a treatment decision. In this example, treatment follows the National Osteoporosis Foundation guidelines[NOF13]; these strategies are discussed further in Section 2.2.4.

Figure 2.5.: Basic microsimulation model structure. Figure 2.5a shows the high-level structure, and Figure 2.5b provides more detail on the estimation and modification of transition probabilities.
With such a large amount of connected data, we were able to assess correlations and generate data in a way that accurately reflects the population of the United States. Generous coverage and widespread availability led the creators of FRAX to standardize BMD measurements into $t$-scores using the mean and variance from NHANES III [DHLT$^+$10]. While NHANES is still being conducted, more recent surveys did not include bone mineral density measurements. The only required attribute that was not directly available in NHANES III was glucocorticoid use$^3$, which we obtained from the Medical Expenditure Panel Survey (MEPS).

Each time FOCUS creates a new patient, it first samples age, race, and sex from the empirical distribution found in the 2010 Census; this distribution is shown in Figure 2.6. Starting ages can range from 50 to 99, and races are those used by FRAX$^TM$: Caucasian, Black, Hispanic, and Asian. In addition to demographics, FRAX$^TM$ requires the assignment of seven clinical risk factors, which are summarized later in the first seven rows of Table 2.5. We created a set of models that considered the simulated patients’ age, race, and sex when assigning the factors—ensuring that the simulated distribution matched the empirical distribution when adjusting for these factors. Rheumatoid arthritis, smoking status, drinking status, bone mineral density, and glucocorticoid use were available as standalone questions in either NHANES III or MEPS, making their estimation straightforward. Secondary osteoporosis has a complex definition (the presence of seven potential conditions$^4$) [KJO$^+$08], and was assigned to 5% of the population at random to reduce complexity. Parental hip fracture was not available directly, and was instead obtained by scaling race-specific estimates of maternal hip fractures to reflect broader fracture prevalence (see Appendix B.1 for details).

$^3$NHANES does ask about medications taken, but translating to a list of medications to a “glucocorticoid use” indicator may introduce errors.

$^4$The following are identified by FRAX$^TM$ as “strongly associated with osteoporosis: type I (insulin dependent) diabetes, osteogenesis imperfecta in adults, untreated long-standing hyperthyroidism, hypogonadism or premature menopause (<45 years), chronic malnutrition, or malabsorption and chronic liver disease.” [KJO$^+$08]
Specifically, we estimated each factor using a linear model of race, age, age$^2$, and sex as covariates, using the sample of patients aged 50 and up. For bone mineral density, we utilized an ordinary least squares model, and logistic regression for all others. The results of these regressions are available in Appendix B.2.

Because fracture history is also an important component of fracture risk, we initialized all individuals at age 50, then simulate each with a treatment strategy that is similar to current practice until the sampled age. If the patient dies before reaching the desired age, the history is discarded and that patient is re-simulated with the same characteristics. This is required because we sample from the joint distribution of age, race, and sex, but simulated patients that start at age 50 may “die” before the age at which they should begin the simulation. As an approximation of the current practice, 35.7% of patients with osteoporosis ($BMD_t \leq -2.5$) were treated $[SMT+14]$ during this initialization. While treatments chosen vary in practice, only alendronate was used here, as it is the most widely used by a large margin (see Figure 2.9 for details). This approach allows for reasonable fracture histories to be generated and retains the ability to start with a representative cohort of the American elderly population.

### 2.2.2. Changes During Model Execution

To simplify the model, the majority of the characteristics do not change as the simulation runs. The exception to this is bone mineral density, as it plays a critical role in the increase in fracture risk as age increases. To model this change, a linear regression model was estimated on NHANES III data, taking the form

$$BMD_i = \beta_0 + \beta_1 gender_i + \beta_2 age_i + \beta_3 1\{race_i = \text{black}\} + \beta_4 1\{race_i = \text{hispanic}\} + \beta_5 1\{race_i = \text{asian}\} + \beta_6 gender_i age_i + \varepsilon_i$$

While it is unclear if such a cross-sectional model would yield consistent estimates given possible selection due to mortality, [WHC02] found that estimates of hip BMD trends were similar when using either a cross sectional or a panel approach.

To enable comparability with the published literature [WHC02] and prevent negative densities, we converted the output from changes in levels to percent changes. To do this, we utilized the bootstrap procedure [Efr79] to estimate the model, predict the percent change for every individual in the sample, then average those percent changes by gender. Women in the sample lose an average of 0.84% BMD each year, while men lose only 0.59%. When the simulation is run, we change the BMD of each patient by a percentage drawn from a distribution specified by this mean and its bootstrapped standard error.

### 2.2.3. Transition Probabilities

The core of the fracture probability calculator is the FRAX™ algorithm, which was developed by the World Health Organization and has been tailored to a large number...
of countries. It takes as inputs femoral hip Bone Mineral Density (BMD), age, race, gender, and seven clinical risk factors (CRFs). These risk factors are glucocorticoid use, smoking, drinking, prior fractures, parent hip fracture, rheumatoid arthritis, and secondary osteoporosis[KJO+08]. While the full algorithm treats each CRF separately, the FRAX™ website provides tables with ten-year fracture probabilities by race, age, gender, BMD, and number of CRFs; this was utilized to allow for rapid calculation of millions of risk assessments. As outputs, FRAX™ provides the 10-year risk of having a hip fracture as well as the 10-year risk of having any major osteoporotic fracture5.

In order to make the scores usable in a one-year time step simulation, we require two types of disaggregation. First is the conversion of ten-year probabilities into one-year probabilities, and second is the separation of major fractures into major, hip, forearm, vertebral, and other fractures. A simple approach for time would divide the ten-year risk by ten, but this will underestimate risk because multiple fractures could take place in a ten-year time. Instead, by inspecting Figure 2.7, which shows the relationship between age and fracture incidence, we observe that an exponential function may fit the data quite well. By using the simplifying assumption that fractures over time are independent given age, the ten-year risk can be expressed as

\[
P(Fracture_{10yr}) = 1 - \prod_{i=1}^{10} (1 - P(Fracture_i))
\]

Because the data appear to vary exponentially with age, we express each year’s fracture

\(^5\)A major osteoporotic fracture is defined as a fracture of the hip, clinical vertebral, forearm, or proximal humerus.
2.2 Flexible Osteoporosis Cost and Utility Simulation (FOCUS)

<table>
<thead>
<tr>
<th>Sex</th>
<th>Hip Fractures</th>
<th>Major Osteoporotic Fractures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>0.1185</td>
<td>0.0716</td>
</tr>
<tr>
<td>Female</td>
<td>0.1295</td>
<td>0.0639</td>
</tr>
</tbody>
</table>

Table 2.2.: Exponential coefficients obtained by performing least squares estimation on fracture incidence data from [EBDH+10]

**Algorithm 2.1 Temporal disaggregation of FRAX™ scores**

Require: Ten-year fracture probability $P_{10\text{yr}}$, age, aggregate incidence $X$

* $\text{baseAge} \leftarrow \left\lfloor \frac{\text{age}}{5} \right\rfloor \times 5$
* $\beta \leftarrow \arg \min_{\alpha, \beta} \sum_{i=50}^{85} (\alpha \exp(\beta \times i) - x_i)^2$
* $\rho \leftarrow \arg \min_{\rho} \left| P_{10\text{yr}} - \prod_{i=1}^{10}(1 - \rho \exp(\beta \times i)) \right|$
* $P(\text{Fracture}) \leftarrow \rho \exp(\beta \times (\text{age} - \text{baseAge}))$

Return: $P(\text{Fracture})$

probability using the parametric form

$$P(\text{Fracture}_i) = \rho \exp(\beta \times i)$$

To estimate $\rho$ and $\beta$, we carry out a two-step procedure. Because $\beta$ determines the shape of the exponential, it is reasonable that it will be similar across patients of a given gender. Accordingly, we fit an exponential curve to incidence data by gender from [EBDH+10] using least squares; the coefficients $\beta$ are shown in Table 2.2. Hip fractures are responsible for a majority of the exponential increase, an observation that also holds in Figure 2.7. Women also have a stronger increase in hip fracture incidence with age than men—the rates are similar when both are younger than 60, but according to the National Hospital Discharge Survey, approximately 70% of all hip fractures occur in women.

With the fixed value of $\beta$, each time a prediction is required, we perform a univariate optimization that finds $\rho$ such that

$$P(\text{Fracture}_{10\text{yr}}) = 1 - \prod_{i=1}^{10}(1 - \rho \exp(\beta \times i))$$

It is necessary to perform this optimization with each new value of $P(\text{Fracture}_{10\text{yr}})$, because $\rho$ scales the exponential function. The one-year probability can then be determined by plugging in the appropriate value for $i$ based on the patient’s age, utilizing the estimated $\rho$ and $\beta$. This procedure is depicted in Algorithm 2.1, and repeated for hip fractures and major fractures separately.

Once single-year probabilities have been obtained, we separate them into the desired fracture types. The probability of a major osteoporotic fracture, $P_{maj}$, and the probability
Chapter 2 Estimating the Value of Systematic Reviews: Osteoporosis Case Study

of a hip fracture, $P_{\text{hip}}$, are given. Hip fracture is one component of major fractures, thus

$$P_{\text{maj}} = P_{\text{hip}} + P_{\text{non-hip}} - P_{\text{hip}}P_{\text{non-hip}}$$

$$P_{\text{non-hip}} = \frac{P_{\text{maj}} - P_{\text{hip}}}{1 - P_{\text{hip}}}$$

under the assumption that hip and non-hip fractures are independent. From data provided in [EBDH+10], we derive shares of each type of non-hip fractures in each five-year age range, by gender. The probability that a fracture $T$ is of type $t$ is then

$$P(T = t) = P_{\text{non-hip}} \times \frac{I_t}{\sum_{j \neq \text{hip}} I_j}$$

where $T$ is a random variable reflecting the fracture type, $I_j$ is the incidence of fractures of type $j$ given the age group and sex of the patient.

While this provides a foundation, prior fractures are an exceptionally strong predictor of future fractures, and these risks are largely specific to a given fracture site [KRL+00]. To make use of the relative risk (denoted $RR$) estimates available in sources like [KRL+00], we follow the derivation in Appendix B.3, which shows that conditional probabilities can be derived from unconditional probabilities and relative risk as

$$P(A|B) = \frac{P(A) \cdot RR}{RR \cdot P(B) + P(\neg B)}$$

In this case, the event $A$ is the occurrence of a specific fracture type, and $B$ is the presence of a prior fracture—its prevalence. While incidence is often published in the literature, prevalence has not been published, so we ran FOCUS to generate these estimates. For each age, gender, and fracture type, the percentage of individuals with a prior fracture was collected in one large model run. In future runs, these were used to adjust fracture risk on the basis of prior fractures. Note that because studies that investigate fracture risks are unable to randomly assign fractures, the difference in risk may contain two components. First, the fracture itself may cause trauma that directly makes future fractures more likely. Alternatively, the occurrence of a fracture provides a signal that the patient’s bone micro-architecture is defective [KRL+00]. FOCUS is agnostic towards these two components and only considers the total risk increase.

With probabilities estimated for each of the four fracture states, death and non-fracture probabilities must make up the remaining probability mass. The Social Security Administration’s Office of the Chief Actuary [SSA09] provided raw mortality rates by age and gender. In cases where a patient has suffered a hip or clinical vertebral fracture, mortality increases sharply in the short term, and decreases but remains elevated for ten years [HMCE+10]. Using the relative risks published in [HMCE+10], we adjusted those mortality probabilities in the same fashion as adjusting fracture risks for prior fractures. After calculating the four fracture probabilities and death probability, all remaining probability mass is assigned to the “no fracture” state. A summary of this entire process of estimating transition probabilities is shown in Figure 2.8.
2.2.4. Treatment Strategies

The primary innovation in FOCUS is its ability to quickly evaluate policies that vary the way physicians provide treatment to prevent fractures. This includes which patients receive treatment, which treatment the patient receives, and the efficacy of the provided treatment. Central to the current analysis, we implement beliefs about efficacy through the relative risks that modify fracture probabilities. If an individual meets the specified criteria for treatment, his or her fracture probabilities are decreased according to relative risks estimated in established literature. Those relative risks can reflect the best current evidence or can be limited to earlier studies; this permits understanding of rational behavior given the knowledge available at different points in time.

Treatment strategies make two primary decisions:

- **Who should receive treatment?** A patient can be given a pharmaceutical agent when any set of criteria are met. As an example, Figure 2.5b shows the decision tree reflecting the current guidelines suggested by the National Osteoporosis Foundation[NOF13]. All patients with a prior hip fracture, prior clinical vertebral fracture, or $BMD_t \leq -2.5$ are given treatment. If the patient has $-2.5 < BMD_t \leq -1.0$, they only receive treatment with a sufficiently large FRAX™ score.

- **What are the effects of treatment?** When given alendronate, it causes multiple important effects–fracture risk is reduced, adverse effects may occur, and drug costs are incurred. The fracture risk reduction is uncertain, but FOCUS makes it trivial to incorporate any treatment whose efficacy can be expressed in terms of relative risks. The adverse effects associated with alendronate are generally mild or very rare[CND+12], so we do not include them, but this could easily be incorporated into the treatment strategy component.
In order to retain maximum flexibility, we define a treatment strategy as any function that accepts a patient with transition probabilities and returns a new set of transition probabilities. The simplest strategy implemented is that of no treatment, wherein the original probabilities are returned without modification. When a more complex strategy determines whom to treat, the strategy is permitted to utilize any knowable characteristic. This includes the patients’ FRAX™ scores, but should not be too complex for providers to implement. The drug chosen for treatment and the specified beliefs about the effects of that treatment modify the transition probabilities. Relative risks of fracture of various types for the given treatment vs. standard care is obtained from the literature (generally randomized controlled trials), either from a single study or a meta-analysis.

Because the focus of the model is not to compare various treatments but to evaluate the best treatment under different information environments, alendronate is the only drug investigated. It is the most popular medication for prevention and treatment of osteoporosis (see Figure 2.9), and has been approved by the FDA since 1995[Mer12]. Despite having a relatively large number of clinical trials, evidence of superiority of any one pharmaceutical agent over another is lacking[CND+12]. Given this lack of evidence to support alternatives, restricting the analysis to the most popular treatment with the largest evidence base is a reasonable simplification.

It is worth noting that parameters correspond with branded alendronate, rather than the generic version. This is most evident in the cost of the medication, with the branded drug costing nearly 14 times the generic price in 2010[NRG11]. The primary reason for investigating the branded drug is that there are demonstrated differences in efficacy, adherence, and adverse effects [BDO+13] between the branded agent and the generic version. Given that the trials synthesized all studied Fosamax, the branded drug, it would be unwise to assume that cost is the only difference.
2.2 Flexible Osteoporosis Cost and Utility Simulation (FOCUS)

2.2.5. Utility and Costs

The primary outcomes for FOCUS are cost and utility, which we obtain from a strong evidence base. In each period, patients accrue utility and costs, and these depend on both the current state and the history of the patient. Utility calculations use a multiplicative model as in [HEB+09], with parameters drawn heavily from a systematic review of osteoporosis utility values [HERR08]. Forearm and other fractures cause a single year of disutility, while hip and vertebral fractures cause some disutility in all future years, in addition to a higher utility loss in the fracture year.

Costs comprise both medical costs after fractures and cost of medications for treatment. The most rigorous and recent estimates of fracture costs were found in [KCD+13], which looks at expenditures directly related to fractures using two approaches. One treats pre-fracture expenditures as a control, and the other uses a cohort matched on age, race, and sex drawn from the non-fracture population as controls. Kilgore defines expenditures that are directly related as comprising “fracture care, rehabilitation, or complications arising from fracture or treatment” [KCD+13], a definition which we utilize as well. Because this does not consider non-fracture expenses, it does not take a true societal viewpoint. When a patient is kept alive because treatment prevents a hip fracture, the medical system will have additional costs associated with the individual’s medical care. By ignoring these expenses, we simplify the problem, but acknowledge the omission of this potentially significant cost category.

The primary limitation of [KCD+13] for the purpose of this model is that it does not specify the costs of “other” fractures, but relationships between the costs of various fracture types are provided by [MGC+03]. We aggregated the costs of all fracture types other than hip, vertebral, and forearm using an incidence-weighted average, and found that these “other” fractures to incur approximately 7% of hip fracture costs.

FOCUS does not include any long-term costs, particularly long-term care, which differs from many other models [HEB+09, PKY+13]. This was a conscious decision with a large effect on results, evidenced by the fact that approximately 60% of total costs in [PKY+13] are due to long-term care. While it is true that a majority of hip fracture patients (58%) are discharged to a nursing home [BLO+09], their stays are often relatively short. The average stay of a nursing home patient with hip fracture is shorter than their peers without fractures [Con94]. More importantly, only 1.4% of hip fracture patients that were living in the community at the time of fracture and are alive 180 days after the fracture are living in a nursing home [BYK+10]. Some of these patients would have moved to a nursing home without fractures simply because of age; the number of individuals that move to and remain in a nursing home as a result of a hip fracture is very small. Previous studies [HEB+09, PKY+13] used the high discharge rate to nursing homes as rationale for assuming that patients spent all remaining years in nursing homes—an assumption that is simply not supported by more recent evidence in the United States [BYK+10]. The decision to omit these costs leads to a small underestimation of the value of hip fracture prevention, but this slight conservatism is preferable to aggressive assumptions.
Chapter 2  Estimating the Value of Systematic Reviews: Osteoporosis Case Study

<table>
<thead>
<tr>
<th>Name</th>
<th>Source</th>
<th>Distribution</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fracture Risk Ratios</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RR Hip on future Hip</td>
<td>[HEB+09]</td>
<td>Uniform</td>
<td>2.3</td>
</tr>
<tr>
<td>RR CV on future CV</td>
<td>[HEB+09]</td>
<td>Uniform</td>
<td>4.4</td>
</tr>
<tr>
<td>RR Forearm on future Forearm</td>
<td>[HEB+09]</td>
<td>Uniform</td>
<td>3.3</td>
</tr>
<tr>
<td>RR Other on future Other</td>
<td>[HEB+09]</td>
<td>Uniform</td>
<td>1.9</td>
</tr>
<tr>
<td>RR CV on future Hip</td>
<td>[HEB+09]</td>
<td>Uniform</td>
<td>2.3</td>
</tr>
<tr>
<td><strong>Utility</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hip fracture, first year</td>
<td>[HEB+09]</td>
<td>Beta</td>
<td>0.797</td>
</tr>
<tr>
<td>Hip fracture, subsequent years</td>
<td>[HEB+09]</td>
<td>Beta</td>
<td>0.899</td>
</tr>
<tr>
<td>CV fracture, first year</td>
<td>[HEB+09]</td>
<td>Beta</td>
<td>0.720</td>
</tr>
<tr>
<td>CV fracture, subsequent years</td>
<td>[HEB+09]</td>
<td>Beta</td>
<td>0.931</td>
</tr>
<tr>
<td>Forearm fracture, first year</td>
<td>[HEB+09]</td>
<td>Beta</td>
<td>0.94</td>
</tr>
<tr>
<td>Other fracture, first year</td>
<td>[HEB+09]</td>
<td>Beta</td>
<td>0.91</td>
</tr>
<tr>
<td><strong>Costs</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hip fracture</td>
<td>[BMTB12]</td>
<td>Uniform</td>
<td>$27,697</td>
</tr>
<tr>
<td>CV fracture</td>
<td>[BMTB12]</td>
<td>Uniform</td>
<td>$8,433</td>
</tr>
<tr>
<td>Forearm fracture</td>
<td>[BMTB12]</td>
<td>Uniform</td>
<td>$6,201</td>
</tr>
<tr>
<td>Other fracture</td>
<td>[BMTB12, MGC+03]</td>
<td>Uniform</td>
<td>$1,956</td>
</tr>
<tr>
<td>Alendronate (generic, annual)</td>
<td>[PKY+13]</td>
<td>Uniform</td>
<td>$98</td>
</tr>
<tr>
<td><strong>Other</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male annual % BMD change</td>
<td>NHANES</td>
<td>Normal</td>
<td>-0.84%</td>
</tr>
<tr>
<td>Female annual % BMD change</td>
<td>NHANES</td>
<td>Normal</td>
<td>-0.59%</td>
</tr>
</tbody>
</table>

**Table 2.3.**: Distribution-based Parameters

### 2.2.6. Summary of Architecture and Parameters Used

To provide structure to the large number of parameters, we break them into logical groups related to the way the values are generated. Distribution-based Parameters in Table 2.3 are random variables that apply to the system at large rather than the individual. Deterministic table parameters shown in Table 2.4 are fixed values that look up a patient by age and gender, returning a tailored but static value. This can be specific to an exact age, or use ranges. Initialization parameters determine the starting distribution of various characteristics in the model, and were obtained in two different ways. Race, age, and gender use a table based on Census data, while the remainder use regression models of race, gender, and age to predict each characteristic. The inputs, intermediate outcomes, and outputs, for each patient are shown in Figure 2.10.

Finally, the parameters must be put into the context of FOCUS to be meaningful. Algorithm 2.2 presents an alternative depiction of the model’s architecture. Patients do not interact with each other—they are realistically initialized, then stepped through the simulation, accruing costs and utility until their death.
Table 2.4.: Deterministic table parameters, which vary by age and gender

<table>
<thead>
<tr>
<th>Name</th>
<th>Source</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>CV fracture share</td>
<td>[EBDH+10]</td>
<td></td>
</tr>
<tr>
<td>Forearm fracture share</td>
<td>[EBDH+10]</td>
<td></td>
</tr>
<tr>
<td>Other fracture share</td>
<td>[EBDH+10]</td>
<td></td>
</tr>
<tr>
<td>Mortality rate</td>
<td>SSA[SSA09]</td>
<td></td>
</tr>
<tr>
<td>Hip fracture prevalence</td>
<td>Derived from this model</td>
<td></td>
</tr>
<tr>
<td>CV fracture prevalence</td>
<td>Derived from this model</td>
<td></td>
</tr>
<tr>
<td>Forearm fracture prevalence</td>
<td>Derived from this model</td>
<td></td>
</tr>
<tr>
<td>Other fracture prevalence</td>
<td>Derived from this model</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.5.: Initialization parameters

<table>
<thead>
<tr>
<th>Name</th>
<th>Source</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rheumatoid Arthritis</td>
<td>NHANES</td>
<td>Logistic Regression</td>
</tr>
<tr>
<td>Smoking Status</td>
<td>NHANES</td>
<td>Logistic Regression</td>
</tr>
<tr>
<td>Drinking Status</td>
<td>NHANES</td>
<td>Logistic Regression</td>
</tr>
<tr>
<td>Secondary Osteoporosis</td>
<td>None</td>
<td>Beta ($\alpha = 4, \beta = 96$)</td>
</tr>
<tr>
<td>Bone Mineral Density</td>
<td>NHANES</td>
<td>Linear Regression</td>
</tr>
<tr>
<td>Glucocorticoid Use</td>
<td>MEPS[MEP12]</td>
<td>Logistic Regression</td>
</tr>
<tr>
<td>Parent Hip Fracture</td>
<td>[CDC10],NHANES</td>
<td>Logistic Regression</td>
</tr>
<tr>
<td>Age</td>
<td>2010 Census</td>
<td>Empirical Distribution</td>
</tr>
<tr>
<td>Race</td>
<td>2010 Census</td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>2010 Census</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2.10.: Inputs, intermediate outcomes, and outputs for each patient in FOCUS
Algorithm 2.2 Flow of one patient through FOCUS
* Sample characteristics to initialize the patient
* Simulate from age 50 to the desired starting age
* Reset cost and utility
while state ≠ Death
* Calculate transition distribution D
* Sample state from D
* Accrue cost, utility
endwhile
Return: cost, QALYs, age at death, number of fractures by type

2.3. Identifying Optimal Treatment Strategies

With the model established, we first identify the best treatment strategy under two different information environments. Under different assumptions, a model with the same structure can yield vastly different results. By altering the decisionmakers’ beliefs about the efficacy of the treatment, we are able to recreate situations that realistically reflect an environment where the systematic review is unavailable.

Specifically, we begin with a set of potential strategies $S$ that provide alendronate to patients using different rules as enumerated in Appendix B.4. The simplest strategy never provides treatment, and the most complex reflects the guidelines of the National Osteoporosis Foundation, which uses fracture history, bone mineral density, and FRAX™ scores.

Before the systematic review, we assume that only data from the Fracture Intervention Trial [CBT+98, BCK+96] are available. Because this is the largest alendronate trial, we assume that physicians recommending treatment for low bone density would be aware of its primary results—even in the absence of any type of review or synthesis. For the post-systematic review case, all trials [APO+95, LWB+95, BCK+96, BDT+97, CBT+98, HCC+98, GPF+98, MCD+98, PFH+99, OEW+00, BGM+00, SFB+00, DDY01, HAF+03, GRP03, RFDI04, QTS+05, ZJC+05, SIKS06, MLC+06, dNJL+06, RFSR07, PKF+08] identified in the 2011 update to AHRQ’s Comparative Effectiveness Report [CND+12] were synthesized using a standard inverse-variance random-effects model [HG11]. The forest plots from this process are shown in Figure B.1 in Appendix B.5, and results are summarized in Figure 2.11. While we believe that these are reasonable approximations of the information environment, there is considerable uncertainty in this assumption, and the varied knowledge across physicians is not explicitly modeled.

Under each information scenario and strategy, we simulated 4,000,000 patients until they reached the “death” state, and recorded the cost of medication and increased health system utilization, QALYs, age at death, and number of fractures by type. Within each information setting, the optimal strategy is defined as

$$s_{info}^* = \arg \max_{s \in S} [100,000 \times QALY(s|info) - Cost(s)]$$
Stated plainly, $s_{ino}^*$ achieves the largest net benefit when valuing each QALY at $100,000. In order to reduce potential sources of randomness unrelated to the review, only the relative risk attributable to the treatment is varied between the pre-review and post-review states.

As shown in Table 2.6, the post-review optimal strategy involves treating more individuals than the pre-review case: treating when $BMD_t \leq -2.5$ was optimal before the review, while a threshold of -2.0 is superior with the updated parameters. It may be surprising that the differences between strategies, particularly in terms of health outcomes, are very small. This is because a large number of individuals with very low fracture risks are simulated. Most other models only generate patients that have a higher-than-average risk of fracture[HEB+09, LBS+08, PDW+11, PKY+13, TBML08], which is logical when the question is one of appropriateness for a specific population. In this study, population-level benefits are of primary concern, and the treatment thresholds are the primary decision. While this makes FOCUS require more computational resources to yield precise estimates, it enables the flexibility required for further analyses.

2.4. Costs and Utility Effects of Individual Fractures

Thus far, we have covered the important components of FOCUS and the first step in the outline provided in Figure 2.1. Because fractures have lingering effects, and are rare events experienced by a large population, estimating the population value of information is more challenging than for many other conditions. In order to utilize the framework presented in [HSBM12] to estimate this value, we modify FOCUS slightly to provide estimates of the long-term cost and quality effects of each fracture type. With these figures, we are able to treat a fracture in the same manner as an acute event. Formally,
Chapter 2  Estimating the Value of Systematic Reviews: Osteoporosis Case Study

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Pre-Review</th>
<th>Post-Review</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cost</td>
<td>QALY</td>
</tr>
<tr>
<td>No Treatment</td>
<td>$1,416</td>
<td>18.08</td>
</tr>
<tr>
<td>BMD ≤ -2.5</td>
<td>$1,419</td>
<td>18.08</td>
</tr>
<tr>
<td>BMD ≤ -2.0</td>
<td>$1,732</td>
<td>18.08</td>
</tr>
<tr>
<td>BMD ≤ -2.5 or Fx</td>
<td>$2,071</td>
<td>18.07</td>
</tr>
<tr>
<td>BMD ≤ -2.0 or Fx</td>
<td>$2,377</td>
<td>18.08</td>
</tr>
<tr>
<td>BMD ≤ -1.5 or Fx</td>
<td>$6,011</td>
<td>18.10</td>
</tr>
<tr>
<td>BMD ≤ -1.0 or Fx</td>
<td>$12,648</td>
<td>18.14</td>
</tr>
<tr>
<td>BMD ≤ -0.5 or Fx</td>
<td>$19,844</td>
<td>18.15</td>
</tr>
<tr>
<td>NOF</td>
<td>$6,092</td>
<td>18.10</td>
</tr>
</tbody>
</table>

Table 2.6.: Per-Person Cost-Effectiveness Model Results: Using the pre-review parameters and a value of $100K per QALY, treating when BMD ≤ -2.5 was optimal before the review, while a threshold of -2.0 is better with the improved parameters. The net benefit (NB) columns are compared to the “No Treatment” baseline, and use a value of $100K per QALY. “Fx” refers to having a prior hip of clinical vertebral fracture.

The population expected value of information is given in [HSBM12] as:

\[
pEV \text{I} = \sum \beta^t \times Dur_t \times Pop_t \times \left( \sum_j Im_{p_{j|I}} \times E_{\theta|I}NB(j, \theta) - \sum_j Im_{p_{j|I}} \times E_{\theta}NB(j, \theta) \right)
\]  

(2.1)

The main components are durability, population, implementation probability, and expected value of information for an individual. \( \beta \) is a discount rate, and \( Dur_t \) is durability, which refers to the likelihood that information will remain valid in period \( t \). Populations in period \( t \) is denoted \( Pop_t \), and \( Im_{p_{j|I}} \) is the implementation probability of alternative \( j \) in period \( t \) given the information \( I \). \( E_{\theta|I}NB(j, \theta) \) is the expected net benefit of treatment \( j \), given parameters \( \theta \) and information \( I \).

To illustrate their approach, [HSBM12] provide an example of noninvasive positive pressure ventilation for acute respiratory failure. Meta-analysis indicates that this treatment is effective for reducing in-hospital mortality. Given that this is the primary result of the treatment, a clear population exists—individuals admitted to the hospital with acute respiratory failure.

In the case of osteoporosis, the primary outcome of interest is the occurrence of fractures, and these can happen for individuals that may not have osteoporosis. Furthermore, occurrence of one fracture increases the risk of future fractures, making it important to consider long-term consequences of treatment decisions. These factors complicate the estimation of durability and population size.

In order to address these issues, we utilized FOCUS to calculate the expected long-term incremental cost and QALY impact of each fracture type. Instead of simulating these...
2.4 Costs and Utility Effects of Individual Fractures

Figure 2.12.: A simplified illustration of the state space for which an expected value must be calculated

outcomes as in Section 2.2, we calculate the expected value (for both costs and QALYs) of each possible first period transition by computing the probability, cost, and health outcome of each possible state until an age of 105 or a minimum probability of $10^{-8}$ is reached. To illustrate the necessity of this approach over a simpler model, consider the effects of having a clinical vertebral fracture. The most obvious are that medical costs are incurred and pain reduces the quality of life for the current and remaining years. Less directly, it increases the risk of future hip and vertebral fractures, which have their own cost and utility effects. Taking this one step further, if a hip fracture is incurred because of the increased probability, mortality increases as a result of that fracture. Furthermore, because utilities are specified in multiplicative terms and lifespans vary, the nominal QALY effect is far from straightforward.

A highly simplified version of the process is illustrated in Figure 2.12. Instead of six states, this shows only three, with all fracture states combined. Only three time steps are shown, while the true process has up to 55 steps. In this illustration, time step $n$ contains $3 \times 2^{n-1}$ possible states. The full model has $6 \times 5^{n-1}$, which means that if a patient starts at age 50, the last level could require computing outcomes in up to $3.3 \times 10^{38}$ unique states.

To calculate these expectations, we utilize a recursive algorithm, which is made feasible by the use of the minimum probability of $10^{-8}$. This enables FOCUS to exclude paths where, for example, 10 hip fractures happen in consecutive years. In each time step, the program adds state-specific values for cost and QALYs to a counter; future values are filled in by recursing down the tree. Once the termination condition is reached (either through a maximum age or minimum probability), the expected values are passed back up the tree, which finishes with a cost and QALY estimate of entering each of the 6 possible states in the next period. By taking the difference between each state and the no fracture state, we obtain an incremental cost of entering each state.

Because future fracture risks and life expectancies vary greatly across individuals, the
Chapter 2  Estimating the Value of Systematic Reviews: Osteoporosis Case Study

<table>
<thead>
<tr>
<th>Fracture Type</th>
<th>Direct Cost</th>
<th>$E(\text{Cost})$</th>
<th>$E(\text{QALY})$</th>
<th>vs. No Fracture</th>
<th>Net Loss ($\text{K}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Fracture</td>
<td>-</td>
<td>$873$</td>
<td>13.18</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Death</td>
<td>-</td>
<td>$0$</td>
<td>0.00</td>
<td>$-873$</td>
<td>$-13.18$</td>
</tr>
<tr>
<td>Forearm</td>
<td>$6,201$</td>
<td>$7,828$</td>
<td>13.05</td>
<td>$6,955$</td>
<td>-0.13</td>
</tr>
<tr>
<td>Other</td>
<td>$1,956$</td>
<td>$3,107$</td>
<td>13.07</td>
<td>$2,235$</td>
<td>-0.11</td>
</tr>
<tr>
<td>Vertebral</td>
<td>$8,433$</td>
<td>$11,107$</td>
<td>12.63</td>
<td>$10,234$</td>
<td>-0.55</td>
</tr>
<tr>
<td>Hip</td>
<td>$27,697$</td>
<td>$28,908$</td>
<td>8.59</td>
<td>$28,035$</td>
<td>-4.59</td>
</tr>
</tbody>
</table>

Table 2.7.: Long-term average cost and QALY impact of a single fracture of each type.

The incremental effect of a fracture will also differ greatly. The QALY impact of a fracture at age 100 is much less than that of age 50, because fewer healthy years are expected. In order to make these estimates reflect population averages, we sampled relevant patient characteristics and fracture histories from distributions derived from the US population at least age 50—the same initialization procedure described in Section 2.2.1.

Table 2.7 shows that the monetary cost of a fracture is generally not much more than the modeled cost, but the QALY impact can be large. At $100,000 per QALY, the quality of life decrease associated with a hip fracture is 16 times the monetary cost—a cost which is already considerable. Some of these figures may be slightly overestimated, because a portion of the increased risk of subsequent fractures is not directly caused by the fracture but provides a information about suboptimal bone micro-architecture [KRL+00]. Disentangling these two components would be valuable future research; it is beyond the scope of this thesis. While it is not the primary focus of this work, these estimates allow us to estimate of the total societal burden of various fracture types. According to the National Hospital Discharge Survey, 258,000 patients were admitted to hospitals for hip fractures in 2010. This corresponds to a direct medical cost of $7.2B and, at $100K per QALY, a utility loss of $118.4B per year.

2.5. Combining the Pieces to Calculate the Population Value of Information

With all of the components established, we can now calculate the populations value of information. By running the two optimal strategies through FOCUS for only one year, we obtain the fracture outcomes for each year that the information from the review is valid. In order to reduce selection effects where random variation in the runs that determine optimal strategies makes the current strategy appear better than it is, we perform a new simulation with 20,000,000 patients. The fracture outcomes are converted to dollars and Quality-Adjusted Life-Years, which we then convert to dollars using the same standard value of $100K. Taking the difference in monetized utility between the strategies provides an estimate of the annual benefit of better treatment decisions attributable to
2.5 Combining the Pieces to Calculate the Population Value of Information

Figure 2.13.: Bootstrapped distribution of estimated per-capita benefits from the improved treatment strategy.

the information made accessible by the systematic review. Finally, we scale this estimate to capture the population value of information.

In this run, we found that using the optimal post-review vs. pre-review strategy decreased annual hip fractures per 10,000 people by 0.43, vertebral fractures by 0.23, forearm fractures by 0.50, other fractures by 0.14, and increased medication cost by $138,634 per 10,000 individuals. Utilizing the results from the previous section to capture the long-term costs and utility effects of each fracture, this change in treatment increases monetary costs by $120,468 and increases QALYs by 2.17 per 10,000 individuals per year, equivalent to a cost per QALY of $2,975. At $100K per QALY, this yields a net benefit of $9.63 per person per year. These results are summarized in Table 2.8. To assess the variance present in this estimate, we drew 2,500 bootstrap samples from both the cost-per-fracture results and changes in number of fractures. The distribution of these results is shown in Figure 2.13, where we see that the variation is quite large. The 90% confidence interval of this estimate is [$0.67, $18.22], and 96.3% of bootstrap samples are greater than zero.

Inserting the results from Table 2.8 and the assumptions from Table 2.9 into Equation 2.1, we obtain a value of approximately $1.5B for the improved treatment decisions attributable to information made accessible by the systematic review. This includes discounting, a decreasing likelihood that the information will remain current as time passes, implementation difficulties, and accounts for the entire population of the United States above age 49. Major assumptions surround durability and implementation probability. Here, durability is a kinked function, where the information retains all of its value for 5 years, then depreciates linearly to zero over the next 5 years. This is approximately in line with [Fle07], which assumes a nine-year lifetime for information from osteoporosis.
<table>
<thead>
<tr>
<th>Annual difference per 10,000</th>
<th>Number</th>
<th>Cost</th>
<th>QALY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hip Fractures</td>
<td>-0.43</td>
<td>-$11,984</td>
<td>1.96</td>
</tr>
<tr>
<td>Vertebral Fractures</td>
<td>-0.23</td>
<td>-$2,394</td>
<td>0.13</td>
</tr>
<tr>
<td>Forearm Fractures</td>
<td>-0.50</td>
<td>-$3,470</td>
<td>0.06</td>
</tr>
<tr>
<td>Other Fractures</td>
<td>-0.14</td>
<td>-$316</td>
<td>0.02</td>
</tr>
<tr>
<td>Medication Cost</td>
<td></td>
<td>$138,634</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2.8.: Summary of results driving the population expected value of information results. Entries are the results of simulating 20,000,000 patients for one year with the optimal pre-review and post-review strategies. Cost and QALY impacts were obtained by scaling the difference in number of fractures by the values obtained in Table 2.7.

trials. Upcoming changes in the pharmaceutical landscape also influence this choice of durability function. According to recent synthesis, teriparatide has the highest probability of being the most effective medicine for reducing fractures[MDM+12], but this has considerable uncertainty and the drug is still quite expensive (7.5x the cost of branded alendronate in 2006[LMN+06]). The patent on teriparatide will expire in 2018[CH04], at which point a new manufacturer will introduce a cheaper generic formulation, which could change the ideal treatment strategies considerably.

The primary shortcoming of this estimate of $1.5B is that the value identified is created by a system including the primary research, systematic reviews, modeling, and guideline creation. The original research costs a substantial amount of money and is certainly of some value—even if not synthesized by a systematic review. Cost-effectiveness models and practice guidelines that they may inform are an important driver of changes in clinical practice. Because no established estimates are available for this portion, we are restricted to crude estimates. Original research is the most resource-intensive segment and is where the knowledge is created; it should claim the largest share of the total benefit. Thus, we assume that 50% of the $1.5B is attributable to the primary research. Given well-synthesized literature, models and guidelines can be less intensive than the systematic review, so 20% is assumed for that portion, leaving 30% for the review itself. In total, this yields an estimate of $450M for the improved outcomes directly attributable to the systematic review.

2.5.1. Limitations

This approach is a flexible one, and the majority of parameters are supported by a strong base of evidence. However, it relies strongly on assumptions about the information environments with and without the systematic review. It also summarizes away the nuances of each individual trial. The fact that [BCK+96] had patients with higher fracture risk than [CBT+98] could potentially be used to tailor treatment effect estimates based on risk, but that was not done in this analysis.

An important uncertainty that affects the model is the long-term effects of bisphos-
2.5 Combining the Pieces to Calculate the Population Value of Information

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Baseline Value</th>
</tr>
</thead>
</table>
| Durability(t)   | $D(t) = \begin{cases} 
1 & \text{if } t \in [1, 5] \\
(10 - t) \times 0.2 & \text{if } t \in (5, 10] \\
0 & \text{if } t > 10 
\end{cases}$ |
| Population(t)   | $98,777,136$   |
| $P(\text{Implementation})$ | $25\%$ |
| $\beta$         | $\frac{1}{1.03}$ |

Table 2.9: Summary of assumptions driving the population expected value of information

Phosphonate use and the ideal treatment duration. According to a recent systematic review, “the evidence base here is especially thin”[CND+12]. Most trials are for a fixed, relatively short period of time, but the model must simulate some patients for upwards of thirty years. Some evidence suggests that treatment longer than five years does not reduce non-vertebral fractures, but this finding was not consistent in all subgroups[CND+12]. If the benefit does persist after stopping, its duration is unclear. These issues are complicated by the discovery that changes in BMD while on treatment account for only a small portion of changes in fracture risk[CND+12], so long-term studies would need to monitor fractures as the key endpoint to answer this open question. In order to be conservative, we assumed that if the conditions were met for a given treatment strategy, treatment was provided—regardless of how long the patient had been on the medication. If benefits are fully realized after five years and persistent, the optimal treatments would treat more people, because the effective cost of fracture reduction would decrease.

Some portions of the health system that bear on this model are not included. One of these is screening costs: we assume that bone mineral density is known, and do not include a cost of obtaining this information. Additionally, no adverse effects associated with the treatment are included. While the side effects of alendronate are generally mild[CND+12], they may lead to fewer treatments being given in reality. The most important of these may be atypical subtrochanteric fractures, which are very rare but appear to have a large increase in relative risk for long-term users of oral bisphosphonates[CND+12].

Finally, we do not explicitly model the creation and interaction of multiple reviews on similar or identical topics. In the case of low bone density, AHRQ sponsored a large review and corresponding update[MAC+07, CND+12], the Cochrane Collaboration produced a similar review looking only at alendronate[WCP+11], and the UK’s Health Technology Assessment program conducted a large systematic review and cost-effectiveness analysis[SJD+05], and others were also conducted. Each new review can add value by either supporting previous work, identifying new studies, or reaching a new population of readers, but it is unlikely that the value is as large as if no other reviews had been conducted on the topic.
2.6. US Benefits of Computer Aided Screening

With a reasonable estimate of the societal benefit provided by one systematic review, we are able to crudely estimate the total benefit created by widespread use of computer aided screening. Work reported in Chapter 4 demonstrated that it is possible to avoid the screening of approximately 70% of the articles returned by a search, while maintaining acceptable levels of sensitivity. Because it is the largest funding source for systematic reviews in the United States, the AHRQ Evidence-Based Practice Center (EPC) program is a reasonable target for adoption of these tools. The EPC program completes approximately 15 reviews per year. By avoiding the screening of 70% of the articles on an average review with 9,000 citations (see Appendix B.6 for details), direct savings per year on AHRQ EPC reports is

$$15_{\text{reviews}} \times 0.7 \times 9000_{\text{citations}} \times time_{\text{PerCitation}} \times 2_{\text{reviewers}} \times reviewerWage$$

At approximately 1 minute per citation, and a burdened labor cost of $200 per hour, the total savings is $630K per year. These savings could be used to reduce the program budget or to fund additional reviews.

Unfortunately, detailed cost estimates of the pieces of systematic reviews, which would allow us to put the cost savings in the context of the AHRQ Evidence-Based Practice Center (EPC) budget, are not available. Screening is a time-consuming part of the review, but reviewing the key questions, devising a search strategy, synthesizing the information, and writing are also important and time-consuming. The total savings of $630K is approximately equal to the total cost of one review; thus we assume that one additional review could be created each year if this screening approach is used across the EPC program.

Calculating the societal benefits from funding additional reviews is much less straightforward given the availability of only a single point estimate. Using the aggressive assumption that these reviews create as much value as the osteoporosis review, computer aided screening generates $450M in societal benefit each year. Another possibility is that the current prioritization practice is highly effective, and that this case study was for a particularly valuable review. Under the assumption that the value of each review is 80% of the value of the review prioritized above it, the 16th review in a year creates 3.5% of the value created by [CND+12], or $16M.

It is important to note that AHRQ is far from the only sponsor of systematic reviews. The Cochrane Collaboration creates a large number of reviews, as do the UK’s Heath Technology Assessment program and many others. The methods presented in this thesis are in no way limited to AHRQ reviews. Furthermore, the screening time improvements are likely to increase as more reviews utilize the tools, as the algorithms can be refined and prior reviews can guide the initialization. The benefits presented also only consider the population of the United States, but the efficacy of particular pharmaceuticals is likely to be quite consistent across countries. These factors could combine to greatly increase the value of providing computer-assistance in the systematic review process.
2.7. Conclusion

We provided the first estimates of the value of a systematic review, finding that the value of information made accessible by the systematic review of low bone density corresponds to a societal benefit of $1.5B, with $450M being attributable to the review itself. Furthermore, presented a flexible new cost-effectiveness model dubbed FOCUS, and used that model to find optimal treatment strategies and long-term costs of osteoporotic fractures. If the computer-assisted screening tools presented in Chapters 4 and 5 are applied to all AHRQ EPC systematic reviews, $630K per year could be saved. Alternatively, if these savings are applied to new reviews, we estimate that the societal benefit falls between $16M to $450M each year. Expanding the use outside of the United States would lead to even greater benefits.
3. Screening Times in Systematic Reviews

3.1. Introduction

Despite intense investigation into computer-assisted methods for abstract screening in systematic reviews [CHB04, CAM09, CAM10, Coh08a, CAM12, DSH+13, FIM10, FIM+11, HSS+12, KMI+09, MKI+10, WSB+12a, WSB+10, WTL+10, WSB+12b, CHPY06, DSH+12, WSBT10], no existing work has published detailed timing statistics on the various phases of the screening process. Such research is in fact possible; using data extracted from reviews conducted by the Southern California EPC using DistillerSR\(^1\) (a web-based screening tool), the time spent in each level of screening can be analyzed. We analyzed these data for seven large systematic reviews, each of which utilized between three and ten different screeners. Each review contained up to four different screening levels: title, abstract, full-text, data abstraction. Title screening uses the paper title alone to judge whether the paper may be relevant. Abstract screening provides more information, and will often answer additional questions regarding study design, country, population etc., thus giving a better idea if the paper is relevant to the search. When the full text is reviewed, a comprehensive understanding of the paper is gained, and detailed questions answered. If the study will be included in the review, study data are entered into a common format that will allow comparison between studies–this is the data abstraction phase.

When most machine learning results are reported, they report measures that treat all articles equally. This work finds that skipping an article that would be rejected in a batch at the title screening phase would save only 1.4 seconds on average, while avoiding an article that would have passed through an individual title screening and abstract screening would save an average of 73.4 seconds. Appropriately pruning an article that would have required title, abstract, and full text screening would save 370 seconds. By presenting analysis of a large number of ratings, we hope researchers can better target their effort and assess the impact of innovations.

\(^1\)http://systematic-review.net/
### 3.2. Background

In addition to aiding estimates of work savings attained by techniques that avoid screening certain articles, this analysis provides valuable behavioral data about a repetitive task with significant cognitive requirements. We hypothesize that both experience and focus will be related to screening times. Whilst learning curves have been studied extensively, session dynamics remain hitherto little understood. The most useful body of literature appears in studies investigating interruptions on knowledge workers, with particular emphasis on computer programmers. Early work on instant messaging\[CCH01\] found a high cost of interruptions for tasks involving “fast searches.” These findings could easily be applied to the systematic review process, as title and abstract screening are similar to “fast searching” in terms of the cognitive processes involved. In a diary study of knowledge workers, tasks that were interrupted took more than twice as long as other tasks[CHW04], furthermore, [MGH05] found that interrupted tasks took more than 40% longer than uninterrupted tasks.

Not all studies have found interruptions to have strong deleterious effects. Monk\[Mon04\] found the unexpected result that frequent interruptions did not lead to increased time requirements for the primary task when compared to infrequent interruptions, hypothesizing that workers are more aggressive in goal setting when they are likely to be interrupted. This agrees with the later finding in [MGK08] that interruptions did not slow task completion time, but did increase stress in the subjects. Unfortunately, the tasks used in these studies are considerably different from citation screening; in [Mon04], the task was to program a VCR to record a show, and subjects were interrupted either every 30 (infrequent) or 10 seconds (frequent). As this is not a repetitive task like citation screening, there is less benefit from building and refining mental models as users make progress. The task in [MGK08] involved responding to varied emails as a Human Resources employee returning from vacation. As the subjects of the different emails varied considerably, there is little benefit to building a mental model of the task.

The literature on interrupted programmers, however, provides more relevant findings. While programming tasks are generally much less repetitive than citation screening, both require building and maintaining a mental model of the task. Title screening has a relatively small memory requirement, but quick identification of information relevant to multiple questions in the abstract or full-text phases utilize large amounts of attentive, associative, and conceptual memory[Par13]. There are indeed certain programming tasks (for example refactoring), that require similar changes across many pieces of code[Par13]. In an observational study of interruptions on 86 programmers and a survey of 414, 10-15 minutes were taken to resume coding after an interruption, and it was rare to have more than one two-hour uninterrupted session per day[Par13].
### 3.3 Data

The data from seven different systematic reviews are presented below; they comprise a total of 110,319 ratings of 42,693 articles. The reviews covered the topics of Bariatric Surgery\[MGL^+13\], Health IT (forthcoming), Mass Casualty Events\[TRF^+13\], Palliative Care (unpublished work), Smoking Cessation (forthcoming), Quality Improvement (forthcoming), and Vitamin D (forthcoming). The number of articles rated, and the total number of ratings by review and level within DistillerSR are presented in Tables 3.1 and 3.2.

As the review progresses, some articles are removed from consideration, and the remainder proceed to a more time-intensive screening process. Although this initial title screening is rapid, it is performed for the largest number of articles and thus is also time consuming. Because the judgments of a single reviewer are imperfect, it is prudent for articles to be screened by two separate reviewers at each level. Initially, there may also be a period where articles are screened more than twice to ensure that all reviewers render similar judgments.

Whilst timing data from DistillerSR are a valuable resource, they are not without draw-
Figure 3.1.: Fit of lognormal distribution to abstract screening time for the Palliative Care review. The dotted vertical lines show the data that were used to fit the distribution, and the red marks are observations tagged as outliers. As expected, the outliers correspond to very large recorded rating times.

backs. The most notable issue is that screening times can only be inferred by taking the difference between the time one web page is loaded and the time at which is the result is submitted. If a reviewer leaves the screen open whilst engaging in other activities, DistillerSR would record an inaccurately long review time. Fortunately, the large number of articles screened at each level is quite large, enabling data-driven outlier detection. To do this, we fit a distribution to a portion of the data, and used that to infer cases where the reviewer was likely engaged in other actions while the screen was loaded.

We determined that many of the screening times were very closely approximated by a lognormal distribution. An illustration of the observed values vs. their predictions from a lognormal distribution is shown for the abstract-level screening on the smoking cessation review in Figure 3.1. Outliers were identified using an approach from [vdL10], which is detailed in Algorithm 3.1. For each review-screening level, we sorted the screening times and fit a lognormal distribution to the data between the 5th and 90th percentiles. Rating times where the residual (based on predictions from the fitted distribution) does not fall in a 98% confidence interval were tagged as outliers. For the purpose of regression models, these were excluded; however, they were replaced with the median for estimating total screening times. The parameters were selected using expert judgment based on experience with the process by which they were generated.

We hypothesized based on personal experience performing article screening that time within a given rating session would be related to screening time. Reviewing these articles, particularly beyond the title stage, require the construction of complex mental models, which must be recalled or reconstructed each time a judgment is given. Longer
Algorithm 3.1 Assignment of Outlier Status from \[vdL10\]

Require: Ordered rating submission times for one reviewer $T$; number of ratings $N$; inverse CDF of hypothesized distribution family $F^{-1}$; $F_{\text{min}}, F_{\text{max}}$ portion of data to use in fitting; $\alpha \pm$ confidence level for denoting outliers

* $X \leftarrow \text{Sort}(T[1 : N - 1] - T[0 : N - 2])$
* $\hat{F}_i \leftarrow \frac{i}{N+1}$
* $\Lambda \leftarrow \{ i \in \{1, 2, \ldots, N\} | F_{\text{min}} \leq \hat{F}_i \leq F_{\text{max}} \}$
* $\hat{\theta} \leftarrow \text{arg min}_{\theta} \sum_{i \in \Lambda} [\ln(X_i) - \ln(F^{-1}(\hat{F}_i|\theta))]^2$
* $Y \leftarrow 0 \text{ array of length } N \text{ holding outlier assignments}$
* $\varepsilon_i \leftarrow \ln(X_i) - \ln(F^{-1}(\hat{F}_i|\hat{\theta}))$
* $\hat{\sigma}_E^2 \leftarrow |\Lambda|^{-1} \sum_{j \in \Lambda} \varepsilon_j^2$, Assuming residuals $\varepsilon_i \sim \mathcal{N}(0, \sigma_E^2)$
* $\ell_{\alpha}^\pm \leftarrow \sqrt{2\hat{\sigma}_E^2} \text{erf}^{-1} \pm \{(1 - 2\alpha \pm)\}$ rejection thresholds
* $i \leftarrow 0$
while $i < N$:
if $i \leq \min(\Lambda)$ and $\varepsilon_i \leq \ell^-$ and ($i = 1$ or $Y_{i-1} = 1$):
* $Y_i \leftarrow 1$
elseif $i \geq \max(\Lambda)$ and $\varepsilon_i \geq \ell^+$ and ($i = N - 1$ or $Y_{i+1} = 1$)
* $Y_i \leftarrow 1$
endif
endwhile
Return: Outlier assignments $Y$
Algorithm 3.2 Assignment of Session Identifiers

Require: Temporally ordered rating submission times for one reviewer $X$, with outliers labeled; number of ratings $N$
* $S ← 1$ array of length $N$ holding session assignments
* $i ← 1$
* $s ← 1$
while $i < N$:
if $X_i - X_{i-1} > 5$ minutes or $X_i$ is an outlier:
* $s ← s + 1$
endif
* $S_i ← s$
endwhile
Return: Session labels $S$

Sessions may lead to reduced rating times, as the expert may recall the information more quickly. However, if the session persists for an extended period of time, the expert may become fatigued—increasing rating times. While “sessions” are not tagged in the data, we generated session identifiers using a simple algorithm. A session is defined as a sequence of ratings that continues as long as (1) no outliers are present, and (2) each action was preceded by another rating within five minutes. The first condition is of particular importance, as ratings identified as outliers correspond to times when the expert left the task and engaged in another activity, thus resetting working memory.

3.4. Descriptive Statistics

In order to determine where future effort should be allocated, it is useful to identify the time taken by various phases of the systematic review procedure. Table 3.3 presents these figures for our seven systematic reviews. Additionally, while estimating labor costs and potential savings, it is helpful to understand the average time taken to rate single articles at different stages.

One interesting tool only available during the title screening phase is the ability to view multiple titles on the page at once. If many are easily identified as irrelevant, it may be faster to review multiple articles at once. This not only reduces loading times, it also allows the screener to more efficiently scan the list for relevant citations. In a “one-per-page” scenario, the user reads a title, selects the appropriate button, and submits the form. The loading time along with the act of scanning the page for the first title and the submission buttons are fixed costs, and the batching approach can amortize those costs. A graphical depiction of the observed relationship is shown in Figure 3.2, which strongly supports the hypothesis that batching titles is more efficient; articles screened individually took an average of 6.6 seconds each, while batched title screening took a mere 0.7 seconds per title.
3.4 Descriptive Statistics

<table>
<thead>
<tr>
<th>Review</th>
<th>Title</th>
<th>Abstract</th>
<th>Full-text</th>
<th>Data Abstraction</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bariatric Surgery</td>
<td>N/A</td>
<td>2.5</td>
<td>N/A</td>
<td>38.5</td>
<td>41.0</td>
</tr>
<tr>
<td>Health IT</td>
<td>1.0</td>
<td>29.6</td>
<td>59.5</td>
<td>N/A</td>
<td>90.1</td>
</tr>
<tr>
<td>Mass Casualty</td>
<td>1.7</td>
<td>107.3</td>
<td>122.0</td>
<td>101.5</td>
<td>332.6</td>
</tr>
<tr>
<td>Palliative Care</td>
<td>5.4</td>
<td>71.4</td>
<td>62.3</td>
<td>105.4</td>
<td>244.5</td>
</tr>
<tr>
<td>Smoking Cessation</td>
<td>15.2</td>
<td>48.0</td>
<td>40.8</td>
<td>25.1</td>
<td>129.2</td>
</tr>
<tr>
<td>Quality Improvement</td>
<td>-</td>
<td>66.3</td>
<td>119.6</td>
<td>391.7</td>
<td>577.6</td>
</tr>
<tr>
<td>Vitamin D</td>
<td>-</td>
<td>64.8</td>
<td>44.7</td>
<td>86.1</td>
<td>195.6</td>
</tr>
</tbody>
</table>

**Table 3.3.:** Hours spent in each review-level. “N/A” occurs when screening was conducted outside of DistillerSR, and “-” occurs when the researchers did not utilize a separate title screening stage. Note that not all screening was done in DistillerSR, so this is an estimate of the lower bound on screening time.

<table>
<thead>
<tr>
<th>Level</th>
<th>Seconds per Article</th>
<th>Articles Screened per Hour</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>1.77</td>
<td>2,033</td>
</tr>
<tr>
<td>Abstract</td>
<td>30.1</td>
<td>120</td>
</tr>
<tr>
<td>Full Text</td>
<td>153</td>
<td>23.5</td>
</tr>
<tr>
<td>Data Abstraction</td>
<td>1,666</td>
<td>2.16</td>
</tr>
</tbody>
</table>

**Table 3.4.:** Average Rating Times by Level. Both columns show the same data, but are both valuable due to the large difference in magnitudes.

**Figure 3.2.:** Relationship between batch size and average screening time per title
Table 3.5.: Average Screening Time by Level and Review

<table>
<thead>
<tr>
<th>Review</th>
<th>Title (sec)</th>
<th>Abstract (sec)</th>
<th>Full Text (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bariatric Surgery</td>
<td>-</td>
<td>20.19</td>
<td>N/A</td>
</tr>
<tr>
<td>HIT</td>
<td>0.35</td>
<td>46.56</td>
<td>5.95</td>
</tr>
<tr>
<td>Mass Casualty</td>
<td>1.11</td>
<td>77.87</td>
<td>2.57</td>
</tr>
<tr>
<td>Palliative Care</td>
<td>0.81</td>
<td>40.08</td>
<td>1.46</td>
</tr>
<tr>
<td>Quality Measures</td>
<td>-</td>
<td>12.85</td>
<td>3.70</td>
</tr>
<tr>
<td>Smoking Cessation</td>
<td>6.61</td>
<td>39.99</td>
<td>3.69</td>
</tr>
<tr>
<td>Vitamin D &amp; Calcium</td>
<td>-</td>
<td>22.76</td>
<td>1.42</td>
</tr>
</tbody>
</table>

Figure 3.3.: Distribution of articles rated per session. As expected, the more intensive stages generally involve screening fewer articles in a session, with Data Abstraction having a large probability mass at 1.

It is worth noting that there is considerable variability in the rating times between reviews, which can be partially explained by the review process. Smoking Cessation did not batch articles in the title screening phase, and required significantly more time per title. Reviews utilizing a separate title screening stage showed increased time spent per article in the abstract screening stage; this increase is presumably because the remaining articles were more likely to be relevant—obviously irrelevant articles would have already been removed. Finally, it is also useful to inspect the distribution of session length, shown in Figure 3.3, in order to assess the reasonability of the session assignment algorithm and understand screener behavior.
3.5. Factors Associated with Screening Times

To investigate the relationships between various observable factors and the time taken to screen an article, we fit an ordinary least squares model of the form:

\[
\text{seconds}_{isr} = \alpha + \beta_0 \text{passedLevel}_{isr} + \beta_1 \text{numWithinSession}_{isr} + \beta_2 \text{numWithinSession}^2_{isr}
\]

\[
+ \beta_3 \text{numWithinLevel}_{isr} + \beta_4 \text{numWithinLevel}^2_{isr}
\]

\[
+ \beta_5 \text{numWithinReview}_{isr} + \beta_6 \text{numWithinReview}^2_{isr} + \gamma_r + \nu_{rs} + \epsilon_{isr}
\]

Here, \(\text{seconds}\) is the time taken to screen the article, the \(i\) subscript denotes a specific rating, \(s\) is a particular screener, and \(r\) is a particular review. The inclusion of squared terms allows for nonlinear effects while retaining a simple interpretation. For title screening only, linear and quadratic terms for the number of articles in a batch were also included. The \(\gamma_r\) and \(\nu_{rs}\) are review and review-screener fixed effects. While inclusion of screener-specific fixed effects would be preferable, the screeners’ identities were anonymized and therefore could not be linked across reviews.

The identity of the screener and the review being conducted will almost certainly be related to screening times, as reading speeds and attention spans differ, as do the complexity of review topics. More broadly, we believe that the number of articles screened in the review, level, and session may relate to screening times consistently across screeners and reviews. The number of articles screened for a given review and level will measure experience with the topic and specific questions respectively; possible session effects have been addressed in Section 3.3.

Table 3.6 shows the regression coefficients, with Table 3.6a displaying the coefficients in natural units and Table 3.6b showing the coefficients standardized as in [Gel08]. In this standardization procedure, outcome variables and continuous explanatory variables are all de-meaned and divided by two standard deviations, which makes coefficients comparable across screening levels and within each model.

Perhaps the most interesting results concern the relationship between rating time and the experience-related variables. The coefficients on experience within a level and review behave largely as expected, and session-related effects are strong. For title screening, batch size is far more important than any of these learning or attention effects.

3.6. Application of Results

To compare metrics based on time savings with those that implicitly assume screening times to be uniform, we consider the case of a soon-to-be-published update to a systematic review on low bone density [MAC+07, CND+12]. In this search, 10,780 articles were retrieved with three topics of interest: efficacy, adverse events, and adherence. DistillerSR was not used, so actual timing is not available. Instead, we will use the averages and coefficients from the above analysis to better understand the savings of the machine learning approach.
(a) Raw regression coefficients showing the relationship between screening time and observable factors.

<table>
<thead>
<tr>
<th></th>
<th>Title (ms)</th>
<th>Abstract (ms)</th>
<th>Full Text (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>passedLevel</td>
<td>210.42*</td>
<td>43,88.34*</td>
<td>22.31*</td>
</tr>
<tr>
<td>numWithinSession</td>
<td>-3.90*</td>
<td>-138.20*</td>
<td>-2.28*</td>
</tr>
<tr>
<td>numWithinSession^2</td>
<td>0.01*</td>
<td>0.40*</td>
<td>0.03*</td>
</tr>
<tr>
<td>numWithinLevel</td>
<td>-0.26</td>
<td>-10.45*</td>
<td>-0.19*</td>
</tr>
<tr>
<td>numWithinLevel^2</td>
<td>0.00</td>
<td>0.00*</td>
<td>0.0002*</td>
</tr>
<tr>
<td>numWithinReview</td>
<td>0.20</td>
<td>7.56*</td>
<td>-0.05*</td>
</tr>
<tr>
<td>numWithinReview^2</td>
<td>0.00</td>
<td>-0.001*</td>
<td>0.000006*</td>
</tr>
<tr>
<td>batchSize</td>
<td>-77.79*</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>batchSize^2</td>
<td>0.76*</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

(b) Standardized regression coefficients. All continuous inputs are divided by two times their standard deviation to enable natural comparability with the binary predictors [Gel08]. Screening times are also standardized to allow comparison across levels. Coefficients are then multiplied by 10 for ease of display.

Table 3.6.: Regression model results: * denotes $p < .05$. 
3.6 Application of Results

<table>
<thead>
<tr>
<th>Round</th>
<th>Number Reviewed</th>
<th>Passed Title</th>
<th>Passed Abstract</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>495</td>
<td>127</td>
<td>81</td>
</tr>
<tr>
<td>2</td>
<td>504</td>
<td>130</td>
<td>64</td>
</tr>
<tr>
<td>3</td>
<td>500</td>
<td>71</td>
<td>32</td>
</tr>
<tr>
<td>4</td>
<td>500</td>
<td>53</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>514</td>
<td>16</td>
<td>10</td>
</tr>
<tr>
<td>Total</td>
<td>2,511</td>
<td>397</td>
<td>203</td>
</tr>
</tbody>
</table>

Table 3.7.: Literature Flow through the machine learning algorithm on the 2014 update for the low bone density review.

Because a production system was not available, batches of articles were selected to be sent to the researchers for review. This process was repeated five times, with approximately 500 articles in each batch and 2,511 articles being rated in total. The complete statistics are shown in Table 3.7. In order to reduce “noise” in the ratings, articles were rated at the title level, reconciled, and screened at the abstract level before ratings were utilized by the machine learning algorithm. Out of 200 relevant articles identified, 79 were found in the first batch, and 64 in the next, while only 10 out of 514 articles in the fifth batch were deemed relevant. This indicates that the initialization from a prior review and the learning from these ratings were quite successful. Given that only 2,511 of 10,780 were manually screened, the reviewers were able to omit 8,269 articles from the screening load. Since this review utilized duplicate screening, this is a saving of 16,538 ratings. Because title screening was used, the majority of the remaining citations have most likely been removed at this inexpensive level. If we assume a total of 20 articles avoided would have passed the title stage\(^2\), this is an additional 40 screenings omitted at the abstract level. The average time spent per title that is judged irrelevant is 1.37 seconds, and for those judged relevant is 2.70 seconds, while irrelevant abstracts required 26.4 seconds on average and relevant ones 43.0 seconds. The total savings is therefore \(16,498 \times 1.37 + 40 \times 2.7 + 40 \times 26.4 = 23,766\) seconds, or 6.6 hours.

If we were to conduct a similar analysis using the number or percentage of articles avoided only, our estimates would be very different. By standard machine learning metrics, this review was a resounding success—only 23.3\% of articles were rated, which would reflect a work savings of 76.7\%. Given that the smoking cessation review was slightly smaller and took approximately 130 hours, it would be reasonable to estimate savings of approximately 100 hours. Unfortunately, the majority of screening time is spent on articles that are truly relevant, the room for savings is less than previously imagined.

\(^2\)It is difficult to estimate the number of irrelevant articles that would have passed the first phase, as our sample is heavily biased towards articles likely to be relevant.
3.7. Conclusion

The triage process used in systematic reviews is effective. Time requirements at early stages of a review are small, but grow quickly; more time is spent per article on relevant documents than those that are irrelevant. The portion of time spent in abstract, full-text, and data abstraction varies considerably across reviews, but relationships between levels are consistent.

Alternative structuring of the screening process can lead to appreciably different time requirements. Primary controllable drivers identified are batching at the title phase and session length in all phases. Presenting multiple titles on screen at once is associated with an order of magnitude decrease in screening time. At the abstract level, articles rated near the ideal session length required an average of 12 seconds less than at the start of a session— a decrease of more than \( \frac{1}{3} \) of the average rating time.

These findings have two important implications for improving the speed and cost of systematic review preparation. Firstly, user interface improvements have the potential to produce considerable time savings. For example, the smoking cessation review, which did not batch titles, spent 15 hours in title screening. Calculated using the average batched time, this may have required only 98 minutes. Furthermore, the reviews on Quality Improvement and Vitamin D did not utilize title screening at all, and had longer abstract screening phases as a result. This finding is somewhat weakened by the observation that the time per abstract is smallest in reviews lacking a title screening phase. It is also possible that batching at the abstract phase would be beneficial, or that a combined batched title/abstract phase would be ideal. Figure 3.4 shows an example of how such an interface could be constructed. Using a keyboard to navigate through articles would enable question answering about that article, and would display the selected abstract below the title. If the title is enough to accept or reject the article, the abstract can be ignored. This interface could realize the gains of rapid title screening and batching without the rework associated with having to evaluate the same article separately at both the title and abstract phase.

To obtain realistic views of the time saved by machine learning methods, it is important to consider a baseline that incorporates the reality of a tiered screening system. Using an upcoming update to a review on low bone density as an example, work savings calculated naïvely and the estimated true work savings differ by a factor of 17.

Future work could continue this research in a number of productive directions. Creating streamlined interfaces and conducting experiments utilizing them could potentially reduce time requirements dramatically. Many refinements to the regression analysis of section 3.5 could enhance understanding. The number of questions on each form could be included, along with the length of the text being reviewed. Because we found that the time is related to relevance, it may be useful to use the presence of a separate title screening phase as an exogenous shock in an instrumental variable approach. This would allow for causal estimates of the change in time requirements as additional irrelevant articles are screened without requiring experimentation. Armed with this information,
3.7 Conclusion

**Figure 3.4.:** An illustration of an interface that would enable more rapid title and abstract screening. Obviously irrelevant articles can be skipped, and more detail is available by looking down at the abstract.

Researchers could make well-informed decisions about search criteria that balance the cost of screening additional articles against the probability that valuable information will be gained from using a wider search.
4. De Novo Reviews

4.1. The Systematic Review Process

A systematic review is a very large and potentially costly undertaking. The annual budget for the AHRQ Comparative Effectiveness program was $15M per year between 2005 and 2007. In 2007, the program released only four systematic reviews, eight summary guides, and four research reports[Ag11]. Given these outputs, in addition to the nomination, prioritization, selection, and management of research, it is likely that the cost per review was approximately $500K. The systematic review process includes searching, screening, data abstraction, and analysis and whilst each process presents its own unique challenges, this paper focuses on the screening phase, as we believe that it has the greatest potential to reduce the research burden.

The search phase outputs a list of documents that may be relevant to the given systematic review; depending on the size and subject matter of the review this list can contain any number of articles, from a few hundred to tens of thousands. In the screening phase, expert reviewers (often with an MD or PhD) review the title and abstract of each article to determine if it meets the predefined inclusion criteria. Customarily, two reviewers screen each article, and if the article is deemed relevant after the title and abstract are read, the article is ordered and the entire article reviewed. Articles that are judged as relevant after full text screening then proceed to the next phase of the review, where researchers use them to answer the questions of interest.

Because it can be challenging to determine if a study should be included, even these experts reviewers do not always agree on inclusion decisions. Cohen’s Kappa is a widely used measure of agreement beyond what would happen by random chance\(^1\), and we find that values of 0.6-0.7 are common; this indicates “substantial agreement”[LK77].

Underlying attempts to provide computer assistance to researchers conducting systematic reviews is the belief that information contained in early inclusion decisions can be applied to other articles under consideration. For example, by marking articles utilizing animal studies as irrelevant, the tool will determine that other animal studies are unlikely to meet the inclusion criteria. With this assistance, requiring duplicate reviews, or even single reviews of every article may be excessively conservative.

\(^1\)Cohen’s Kappa is defined as \(\kappa = \frac{P(\text{agreement}) - P(\text{chance agreement})}{1 - P(\text{chance agreement})}\)
4.2. Domain-Specific Concerns

At first glance, the use of machine learning to determine the relevance of articles appears to be a standard text classification problem. Accurate algorithms for text classification have existed for decades [SD98], but a number of issues make their application to this field challenging. Most notably, the exclusion of false negatives is far more costly than the inclusion of false positives. The dearth of positive examples exacerbates this issue; many reviews result in far fewer than ten percent of retrieved articles being deemed relevant, with some reaching less than one percent that meet all the inclusion criteria. Each false positive (when an irrelevant article is mistakenly deemed relevant) requires an expert review of the article, which may take up to 30 seconds. False negatives, conversely, have the potential to jeopardize the validity of the review. If readers of the review are aware of a critical study that has been omitted, it undermines the conclusions made in the review.

The majority of machine learning algorithms minimize a convex approximation of the 0-1 loss function, a process which attempts to maximize accuracy. This is problematic in the case of systematic reviews because it treats false positives and false negatives identically. For example, if only 5% of articles are relevant, a classifier that defines 100% of articles as irrelevant will achieve 95% accuracy while providing no practical value. Simple methods of cost-sensitive classification exist, such as weighting positive examples more heavily [VCC99], and Section 4.6 shows that this works quite well in practice. When combined with active learning, which finds relevant articles more quickly than random sampling, benefits improve markedly. However, despite the success of this simple approach, it is valuable to have a repository of multiple successful approaches, as the optimal approach for one dataset may be inferior for another.

An additional noteworthy characteristic of systematic reviews is that the set of documents is fixed. Our goal is therefore not to build a model that can accurately classify articles into groups of relevant and irrelevant articles, but rather to classify a fixed set of documents. Following the principle of transduction outlined by Vapnik, our approach should find the desired answer directly, rather than first solving a more general problem [Vap00]. If an example is particularly difficult for our classifier, perhaps because it contains many unseen words, an expert should screen that example rather than risk a classification error. This implies that there must be a feedback mechanism between the classifier and the expert screener.

Given that such feedback should exist, classifiers that are computationally expensive to train are infeasible. Any time spent on computation after receiving classification from the expert but before the expert is given additional articles to screen must be viewed as wasted time. Coupled with the high dimensionality of text in the common vector space model, this places non-trivial restrictions on the models available. It excludes algorithms that require extensive hyperparameter tuning, and is difficult to train nonlinear models within the time requirements. Algorithms that can incorporate new classifications without retraining the entire model, and those that provide a natural measure of uncertainty are therefore especially attractive.
Additionally, considerable work has been done to index and categorize medical text; categorization on the basis of study type or Medical Subject Headings (MeSH) can be very useful for classification [Coh08a]. However, high-quality systematic reviews must reference multiple databases in order to reduce the potential for bias [Age08], as some databases do not perform these categorizations, and no two databases categorize in the same manner. Training a model on these database-specific features would result in a model that performs very poorly with other sources, which could bias the synthesis. Similarly, features like the impact factor of the journal in which an article was published may be highly predictive, but would bias the model towards popular journals. Exclusion of articles for unprincipled reasons such as these is undesirable to a systematic review and as a result, we limit ourselves to using only the articles’ title and abstract, which are available from any database.

Finally, though it will not be addressed in detail here, the true classes of interest are often more complex than “relevant” and “irrelevant.” Systematic reviews address a number of key questions, and associating articles with the appropriate key question is desirable. One article can be relevant to multiple questions; this is an instance of multi-label classification rather than multi-class classification. Furthermore, articles can be excluded for multiple reasons: some may not address the topic, some may be the wrong type of study, others may lack data or be poorly implemented. It is desirable, and in some cases required, to classify the excluded articles by their reason for exclusion. A number of multi-label classification techniques have been proposed, including an array of problem transformation approaches [TK07], supervised topic models [RCSS11], and tree-based modifications to support vector machines [CL11a], and application of these techniques for systematic reviews would be a valuable avenue for future research.

### 4.3. Proposed Approach

#### 4.3.1. Notation and Background

Consider a set of documents $\mathcal{D}$, where each element $x_i \in \mathbb{R}^d$ is the feature vector for an article, dominated by its words represented using the vector space model of text. For a cost, we are able to request a label $y_i \in \{+1, -1\}$ for any document in unlabeled set $\mathcal{U}$, transferring it to the set of labeled documents $\mathcal{L}$. The set of all relevant documents is $\mathcal{D}^+$, and the irrelevant documents are $\mathcal{D}^-$, and the same superscript applies to $\mathcal{U}$ and $\mathcal{L}$. We train a classifier that outputs a score $s(x) : \mathbb{R}^d \rightarrow \mathbb{R}$ and a threshold $t$ such that a document $x$ is predicted to be relevant if $s(x) \geq t$. The goal is to identify all relevant articles at the lowest possible effort:

$$\min |\mathcal{L}| \text{ s.t. } \mathcal{D}^+ = \mathcal{L}^+$$

A simple tabular illustration of possible outcomes for a single prediction is shown in Table 5.2. A wide array of performance measures can be derived from this table. The
most commonly used metric is accuracy, which is $\frac{tp + tn}{|D|}$. When false positives and false negatives have different costs, it is often desirable to create measures that reflect those concerns. Within information retrieval, two of the most commonly used are precision and recall, which are defined as

$$\text{Precision} = \frac{tp}{tp + fp}$$
$$\text{Recall} = \frac{tp}{tp + fn}$$

Precision can be understood as the portion of articles retrieved that are relevant, and recall as the portion of all relevant articles that are retrieved.

An alternative way to understand the performance of a binary classifier without specifying a threshold is the Receiver Operator Characteristic (ROC) curve, which plots the true positive rate vs. the false positive rate as the threshold $t$ is varied. A common measure of the discriminative performance of a classifier is the area under the ROC curve (AUC). If the prediction $s(x)$ is assigned at random, the true positive and false positive rates will be approximately equal, and AUC will be close to 0.5. A perfect classifier can obtain a true positive rate of 1.0 with a false positive rate of 0, yielding an AUC of 1.0. This measure can also be interpreted as the probability that a randomly chosen relevant article will be more highly rated than a randomly chosen irrelevant article [Han82].

A popular statistical learning algorithm is the Support Vector Machine (SVM), which aims to find a hyperplane that separates the two classes (relevant and irrelevant) with the maximum margin [Vap00]. In this setting, the margin is the distance between the hyperplane and the nearest training instances. Formally, the optimization problem carried out for a linear, soft margin SVM with no bias term is

$$\min_{w, \xi \geq 0} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{|\mathcal{L}|} \xi_i$$
$$\text{s.t. } y_i w^T x_i \geq 1 - \xi_i, \ i = 1, \ldots, |\mathcal{L}|$$

where $w$ is a weight vector, $C$ is a tunable parameter that trades off training and generalization error, and $\xi$ is a slack variable reflecting the SVM margin [HM13]. This technique has strong theoretical properties [Vap00] and empirical performance, but can perform sub-optimally when the number of positive instances is very different from the number of negative instances [HM13]. The following sections present two methods that have shown to be successful at maintaining computational efficiency while performing well under this class imbalance.
4.3 Proposed Approach

4.3.2. Differential Error Costs

The simplest approach that our experiment identifies as successful uses differential error costs to increase the importance of relevant articles. In the above formulation of the SVM, the constant \( C \) penalizes the misclassification of a relevant article identically to the misclassification of an irrelevant article. Veropoulos et. al. introduced the ability to differentially penalize each type of error [VCC99] by instead carrying out the following optimization:

\[
\min_{w, \xi \geq 0} \frac{1}{2} \|w\|^2 + C^+ \sum_{\{i|y_i=+1\}} \xi_i + C^- \sum_{\{i|y_i=-1\}} \xi_i \\
\text{s.t.} \ y_i w^T x_i \geq 1 - \xi_i, \ i = 1, \ldots, |L|
\]

While the empirical performance has been mixed [VCC99, AKJ04], it is an extremely simple modification that adds little to computational time. Because we follow the heuristic of setting \( \frac{C^+}{C^-} = \frac{\hat{C}^+}{\hat{C}^-} \) [AKJ04], it introduces no additional hyperparameters.

4.3.3. Classification as Bipartite Ranking

A slightly more complex approach to increasing the sensitivity of the classifier involves converting the classification problem into one of ranking. Binary classification seeks to learn a decision function \( c : X \rightarrow \{-1, +1\} \), where \( X \) is the observable feature space. Letting \( X^{(2)} \) be the feature space over all observable pairs, a ranker learns a preference function \( r : X^{(2)} \rightarrow \{-1, +1\} \), where +1 denotes that the first element is preferred, and −1 that the second is preferred [BBB+08]. Because this definition only considers pairs of instances rather than list-wise preferences, it is sometimes referred to as bipartite ranking.

In statistical learning theory, a reduction is the conversion of a complex problem into a simpler one. Binary classification is the most basic learning problem, thus many reductions transform complex goals into this class. A common measure of the impact of such a reduction is regret, which is “the difference between the incurred loss and the lowest achievable loss on the problem” [BBB+08]. Analyzing regret instead of error can be beneficial because it decouples noise inherent in the problem from errors caused by the specific solution to the problem. As Balcan et. al. demonstrate [BBB+08], there exists a simple reduction from bipartite ranking to binary classification that has good performance guarantees. Specifically, using a binary classifier with regret \( r \) on the classification problem induced by classifying all pairs of articles gives an upper bound on AUC regret of \( 2r \).

The advantage of this result is that it provides a way to use any binary classifier to maximize AUC by transforming the inputs. Since, as mentioned above, AUC bears closer resemblance to the true goal in systematic reviews than accuracy, this allows
the considerable research in binary classification to be applied to optimizing a more natural object function.

Mechanically, the reduction involves training a binary classifier on pairs of documents. A problem with this reduction is that it involves transforming the inputs from $\mathcal{R}^{n \times d}$ to $\mathcal{R}^{n^2 \times d}$. While this may not be problematic in all applications, recall that it is valuable to retrain the model in response to expert feedback, and as such, long training times incur a real cost. Depending on the size of the review, two options are promising. When the size of the data is modest, the algorithm presented in [Joa05] obtains a high quality solution with complexity $O(n \log n)$. If the review is exceedingly large (for example, the approximately 50,000 article review in [FIM+11]), a stochastic pairwise solution can be used to enable online learning.

### 4.3.3.1. Two Useful Optimization Algorithms

The proposed ranking algorithm, like the differential cost approach above, is based on support vector machines [Joa02, Joa05, Scu09]. A highly flexible modification of the basic SVM problem, presented in [Joa05], implicitly transforms the problem into one that allows for a variety of multivariate loss functions $\Delta$:

$$
\min_{w, \xi \geq 0} \frac{1}{2} \|w\|^2 + C\xi \\
\text{s.t.} \forall \bar{y}' \in \bar{Y} \setminus \bar{y} : w^T[\Psi(\bar{x}, \bar{y}) - \Psi(\bar{x}, \bar{y}')] \geq \Delta(\bar{y}', \bar{y}) - \xi
$$

Where $w$ is a weight vector, $C$ is a tunable parameter that trades off misclassification error and SVM margin, $\Psi$ is a function assessing the agreement between $x$ and $y$, $\xi$ is a slack variable, and $\Delta$ is a multivariate loss function [Joa05]. In the case of the ranking SVM, $\Delta$ is $1 - \text{ROC Area}$, which is proportional to the number of swapped pairs. More generally, this formulation can be used when the loss function being optimized cannot be calculated at the level of a single training instance, enabling metrics including $F_1$, precision-recall breakeven point, precision@k, and recall@k [Joa05].

When used with $\Delta = 1 - \text{ROC Area}$, this formulation seeks to minimize the number of times an irrelevant article is rated above a relevant article and minimize the magnitude of the weight vector $w$, using $C$ to trade off between these two goals. One way to solve this optimization problem is to use the 1-slack formulation found in [JFY09], which is a cutting plane algorithm. This is an efficient batch algorithm with good practical performance and is used in the results below. While there is no incremental version of this algorithm, when we train on batch $t + 1$, the weight vector $w$ can be initialized with the weights learned from batch $t$. This “warm start” can reduce the number of iterations required to reach the desired solution quality.

There is one important distinction to note: in the differential error cost approach, the loss can be calculated at the level of a single document, while that is not possible in the 1-slack formulation. Although that simplicity enables faster training, the multivariate loss formulation adds considerable flexibility. Consider, for example, a case where the
search terms are very weak indicators of relevance, causing a very large number of articles to be returned, a very small portion of which can be considered relevant. Furthermore, different reviewers have different preferences regarding batch size. Using a multivariate loss function, we are able to train an SVM to optimized precision@k, where k is the batch size for each rater. By using a tailored measure, importance is placed at the appropriate point in the ranked list, while the simpler DEC approach will emphasize errors that would never be considered for inclusion in the next batch.

For cases where this algorithm requires excessive processing time, the Stochastic Pairwise Descent algorithm of [Scu09] can provide considerable improvement. It can be run without reading all examples into memory, and the estimator can be utilized before training completes—which may be useful for ensuring that the user does not wait beyond a given computational time allowance. While the 1-slack algorithm utilizes the framework of structural support vector machines to optimize non-standard performance metrics, this approach explicitly samples pairs of documents, and at each iteration, takes a step in the direction of the gradient specified by the following optimization problem

$$\min_w \frac{C}{2} \|w\|^2 + \frac{1}{|P|} \sum_{(a,b,y_a,y_b) \in P} \max(0, 1 - \langle w, \text{sign}(y_a - y_b)(a - b) \rangle)$$

Here, w and C are as above, a and b are individual articles in D, and P is the set of all pairs of articles. This algorithm involves randomly sampling pairs of documents, and taking the appropriate gradient step. Because this amounts to learning a binary classifier over pairs of examples, the convergence properties of gradient descent are maintained[Scu09]. Given that a user will often wait on computations, this is a significant practical improvement in cases with many ratings. Another benefit of this approach is that any gradient-based update rule can be used. In particular, the normalized adaptive gradient descent algorithm of [RML13] was nearly always superior to Pegasos, Passive-Aggressive, and standard gradient descent.

### 4.3.4. Active Learning

Active learning is the practice of allowing a classifier to choose the articles for which labels are requested, ideally improving performance with fewer training instances[Set10]. This is frequently useful, as obtaining labeled data can be costly. In the case of systematic reviews, the experts conducting the review are expensive; the average salary of early-career physician-researchers exceeded $150,000 in 2010-2011[JGS+13]. In an effort to reduce these costs, the field of active learning has proposed a large number of algorithms by which instances can be selected, however there appears to be little empirical difference between simple and complex methods[Scu07].

The most common type of active learning is to select the instances about which the classifier is most uncertain[Set10]. In our case, the learner outputs a score $$s(x) = \langle w, x \rangle$$. A
threshold \( t \) is then selected, above which we predict that an instance is relevant\(^2\). Therefore, the most uncertain instances are those close where \( |\langle w, x \rangle - t| \) is small. Theoretically, in the case of support vector machines, this is an effective approach, as it corresponds to the optimal decrease in the area of the version space\(^{[TK02]}\), defined as the set of hyperplanes that separate the classes. When the version space is large, the classifier is highly uncertain about the true decision boundary. Other commonly used techniques in active learning include query-by-committee, expected model change, Fisher information ratio, and density weighted methods. The Query-by-committee technique involves constructing a number of classifiers that are consistent with the current training data, then selecting instances based on the disagreement of the component classifiers\(^{[Set10]}\). Expected model change selects instances where the decision boundary of the classifier is likely to change if the additional label is obtained. This is quite clearly illustrated by the “expected gradient length” approach of \(^{[SC08]}\). When training using gradient descent, the gradient is \( \nabla \ell(L; w) \), and with \( \langle x, y \rangle \) added becomes \( \nabla \ell(L \cup \langle x, y \rangle ; w) \). For binary classification, the greatest model change will be given by

\[
\arg \max_x P(y = 1|x) \| \nabla \ell(L \cup \langle x, y \rangle ; w) \| + P(y = -1|x) \| \nabla \ell(L \cup \langle x, y \rangle ; w) \| 
\]

Because \( \nabla \ell(L; w) \) will be small, as the learner is optimized to convergence,

\[
\| \nabla \ell(L \cup \langle x, y \rangle ; w) \| \approx \| \nabla \ell(\langle x, y \rangle ; w) \|
\]

, is the gradient that would be computed for updates. Essentially, this finds the instance resulting in the largest change in the weight vector, weighting potential changes by their probability. The Fisher information ratio approach utilizes the likelihood function to select instances that contain the variables contributing to model uncertainty for \( U \)[Set10].

The Fisher information matrix is the variance of the derivative of log-likelihood

\[
\mathcal{I}(\theta) = -\int_x P(x) \int_y P(y|x) \frac{\partial^2}{\partial \theta^2} \log P(y|x)
\]

Because this is a matrix, one way to find the optimal instance is to minimize the trace norm of the Fisher Information Ratio

\[
\arg \min_x \text{tr}(\mathcal{I}_x(\theta)^{-1}\mathcal{I}_U(\theta))
\]

where \( \mathcal{I}_x(\theta) \) measures the Fisher information with respect to one unlabeled instance, and \( \mathcal{I}_U(\theta) \) remains the same over all unlabeled instances. Though theoretically strong, this approach has considerable time and space complexity, as \( \mathcal{I}(\theta) \in \mathbb{R}^{d \times d} \), and \( d \) is large for our textual data. The final approach discussed here is the class of density-weighted methods. These seek to balance informativeness (as assessed by a method such as uncertainty sampling) and representativeness in order to avoid querying outliers. Whilst the Fisher information ratio approach naturally avoids outliers, other methods

\(^2\)This is equivalent to learning a bias term on a model and using a threshold of 0; however, we separate the steps to make the control of sensitivity more straightforward.
4.3 Proposed Approach

may select instances that will change the model considerably but only be relevant for a very small subset of the unlabeled instances [Set10].

Much of the literature on active learning focuses on sequential learning, where a single instance is labeled, the classifier is updated, and a single additional instance is presented to the screener. However, because the fraction of relevant articles is often very low, many screeners prefer to view multiple articles at the same time. In a later chapter, we find that this approach can decrease title screening time seven-fold. However, if many unlabeled articles are similar, selecting a batch of the most uncertain articles may be far from optimal, as one article may resolve the uncertainty. Various approaches to this problem have been proposed, ranging from heuristics to formal optimization problems (see [Set10] for an overview). However, the empirical change in performance between a simple approach and more careful batch construction appears to be small [HJZL06].

It is worth noting that the benefits provided by utilizing active learning for systematic reviews differ slightly from those customarily desired. Our goal is to discover all of the relevant articles with the smallest possible labeling effort. By carefully selecting documents for labeling, we allow the classifier to be improved quickly, following the standard justification for this tool in addition to also obtaining labels for the difficult-to-classify articles. Since these are no longer in the set of documents on which an error can be made, performance is further increased. Finally, because relevant articles must be reviewed at some point, the effective cost of labeling a relevant document is zero. As such, the optimal strategy for this problem will sample from areas that have a higher score than a strategy that simply seeks to construct a maximally accurate classifier.

Formalizing that notion, let the utility of reviewing an article be given by

\[
U(x_i) = \begin{cases} 
1 & \text{if } y_i = +1 \\
-\alpha & \text{if } y_i = -1 
\end{cases}
\]

where \(\alpha < 1\) is the cost of reviewing an irrelevant article, then our goal is the maximize the total utility obtained. While SVM outperformed logistic regression in our experiments, to motivate the instance selection, assume that \(P(y_i = +1|x, w) = \frac{1}{1 + \exp(-w^T x_i)}\), with prior beliefs about the weights \(w\) as \(w_i \sim \mathcal{N}(0, \sigma^2)\). Each time a batch is requested, we would ideally perform the following optimization

\[
\max_{\mathcal{X}_A} \left( \sum_{x \in \mathcal{X}_A} U(x) \right) + \beta V(\mathcal{X}_A, w_t)
\]

where \(V(\mathcal{X}, w)\) represents the best possible utility given the unlabeled documents in \(\mathcal{X}\) and the weight vector \(w\), \(0 < \beta < 1\) is the discount rate, and we seek to select a set \(\mathcal{X}_A\) for expert review. An exact solution to this problem is difficult to obtain–\(V(\mathcal{X}, w)\) depends upon quality of the model encoded in \(w\) and both the relevance of the articles remaining in \(\mathcal{X}\) and the relationships between their features and unobserved labels. Dynamic programming is a common approach to solving such a problem [SB98],
but in this case is infeasible because of the combinatorial nature of set selection. While a more sophisticated approximation or heuristic would be valuable, the present work utilizes a greedy heuristic with forced exploration. Inspired by the success of the $\epsilon$-greedy heuristic on the problem of multi-armed bandits [CL11b], we implement a similar approach to instance selection. In this strategy, the most likely article to be relevant is chosen with probability $1 - \epsilon$, and a random article is chosen with probability $\epsilon$. Intuitively, the greedy portion of the algorithm optimizes $U(x)$, while the exploration present in the random selection improves $V(X_-, w_t)$. This occurs both by improving the quality of $w_t$ as well as increasing the expected number of relevant articles in $X_\sim A$. In the batch setting, we first drew a number of articles as “exploration” from a $n_{\text{explore}} \sim \text{Binomial}(n, \epsilon)$, where $n$ is the batch size: these articles were randomly selected from the unlabeled articles. The remaining articles were the $n - n_{\text{explore}}$ with the highest $s(x)$: those that were most likely to be relevant. When no articles had been rated, the batch was chosen at random.

### 4.4. Performance Evaluation

Ideally, we would like to carry out the following optimization problem, capturing all relevant articles at the lowest possible labeling cost:

$$\min |L| \quad \text{s.t.} \quad L^+ = D^+ \quad (4.1)$$

However, directly attempting to carry out this optimization has a number of problems. Most importantly, it opaquekly combines three related problems: ranking, thresholding, and stopping. Ranking and thresholding have been addressed in previous sections, but development of a stopping criterion will be expanded upon here. In order for this approach to be useful, the experts will stop rating documents before all have been rated, however, in the traditional document classification problem, the training phase and testing phase are distinct. When learning incrementally, we can transfer documents from $U$ to $L$ as long as it is beneficial, but doing so has a cost. Additionally, all of the documents in $U$ that are predicted to be relevant must be reviewed.

Noise in the labels will create undesirable results. While the 0-1 loss function is quite tolerant of noise, this is not the case with alternative loss functions [MS13]. Given that 0-1 loss, the optimization of which is equivalent to maximizing accuracy, is unhelpful for the systematic review problem, possession of high-quality labels is critical. The optimization specified in Equation 4.1 is especially sensitive to noise, as it includes no capability to omit a single positive article from the set of labeled documents.

Because of the difficulty of measurement, we rely on proxy metrics related to our true interests. Cohen [CHPY06] presented the first metric that was specifically tailored to systematic reviews: Work Saved over Sampling (WSS), which is defined as

$$WSS = \frac{tn + fn}{|D|} - 1 + \frac{tp}{tp + fn}$$
In published literature, Work Saved over Sampling at 95% Recall (WSS@95) has been reported, and is similarly defined as

\[ WSS@95 = \frac{tn + fn}{|D|} - 0.05 \]  (4.2)

This measures the reduction in workload obtained versus what would be obtained by random sampling at the level of recall provided by the classifier. While work reduction is the goal, this metric is not ideal, as false negatives here result in better values, despite the fact that false negatives are the most costly error type. It also is evaluated only on a testing set, which leads to overestimation of the performance. To illustrate this, consider a case where the stopping rule is “label half of the documents.” An estimate of WSS will be obtained, but would require a manual review of half of the documents, so the true work savings is only 50% of the estimated WSS. Despite this flaw, we present this metric in the following section for ease of comparison with published results.

An attempt to rectify this issue of ignoring effort on the training set is presented in [WTL+10], where the complementary metrics of yield and burden are presented. These are essentially adaptations of sensitivity and specificity that incorporate the number of instances labeled directly into the metrics. Letting a superscript denote performance on the unlabeled set and \( L \) on the labeled set, they are defined as\(^3\)

\[
Yield = \frac{tp^C + tp^U}{tp^C + tp^U + fn^U} \\
Burden = \frac{|L| + tp^U + fp^U}{|D|}
\]

While useful, practitioners are unlikely to explicitly consider a tradeoff between these two metrics, making the presentation of them unnecessarily complex. Using a fixed recall threshold, one can estimate the work savings as simply

\[
Work Savings = \frac{|U|}{|D|}
\]

This measure is very similar to \((1 - Burden)\), but makes the assumption that the documents remaining in the unlabeled set will remain unlabeled. This simple metric is used in our results that utilize an incremental experimental protocol.

**4.5. Data**

Because each review is different, and the tool must perform consistently, it was important to compare potential approaches across many datasets. A set of 15 drug reviews

---

\(^3\)The terms here refer to the entries in a binary confusion matrix: true positive(tp), true negative(tn), false positive(fp), and false negative(fn)
(Which we refer to as OHSU after the releasing university) was published for the research community in [Coh08b], and has been used in multiple publications [CHPY06, MKI+10, Coh08a, CAM09]. In all of the publications utilizing the OHSU dataset, five repetitions of 2-fold cross-validation were used, with no concern given to the sequential nature of systematic reviews. One important limitation of these data is that, because articles were matched with ratings after the reviews took place, only 30%-50% of the articles rated by the reviewers were included in the public data. The learning task is more difficult at early stages of the review, thus work savings on these cases are underestimated. To remedy this, we also present the results of a large review of Low Bone Density (LBD)[MAC+07, CND+12] that will be revisited in the future chapters of this work. Unlike the OHSU data, this had labels by different questions of interest: efficacy, adherence, and adverse effects.

The descriptive statistics of the OHSU datasets are shown in Table 4.2 and the LBD reviews in Table 4.3. The OHSU data were beneficial because they covered a variety of topics and has been used as a benchmark, while the LBD data were more realistic. All text was prepared using the same process. First, the title and abstract were concatenated, then tokenization and stemming were performed using the method proposed in [JZ07], which is tailored to biomedical text. Stop words were removed, and numbers replaced with a special numeric token. A small number of “meta-features” were added: the number of words, characters, maximum and average word length, number of p-values, and number of numbers. The Term Frequency-Inverse Document Frequency (TF-IDF) representation of the text was calculated and normalized to length one and the meta-features were individually scaled to fall between 0 and 1. The final feature vector was the concatenation of this TF-IDF vector and the meta-features. In future work, it may be both useful and feasible to include other meta features, such BM-25 score[S.E95] for the document and the query specified in the search strategy.

4.6. Results

Two sets are results are presented below. Firstly, we used five replications of two-fold cross-validation, which allowed for simple comparisons with previously published literature. Secondly, a sequential experiment was conducted, which used the $\epsilon$-greedy approach to select articles for rating.

4.6.1. OHSU Experiments

The metric presented in Table 4.4 is WSS at 95% recall, as defined in 4.2. This metric measured the reduction in workload versus what would be obtained by random sampling at the level of recall provided by the classifier, and has been reported by two other groups investigating the same problem[MKI+10, Coh08a]. In Table 4.4, Cohen 2008 and Matwin 2010 previously published results that utilized title, abstract, MeSH headings,
### Table 4.2:
Summarized literature flow from the OHSU reviews. Note that this dataset includes only those that could be matched to PubMed by the original authors; the original studies reviews were larger. N refers to the number of articles—either in the full review, or were deemed relevant at the abstract or full-text levels.

<table>
<thead>
<tr>
<th>Review</th>
<th>N</th>
<th>Abstract % Inc</th>
<th>N</th>
<th>% Inc</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACE Inhibitors</td>
<td>2,543</td>
<td>183</td>
<td>41</td>
<td>2%</td>
</tr>
<tr>
<td>ADHD</td>
<td>851</td>
<td>84</td>
<td>20</td>
<td>2%</td>
</tr>
<tr>
<td>Antihistamines</td>
<td>310</td>
<td>92</td>
<td>16</td>
<td>5%</td>
</tr>
<tr>
<td>Atypical Antipsychotics</td>
<td>1,120</td>
<td>363</td>
<td>146</td>
<td>13%</td>
</tr>
<tr>
<td>Beta Blockers</td>
<td>2,072</td>
<td>302</td>
<td>42</td>
<td>2%</td>
</tr>
<tr>
<td>Calcium Channel Blockers</td>
<td>1,218</td>
<td>279</td>
<td>100</td>
<td>8%</td>
</tr>
<tr>
<td>Estrogens</td>
<td>368</td>
<td>80</td>
<td>80</td>
<td>22%</td>
</tr>
<tr>
<td>NSAIDs</td>
<td>393</td>
<td>88</td>
<td>41</td>
<td>10%</td>
</tr>
<tr>
<td>Opioids</td>
<td>1,915</td>
<td>48</td>
<td>15</td>
<td>1%</td>
</tr>
<tr>
<td>Oral Hypoglycemics</td>
<td>503</td>
<td>139</td>
<td>136</td>
<td>27%</td>
</tr>
<tr>
<td>Proton Pump Inhibitors</td>
<td>1,333</td>
<td>238</td>
<td>51</td>
<td>4%</td>
</tr>
<tr>
<td>Skeletal Muscle Relaxants</td>
<td>1,643</td>
<td>34</td>
<td>9</td>
<td>1%</td>
</tr>
<tr>
<td>Statins</td>
<td>3,465</td>
<td>173</td>
<td>85</td>
<td>2%</td>
</tr>
<tr>
<td>Triptans</td>
<td>671</td>
<td>218</td>
<td>24</td>
<td>4%</td>
</tr>
<tr>
<td>Urinary Incontinence</td>
<td>327</td>
<td>78</td>
<td>40</td>
<td>12%</td>
</tr>
</tbody>
</table>

### Table 4.3:
Summary of the Low Bone Density reviews[MAC+07, CND+12]. The “any inclusion” row is less than the sum of other inclusions because some articles can be relevant for multiple categories.

<table>
<thead>
<tr>
<th>Inclusions</th>
<th>2007</th>
<th>2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>Articles</td>
<td>14,700</td>
<td>7,089</td>
</tr>
<tr>
<td>Efficacy Inclusions</td>
<td>218</td>
<td>55</td>
</tr>
<tr>
<td>Adverse Effect Inclusions</td>
<td>279</td>
<td>101</td>
</tr>
<tr>
<td>Adherence Inclusions</td>
<td>-</td>
<td>72</td>
</tr>
<tr>
<td>Any Inclusions</td>
<td>382</td>
<td>188</td>
</tr>
</tbody>
</table>
Table 4.4.: Work saved over sampling at 95% recall results on a benchmark dataset indicate state-of-the-art performance despite having access to less powerful features. The first two columns are previously published studies, while the third has not before been used in bioinformatics. Results in the first three columns are on five replications of 2-fold cross validation. The best score on each dataset is in bold, though ties cannot be depicted because confidence intervals are not presented in [Coh08a, MKI+10].

<table>
<thead>
<tr>
<th></th>
<th>Cohen 2008</th>
<th>Matwin 2010</th>
<th>SVM(^W)</th>
<th>Rank SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACE Inhibitors</td>
<td>73%</td>
<td>52%</td>
<td>81%</td>
<td>79%</td>
</tr>
<tr>
<td>ADHD</td>
<td>53%</td>
<td>62%</td>
<td>69%</td>
<td>72%</td>
</tr>
<tr>
<td>Antihistamines</td>
<td>24%</td>
<td>15%</td>
<td>17%</td>
<td>24%</td>
</tr>
<tr>
<td>Atypical Antipsychotics</td>
<td>17%</td>
<td>21%</td>
<td>15%</td>
<td>15%</td>
</tr>
<tr>
<td>Beta Blockers</td>
<td>47%</td>
<td>37%</td>
<td>57%</td>
<td>54%</td>
</tr>
<tr>
<td>Calcium Channel Blockers</td>
<td>43%</td>
<td>23%</td>
<td>37%</td>
<td>38%</td>
</tr>
<tr>
<td>Estrogens</td>
<td>41%</td>
<td>38%</td>
<td>36%</td>
<td>32%</td>
</tr>
<tr>
<td>NSAIDs</td>
<td>67%</td>
<td>53%</td>
<td>69%</td>
<td>68%</td>
</tr>
<tr>
<td>Opioids</td>
<td>36%</td>
<td>55%</td>
<td>50%</td>
<td>48%</td>
</tr>
<tr>
<td>Oral Hypoglycemics</td>
<td>14%</td>
<td>9%</td>
<td>12%</td>
<td>12%</td>
</tr>
<tr>
<td>Proton Pump Inhibitors</td>
<td>33%</td>
<td>23%</td>
<td>44%</td>
<td>37%</td>
</tr>
<tr>
<td>Skeletal Muscle Relaxants</td>
<td>37%</td>
<td>27%</td>
<td>44%</td>
<td>48%</td>
</tr>
<tr>
<td>Statins</td>
<td>49%</td>
<td>32%</td>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>Triptans</td>
<td>35%</td>
<td>27%</td>
<td>17%</td>
<td>20%</td>
</tr>
<tr>
<td>Urinary Incontinence</td>
<td>43%</td>
<td>30%</td>
<td>46%</td>
<td>43%</td>
</tr>
<tr>
<td>Weighted Average</td>
<td>45%</td>
<td>36%</td>
<td>49%</td>
<td>49%</td>
</tr>
</tbody>
</table>

Table 4.4.: Work saved over sampling at 95% recall results on a benchmark dataset indicate state-of-the-art performance despite having access to less powerful features. The first two columns are previously published studies, while the third has not before been used in bioinformatics. Results in the first three columns are on five replications of 2-fold cross validation. The best score on each dataset is in bold, though ties cannot be depicted because confidence intervals are not presented in [Coh08a, MKI+10].

and publication type information from MEDLINE. The SVM\(^W\) and Rank SVM column reflects this paper’s contribution, and utilized only the title and abstract, without incorporating active learning as mentioned in Sections 4.3.4. Despite intentionally omitting some of the most useful features[Coh08a], our approaches described above achieved the highest work savings on 9 of 15 reviews, versus 5 for [Coh08a] and 2 for [MKI+10], and both approaches were indistinguishable in terms of weighted average savings.

Instead of comparing to published results, Table 4.5 uses identical data for each algorithm, and illustrates the strong performance of SVM approaches that correct for class imbalance. The first three columns contain commonly used machine learning algorithms, though the logistic regression implementation uses the same differential error costs as in the SVM described in Section 4.3.2. The “Bagged Undersampled SVM”, (BU-SVM) is very similar to the approach from [WTL+10]. In this approach, each base predictor is a linear SVM trained on all relevant articles and an equal number of randomly selected irrelevant articles. The predictions from a number of such classifiers were averaged, yielding the final prediction. The original implementation trained a total of 33 classifiers—11 on title words, 11 on abstract words, and 11 on MeSH headings. Our implementation

\(^4\text{SVM}^W\) is used to denote that each class’ errors are “weighted” differently
4.6 Results

differs in that it only trained classifiers on the combined title and abstract text, and used seven base classifiers. Naive Bayes did not perform well, and logistic regression was also regularly outperformed. The comparison between the unweighted SVM and the one with differential costs showed that this simple correction can be highly effective. Using these data, there would be little reason to prefer any one of the three rightmost methods for performance reasons alone.

Finally, Table 4.6 shows the results obtained using the sequential experimental protocol. In this setup, 25 articles were chosen randomly, then the $\epsilon$-greedy strategy carried out with batches of 25 articles and $\epsilon = 0.05$. New articles were labeled until at least 95% of the relevant articles are labeled. The results indicate that work savings are still feasible, and of comparable magnitude to the cross-validation experiments. Furthermore, the classifiers that performed well in earlier experiments also performed well with this more realistic protocol. It is difficult to obtain significant savings for the small reviews (e.g. antihistamines), but the weighted average work savings of our approach was still nearly 50%. It is important to note that, because of the way these data were constructed, only 30% to 51% of the total articles were included. Because the most difficult part of the learning task occurs when few ratings are available, the work savings with 100% of the articles will almost certainly be greater.

4.6.2. Low Bone Density Experiments

Examining the LBD data, relative performance is similar to that described previously, but savings are much greater. The cross-validated WSS@95% ranged from 79% to 90% for the three competitive approaches, with batch sequential work savings that ranged from 82% to 95%. While the individual question level results were not directly of practical importance, they did serve two important purposes. Firstly, they served as additional test cases that had particularly extreme class imbalances (1-2% relevant). It is clear that the 3 SVM-based approaches that adjust for class imbalance, particularly when combined with active learning, had no problem with this skew. Secondly, the relationship between those and the “any inclusion” models is valuable. The “any” results were no better than the poorest performing question-specific model. If the models were of equal quality, one would expect that the “any” performance would be superior, as false positives for one question may be relevant for another question. This outcome indicates that the use of question-level labels can produce a more discriminative model which is reasonable given that all proposed models learn a linear decision boundary. In [CND+12, MAC+07], the adherence articles were largely observational studies, whilst efficacy results were either clinical trials or systematic reviews. When grouped into one “relevant” label, the linear model was unable to learn that observational studies were irrelevant for efficacy, but valuable for adherence. This observation further suggested that multi-label learning algorithms may be useful for modeling the relationships between each question, potentially increasing the savings.
### Table 4.5:

Work saved over sampling at 95% recall results using commonly used classifiers on identically prepared data. Classifiers with a $W$ superscript use weights to allow for differential error costs as in [VCC99], where the importance of each instance is the inverse of its class proportion. The three SVM-based approaches that account for class imbalance are all competitive.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Naive Bayes</th>
<th>Log Reg$^W$</th>
<th>SVM</th>
<th>SVM$^W$</th>
<th>BU-SVM</th>
<th>RankSVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACE Inhibitors</td>
<td>53%</td>
<td>81%</td>
<td>73%</td>
<td>81%</td>
<td>80%</td>
<td>79%</td>
</tr>
<tr>
<td>ADHD</td>
<td>56%</td>
<td>29%</td>
<td>54%</td>
<td>69%</td>
<td>54%</td>
<td>73%</td>
</tr>
<tr>
<td>Antihistamines</td>
<td>23%</td>
<td>29%</td>
<td>30%</td>
<td>11%</td>
<td>15%</td>
<td>15%</td>
</tr>
<tr>
<td>Atypical Antipsychotics</td>
<td>33%</td>
<td>34%</td>
<td>30%</td>
<td>40%</td>
<td>36%</td>
<td>37%</td>
</tr>
<tr>
<td>Beta Blockers</td>
<td>27%</td>
<td>32%</td>
<td>42%</td>
<td>19%</td>
<td>15%</td>
<td>15%</td>
</tr>
<tr>
<td>Calcium Channel Blockers</td>
<td>33%</td>
<td>34%</td>
<td>32%</td>
<td>30%</td>
<td>36%</td>
<td>32%</td>
</tr>
<tr>
<td>Estrogens</td>
<td>48%</td>
<td>46%</td>
<td>66%</td>
<td>63%</td>
<td>69%</td>
<td>68%</td>
</tr>
<tr>
<td>NSAIDs</td>
<td>12%</td>
<td>28%</td>
<td>29%</td>
<td>12%</td>
<td>12%</td>
<td>12%</td>
</tr>
<tr>
<td>Opioids</td>
<td>48%</td>
<td>46%</td>
<td>66%</td>
<td>63%</td>
<td>69%</td>
<td>68%</td>
</tr>
<tr>
<td>Oral Hypoglycemics</td>
<td>12%</td>
<td>14%</td>
<td>43%</td>
<td>43%</td>
<td>44%</td>
<td>44%</td>
</tr>
<tr>
<td>Proton Pump Inhibitors</td>
<td>16%</td>
<td>16%</td>
<td>36%</td>
<td>36%</td>
<td>37%</td>
<td>37%</td>
</tr>
<tr>
<td>Skeletal Muscle Relaxants</td>
<td>32%</td>
<td>41%</td>
<td>32%</td>
<td>27%</td>
<td>41%</td>
<td>44%</td>
</tr>
<tr>
<td>Statins</td>
<td>4%</td>
<td>4%</td>
<td>49%</td>
<td>49%</td>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>Urinary Incontinence</td>
<td>49%</td>
<td>49%</td>
<td>49%</td>
<td>49%</td>
<td>49%</td>
<td>49%</td>
</tr>
<tr>
<td>Weighted Average</td>
<td>35%</td>
<td>42%</td>
<td>47%</td>
<td>47%</td>
<td>47%</td>
<td>47%</td>
</tr>
</tbody>
</table>

**Table 4.5:** Work saved over sampling at 95% recall results using commonly used classifiers on identically prepared data. Classifiers with a $W$ superscript use weights to allow for differential error costs as in [VCC99], where the importance of each instance is the inverse of its class proportion. The three SVM-based approaches that account for class imbalance are all competitive.
4.6 Results

<table>
<thead>
<tr>
<th>Drug Class</th>
<th>Log Reg&lt;sup&gt;W&lt;/sup&gt;</th>
<th>SVM&lt;sup&gt;W&lt;/sup&gt;</th>
<th>BU-SVM</th>
<th>RankSVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACE Inhibitors</td>
<td>82%</td>
<td>82%</td>
<td>83%</td>
<td>83%</td>
</tr>
<tr>
<td>ADHD</td>
<td>59%</td>
<td>54%</td>
<td>67%</td>
<td>52%</td>
</tr>
<tr>
<td>Antihistamines</td>
<td>3%</td>
<td>9%</td>
<td>11%</td>
<td>3%</td>
</tr>
<tr>
<td>Atypical Antipsychotics</td>
<td>26%</td>
<td>17%</td>
<td>17%</td>
<td>26%</td>
</tr>
<tr>
<td>Beta Blockers</td>
<td>65%</td>
<td>63%</td>
<td>68%</td>
<td>61%</td>
</tr>
<tr>
<td>Calcium Channel Blockers</td>
<td>46%</td>
<td>47%</td>
<td>49%</td>
<td>42%</td>
</tr>
<tr>
<td>Estrogens</td>
<td>32%</td>
<td>27%</td>
<td>39%</td>
<td>34%</td>
</tr>
<tr>
<td>NSAIDs</td>
<td>63%</td>
<td>63%</td>
<td>63%</td>
<td>63%</td>
</tr>
<tr>
<td>Opioids</td>
<td>22%</td>
<td>24%</td>
<td>27%</td>
<td>13%</td>
</tr>
<tr>
<td>Oral Hypoglycemics</td>
<td>14%</td>
<td>14%</td>
<td>14%</td>
<td>13%</td>
</tr>
<tr>
<td>Proton Pump Inhibitors</td>
<td>40%</td>
<td>44%</td>
<td>39%</td>
<td>45%</td>
</tr>
<tr>
<td>Skeletal Muscle Relaxants</td>
<td>24%</td>
<td>41%</td>
<td>39%</td>
<td>39%</td>
</tr>
<tr>
<td>Statins</td>
<td>44%</td>
<td>54%</td>
<td>44%</td>
<td>50%</td>
</tr>
<tr>
<td>Triptans</td>
<td>7%</td>
<td>27%</td>
<td>36%</td>
<td>27%</td>
</tr>
<tr>
<td>Urinary Incontinence</td>
<td>31%</td>
<td>37%</td>
<td>39%</td>
<td>39%</td>
</tr>
<tr>
<td>Weighted Average</td>
<td>44%</td>
<td>48%</td>
<td>48%</td>
<td>46%</td>
</tr>
</tbody>
</table>

Table 4.6.: Work savings at 95% recall results using an batch-sequential experimental protocol. Each classifier is used in an ϵ-greedy approach with 25 articles per batch, and continues until at least 95% of relevant articles have been labeled. Note that random sampling would obtain 5% savings on average. Again, the three SVM-based methods perform well.

<table>
<thead>
<tr>
<th>Year</th>
<th>Adverse Effects</th>
<th>Naive Bayes</th>
<th>Log Reg&lt;sup&gt;W&lt;/sup&gt;</th>
<th>SVM</th>
<th>SVM&lt;sup&gt;W&lt;/sup&gt;</th>
<th>BU-SVM</th>
<th>RankSVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>Adverse Effects</td>
<td>79%</td>
<td>84%</td>
<td>84%</td>
<td>86%</td>
<td>86%</td>
<td>85%</td>
</tr>
<tr>
<td>2012</td>
<td>Adverse Effects</td>
<td>72%</td>
<td>76%</td>
<td>73%</td>
<td>79%</td>
<td>78%</td>
<td>79%</td>
</tr>
</tbody>
</table>

Table 4.7.: Cross-validated WSS@95% Recall results for the LBD reviews. Again, multiple approaches are competitive, and the SVM-based methods that account for class imbalance work well.
### Table 4.8.

Incremental work savings at 95% recall for the LBD reviews. All methods generated substantial work savings.

<table>
<thead>
<tr>
<th></th>
<th>Log Reg&lt;sup&gt;W&lt;/sup&gt;</th>
<th>SVM&lt;sup&gt;W&lt;/sup&gt;</th>
<th>BU-SVM</th>
<th>RankSVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adverse Effects</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>88%</td>
<td>89%</td>
<td>89%</td>
<td>89%</td>
</tr>
<tr>
<td>2012</td>
<td>79%</td>
<td>82%</td>
<td>82%</td>
<td>82%</td>
</tr>
<tr>
<td>Efficacy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>91%</td>
<td>90%</td>
<td>91%</td>
<td>91%</td>
</tr>
<tr>
<td>2012</td>
<td>82%</td>
<td>84%</td>
<td>89%</td>
<td>85%</td>
</tr>
<tr>
<td>Adherence</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>94%</td>
<td>94%</td>
<td>95%</td>
<td>95%</td>
</tr>
<tr>
<td>Any Inclusion</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>90%</td>
<td>90%</td>
<td>90%</td>
<td>90%</td>
</tr>
<tr>
<td>2012</td>
<td>84%</td>
<td>85%</td>
<td>86%</td>
<td>86%</td>
</tr>
<tr>
<td>Average</td>
<td>87%</td>
<td>88%</td>
<td>89%</td>
<td>88%</td>
</tr>
</tbody>
</table>

**4.7. Conclusions**

In this chapter, we have explained characteristics of the article screening problem for systematic reviews and presented two approaches that yield significant work savings. This allows for high-quality synthesis of medical literature with considerable cost savings. The results are generally superior to state-of-the-art methods while respecting the requirements that a computer-aided system would face (primarily computational time and lack of indexed features). In addition to demonstrating savings, the explication of the unique aspects of this problem will serve as building blocks for future research. In the following chapters, we expand this approach to account for updating existing reviews, evaluate the time savings, and identify the effects of this innovation on the health system more broadly.
5. Updating Existing Reviews

5.1. Introduction

Systematic reviews are utilized for a variety of purposes, including clinical practice guideline development, cost-effectiveness modeling, research prioritization, and understanding the strengths and weakness of existing research. For the majority of these applications, it is critical that the findings remain up-to-date. An outdated review could potentially be of less value than a nonexistent review, as outdated evidence may provide practitioners with a false sense of confidence. Some experts recommend that reviews be updated every two years [The11, MTT+08, GTT+10], but this rarely happens in practice [JCJ+98, MTT+07]. Furthermore, it is difficult to determine when a review requires an update, as the determination depends on whether new or differing evidence has been produced.

In 2007, a quasi-random sample of 100 systematic reviews published between 1995 and 2005 was analyzed in [SSA+07]. Defining important changes as those that change the estimated effect sizes by at least 50% or statistical significance, they found that the median time period between publication and important changes occurring to be 5.5 years. More importantly, although 25% of reviews were up-to-date after 10 years, 7% of reviews’ conclusions were out of date by the time the report was published [SSA+07]. Given the rate at which trials and systematic reviews are published has approximately doubled since 2007, this presents a troubling issue. Time since publication can be easily understood by an end user, but the large variance in information durability makes it difficult to ascertain the validity of the review; for example, a systematic review published four years ago could be absolutely valid or completely outdated. The harm of out-of-date reviews is not only that they lack the most recent information for the topic they discuss, but that they may reduce users’ trust in the review process as a whole. One may use the analogy of car purchasing; the uncertainty surrounding the quality of used cars is harmful to that market as whole [Ake70].

From a machine learning perspective, the task of updating reviews is an attractive target because examples of relevant and irrelevant articles already exist. Additionally, because the original screening process most likely included screening by two reviewers and ratings are reconciled, the labels are more likely to be correct. However, for cases such as the emergence of a new treatment, or a novel adverse effect, the questions that a review seeks to answer may change by the time of updating. In these cases, words that were previously related to irrelevance may now indicate relevance. The key to saving work in this area is leveraging the existing data while allowing for adaptation to new concepts.
5.1.1. Prior Work

While considerable work has been done on machine learning for citation screening in the systematic review process[CHB04, CAM09, CAM10, Coh08a, CAM12, DSH+13, FIM10, FIM+11, HSS+12, KMI+09, MKI+10, WSB+12a, WSB+10, WTL+10, WSB+12b, CHPY06, DSH+12, WSBT10], less has been conducted on the handling of prior review data when conducting updates. Thus far, three approaches have been either proposed or used as the de facto standard. The most basic of the approaches involves simply using the original data to predict relevance of the documents retrieved in the update search[CAM09]. This is the simplest approach, and does not allow the model to incorporate changes in the questions of interest or the literature produces since the original review was published. Another promising approach utilizes domain knowledge to create a set of hand-engineered features that could predict relevance[DSH+13]. Whilst this approach also does not attempt to learn directly from the decisions on the update, the domain knowledge generalizes the features such that changes in vocabulary do not cause a loss in predictive power. For example, when studying Low Bone Density (LBD), the terms “alendronate” and “ibandronate” are mapped to “intervention.” If a new intervention is investigated in the update, it will also be mapped to “intervention,” and benefit from inclusion decisions about alendronate and ibandronate. In experiments, this approach generated >50% work savings with at least 99% sensitivity on the two tested reviews[DSH+13].

The third approach differs from the preceding two in that it has been proposed in the context of continually updated reviews[WDS+13]. It can be used to predict the relevance of any new article that matches the search terms as it is published. When the relevance decision is made, the models are retrained with new data. In the four reviews used for experimental evaluation, work savings of between 66% and 92% were realized[WDS+13]. Despite these impressive savings, this model is not universally practicable, as very few publishers have the budget for continual updating. Because literature tends to change gradually, savings will be smaller when the update occurs many years after the original publication.

5.1.2. Different Types of Updates

Depending on the circumstances behind a given review, the procedure for updating may vary considerably. If a funding agency such as AHRQ determines that an update is required, a full update may be undertaken. This may involve utilizing the same search strategy, inclusion criteria, and key questions used originally, or it may update one or more of these components. For example, in the case of a comparative effectiveness report on LBD[MAC+07], the update searched for additional data regarding the original questions, but also investigated the efficacy of one novel therapy (denosumab) and any new adverse effects related to bisphosphonates[CND+12]. These “full updates” are the most involved update type, and are often nearly as intensive as a de novo review.
5.2 Surveillance Updates

Before a full update is conducted, it may be beneficial to perform a “surveillance update” to assess freshness of the original findings. This involves executing the original search, but limiting results to five respected general medical journals and five specialist journals\[ANM^+13\]. Because the pool of potential articles is much smaller, the researchers are able to complete the review in a much shorter time frame. The goal of this surveillance update process is not to produce a new systematic review, but to aid the decision of when to fund in depth updates. As such, there is a higher tolerance for error if the resource savings are favorable and the organization tasked with performing surveillance is often not the group that conducted the original review. Because there is no central repository of review screening decisions, it is more difficult to obtain training data from the original review. In response to this lack of availability, we have developed a process that scans electronic versions of the original review and extracts enough information to recreate a portion of the original training data. Once extracted, the same machine learning algorithms can then be applied to this data.

A final update method involves lower acute resource expenditure by continuously keeping the report updated. As new studies matching the search criteria are published, they are screened by the research team and findings updated if required. These continually updated syntheses are sometimes referred to as “live synopses”\[WDS^+13\]. Thus far, these products have been used most heavily in genetics, prominent examples of which include AlzGene, SzGene, COPDGene, and PDGene\[CCC^+10, ABM^+08, BMM^+07, LRM^+12\].

5.2 Surveillance Updates

5.2.1. Approach

Because one group performs surveillance updates for many review topics at once, the updates were not conducted by the researchers of the original reviews. The primary contribution of our approach was the automated extraction of labeled data from existing reviews. In order to extract the data, we first downloaded the original systematic review in HTML format. After scanning the results section and evidence tables for references, we programatically extracted both the references and a small amount of context. A reviewer then scanned the context to eliminate references that discussed background information of the condition or treatment without providing new data, and remaining articles were labeled as relevant. All other articles that matched the review-described search were labeled as irrelevant.

When a small number of articles is present in the resulting update search, obtaining feedback for the learning algorithm becomes difficult. As a result, we simply treated the original review as the training set and the update as a test set. Due to its success on \textit{de novo} systematic reviews in the prior chapter, we use a Support Vector Machine with
Algorithm 5.1 Determination of the threshold for surveillance updates

Require: Number of iterations $N$, training data $(X,Y)$, recall goal $r$

* $i \leftarrow 0$

while $i < N$:

* Randomly partition 66% of the data to $(X_0,Y_0)$; remaining 33% to $(X_1,Y_1)$

* Fit the SVM with DEC to $(X_0,Y_0)$

* Predict $\hat{Y}_1$

* $t_i \leftarrow \max_t \frac{\left| \{ y : (\hat{y}_i \geq t) \land (y_i = 1) \} \right|}{\left| \{ y : \hat{y}_i \geq t \} \right|} \geq r$

* $i \leftarrow i + 1$

endwhile

Return: average optimal threshold $\frac{1}{N} \sum_i t_i$

Differential Error Costs, which carries out the following optimization

$$\min_{w,\xi \geq 0} \frac{1}{2} \|w\|^2 + C^+ \sum_{\{i|y_i=+1\}} \xi_i + C^- \sum_{\{i|y_i=-1\}} \xi_i$$

$$s.t. y_i w^T x_i \geq 1 - \xi_i, \ i = 1, \ldots, |\mathcal{L}|$$

where $\mathcal{L}$ is the set of labeled documents, $x_i$ is the feature vector associated with input $i$, $y_i \in \{-1, +1\}$ is the relevance judgment for article $i$, $w$ is a weight vector, $C$ is a tunable parameter that trades off misclassification error and SVM margin, and $\xi$ is a slack variable reflecting the SVM margin. As before, to eliminate hyperparameters, we set $\frac{C^+}{C^-} = \frac{\mathcal{L}^+}{\mathcal{L}^-}$ [AKJ04] (for a more thorough treatment of the feature representation and learning algorithm, refer to the previous chapter). Unlike the situation where ratings are obtained incrementally, here we made a one-time decision about the number of articles for which to request ratings. To do this, we repeatedly randomly partitioned the original review data into training and test sets and found the largest threshold that reached the specified recall goal, as illustrated in Algorithm 5.1.

5.2.2. Experiments

To demonstrate the value of the proposed approach to surveillance updating, we investigated its performance on ten surveillance reviews, summarized in Table 5.1. There was considerable variation in both the fraction of relevant articles and the size of the reviews; more targeted searches often result in surveillance updates with a higher percentage of relevant articles.

Understanding the performance requires analysis of both the fraction of relevant articles captured and the work savings. Given the classifier’s output $s(x_i)$ and the threshold $t$ determined by Algorithm 5.1, Table 5.2 shows the contingency table of potential outcomes. Ideal performance would result in zero for both the $fp$ and $fn$ categories, but
## 5.2 Surveillance Updates

<table>
<thead>
<tr>
<th>Review</th>
<th>Original</th>
<th>Update</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Relevant</td>
</tr>
<tr>
<td>ADHD</td>
<td>13,736</td>
<td>1%</td>
</tr>
<tr>
<td>Autism Spectrum Disorders</td>
<td>3,068</td>
<td>6%</td>
</tr>
<tr>
<td>Breast Cancer Prevention</td>
<td>4,320</td>
<td>3%</td>
</tr>
<tr>
<td>Cystic Fibrosis</td>
<td>1,215</td>
<td>7%</td>
</tr>
<tr>
<td>Hip Fracture</td>
<td>3,823</td>
<td>3%</td>
</tr>
<tr>
<td>Prostate Cancer</td>
<td>18,789</td>
<td>1%</td>
</tr>
<tr>
<td>Sleep Apnea</td>
<td>14,507</td>
<td>2%</td>
</tr>
<tr>
<td>Surgical Biopsy</td>
<td>5,410</td>
<td>3%</td>
</tr>
<tr>
<td>Traumatic Brain Injury</td>
<td>1,527</td>
<td>9%</td>
</tr>
<tr>
<td>Treatment-Resistant Depression</td>
<td>1,533</td>
<td>5%</td>
</tr>
</tbody>
</table>

Table 5.1.: Surveillance updating datasets used for evaluation

\[
y_i = 1 \quad y_i = -1
\]

\[
s(x_i) \geq t \quad tp \quad fp
\]

\[
s(x_i) < t \quad fn \quad tn
\]

Table 5.2.: Binary classification contingency table; \(t\) denotes true, \(f\) false, \(p\) positive, and \(n\) negative.

This is unlikely to be attainable in practice. Because missing a relevant article is a more significant issue than requiring additional screening, false negatives are far more costly than false positives. To capture both of these, we report sensitivity and specificity, which are defined as:

\[
Sensitivity = \frac{tp}{tp + fn}
\]

\[
Specificity = \frac{tn}{tn + fp}
\]

Sensitivity of 100% can only be achieved when no relevant articles are missed, and specificity provides a measure of the work savings.

The results in Table 5.3 demonstrate that while the work savings were generally modest, it was possible to train a useful model on data extracted from published systematic reviews. Hip Fracture is a relatively uninformative topic, as it had only a single relevant article in the update. While the approach achieves moderate savings and acceptable recall overall, the variability is problematic. With a larger collection of data, future research may be able to identify characteristics predictive of poor performance, which could allow researchers to screen such topics manually, or tailor machine learning algorithms to their unique challenges.
Table 5.3: Detailed performance on the surveillance updates. Note that the abbreviated procedure makes time savings more difficult, but we are still able to obtain improvements over the status quo.

5.3. Full Updates

5.3.1. Approach

Multiple approaches to handling this challenge exist, ranging from simple approaches such as ignoring the original data to complex transfer learning algorithms[PY10, PTKY11, XPP+11, DYXY07, EDL08, BH03]. In this work, we restrict analysis to approaches that can be formulated as data modifications. Two approaches are simple enough that they form a natural baseline; proposed solutions should perform better than these trivial approaches. In one, which we refer to as “combine,” data from the original review and update are combined, with no acknowledgment of the source. The other baseline, termed “update only”, treats the update as if it were an entirely new review, ignoring data from the original review.

Earlier research also found that an alternative, simple approach, called EasyAdapt, can produce effective results in practice[DM07]; we investigated its performance on this task. EasyAdapt expands the feature space by adding an interaction term between the document features and an indicator for being part of the update. Denoting a document’s feature vector as $x_i$, after preprocessing it becomes

$$(x_i, 1\{x_i \in \text{original}\} \times x_i, 1\{x_i \notin \text{original}\} \times x_i)$$

This results of this approach are presented visually in Table 5.4, where a common set of features is denoted $d_C$, factors specific to the original review are in $d_O$, and the update-specific factors are captured in $d_U$. Letting $|V|$ be the size of the vocabulary such that $x_i \in \mathcal{R}^{|V|}$, EasyAdapt transforms the update data to $(x_i, 0, x_i) \in \mathcal{R}^{3 \times |V|}$. Theoretically, this has been shown to be equivalent to the inclusion of a Bayesian hierarchical prior.
### 5.3 Full Updates

<table>
<thead>
<tr>
<th></th>
<th>$d_C$</th>
<th>$d_O$</th>
<th>$d_U$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>$X_O$</td>
<td>$X_O$</td>
<td>$0$</td>
</tr>
<tr>
<td>Update</td>
<td>$X_U$</td>
<td>$0$</td>
<td>$X_U$</td>
</tr>
</tbody>
</table>

**Table 5.4.** Graphical depiction of EasyAdapt [DM07]

<table>
<thead>
<tr>
<th>Approach</th>
<th>Original</th>
<th>Update</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combine</td>
<td>$x_i$</td>
<td>$x_i$</td>
</tr>
<tr>
<td>EasyAdapt</td>
<td>$(x_i, x_i, 0)$</td>
<td>$(x_i, 0, x_i)$</td>
</tr>
<tr>
<td>Offset</td>
<td>$(x_i, 0)$</td>
<td>$(x_i, x_i)$</td>
</tr>
<tr>
<td>Update Only</td>
<td>$\emptyset$</td>
<td>$x_i$</td>
</tr>
</tbody>
</table>

**Table 5.5.** Feature Representation for the four approaches evaluated herein. Each feature vector is transformed according to the table, where $0$ is a zero vector of the same dimension as $x_i$ and $\emptyset$ means that the instances are not used.

[FM09]. In practice, however, it is implemented as a simple data preprocessing step. The Python code carrying out this transformation, which consumes only two lines, is available in Listing 5.1. Utilizing this expanded feature representation, the classifier can learn features that are common to both reviews, as well as features that are unique to each review. While there is no guarantee that performance with this approach will be superior, it has been very successful in prior evaluations [DM07].

Inspired by EasyAdapt, we developed an alternative approach, referred to herein as “Offset,” which also operates through feature space expansion. While EasyAdapt learns three sets of coefficients, Offset only learns two: the baseline and “offsets” from that baseline. We simply augment the feature vector $x_i$, to become $(x_i, 1\{x_i \notin \text{original}\} \times x_i)$. Again, this code is presented in Listing 5.1, and is slightly shorter than EasyAdapt. Because all of the approaches proposed herein can be expressed in terms of feature-space modifications, Table 5.5 provides a concise depiction of the methods under investigation.

In an experiment of news article recommendation, Chapelle and Li found that learning from incremental data with a small delay yields significant benefits over updating models.

**Listing 5.1:** Python code implementing EasyAdapt and Offset

```python
import scipy.sparse as sps

# EasyAdapt
X0 = sps.hstack((X0, X0, sps.csr_matrix(X0.shape)), format='csr')
X1 = sps.hstack((X1, sps.csr_matrix(X1.shape), X1), format='csr')

# Offset
X0 = sps.hstack((X0, sps.csr_matrix(X0.shape)), format='csr')
X1 = sps.hstack((X1, X1), format='csr')
```

89
### Table 5.6.

<table>
<thead>
<tr>
<th>Review</th>
<th>Original</th>
<th>Update</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N Relevant</td>
<td>N Relevant</td>
</tr>
<tr>
<td>Low Bone Density (LBD)</td>
<td>14,700</td>
<td>7,089</td>
</tr>
<tr>
<td>Atypical Antipsychotics (AAP)</td>
<td>1,307</td>
<td>3,591</td>
</tr>
<tr>
<td>Health IT (HIT)</td>
<td>4,633</td>
<td>7,137</td>
</tr>
</tbody>
</table>

Datasets used for evaluating this approach; N refers to the number of articles.

in large batches[CL11b]. However, because our current approach has not been deployed in a production system, realtime feedback is infeasible; in our research prototype we request ratings in batches. As with de novo reviews, it is valuable to balance the selection of articles that are more likely to be relevant with exploration of the space of potentially relevant articles. We utilize the same $\epsilon$-greedy strategy as on de novo reviews: prefer sampling articles most likely to be relevant, but force exploration by selecting $\epsilon = 0.05$ of each batch randomly. All experiments used the same linear support vector machine (SVM) with differential error costs described in Section 5.2.

### 5.3.2. Experiments

We conducted experiments on three full review updates of various sizes: Low Bone Density (LBD), Atypical Antipsychotics (AAP) and Health IT (HIT), whose descriptive statistics are shown in Table 5.6. As in the prior chapter, some reviews have multiple outcomes of interest. For both the LBD and AAP full reviews, the reviewers judged relevance for efficacy and adverse effects (AE) separately. In the “Any” category below for these reviews, relevance was taken to mean that the article was included for either efficacy or adverse effects (or both). While three is a relatively small sample, the variety in these reviews made them attractive targets of inquiry; LBD is a large review with multiple outcomes, AAP is unique in that the original search returned fewer articles than the update, and HIT lacks a concise set of interventions to form the subject of the review.

We hypothesized that the time lag between the original review and the update could have strong effects of performance. For LBD, the original systematic review was published in 2007[MAC+07], while the update was published in 2012[CND+12], the AAP original and update occurred in 2007[SMB07] and 2011[MT12], respectively, and the original HIT review was conducted in 2011 whilst the update occurred in 2012. We thus expect HIT to benefit most from original data.

Obtaining the data for the full reviews is often relatively straightforward as, unlike surveillance updates, the organization producing a full review update is usually the same organization that conducted the original review. Results used a batch sequential experimental protocol, where 25 articles were chosen randomly and processed using an
5.3 Full Updates

<table>
<thead>
<tr>
<th>Review</th>
<th>Total Articles</th>
<th># Reviewed</th>
<th># Avoided</th>
<th>% Avoided</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAP</td>
<td>3,591</td>
<td>1,015</td>
<td>2,576</td>
<td>72%</td>
</tr>
<tr>
<td>Health IT</td>
<td>7,137</td>
<td>1,422</td>
<td>5,714</td>
<td>80%</td>
</tr>
<tr>
<td>LBD</td>
<td>7,089</td>
<td>987</td>
<td>6,101</td>
<td>86%</td>
</tr>
</tbody>
</table>

Table 5.7: Summary of performance on the full update reviews. This uses a linear SVM with differential error costs and the “Offset” strategy described in Section 5.3.1.

<table>
<thead>
<tr>
<th>Review</th>
<th>Combine</th>
<th>EasyAdapt</th>
<th>Offset</th>
<th>Update Only</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AE</td>
<td>71%</td>
<td>73%</td>
<td>71%</td>
<td>72%</td>
</tr>
<tr>
<td>Efficacy</td>
<td>81%</td>
<td>81%</td>
<td>82%</td>
<td>81%</td>
</tr>
<tr>
<td>Any</td>
<td>70%</td>
<td>72%</td>
<td>72%</td>
<td>71%</td>
</tr>
<tr>
<td>Health IT</td>
<td>79%</td>
<td>79%</td>
<td>80%</td>
<td>76%</td>
</tr>
<tr>
<td>LBD</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AE</td>
<td>85%</td>
<td>86%</td>
<td>85%</td>
<td>82%</td>
</tr>
<tr>
<td>Efficacy</td>
<td>85%</td>
<td>86%</td>
<td>86%</td>
<td>83%</td>
</tr>
<tr>
<td>Any</td>
<td>85%</td>
<td>86%</td>
<td>86%</td>
<td>86%</td>
</tr>
</tbody>
</table>

Table 5.8: Detailed comparative performance on the full update data.

$\epsilon$-greedy strategy ($\epsilon = 0.05$); new articles were then similarly processed until at least 95% of the relevant articles were labeled.

In the context of full reviews, we were able to demonstrate significant reductions in workload, summarized in Table 5.7. In Table 5.8, alternative approaches are compared using the batch sequential protocol; here we can see that EasyAdapt and Offset are generally better than baseline approaches. Interestingly, in some cases, data from the original review were not useful, for example “LBD Any.” To obtain deeper understanding of this, Figure 5.1 plots the learning curves for each of the four approaches on the three full reviews. Overall, the differences between approaches were minimal, with the exception that discarding the original data was harmful for HIT. It appears that a high quality model can be trained with relatively few ratings, which reduces the benefit of even very closely related data when subject to the high sensitivity requirements of systematic reviews.

It is reassuring that the relative performance between the various approaches can be explained by the differences between the update and the original systematic reviews. The HIT update was most similar in time and scope to its original, and benefitted from original data more than others. For LBD and AAP, either the evidence base or the focus shifted, and the benefit of using original data was smaller. These findings suggest that future research may profit from utilizing original data only in the early stages of the update.
Figure 5.1.: Learning curves for various approaches on the three investigated reviews
5.4. Discussion

In this work, we showed that the proper utilization of data from a previous review can reduce the expert workload when updating systematic reviews, in excess of workload reductions attainable on *de novo* reviews. Benefits are reduced when the evidence or focus of the review changes, and are strongest at early stages of the review. In some cases, it can be beneficial to obtain examples of new literature early in the process, and the machine learning approach proposed herein enables that goal.

We also demonstrated that extraction of data from published reviews is not only possible, but can be of benefit when assessing the need to update a systematic review. Because a predictive model can be built from the published systematic review and trained to predict the relevance of all newly published articles, this could enable efficient, near-real-time updating of systematic reviews. It could also be used to revise the approach used when conducting surveillance updates, altering the approach from one where only the most highly cited journals are searched to one where only articles with high probability of relevance are reviewed. While the benefits are smaller in a narrowly targeted search, this remains an interesting avenue for future research.
6. Conclusion

By providing context, structure, and explication of the important aspects of systematic reviews, this thesis laid a foundation for widespread adoption of techniques that expedite review creation. Chapter 2 demonstrated that systematic reviews provide a substantial societal benefit, with a review on Low Bone Density (LBD) generating a net benefit of approximately $450M. In Chapter 3, we provided the first data on the time taken to screen articles in systematic reviews. User interface changes can result in very significant reductions in resource requirements: presenting multiple articles on one screen while reviewing titles leads to a seven-fold reduction in time taken per article. Experience and mental state are also related to screening times, with abstracts reviewed at ideal session lengths requiring 33% less time than those at the beginning of a session. Our approach utilizing machine learning to aid in review preparation is described in Chapter 4, where we outperformed other published approaches and avoided screening 80% of articles on large systematic reviews while capturing 95% of relevant articles. Finally, Chapter 5 demonstrated that data from an existing review can be used to perform updates more efficiently than for de novo reviews, though the incremental benefit is small when the key questions or treatments of interest have changed.

6.1. Implementation Considerations

The tools presented in this thesis will only benefit society if they are implemented. A number of technical and organizational features have the potential to make implementation difficult, but opportunities exist to reap benefits far beyond those previously mentioned.

6.1.1. Challenges

With research on computer assisted screening published as early as 2006[CHPY06], one may be surprised to find that these techniques are not widely used in practice in 2014. We believe that this is the broadly due to (1) a disconnection between academic research and practice, and (2) institutional factors unique to systematic reviews.

In the first category, deviating from standard experimental protocols has been challenging. As mentioned in Section 4.4, early research utilized cross-validation to demonstrate the potential for learning. While we showed that relative performance of cross-validation
is similar to a more realistic protocol in Section 4.6, this finding was not knowable *a priori*. A practical consideration which has yet to be satisfactorily addressed surrounds the determination that screening is complete. This was first addressed by [WTL+10], who found that a simple heuristic of screening 50% of all articles was sufficient to achieve 100% recall. This rule, while a useful starting point, is not based on large-scale evaluation, and has no connection to the data of a particular review. When applying our proposed techniques to legitimate reviews, such as the LBD update described in Section 3.6, we utilized heuristics based on changes in predictions and the number of relevant articles recently identified. While this allowed for significant cost savings, it is difficult to justify an approach that lacks theoretical underpinnings and evaluation on a broad class of systematic reviews. The heavy utilization of MeSH indexing in published studies also adds significant practical difficulties; this has not been mentioned in any study using these features.

Additionally, systematic reviews employ multiple screeners that may work at varied speeds, and their ratings are not always consistent. In the traditional process, discrepancies are reconciled when the screeners affected are available. When the system is designed to quickly respond to a user’s feedback, this reconciliation is not possible. All results that we have presented utilized reconciled judgments, as have all similar publications of which we are aware. The optimal article selection procedure with rating noise is far from obvious. Comparing ratings regularly would enable the algorithm to ascertain likely labeling errors, but would remove some of the benefit of utilizing independent screeners.

Because the research on computer aided screening has been done by multiple groups, the tools utilized by each group are different. Repeating a previously published procedure will not result in publications, despite aiding practitioners in assessing the validity of that approach. When comparing to a methodological choice such as restricting searches to English [JHS+02, MPK+00, MPLK03] or to articles indexed by MEDLINE [SBM+03], this variability raises valid concerns. Success of the approach proposed by researchers at OHSU [CHPY06] says very little about the effect of following a procedure proposed by researchers at Tufts [WTL+10]. Until one approach has been formally described, implemented in a production-tested system, and undergone large-scale evaluation, adoption of the tools carries excess risk.

Making deployment more challenging, myriad tools are currently used for screening. At the Southern California EPC alone, we are aware of early screening stages that have been conducted using paper and highlighters, EndNote, an internally developed system, and DistillerSR. This fragmentation demonstrates that the researchers conducting the screening have strong preferences and established workflows; it therefore is unlikely that a single approach will satisfy all needs.

Coupled with the published literature’s lack of focus on these practical considerations, institutional and market characteristics have slowed adoption of computer assistance. Because organizations funding and publishing systematic reviews expect that all relevant studies are included, the justification of practices that may omit articles is challenging. We hope that by focusing on the ultimate uses of the information, policymakers will be
able to move past this dogma into informed decisions regarding trade-offs. In the short term, widespread adoption of these tools would require investment in development, scaling, testing, and training. After these initial costs, the technology platform will require maintenance and user support. Though these would be primarily one-time costs, it can be difficult to justify such capital investments in a resource-constrained environment.

Adoption would have fewer barriers if it presented an attractive target for private investment. Machine learning, data management, and user interface design are all commonly needed skills in the technology sphere, and form the core requirements for successful deployment of the tools presented in this thesis. Despite the match in skills, the market itself is only moderately attractive. Because the current practice has minimal IT requirements, it would be difficult for groups conducting systematic reviews to justify large software expenditures. A relatively small number of organizations conduct large systematic reviews, further limiting the market. While the largest reviews are undertaken by these specialized groups, other systematic reviews are conducted by a broad research base. The Cochrane Collaboration, a non-profit organization that promotes evidence-based medicine internationally, has 31,000 individuals contributing to its mission[The14].

6.1.2. Opportunities

In spite of these challenges, a modernized systematic review process will enable innovations beyond those proposed in this thesis. Because the implementation challenges are significant, we focused enabling the output of systematic reviews using our tools be essentially identical to those produced using the current process. Future work that does not abide by that restriction can create impressive results in the longer term. An example of one possible innovation would combine components identified in this paper to provide a real-time indicator of the “freshness” of a systematic review. Recall from the surveillance updating presentation in Section 5.2 that we were able to extract data to inform a machine learning algorithm from a published review, with little human intervention. With additional research, the incremental human effort could be driven to zero, enabling automated extraction of labeled data from published reviews.

Armed with this data, our system could be used to train a classifier that identifies relevant articles from all of MEDLINE. While the predictions would not be perfect, our unpublished experiments have shown that the articles identified as most relevant are very frequently correct; precision is high when considering only a small number of candidate articles. This small list could provide users with information that enables a user to quickly estimate the likelihood that more recent information jeopardizes conclusions from the review. Extending the system further, a machine learning system could be built to predict a probability that the review’s conclusions are current, using both characteristics of the review and characteristics of the articles likely to be relevant. Displaying this indicator on PubMed or on the websites of publishers would enable users to have increased confidence that the systematic review accurate, or better understanding of the reasons that it may not be current.
Further into the future, it may be possible to synthesize evidence on the fly, replacing large portions of the systematic review process with machine learning systems. Natural language could be converted into database queries comprising the search strategy, just as search engines such as Google have enabled natural language search of the internet. Key questions would not need to be determined with a high degree of rigor, because they could be reformulated in response to outputs. Meaningful synthesis of the relevant information would remain challenging, though information extraction tools could simplify comparisons across articles.

To have a significant effect on population health, systematic reviews need not lead to large changes in beliefs about different treatments. The analysis of osteoporotic fractures in Chapter 2 did not modify the choice of drug, but only the treatment eligibility criteria. Modest changes in beliefs regarding the efficacy of alendronate led to a change in ideal treatment decisions, and this resulted in an estimated societal benefit of approximately $1.5B. Expenditures on healthcare are sufficiently large that even small improvements in care can have societal benefits. It is simple to share the information resulting from a systematic review with a wide audience, and incremental costs are essentially nonexistent. Benefits to each new person exposed can be as large as previous readers. With this structure, lowering the cost can increase the ideal number of reviews undertaken—allowing more people to benefit from better treatment decisions.

While we have only discussed the application to systematic reviews, the tools proposed in this thesis are valuable in a number of other contexts. Systematic reviews and meta-analyses are used in domains beyond healthcare: foreign aid [Viv14], education [CMHE04], and criminology [Bra05] comprise a small sample of uses outside of medicine. Computer-aided screening would reduce effort and increase timeliness for these systematic reviews as well. Furthermore, the vast majority of scholarly research conducts a literature review as part of its background research. In these cases, the quality of literature reviews could be increased without additional resource expenditures. Because these searches do not utilize a systematic screening protocol, using such a structure is likely to be more comprehensive and more reproducible. As relevant articles are identified most quickly early in the process (see Figure 5.1) researchers can easily balance timing and resource limitations with the desire to conduct a comprehensive survey.

### 6.2. How This Work Could Affect Policy Development

The most direct impact of this work will be the adoption of the proposed techniques to reduce the time required to produce systematic reviews. The machine learning approach reduces the resource requirements for one segment of the bench-to-bedside process, but is only a small portion of the work presented herein. More than anything else, we hope that this work leads researchers and policymakers to view the creation, synthesis, and use of medical evidence as tools for improving health outcomes. Systematic reviews are one tool that facilitate the flow of information, but all synthesis need not adhere to the unbending requirements of the systematic review.
Because the cost of evidence synthesis generally exceeds its value to one individual or organization, single providers will not synthesize information directly. Publication of the review creates positive externalities for other potential users—benefits are created that are not captured by the individual paying for the product. Per economic theory, goods with such market failures are attractive targets for government funding, but the specific mechanisms will have strong effects on the way these products are provided [WV05]. With the policy of government provision of medical evidence taken as given, the following discusses ways that reframing of the process and tailoring the organizational structure to those goals will benefit the health system.

### 6.2.1. Reframing the Goals of Evidence Creation and Synthesis

The systematic review process is currently quite rigid. While this standardization creates benefit for quality control and assessment, it does not provide flexibility to tailor the process to differing requirements. For example, in the case of rare diseases, it may be preferable to undertake an abbreviated process of evidence synthesis; the evidence will be less, and the number of individuals utilizing the evidence will be smaller.

One concrete policy option in this vein is to issue a contract or set of contracts for updating systematic reviews, where this contract would be separate from the original preparation. With a small number of firms responsible for updating, incentives to streamline the process will be strengthened. Fixed price contracts could strengthen these incentives further. While some may argue that this would lower quality standards, the structure of the report is set in the original systematic review. Search strategies, key questions, and the structure of all evidence tables need not change. With an efficient process, it would be reasonable to expect that new studies could be incorporated within one month of publication.

Ultimately, the problem that evidence-based medicine attempts to solve is that health-care is too expensive, particularly for the quality of outcomes attained. One of many solutions instituted is to create an organization (AHRQ) with the mission of producing evidence that supports system-wide health improvement. With the reworking of its mission statement, AHRQ is moving in the correct direction.

Where this work identifies gaps the current approach is on investments that can reduce inefficiency in the process of creating, synthesizing, and utilizing medical evidence, which was illustrated in Figure 1.3 of the introduction. This flow of information can be improved in two ways: increasing efficiency of component processes (such as systematic review preparation), or reducing friction between those components. The first way will be more comfortable, and is thus a recommended first step. The computer-assisted screening tool discussed in this work would aid that goal, as would the proposed issuance of maintenance contracts for systematic reviews.

Reducing friction between pieces in the chain is more challenging, but potentially bears greater rewards. AHRQ is not directly responsible for basic biomedical research or for implementation, though its products are useful for both. Without protecting the value
of ensuring the products are utilized, it would be easiest for AHRQ to focus only on producing evidence. In order to make the research products create an impact, they must work closely with those responsible for implementation. They must also ensure that strength-of-evidence assessments are produced at the appropriate level of detail and that they are disseminated to organizations like NIH that make funding decisions on future research. Stated bluntly, systematic reviews are not worthy of governmental support based on intrinsic merit—they are valuable because of improved treatment decisions, better research prioritization, or reduced effort on other tasks required by the health system.

Viewing systematic reviews in this light enables policymakers to make better decisions about policies relating to the production and utilization of evidence in the health system. Budgetary decisions must not be made because a certain number of reviews “should” be produced each year, but based on a holistic view of the evidence needs of the health system. Prioritization approaches[HSBM12] are strong steps in this direction, but we encourage a broader view. The approach proposed in [HSBM12] involves estimating the expected value of a systematic review, and Chapter 2 borrowed from that framework. A beneficial side-effect is that various components of the equation are made clear. In the example of osteoporotic fractures in Chapter 2, the annual disease burden is staggering; it is an attractive research target. The expected difference between treatments, however, is small. Short term head-to-head trials of existing drugs are therefore unlikely to warrant funding. The primary uncertainties regarding the adverse effects of bisphosphonates surround osteonecrosis of the jaw and atypical subtrochanteric fractures[CND+12]. Because these are very rare events[CND+12], it would be irrational to fund clinical trials aimed exclusively at understanding these outcomes. Large scale observational studies of these outcomes are likely to be far more valuable. As the cost of alendronate has fallen since the introduction of a generic formulation, it becomes preferable to treat individuals with increasingly lower risks of fracture[NRG12]. Improved estimates of the harms relating to treatment are therefore extremely beneficial when creating guidance about which individuals should receive treatment. The segment of the bench-to-bedside process that is truly weak in the case of fracture prevention is implementation. Only 35.7 percent of individuals with diagnosed osteoporosis receive treatment [SMT+14], and the analysis in Chapter 2 shows that even under conservative cost assumptions, it is beneficial even to treat individuals with higher bone density than would warrant a diagnosis of osteoporosis.

Resulting from this analysis, for this topic, implementation should receive the largest share of governmental support in this example. Implementation support could come in a number of forms, including development of practice guidelines, creation of performance measures, or information outreach programs. Making this assessment would have been very difficult without the systematic review, but the implications affect allocation of resources across agencies. Because budgets are nearly impossible to transfer between agencies, this valuable finding cannot be acted upon. Budgets could be modified between fiscal years, but meaningful changes are difficult to accomplish in practice.

If the problem can be successfully reframed, the governmental support given to evi-
6.2 How This Work Could Affect Policy Development

dence creation, synthesis, and implementation should be increased. We estimate that the reviews on low bone density [MAC+07, CND+12] created $0.45B in societal benefit. While their cost is unknown, it is likely to be smaller by an order of magnitude. With such benefits possible, system-wide health outcomes would be improved with increased availability of clinically useful evidence.

6.2.2. Organizational Implications

We have discussed a number of individual components of the larger evidence pathway including primary research, systematic reviews, cost-effectiveness analysis, and guideline creation. It is not immediately clear if one organization should bear responsibility for supporting the broad pathway, or separate organizations should govern the individual pieces.

In the first scenario, where one organization has a broad mandate, a primary benefit is that the budget would be somewhat fungible across different products. If a clinical trial shows exceptional promise for some treatment, the best strategy may involve funding additional trials rather than conducting synthesis. Alternatively, if the proposals for primary research are weak in one quarter, funds can be shifted to synthesis. Were the activities governed by multiple organizations, funding could only be shifted between fiscal years, and there would be strong pressure for each organization to attempt to increase its own budget. Over time, this would lead to a suboptimal allocation of resources. If each component were undertaken by a separate organization, it would be easier to instill a sense of mission within each group; the goal would be clearly defined. Personnel decisions would be simplified, because the skills required would be more focused.

Within the United States, the Agency for Healthcare Research and Quality has the most direct responsibility for this process. Its mission prior to FY15 was:

“To improve the quality, safety, efficiency, and effectiveness of health care for all Americans.”

and has recently been changed to:

“To produce evidence to make health care safer, higher quality, more accessible, equitable, affordable, and to work with the U.S. Department of Health and Human Services (HHS) and other partners to make sure that the evidence is understood and used.”

This new mission statement is considerably stronger, as it clarifies the organizational goals in a more concrete manner. At its core, AHRQ produces evidence, which can result from either primary research or synthesis. However, the research AHRQ conducts should focus on “long-term and system-wide improvement of health care quality and effectiveness” [Age15]. Two other organizations within the Department for Health and Human Services also conduct medical research: the National Institutes of Health conducts “[b]iomedical research to prevent, diagnose and treat diseases,” and the Center for
Disease Control conducts “[p]opulation health and the role of community-based interventions to improve health.” [Age15]

While this segmentation is quite clear, it separates the production of primary medical research into multiple organizations. A systematic review could identify gaps in medical evidence, but NIH would be responsible for acting on those gaps. It is far from desirable for action on a valuable result from a systematic review to be the responsibility of an organization that is not the primary audience of the report.

Furthermore, AHRQ does not claim to have expertise or a broad mandate for the implementation of its research[Age15]. This makes implementation a “precarious value”—a goal that is outside the primary mission of the organization[Mar97]. As mentioned previously, the primary goal of AHRQ is to produce evidence. Evidence alone will not improve health outcomes, and a research organization will tend to produce reports in a style that is comfortable to researchers. Practitioners and patients will often not have the patience or background knowledge required to digest hundreds of pages of material on a single clinical question. If implementation is to be emphasized, those efforts must be protected from the pressures exerted by an organization whose primary goal lies elsewhere.

As a higher-level organization, the Department of Health and Human Services (HHS) should centralize some decisionmaking over allocation to specific topics and types of research. This will require additional funding, which must be approved by Congress. For these legislators, the most important lesson from this thesis is that while the current practice borders on unsustainability, EBM provides value and can be conducted meaningfully with the appropriate infrastructure. Making that infrastructure available will require short-term investment into organizations that have not historically made these types of capital expenditures.

An additional policy option, inspired by the aligned incentives in integrated providers [DRKR+12], is to transfer responsibility for production, synthesis, and implementation of medical evidence that concerns the elderly to the Center for Medicare & Medicaid Services (CMS). The burden of osteoporotic fractures referenced frequently in this thesis is borne by individuals covered by Medicare. CMS is a much larger organization than AHRQ: its budget in FY13 was $763.14B versus $0.43B for AHRQ—transferring responsibility to CMS would present both challenges and opportunities. The existence of AHRQ as a standalone organization demonstrates its importance to the national health system, and is likely a point of pride for its employees. That same small scale makes political battles more challenging, and existing as a separate organization increases scrutiny on its budget. Within a budget exceeding $700B, it would be easier to justify spending the modest amount required to develop information systems to streamline the process of creating and updating systematic reviews and practice guidelines. Given that CMS has the ability to structure incentive payments to its providers[KPSW12], it has mechanisms to encourage implementation of guidelines. Benefits from improved care would directly aid CMS, helping to motivate employees and justify investments in evidence creation and synthesis.

Despite the many benefits, this policy has significant drawbacks; primarily regarding
6.3 How Stakeholders Benefit from this Work

the provision of evidence for conditions that do not disproportionately affect the elderly. If resources previously devoted to AHRQ are shifted to CMS, the remaining budget of AHRQ would be small and benefits to scale would be reduced. Capital and methodological investments would be more difficult to justify, and the organization’s mission would be muddled. A more thorough analysis of this type of policy is left to future work.

Most importantly, we challenge governmental health organizations to view systematic reviews, and indeed most healthcare research, not as research with intrinsic value but as tools for improving health outcomes. Individual stakeholders can improve the system in isolation, but significant improvements must result from coordinated focus on the broad problem formulation.

6.3. How Stakeholders Benefit from this Work

Because the process of generating medical evidence to changing clinical practice is so broad, it is not surprising that many stakeholders are involved. This thesis can provide benefits to many of them.

Groups Funding Systematic Reviews  Organizations like AHRQ that fund systematic reviews bear the largest responsibility and opportunity for implementing the findings of this work. Those opportunities fall into two broad categories of streamlining the current systematic review process and focusing on the larger process of evidence creation and synthesis.

By encouraging its contract recipients to utilize computer-assisted screening tools and setting guidelines for their use, AHRQ could quickly reduce the cost and time requirements of each review. In the current environment, because the techniques are not widely adopted, utilization is reported in less detail than other sections of the report; this opacity can slow adoption because readers will not know if the specific approach was sensible. AHRQ should also consider issuing contracts to keep reviews updated. The fact that the structuring has been completed in the original would allow lower cost competitors to bid for these contracts, and sharing costs across many reviews would reduce costs further.

In a similar category, requiring standardized reporting and data formats could lead to longer-term cost savings. Concretely, contracts could require submission of screening data and abstracted data in a machine-readable format. Articles could then populate a publicly available database that allows for reuse. AHRQ has taken a positive step towards this in its Systematic Review Data Repository (SRDR), but this currently contains data for only ten reviews. Requiring use would increase short-term costs, but has potential to facilitate long-run cost savings by leveraging prior work. In the future, it may also serve as a useful dataset for training machine learning models to assist with data abstraction. A shared systematic review screening decision repository would similarly allow for refinement and evaluation of machine learning models. Evaluation on
such a dataset could serve as a validity check for models that are otherwise difficult to evaluate without expertise.

A beneficial short-term action that encourages AHRQ to view its products as tools for improving health outcomes is a prioritization process akin to the proposal of [HSBM12]. Initial topic proposals should be accepted from a wide audience, as early benefit scoping can be undertaken rapidly[HSBM12]. With the prioritized topic list, with components of the prioritization laid out clearly, the organizations could make improved resource allocation decisions. If one review is unlikely to be valuable because changes would not be implemented, resources could be instead devoted to attempts to change the features that make implementation challenging. If the prioritization instead indicates that a review will not provide value because little primary research has been conducted, resources could be allocated to that primary research, or to other reviews.

**Preparers of Systematic Reviews**  Organizations preparing systematic reviews have the most straightforward path to benefiting from the work in this thesis. Adopting the machine learning-based tools in this thesis will reduce time and costs, without requiring any change to the final product. Benefits are likely to grow as these tools are refined, evaluated in additional contexts, and as researchers become more comfortable with the new approach. Additionally, as the use of these tools becomes more widespread, is may be valuable to incorporate reporting requirements into the PRISMA statement[MLTA09], allowing for more rapid quality assessment.

**Other Health Researchers**  Individuals or groups conducting primary biomedical and other interdisciplinary healthcare research can benefit from this thesis by using it as a reminder to consult syntheses when proposing new work. While it is likely that specialists will be familiar enough with their domain of expertise that they are aware of the state of the field, others may not be. For example, researchers with significant experience utilizing large administrative databases could benefit from the identification of gaps in understanding of adverse effects identified by systematic reviews. Health economists, for example, may be able to design mechanisms to improve implementation of published evidence, but may not regularly read systematic reviews.

**Technology Companies**  Private companies can use the problem definition to assess market attractiveness and the skills required to make products that serve the evidence-based medicine community. While it is unlikely to be especially lucrative in its current state, moderate changes have potential to considerably alter the market. For example, if continually updated reviews became more standard, there would be value in a software tool that facilitates regular execution of literature searches, notifications of screeners, and machine learning-based screening assistance. Furthermore, automatically generating web-accessible reports from the structured review data would be nearly required, presenting an additional opportunity.
While early opportunities would focus primarily on data management, transitioning to more innovative product development would be possible after demonstrating success. If a company were to become a dominant provider for these short-term needs, they would possess a great deal of data. Those data could be utilized to streamline the process further, using more advanced tools from machine learning or human-computer interaction.

**Publishers**  
With the existence of established best practices, publishers have little incentive to include systematic reviews or other types of synthesis that do not adhere to those standards. Adherence to guidelines is a useful tool for ensuring quality, but has led to some evidence being outdated even before it is published[SSA+07]. By establishing standards for techniques like computer aided screening and for abbreviated reviews, quality could be maintained. This may require the creation of separate publications for non-systematic reviews, but those reviews could provide multiple benefits. Most obviously, synthesized information would be available more rapidly and updated more often. These abbreviated reviews are often comprehensive enough to aid in cost effectiveness models. Finally, they can serve as building blocks for systematic reviews, aiding in creation, framing, and prioritization for formal systematic reviews. If a machine learning tool is utilized, abbreviated reviews can be used like an original as presented in Chapter 5 to reduce screening requirements. Results of an abbreviated review could also inform the development of search strategies, and be used to estimate sensitivity and specificity of a proposed literature database query. Certain questions could be emphasized or de-emphasized in the full review based on abbreviated findings. Components of the value of information calculations useful for prioritization could also be informed by these products, aiding in the efficient allocation of scarce public resources.

**Large Provider Organizations**  
Because large provider organizations like Kaiser bear the cost and reap the benefits of any synthesis they undertake, they could see significant gains from this work. Individual physicians or small practices have no incentive to create guidelines or synthesize evidence. When an organization like Kaiser Permanente (KP) covers more than 8.6M members[DRKR+12], incentives are very different. Recognizing the value of high-quality, standardized, up-to-date clinical practice guidelines, KP has established a process for incorporating evidence into care[DRKR+12]. When available, they utilize existing guidelines and systematic reviews, but will conduct de novo reviews if necessary[DRKR+12].

This approach is already a potential model for other organizations, because incentives are aligned over the broad process. KP could save money in the short term by eliminating this effort, but values long term improvements in care. Reducing the cost of systematic reviews translates directly into savings that can be reinvested or passed to members. Budgets within a single provider organization are far more fungible than between agencies in the federal government. KP also leverages software tools that assist in data management, data sharing, and updating reports with new data[Doc14].
Patients  By organizing and encouraging researchers to publish summaries of findings and syntheses in plain language, patients can assist in strengthening the pressures for providers to follow evidence-base standards. An example of alternate values being placed on accessibility is found in comparing Cochrane reviews to those of the AHRQ Comparative Effectiveness Review program. Cochrane reviews such as [RGJ13] contain a section called “Practical Evidence About Real Life Situations” (PEARLS) with a standard format and concise summary. The format begins with a “clinical question” and is immediately followed by a “bottom line.” More advanced users can read further to understand details. In AHRQ Comparative Effectiveness Reports, objectives, data sources, and review methodology are shown first, and the results summary discusses the studies included before reporting conclusions [HABW14, MJW13, BEH+14]. These reports do provide “summary guides” for patients[Joh13] and providers[Joh12], but these are more detailed than the PEARLS. When presented in an accessible manner, educated patients can increase physicians’ pressure to utilize the best available evidence in treatment decisions. Even without exerting this pressure, if the process of evidence production, synthesis, and implementation is improved, patients will ultimately experience the rewards.

6.4. Conclusion

We intend this for exposition to encourage proponents of evidence-based medicine to continually drive innovation, both in technology and in the institutions supporting the healthcare system. Information is being created more rapidly in nearly every scholarly domain. EBM practitioners arrived early at the conclusion that medical evidence could be used to promote patient health, but this information is being produced at an increasing rate. By utilizing and furthering the tools presented herein, we hope that the health system will be able to more appropriately make use of that wealth of data to provide timely, accurate, and actionable evidence to a wide range of stakeholders.
Appendix
## A. Acronym List

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAP</td>
<td>Atypical Antipsychotics</td>
</tr>
<tr>
<td>ACE</td>
<td>Angiotensin-Converting-Enzyme</td>
</tr>
<tr>
<td>ADHD</td>
<td>Attention deficit hyperactivity disorder</td>
</tr>
<tr>
<td>AE</td>
<td>Adverse Effect</td>
</tr>
<tr>
<td>AHRQ</td>
<td>Agency for Healthcare Research and Quality</td>
</tr>
<tr>
<td>AUC</td>
<td>Area Under Curve</td>
</tr>
<tr>
<td>BMD</td>
<td>Bone Mineral Density</td>
</tr>
<tr>
<td>$BMD_t$</td>
<td>Bone Mineral Density $t$-score</td>
</tr>
<tr>
<td>BU-SVM</td>
<td>Bagged Undersampled Support Vector Machine</td>
</tr>
<tr>
<td>CDF</td>
<td>Cumulative Density Function</td>
</tr>
<tr>
<td>CENTRAL</td>
<td>Cochrane Central Register of Controlled Trials</td>
</tr>
<tr>
<td>CI</td>
<td>Confidence Interval</td>
</tr>
<tr>
<td>CMS</td>
<td>Centers for Medicare &amp; Medicaid Services</td>
</tr>
<tr>
<td>CRF</td>
<td>Conditional Risk Factor</td>
</tr>
<tr>
<td>CV</td>
<td>Clinical Vertebra</td>
</tr>
<tr>
<td>DEC</td>
<td>Differential Error Costs</td>
</tr>
<tr>
<td>EBM</td>
<td>Evidence-Based Medicine</td>
</tr>
<tr>
<td>EPC</td>
<td>Evidence-Based Practice Center</td>
</tr>
<tr>
<td>FDA</td>
<td>Food and Drug Administration</td>
</tr>
<tr>
<td>FOCUS</td>
<td>Flexible Osteoporosis Cost and Utility Simulation</td>
</tr>
<tr>
<td>HHS</td>
<td>Health &amp; Human Services</td>
</tr>
<tr>
<td>HIT</td>
<td>Health Information Technology</td>
</tr>
<tr>
<td>HTML</td>
<td>Hypertext Markup Language</td>
</tr>
<tr>
<td>IT</td>
<td>Information Technology</td>
</tr>
<tr>
<td>KP</td>
<td>Kaiser Permanente</td>
</tr>
<tr>
<td>LBD</td>
<td>Low Bone Density</td>
</tr>
<tr>
<td>MEPS</td>
<td>Medical Expenditure Panel Survey</td>
</tr>
<tr>
<td>NB</td>
<td>Net Benefit</td>
</tr>
<tr>
<td>NHANES</td>
<td>National Health and Nutrition Examination Survey</td>
</tr>
<tr>
<td>NIH</td>
<td>National Institutes of Health</td>
</tr>
<tr>
<td>NOF</td>
<td>National Osteoporosis Foundation</td>
</tr>
<tr>
<td>OHSU</td>
<td>Oregon Health Sciences University</td>
</tr>
<tr>
<td>PEARLS</td>
<td>Practical Evidence About Real Life Situations</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
</tr>
<tr>
<td>PICOTS</td>
<td>Populations, Interventions, Comparators, Outcomes, Timings, and Settings</td>
</tr>
<tr>
<td>PRISMA</td>
<td>Preferred Reporting Items for Systematic Reviews and Meta-Analyses</td>
</tr>
<tr>
<td>QALY</td>
<td>Quality-Adjusted Life-Year</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver Operating Characteristic</td>
</tr>
<tr>
<td>RR</td>
<td>Relative Risk</td>
</tr>
<tr>
<td>SRDR</td>
<td>Systematic Review Data Repository</td>
</tr>
<tr>
<td>SSA</td>
<td>Social Security Administration</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
</tr>
<tr>
<td>TF-IDF</td>
<td>Term Frequency-Inverse Document Frequency</td>
</tr>
<tr>
<td>VCR</td>
<td>Videocassette Recorder</td>
</tr>
</tbody>
</table>
B. Value of Information and Low Bone Density


While NHANES III contains a staggering amount of data about its participants, it does not inquire about paternal hip fractures. Because FRAX™ asks if either parent has fractured a hip, we first predict maternal hip fracture probabilities by race, and sex, then scale those estimates to reflect the true quantity of interest (hip fracture of either parent). Our underlying assumption is that the probability a patient’s mother has fractured a hip is proportional to the probability that either parent has fractured a hip. While this is a simplification, it allows average rates of parental hip fracture to be accurate and to fit the observed distribution of maternal hip fractures by race and sex. Stated differently, we assume that there is an underlying parental hip fracture risk for individuals of a given race and sex, and treat maternal hip fractures as a strong signal of that risk.

According to the National Hospital Discharge Survey, the percent of hip fractures affecting women is approximately 70%, which we denote $\text{pctWomenFx}$. If the empirical distribution of paternal fracture risk by race is close to that of maternal fracture risk, a reasonable approximation of paternal hip fracture is thus

$$\text{P} (\text{FatherHipFx}) = \text{P} (\text{MotherHipFx}) \times \frac{1 - \text{pctWomenFx}}{\text{pctWomenFx}}$$

which keeps the relative fraction of maternal hip fractures equal to $\text{pctWomenFx}$. As a simple check, assume temporarily that $\text{P} (\text{MotherHipFx}) = 0.2$. We would estimate of

$$\text{P} (\text{FatherHipFx}) = 0.2 \times \frac{1 - 0.7}{0.7} \approx 0.0857.$$  

With equal numbers of men and women, we would expect that women incur \( \frac{0.2}{0.2 + 0.0857} = 0.7 \) of all hip fractures. Because both parents can have a hip fracture, the estimated paternal hip fracture probability is not simply the summation, it is:

$$\text{P} (\text{ParentHipFx}) = \text{P} (\text{MotherHipFx}) + \text{P} (\text{FatherHipFx}) - \text{P} (\text{MotherHipFx} \land \text{FatherHipFx})$$
Making the assumption that the parents’ fractures are independent, this is

\[
P(\text{ParentHipFx}) = P(\text{MotherHipFx}) + P(\text{FatherHipFx})
- P(\text{MotherHipFx}) \times P(\text{FatherHipFx})
\]

Continuing the example, we would estimate the risk of hip fracture of either parent as \(0.2 + 0.0857 - (0.2 \times 0.0857) \approx 0.269\). This illustrates that we can scale race-specific estimates of maternal hip fracture probability to obtain reasonable estimates of the probability that either parent had a hip fracture.

### B.2. Initialization Regression Results

Results of the regression models used to initialize clinical risk factors appear in Table B.1.

### B.3. Group Risks from Relative Risks

Consider two events, A and B (for example, A could be having a fracture, and B the presence of a prior fracture). If only \(P(A)\), \(P(B)\) and relative risk(\(RR\)) are available, it is possible to derive \(P(A|B)\) using the following (noting that \(RR = \frac{P(A|B)}{P(A|\neg B)}\)):

\[
P(A) = P(A|B)P(B) + P(A|\neg B)P(\neg B)
\]

\[
P(A|B) = \frac{1}{P(B)} \left[ P(A) - P(A|\neg B)P(\neg B) \right]
\]

\[
= \frac{P(A)}{P(B)} - P(A|\neg B) \frac{P(\neg B)}{P(B)}
\]

\[
= \frac{P(A)}{P(B)} - \frac{P(A|B)}{RR} \frac{P(\neg B)}{P(B)}
\]

\[
P(A|B) + \frac{P(A|B)}{RR} \cdot \frac{P(\neg B)}{P(B)} = \frac{P(A)}{P(B)}
\]

\[
P(A|B)(1 + \frac{P(\neg B)}{RR \cdot P(B)}) = \frac{P(A)}{P(B)}
\]

\[
P(A|B) = \frac{P(A)}{P(B)} \left( \frac{RR \cdot P(B)}{RR \cdot P(B) + P(\neg B)} \right)
\]

\[
= \frac{P(A) \cdot RR}{RR \cdot P(B) + P(\neg B)}
\]

Therefore, \(P(A|B) = \frac{P(A) \cdot RR}{RR \cdot P(B) + P(\neg B)}\).
B.4. Treatment Strategies Implemented

- No Treatment
- $BMD_t \leq -2.5$
- $BMD_t \leq -2.0$
- $BMD_t \leq -2.5$ or had hip or spine fracture
- $BMD_t \leq -2.0$ or had hip or spine fracture
- $BMD_t \leq -1.5$ or had hip or spine fracture
- $BMD_t \leq -1.0$ or had hip or spine fracture
- $BMD_t \leq -0.5$ or had hip or spine fracture
- According to 2010 National Osteoporosis Foundation Guidelines:
  - Had hip or spine fracture or
  - $BMD_t \leq -2.5$ or
  - $BMD_t \leq -1.0$ and ten-year hip fracture risk $\geq 3\%$ or ten-year major fracture risk $\geq 20\%$

B.5. Meta-Analysis Forest Plots

The results of the meta-analysis are shown in Figure B.1

B.6. Recent AHRQ EPC Evidence Reports

Table B.2 shows the number of articles found in the literature search and days between the end of the literature search and publication for the most recent 15 EPC reports as of April 7, 2014.
Figure B.1.: Results of the meta-analysis using fractures of each type as the outcome of interest.
Table B.1.: Models used to generate clinical risk factors as a function of age, race, and sex. Bone Mineral Density uses linear regression, while the others use logistic regression.
<table>
<thead>
<tr>
<th>Report Num</th>
<th>Title</th>
<th># Found in Search</th>
<th>Days Between Search End and Publication</th>
</tr>
</thead>
<tbody>
<tr>
<td>214</td>
<td>Smoking Cessation Interventions in Pregnancy and Postpartum Care</td>
<td>2,454</td>
<td>390</td>
</tr>
<tr>
<td>213</td>
<td>Communication and Dissemination Strategies To Facilitate the Use of Health-Related Evidence</td>
<td>4,152</td>
<td>250</td>
</tr>
<tr>
<td>212</td>
<td>Oral Contraceptive Use for the Primary Prevention of Ovarian Cancer</td>
<td>6,476</td>
<td>365</td>
</tr>
<tr>
<td>210</td>
<td>Screening and Diagnosing Gestational Diabetes Mellitus</td>
<td>14,398</td>
<td>154</td>
</tr>
<tr>
<td>209</td>
<td>Multigene Panels in Prostate Cancer Risk Assessment</td>
<td>1,998</td>
<td>273</td>
</tr>
<tr>
<td>207</td>
<td>Allocation of Scarce Resources During Mass Casualty Events</td>
<td>5,716</td>
<td>213</td>
</tr>
<tr>
<td>206</td>
<td>Enabling Patient-Centered Care Through Health IT</td>
<td>17,749</td>
<td>700</td>
</tr>
<tr>
<td>205</td>
<td>Diagnosis and Management of Febrile Infants (0–3 Months)</td>
<td>2,676</td>
<td>556</td>
</tr>
<tr>
<td>204</td>
<td>An Evidence Review of Active Surveillance in Men With Localized Prostate Cancer</td>
<td>2,175</td>
<td>122</td>
</tr>
<tr>
<td>203</td>
<td>Enabling Health Care Decisionmaking Through Clinical Decision Support and Knowledge Management</td>
<td>15,176</td>
<td>486</td>
</tr>
<tr>
<td>202</td>
<td>Transition of Care for Acute Stroke and Myocardial Infarction Patients: From Hospitalization to Rehabilitation, Recovery, and Secondary Prevention</td>
<td>5,783</td>
<td>559</td>
</tr>
<tr>
<td>201</td>
<td>Enabling Medication Management Through Health Information Technology (Health IT)</td>
<td>32,785</td>
<td>304</td>
</tr>
<tr>
<td>200</td>
<td>Safety of Probiotics Used to Reduce Risk and Prevent or Treat Disease</td>
<td>11,977</td>
<td>242</td>
</tr>
<tr>
<td>199</td>
<td>Health Literacy Interventions and Outcomes: An Updated Systematic Review</td>
<td>3,569</td>
<td>310</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td></td>
<td><strong>9,077</strong></td>
<td><strong>352</strong></td>
</tr>
</tbody>
</table>

Table B.2.: High-level data on recent EPC reports, with references available at [http://www.ahrq.gov/research/findings/evidence-based-reports/numbered/index.html](http://www.ahrq.gov/research/findings/evidence-based-reports/numbered/index.html)
C. Timing Details

C.1. Empirical Distributions of Screening Times
**Figure C.1.**: Empirical Distribution of Screening Times. Colors represent different forms at the same screening level
Bibliography


[CND+12] Carolyn J Crandall, Sydne J Newberry, Allison Diamant, Yee-Wei Lim, Walid F Gellad, Marika J Suttorp, Aneesa Motala, Brett Ewing, Beth


[MDM+12] Mohammad Hassan Murad, Matthew T Drake, Rebecca J Mullan, Karen F Mauck, Louise M Stuart, Melanie A Lane, Nisrin O Abu Elnour, Patricia J Erwin, Ahmad Hazem, Milo A Puhan, Tianjing Li, and Victor M


[SJD+05] M Stevenson, M Lloyd Jones, E De Nigris, N Brewer, S Davis, and J Oakley. A systematic review and economic evaluation of alendronate, etidronate,


Bibliography


[WDS+13] Byron C Wallace, Issa J Dahabreh, Christopher H Schmid, Joseph Lau, and Thomas a Trikalinos. Modernizing the systematic review process to


