The Development of an Orthopedic Waiting List Algorithm for Elective Total Hip and Total Knee Replacement Surgery

January 2002

M. A. Ashworth
Ms. Carol Brulé
Mr. Andrew Day
Dr. Mark Harrison
Ms. Wilma Hopman
Dr. John Rudan

Funding provided by:
Canadian Health Services Research Foundation
Care Delivery Network of Southeastern Ontario
Principal Investigator:

Dr. M. A. Ashworth

Corresponding Author:

Ms. Wilma Hopman
Director
MacKenzie Health Services Research Department of Community Health and Epidemiology
3rd Floor
Abramsky Hall
Queen’s University
Kingston, ON K7L 2N6

Telephone: 613-545-2548
Fax: 613-545-6686

E-mail: hopmanw@post.queensu.ca

This document is available on the Canadian Health Services Research Foundation web site (www.chrsf.ca).

For more information on the Canadian Health Services Research Foundation, contact the foundation at:
11 Holland Avenue, Suite 301
Ottawa, Ontario
K1Y 4S1
E-mail: communications@chrsf.ca
Telephone: (613) 728-2238
Fax: (613) 728-3527

Ce document est disponible sur le site Web de la Fondation canadienne de la recherche sur les services de santé (www.fcress.ca).

Pour de plus amples renseignements sur la Fondation canadienne de la recherche sur les services de santé, communiquez avec la Fondation à l’adresse suivante :
11, avenue Holland, bureau 301
Ottawa (Ontario)
K1Y 4S1
Courriel : communications@fcress.ca
Téléphone : (613) 728-2238
Télécopieur : (613) 728-3527
The Development of an Orthopedic Waiting List Algorithm for Elective Total Hip and Total Knee Replacement Surgery

Dr. M. A. Ashworth 1
Ms. Carol Brulé 2
Mr. Andrew Day 3
Dr. Mark Harrison 1
Ms. Wilma Hopman 2,3
Dr. John Rudan 1

1 Department of Surgery, Queen’s University
2 MacKenzie Health Services, Department of Community Health and Epidemiology, Queen’s University
3 Clinical Research Unit, Kingston General Hospital

Acknowledgments:

The authors would like to thank all of the patients who gave freely of their time to participate in this study, the orthopedic surgeons who allowed us to access their patients, the secretaries who assisted us in patient identification and the Medical Records staff of both Kingston General Hospital and Hotel Dieu Hospital for their assistance.

In addition, the authors wish to thank the following individuals for their contributions to this work: Ms. Helen Coo, Ms. Kellie Squires, Dr. Alicia Paris, Mr. Claude Brulé, Ms. Dianne Groll, Ms. Cindy Bolton, Dr. Karen Seigel, Ms. Judy Tse, Mr. Kenneth Beckett, Dr. Richard Riopelle, Dr. John Gordon and the late Dr. Thomas Mackenzie.
# Table of Contents

Key implications for Decision Makers ................................................................. i

Executive Summary .......................................................................................... ii

Background ........................................................................................................... 1

Objectives .............................................................................................................. 3

Methods............................................................................................................... 3

Analysis............................................................................................................... 6

Results............................................................................................................... 9
  Descriptive Information .................................................................................. 9
  Changes from Pre-Surgical to 6-Month and 12-Month Post-Surgical Assessment ......................................................... 12
  Differences in Outcome for Hip Versus Knee Surgery ...................... 14
  Change Versus no Change in Outcomes ................................................... 15
  Regression Analyses .................................................................................. 15
  Time to Surgery ......................................................................................... 17
  Pre-Surgical changes While Followed on Waiting List ....................... 19
  Waiting List Algorithm .............................................................................. 24

Discussion and Conclusions ............................................................................ 29

References ......................................................................................................... 32
Key Implications for Decision Makers

- Some surgeons are already effectively and equitably arranging their waiting lists and prioritizing on the basis of disability and functional ability.

- A tested shared waiting list across the five Kingston, Ontario surgeons produced no evidence that sharing the waiting list would decrease the burden of waiting on health-related quality of life for patients.

- A questionnaire could be developed to assist in the prioritization of surgery. This questionnaire would look at lifestyle factors, pain, stiffness and functional ability. The answers should show the optimal time to do the surgery for patients who are expected to have a significant improvement in functional ability.

- The Medical Outcomes Trust 36-item short form (SF-36) questionnaire and the Western Ontario and McMaster Universities Arthritis Index (WOMAC) questionnaire were both used in this study. The SF-36 is more responsive to clinical improvement than the other questionnaires tested, and the WOMAC questionnaire was specifically designed for the evaluation of patients with osteoarthritis of the hip and knee. The WOMAC questionnaire is most useful when determining whether a patient should be put on a waiting list. When it comes to rating the overall impact of surgery, however, the mental and physical scores of the SF-36 are more useful than the WOMAC.

- The researchers developed a program using SAS software (version 8.1) that allowed the user to evaluate the health-related quality of life waiting cost for use in prioritization of patients, then slotted the patients on a waiting list across five surgeons.

- It would be appropriate for a questionnaire determining the prioritization of surgery to find out what patients’ expectations are, discuss them in the light of whether or not they have a significant other illness. It would then be appropriate to use two sets of criteria for those with the one significant illness requiring a hip or knee replacement, or more than one.

- The sample in this study is not guaranteed to be reflective of the general Ontario patient population.
Executive Summary

Elective total hip replacements and total knee replacements are the definitive treatment for the pain and disability caused by arthritic diseases of the hip and knee. Both procedures are increasing in prevalence throughout the developed world due to improved surgical methods and the aging population. These increases have not only had a significant impact on healthcare spending, but have also resulted in lengthy waiting lists for both procedures. There is a clear need to establish objective criteria for case selection and priority, as well as standard measures to assess the urgency of patients’ condition and the extent of expected benefit. Waiting lists should not only be managed on the basis of chronology, but on a number of other factors, including both the clinical and the social impact of delayed surgery.

The objectives of this research were as follows:

1. To assess changes in patient-perceived health-related quality of life and functional status while on the waiting list for hip and knee replacement, and at six months and one year following surgery.
2. To describe and identify factor(s) which predict the post surgical change in health and functional status of patients awaiting replacements.
3. To apply these factor(s) to the actual waiting list for total joint replacement, and compare the changes in health and functioning to those that would have (theoretically) resulted if surgery had been prioritized by the expected improvement as determined by these factor(s).

An observational, prospective outcomes design was used for this study. All patients on the waiting lists of five orthopedic surgeons in Kingston, Ontario between October, 1998 and March of 2001 were invited to participate. All primary hip and knee replacement patients (excluding revisions and fractures) were eligible to participate. Contact information for each eligible patient was obtained from the surgeons, and a survey package was mailed to each, containing a patient information and consent form, the Western Ontario and McMaster Universities Arthritis Index (WOMAC) questionnaire — used to assess quality of life when a patient has osteoarthritis of the hip and/or knee — and the Medical Outcomes Trust 36-item short form (SF-36). The SF-36 assesses eight health concepts: 1) limitations in physical activities because of health problems; 2) limitations in social activities because of physical or emotional problems; 3) limitations in usual role activities because of physical health problems; 4) bodily pain; 5) general mental health (psychological distress and well-being); 6) limitations in usual role activities because of emotional problems; 7) vitality (energy and fatigue); and 8) general health perceptions. Baseline scores were collected at the beginning of the study (for those who were already on the list), or at the time that they were put on the list (for those who were subsequently added to the list during the study period). Questionnaires were sent out every three months before surgery and at six and 12 months after the surgery.
To evaluate the proposed prioritization criteria, the differences between the pre-surgical and post-surgical scores were multiplied by the length of waiting time. It was possible to retrospectively examine the impact on the average 'cost' of waiting (health-related quality of life cost), when patients who experienced large improvements following surgery were operated on before patients who experienced less improvement. Due to the large number of outcomes, we used the SF-36 mental component score and physical component score, as they are widely used general measures of overall health-related quality of life, they incorporate all eight domains of the SF-36 and they measure two distinct and broad components of an individual’s general health state. We therefore restricted our evaluation of the cost of waiting (health-related quality of life cost) to the areas defined by the difference between the pre-surgical and the post-surgical mental and physical scores. In order to evaluate the cost of waiting under various re-ordering schemes, a program was developed using SAS software (version 8.1) that allowed the user to evaluate the waiting costs for any prioritization criteria.

A total of 673 patients were initially enrolled in the study. Of these, 658 were evaluable, 389 had surgery during the study period and 250 had useable six-month follow-up data. When patients were placed on the list, they scored far below the mean for age- and sex-matched peers on the SF-36 measure of health-related quality of life. Moreover, in spite of this poor level of functioning, patients continued to deteriorate while on the waiting list. This was more pronounced in hip replacement patients than in knee replacement patients. However, there is strong evidence to suggest that patients with worse physical functioning and more pain received surgery sooner than those experiencing less pain and impairment.

There were clear and statistically significant post surgical improvements in all domains of the Western Ontario and McMaster Universities Arthritis Index and all domains and summary scores of the SF-36. Hip replacement patients had poorer scores initially than the knee replacement patients did, but achieved better scores at the time of the six-month follow-up, reflecting poorer pre-surgical functioning and more post-surgical gain. There was, however, a small proportion of patients who did not improve following surgery. Regression analyses identified that better baseline scores (lower disease severity), a number of comorbid conditions, anesthetic risk and body mass index were associated with less improvement. These results suggest that some patients may receive surgery too early to achieve any significant benefit.

There were differences between surgeons in how severe their patient's illness was and in health-related quality of life at the time of the initial visit, suggesting that some surgeons place patients on the waiting list sooner than others would. However, these differences were no longer apparent at the time of surgery. Patients with poorer levels of functioning were receiving their surgery sooner than those with higher levels of functioning. This suggests that the surgeons already effectively stratify their patients and prioritize at least in part on the basis of disability and functional status.

Re-ordering of patients on the basis of maximum benefit, defined as maximum change in the physical and mental scores of the SF-36, indicated that while some gains may be possible under optimal circumstances, surgeons performed better than a 'first-come-first-served' and a random allocation. There was no strong evidence to suggest that a shared waiting list would improve health-related quality of life for patients.
A significant limitation of this study was the limited proportion of data available when the database was locked for analysis. Longer follow-up on the enrolled patients would substantially increase the proportion of available data and allow for more reliable results. Also, the sample is limited to five surgeons, which may limit the generalizeability of the results.

This report therefore supports the findings of a number of other researchers, who identified significant gains in quality of life and functional status following total joint replacement. It also finds that patients continued to decline while on the waiting list, but that surgery appears to be allocated in an equitable manner, based at least partially on the surgeons' subjective estimate of disease severity and patient disability.
**Background**

Elective total hip replacements (THR) and total knee replacements (TKR) are the definitive treatment for the pain and disability caused by arthritic diseases of the hip and knee. Both procedures are increasing in prevalence throughout the developed world due to a combination of improved surgical methods and the aging of the population. In the 1994/95 fiscal years, more than 25,000 total joint replacements (TJR) were performed in Canada alone\(^1\). Within Ontario, the rate of THR per 100,000 adults aged 20 years and older increased from 44 in 1981 to 84 by 1994, and the corresponding rate of TKR increased from 14 in 1981 to 80 by 1994\(^2\).

These increases have not only had a significant impact on health care expenditure\(^3\), but have resulted in problems with lengthy waiting lists for both procedures\(^2,5\). Waiting times vary widely from surgeon to surgeon and from hospital to hospital, and the length of waiting time is not always associated with the severity of the condition\(^6\). Often, it is determined by a complex array of factors, including system capacity, the number of patients on the list and the number of emergencies arising while elective cases are waiting\(^6\). One examination of waiting lists in Ontario found that while 14% of those in severe pain wait less than four months for surgery, 40% wait for over a year, suggesting a significant problem with access to this service\(^7\). Similar problems with waiting list management have been documented in the US, Australia, the United Kingdom and New Zealand\(^2,5\).

There is a clear need to establish objective criteria for case selection and priority, as well as standardized measures to assess the urgency of patients’ condition and the extent of expected benefit\(^5,6\). Waiting lists should not only be managed on the basis of chronology, but on a number of other factors, including both the clinical and the social impact of delayed surgery\(^5\). The wait for TJR in Canada has made it the focus, along with four other conditions (MRI scanning, cataract surgery, general surgery procedures and children’s mental health), of the Western Canada Waiting List Project\(^8\), which has attempted to address a number of these issues.
One of the primary tasks of the WCWL was to clearly define the key terms used in describing waiting lists, including severity, urgency, need and priority. They defined severity as the degree or extent of suffering, limitation of activity, or risk of death. Urgency and need were defined as severity plus considerations of the expected benefit and natural history of the condition. Relative priority was defined as urgency with or without consideration of social factors, while expected benefit was defined as the extent to which desired outcomes are likely to exceed undesired outcomes.

Their definition of priority is interesting in that it raises the possibility of social factors in addition to traditional aspects of pain and function. The relevance of non-clinical factors to surgical priority is an important but controversial topic, which does not yet have extensive research support. One study of prioritization for cardiac surgery found that when physicians rated a number of hypothetical cases which were designed to be of equal clinical urgency, non-clinical factors such as demographic information and lifestyle factors had an impact on the priority assigned. In a New Zealand assessment of priority for a number of different conditions, clinicians preferred to have certain non-clinical factors, such as ability to work, ability to live independently, and caring for dependents, incorporated into the criteria. Llewellyn-Thomas et al. recommend that information from the patients themselves also be considered, in addition to the clinical information available.

This is consistent with a number of outcome studies that have used patient perceptions to assess outcomes of surgeries such as TKR and THR, and has represented an important shift in the assessment of post-surgical outcomes. While there is no shortage of evidence that demonstrates the effectiveness of TJR in the management of pain and functional status, there has, until recently, been less research into the patient perspective of surgical success through the measurement of patient-perceived health-related quality of life (HRQOL). Maximizing function and well-being are important goals of health care, particularly in chronic disease. Measurement of HRQOL should therefore be included along with more traditional endpoints. This is particularly true for elective surgery, as there is evidence to suggest that some patients report little or no benefit from the procedure even when the surgery was a success from the perspective of the medical community. More recent work has supported the importance or assessing HRQOL for...
TJR\textsuperscript{8,19-22}. An excellent review of the outcomes literature for TJR is included in the final report of the Western Canada Waiting List Project\textsuperscript{8}.

**Objectives**

1. To assess changes in patient-perceived HRQOL and functional status while on the waiting list for TKR and THR, and at six months and one year following surgery.
2. To describe and identify factor(s) which predict the post-surgical change in health and functional status of patients awaiting TJR.
3. To apply these factor(s) to the actual observed waiting list for total joint replacement, and compare the observed changes in health status and function attained, to those that would have (theoretically) resulted if surgery had been prioritized by the expected improvement as determined by these factor(s).

**Methods**

An observational, prospective outcomes design was used for this study. All patients on the waiting lists of five orthopedic surgeons in Kingston, Ontario between October, 1998 and March of 2001 were invited to participate. All primary THR and TKR patients (excluding revisions and fractures) were eligible to participate. Contact information for each eligible patient was obtained from the surgeons, and a survey package was mailed out to each of them. The package included a covering letter explaining the nature of the study and our research objectives, a consent form, a background information form, the Medical Outcomes Trust 36-item health survey 1.0 (SF-36) as described below, and the Western Ontario MacMaster Universities Osteoarthritis Index (WOMAC), also described below. The background information form asked several questions that could not always be obtained from medical chart review, such as education level, marital status and employment status.

Baseline scores were collected at the beginning of the study (for those who were already on the list), or at the time that they were put on the list (for those who were subsequently added to the list during the study period). Mailed questionnaires were sent out every
three months pre-operatively until the time that the patients received their surgery. The
questionnaires were also sent at 6 and 12 months post-operatively.

Chart review was used to collect sociodemographic and clinical information, as well as
data relating to the surgical procedure. Sociodemographic data included age, gender,
whether they had support at home and whether they had dependents at home. Clinical
data included pre-surgical mobility, presence of osteoarthritis or rheumatoid arthritis in
the operative joint, history of previous joint replacement (revisions were excluded but
patients may previously have had a different joint replaced) and, if so, which joint.
Common comorbid conditions such as angina, chronic obstructive pulmonary disease,
smoking history, hypertension requiring medication, stroke, insulin-dependent diabetes
mellitus and depression were also included. Less frequent but significant comorbid
conditions such as liver disease, chronic renal failure and history of aortic abdominal
aneurysm were collected under a single variable for ‘other significant comorbidity’.
Surgical data included information such as medication use, transfusions, type of
prosthetic device, use of bone cement and hemoglobin levels. Additional data included
the admission and discharge date, length of stay, minutes of surgery and in-hospital
complications.

Health-related quality of life was measured by means of the Medical Outcomes Trust 36-
item health survey (SF-36)\textsuperscript{23}. The SF-36 contains 36 items that, when scored, yield 8
domains. Physical functioning (10 items) assesses limitations in physical activities such as
walking and climbing stairs. The role physical (4 items) and role emotional (3 items)
domains measure problems with work or other daily activities as a result of physical health
or emotional problems. Bodily pain (2 items) assesses limitations due to pain, while vitality
(4 items) measures energy and tiredness. Social functioning (2 items) examines the effect of
physical and emotional health on normal social activities, while mental health (5 items)
assesses feelings of happiness, nervousness and depression. The general health perceptions
domain (5 items) evaluates personal health and the expectation of changes in health\textsuperscript{23}.

All domains are scored on a scale from 0 to 100, with 100 representing the best possible
health state. One additional unscored item compares the respondent's assessment of their
current health to that of one year ago. Summary scores for a Physical Component Score
(physical functioning, role physical, bodily pain and general health perceptions) and a Mental Component Score (vitality, social functioning, mental health and role emotional) can also be derived\textsuperscript{24}.

The reliability and validity of the SF-36 and the two summary components have been well documented by the developers of the instrument\textsuperscript{25-28}. Comparison of a series of generic health status measures indicates that the SF-36 is not only psychometrically sound but is more responsive to clinical improvement than the other instruments tested\textsuperscript{29,30}. Moreover, health functioning changed in the hypothesized direction with increased age, socioeconomic status and disease status in a population-based longitudinal study of SF-36, suggesting that it is sensitive to changes in the health of the general population\textsuperscript{31}. Because the SF-36 focuses on generic health concepts not specific to any age, disease or treatment groups, it is useful for comparing the relative burden of different diseases, as well as for assessing the impact of interventions\textsuperscript{32}. Canadian normative data have recently been developed, providing an opportunity to compare the scores for various patient groups to their age- and gender-matched peers\textsuperscript{33}.

Condition-specific, functional status was assessed by means of the Western Ontario and MacMaster Universities Osteoarthritis Index (WOMAC), which was specifically designed for the evaluation of patients with osteoarthritis of the hip and knee\textsuperscript{34}. The WOMAC contains 24 items, and yields three domain scores. Five items contribute to the pain score, which specifically asks patients about pain they are currently experiencing due to arthritis in their hips and/or knees. Two items contribute to the stiffness score, which specifically asks about stiffness in their hips and knees. Seventeen items contribute to the physical functioning score, which relates to their ability to move around and look after themselves\textsuperscript{34}. Domain scores range from 0 (good) to 4 (poor).

The instrument was initially validated in a double blind study of 57 patients. All three domains met the conventional criteria for face validity, content validity, construct validity, reliability and responsiveness\textsuperscript{34}. Additional studies have also found the WOMAC to be valid, reliable and responsive\textsuperscript{3,35}, and to be an appropriate measure of lower body function in osteoarthritis\textsuperscript{36}. A number of studies that utilized both the SF-36 and the WOMAC found both measures to be responsive and suitable\textsuperscript{3,21,35,37}. The
WOMAC was more responsive than the SF-36 in small samples and in the short term\textsuperscript{38}, and had greater power to detect treatment differences\textsuperscript{37}. However, both the WOMAC and the SF-36 detected significant and meaningful changes in outcome following THR and TKR\textsuperscript{3,38}.

**Analysis**

Data were entered into SPSS 10.0 spreadsheets developed for this project. Both the SF-36 and the WOMAC were scored as per the instructions of the developers\textsuperscript{23,24,34}. Difference scores were developed to assess the changes from the pre-surgical assessment to the 6- and 12-month follow-up. The two questionnaires differ in that a high score is indicative of good function on the SF-36 while it represents poor function on the WOMAC. Because the regression analyses would be easier to interpret if positive difference scores represent positive outcomes, the difference scores for the SF-36 were calculated by subtracting the pre-surgical score from the post-surgical score, as we hypothesized that scores would improve following surgery. The reverse was done for the WOMAC scores, in that the post-surgical score was subtracted from the pre-surgical score. This has no impact on the absolute value of the changes, but facilitates interpretation of the results.

Descriptive statistics such as frequency counts, means and standard deviations were performed in order to assess the sample characteristics. Pearson correlation coefficients were calculated to determine the relationship between the various outcomes (8 domains and two summary scores of the SF-36, and the 3 WOMAC scores). Parametric tests were used, as most of the SF-36 domains tend to be reasonably normally distributed, permitting the use of parametric tests\textsuperscript{23}. Frequency tables were also developed to assess the number of patients who improved, experienced no change or worsened in their WOMAC or SF-36 scores. A significant portion of the analysis was devoted to assessing the comparability of the various groups to determine if they could, in fact, be collapsed or if they needed to be analyzed separately. This involved a number of independent t-tests to assess the differences between outcomes for unilateral and bilateral knee surgery patients, outcomes of hips versus knees and outcomes of those already on the list at the time the study started as opposed to those added later. Dependent t-tests were utilized to compare
the pre-surgical scores to the 6-month and to the 12-month follow-up, as well as to assess additional changes from 6- to 12-months. Stepwise linear regression was utilized to identify factors associated with changes in SF-36 domains and WOMAC scores. Baseline domain scores were included as predictors, but only the relevant domain score for the outcome under examination was included (e.g. baseline physical functioning was included in the regression for change in physical functioning). This was used to adjust for the score at baseline, and was necessary to account for the potential impact of floor and ceiling effects (e.g. those with poorer baseline scores potentially had more to gain). Probability of entry was set to .10 rather than the traditional .05 so as not to miss any important trends.

WOMAC and SF-36 data that were collected while patients were on the list were assessed by examining the changes from one 3-month time point to the next. Average change per month was calculated by dividing the difference between the first and last measure by the duration (in months) between the measures. The multivariate Wilks’ Lambda test was used to assess differences across surgery type (hips versus knees) and across surgeons.

The methodology utilized in the evaluation of the prioritization criteria was based on the fact that, for the majority of patients, there is a substantial improvement in HRQOL following a TKR or THR. This implies that most patients suffer a reduced HRQOL while waiting for surgery, as compared to their eventual post-surgical HRQOL. One method of assessing the cost of waiting for surgery is to examine the difference between the pre-surgical and post-surgical scores, and multiplying this by the length of waiting time. For example, if a patient has a SF-36 PCS of 25 when placed on the waiting list, waits for six months for surgery, and attains a post-surgical PCS of 35, the cost of waiting as measured by the PCS is equal to 60 PCS points*months (10 points multiplied by six months) for that patient. Alternatively, one who attains a 5-point increase in PCS following surgery but waits for 12 months would also have a total cost of 60 PCS points*months (5 points multiplied by 12 months). The cost of waiting is therefore directly proportional to 1) the magnitude of the difference between the pre-surgical and post surgical HRQOL and 2) the time a patient waited at this reduced level of QOL. For the minority of patients who get worse after surgery the cost of waiting is
negative, implying that the cost of waiting for surgery actually decreases as the waiting time increases. Using this method of assessing waiting list management, a waiting list would be optimally managed when the average cost of waiting, measured across all patients on the list, is as small as possible.

There are at least three factors that could potentially be modified to affect the average waiting list costs: 1) more surgeries can be performed so that waiting times are reduced; 2) alternative criteria may be used to enter patient on the waiting list, altering the size and/or patient composition of the waiting lists; or 3) the order in which patients on the list receive surgery may be altered such that patients with the most to gain from surgery are done earlier than those with less expected benefits. The first two factors are considered to be outside of the scope of the algorithm. However, it is possible to retrospectively examine the impact on the average cost of waiting, had those patients who experienced large improvements following surgery been done prior to those who experienced less improvement. In other words, we examined the impact on waiting costs when the patients were re-ordered to the available surgery times according to the magnitude of their post surgical improvements. The composition of the patients on the list and the number of available surgery times may not themselves be modified under our model. Thus, while the surgery times may be re-distributed across patients, the overall average waiting time remains fixed.

A large number of outcomes (three WOMAC domains, eight SF-36 domains and two summary component scores) were assessed in this study, making it very difficult to develop an algorithm that optimizes all outcomes. The WOMAC is an important outcome, in that it assesses pain, stiffness and physical functioning in the affected joint(s). However, the impact of osteoarthritis is also evident in poor pre-surgical HRQOL\textsuperscript{14,15,21}. The WOMAC scores are likely to be more useful than the SF-36 domain scores as criteria for adding a patient to the waiting list. However, the overall impact of surgery is probably better assessed by examining a broader array of outcomes, including the impact of surgery on such areas as social functioning, mental health and vitality, in addition to the more traditional endpoints of pain and physical functioning. For this reason, we chose to use the SF-36 MCS and PCS, as they are widely used general measures of overall health related QOL, incorporate all eight domains of the SF-36 and
measure two distinct and broad components of an individual’s general health state\textsuperscript{24}. We therefore restricted our evaluation of the cost of waiting (HRQOL cost) to the areas defined by the difference between the pre-surgical and the post-surgical MCS and PCS.

In order to evaluate the cost of waiting under various re-ordering schemes, a program was developed in SAS V8.1 that allowed the user to evaluate the waiting costs for any prioritization criteria. The criteria were provided to the program as parameters, and may be in the form of any equation or formula that evaluates to a single value. The program first evaluated the criteria provided, and then ordered the patients by assigning the next available surgery slot to the patient with the maximum criteria value. Timing issues were taken into consideration, so that patients could not be assigned a surgery date prior to the date that they were placed on the list. The program also had the option to optimally re-order the patients within the five surgeons or across all surgeons. Across-surgeon optimization allows slots of one surgeon to be filled by patients from a different surgeon, whereas for within surgeon optimization patients may only take surgery slots available for the surgeon that placed them on the list. The surgeons in Kingston do not share waiting lists at this time, so the within surgeon optimization more closely represents the current practice.

**Results**

**Descriptive Information**

A total of 1,038 patients were identified as potential participants between October, 1998 and March of 2001. Of these, 673 participated in the study and 365 did not. Among those who did not, the following is a breakdown of the reasons for non-participation:

<table>
<thead>
<tr>
<th>Reason</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Declined to participate</td>
<td>207</td>
</tr>
<tr>
<td>Missed (received name after surgery was complete)</td>
<td>38</td>
</tr>
<tr>
<td>Had surgery elsewhere</td>
<td>5</td>
</tr>
<tr>
<td>Revisions (therefore excluded)</td>
<td>18</td>
</tr>
<tr>
<td>Deceased</td>
<td>13</td>
</tr>
<tr>
<td>Canceled their surgery</td>
<td>21</td>
</tr>
<tr>
<td>Moved with no forwarding address</td>
<td>29</td>
</tr>
<tr>
<td>Other (mental illness, unable to speak English, etc.)</td>
<td>34</td>
</tr>
</tbody>
</table>
Of the 1,038 potential participants identified, 158 turned out to be ineligible for a variety of reasons. Of the remaining 880 eligible patients, 207 refused, for a participation rate of 76.5%. Independent t-tests (age) and chi-square tests (gender and type of surgery) indicated that there were no significant differences between those who participated and those who did not. No additional data were available for the non-participants.

The procedure for the 673 participants included 285 total hip replacements, 325 single total knee replacements, 48 bilateral total knee replacements and 5 bilateral hip replacements. This information was missing for 10 patients. Independent t-tests indicated that there were few significant differences in outcome between the single and the bilateral knee replacements. The single replacements had an average improvement of 19.7 for the physical functioning domain of the SF-36, while the bilateral replacement group had an average change of 35.5 (p = .02). The bilateral group also had slightly more improvement in the role physical functioning domain of the SF-36 (45.5) as compared to the single replacement group (21.2, p = .04). Other than these two differences, however, the two groups were comparable and were therefore combined for all subsequent analyses to increase statistical power. Those who were missing the type of surgery were excluded from all surgery-specific analyses, and the 5 bilateral hip replacement patients were also excluded from further analyses, as they represent a subgroup that is different from the others. This left a group of 658 patients who were scheduled for unilateral hip replacement, unilateral knee replacement or bilateral knee replacement and who were evaluable for this analysis. Of these, 22 were removed from the waiting list analysis due to incomplete data, leaving a total of 636 patients (271 THR and 365 TKR patients). Table 2 outlines the rate of surgery within the two groups.

<table>
<thead>
<tr>
<th>Table 2: Rate of Surgery for THR and TKR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hip versus knee</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>hip</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>knee</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Total</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
Rates of THR and TKR varied substantially between surgeons. The rates by surgeon are presented in Table 3. In order to protect the identity of the surgeons, they were randomly assigned a number of 1 through 5.

**Table 3: Rates of THR and TKR by Surgeon**

<table>
<thead>
<tr>
<th>Hip versus knee</th>
<th>Surgeon</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>hip</td>
<td>9</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>20.0%</td>
<td>58.8%</td>
</tr>
<tr>
<td>knee</td>
<td>36</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>80.0%</td>
<td>41.2%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>45</td>
<td>85</td>
</tr>
</tbody>
</table>

Appendix 5.1.1 contains the sample characteristics for the entire group, by surgical group (hips or knees) and for those for whom follow-up data are available. The average age of the sample was 68.4 years. There were no significant differences in the mean age of the hip patients (67.9 years) and the knee patients (69.0 years). There were 292 (43.4%) males and 371 (55.1%) females in the sample, with an additional 10 (1.5%) for whom these data were missing. The differences in the proportion of males and females in the hip versus knee sample approached significance (p = .069), with 48.1% males in the hip sample but only 40.8% males in the knee sample. No other characteristics approached significance.

Appendix 5.1.2 contains additional sample characteristics obtained from the chart review, presented for those who had both their surgery and a 6-month follow-up, as this is the group on which the subsequent regression analyses are based. These results are presented for the entire sample and for THR and TKR separately. There were significant differences between the THR and TKR patients in the number who had osteoarthritis in other joints (17.1% versus 36.5%, p = .017). The difference in time on the waiting list approached significance (9.6 versus 11.2 months, p = .08), as did the presence of other significant comorbidity (7.6% versus 14.9%, p = .08). No other variables approached significance. Appendix 5.1.3 contains additional information collected from the chart, and presented for descriptive reasons, but not used in the regression analyses. They were
not used because we wished to be able to predict outcomes based on the patient characteristics while on the waiting list.

**Changes from Pre-Surgical to 6-Month and 12-Month Post-surgical Assessment**

Appendix 5.2.1 contains the pre-surgical, the 6-month follow-up and the 12-month follow-up scores for the entire sample, as well as for the hips and the knees separately. Figures 1 – 3 contain the domain score changes for the SF-36 scores for the entire sample, total hip patients alone and total knee patients alone for the pre-surgical assessment and the 6- and 12-month follow-up. In addition, these figures contain the age-matched normative data for the SF-36. Figure 4 contains the WOMAC data for the entire sample for the pre-surgical assessment and the 6- and 12-month follow-up, while Figure 5 contains the WOMAC data for the hip replacement and knee replacement patients separately. No normative data are available for the WOMAC. Figure 6 contains the data for an SF-36-based domain, called the lifestyle (LS) domain, which is a combination of role physical, role emotional and social functioning domains. The LS domain was developed on the basis of the information obtained during the assessment of the scores while patients were on the waiting list, and will be described in more detail in the next section. A LS score was not calculated for the 12-month follow-up, as the 6-month and the 12-month follow-up data are very similar for all of the outcomes tested, and as a result only the 6-month follow-up data are used in most of the analyses.

Appendix 5.2.2 contains the results from the one-sample t-tests to examine the changes in outcome from the pre-surgical to the 6-month follow-up, pre-surgical to the 12-month follow-up and 6-month follow-up to the 12-month follow-up. These data support what is already evident in the Figures. There are clear and statistically significant improvements in all WOMAC domains, all SF-36 domains and both summary scores (p < .001 in all cases). The data for the 6-month follow-up, as compared to the 12-month follow-up, indicate that there are few changes after the first six post-surgical months. WOMAC pain scores, which improved greatly in the first six months, showed a small, non-significant worsening (p = 0.31). Three of the SF-36 domains also showed slight deterioration in the 6-month to the 12-month follow-up (general health perceptions, p = 0.37; vitality, p = 0.32; mental health component score, p = 0.61). This pattern was consistent for both the hip and knee subgroups. Because there are no significant changes from 6 to 12 months,
only the changes from the pre-surgical assessment to the 6-month follow-up assessment are examined in detail. Data for the 12-month follow-up were not used for regression analyses for the same reasons, and because the sample size drops significantly for the 12-month follow-up. Appendix 5.2.2 also contains the dependent t-tests for the pre-surgical to the 6-month follow-up for the hips patients separately, and for the knee patients separately. Data for pre-surgical to 12 months, and for 6 to 12 months, are not shown as they are similar to the overall statistics (hips and knees combined), with large significant changes from pre-surgical to 12 months, and little change from 6 to 12 months.

Appendix 5.2.3 contains the Pearson’s correlations between the SF-36 domains and the WOMAC scores for the pre-surgical and the 6-month follow-up values. For the presurgical values, correlations ranged from a low of .251 (between the SF-36 general health perceptions and the SF-36 physical functioning domains) to a high of .848 (WOMAC physical functioning and WOMAC pain). The results for the 6-month follow-up are similar, with correlations ranging from a low of .279 (SF-36 role physical with the SF-36 mental health) and a high of .865 (WOMAC physical functioning and WOMAC pain). All correlations were in the expected directions, keeping in mind that higher SF-36 scores and lower WOMAC scores represent better outcomes, and all were significant (p < .001). All change scores were also significantly correlated, again in the expected direction (data not shown). The weakest correlation was between the change in mental health and the change in role physical functioning (r = .118, p = .05). The majority of change scores were highly correlated with significance levels in excess of p < .001.

Figures 7a – 7n contain a closer examination of the relationship between the pre-surgical scores for all WOMAC and SF-36 domains, and the change in the SF-36 PCS and MCS scores. The PCS and MCS were chosen as they are linear combinations of the eight SF-36 domain scores, and thus represent summaries of the physical and mental outcomes. Each page contains two graphs, with the top one containing the relationship between each of the 14 outcomes with the PCS change, and the lower one displaying the relationship between each of the 14 outcomes and the MCS change. For example, Figure 7a contains the relationship between the pre-surgical WOMAC pain scores and the PCS (top graph) and the MCS (bottom graph). It can be seen that there was fairly consistent improvement in the PCS regardless of the pre-surgical level of WOMAC pain, although there are a few
at the lower WOMAC level (less pain) that did not experience improvement. For the
MCS, however, it can be seen that there only appeared to be improvement at higher
(greater pain) levels of WOMAC pain. A LOWESS (locally weighted scatterplot
smoother)\textsuperscript{39,40} as implemented in S-Plus software, was used to display the trend, which
was often not linear. Most of the figures (7a – 7n) display the same pattern of greater
improvement in the physical summary score than in the mental summary score, with
consistent improvement in PCS scores across the pre-surgical scores, but with
improvement in the MCS only apparent at higher levels of pain and disability. This
pattern indicates that the pre-surgical scores are more predictive of change in the MCS
than in the PCS, even though the changes in the PCS tend to be larger. On several of the
figures (e.g. 7g SF-36 general health perceptions), better levels of pre-surgical
functioning are in fact associated with a decline in the MCS post-surgically, suggesting
again that for some, surgical intervention may have been too early for significant benefit.

**Differences in Outcome for Hip Versus Knee Surgery**

Appendix 5.3.1 contains the data from the independent t-tests that were used to assess the
baseline and six-month follow-up scores for those who had hip surgery (procedure 1) as
compared to those who had knee surgery (procedure 2, single and bilateral procedures
included). It is apparent from the results that the hip replacement patients had poorer pre-
surgical scores in all WOMAC and SF-36 domains. These differences were significant,
or very close to significance, for all but the general health perceptions, role emotional,
mental health and mental component (summary) scores. At the time of the 6-month
follow-up the opposite results could be seen. The hip replacement patients consistently
achieved better (lower) scores in all WOMAC domains, and all but one SF-36 domain
(the 6-month follow-up role physical scores for the two groups were identical). The
differences between the hip and knee groups were significant or close to significance for
all three WOMAC domains, as well as for the bodily pain and general health perceptions
domains of the SF-36. A comparison of the actual change scores (difference from
baseline to 6-months) of the two groups (Appendix 5.3.2) supports the finding of larger
changes in the hip replacement patients as compared to the knee replacement patients.
Change versus No Change in Outcomes

Not all patients experienced improved outcomes. Table 4 outlines the percentage of those who had a worse score, the same score and an improved score from baseline to the 6-month follow-up assessment. Chi-square tests indicated that there were no significant differences in the proportions within each category for the hip and the knee patients, so only the overall numbers are presented.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Worsened</th>
<th>Stayed the Same</th>
<th>Improved</th>
</tr>
</thead>
<tbody>
<tr>
<td>WOMAC Pain</td>
<td>8.2 %</td>
<td>2.9 %</td>
<td>88.9 %</td>
</tr>
<tr>
<td>WOMAC Stiffness</td>
<td>6.9 %</td>
<td>7.9 %</td>
<td>75.2 %</td>
</tr>
<tr>
<td>WOMAC Physical Functioning</td>
<td>9.7 %</td>
<td>0.4 %</td>
<td>89.9 %</td>
</tr>
<tr>
<td>SF-36 Physical Functioning</td>
<td>11.7 %</td>
<td>6.4 %</td>
<td>81.9 %</td>
</tr>
<tr>
<td>SF-36 Role Physical Functioning</td>
<td>6.1 %</td>
<td>44.4 %</td>
<td>49.5 %</td>
</tr>
<tr>
<td>SF-36 Bodily Pain</td>
<td>8.5 %</td>
<td>10.7 %</td>
<td>80.8 %</td>
</tr>
<tr>
<td>SF-36 General Health Perceptions</td>
<td>30.5 %</td>
<td>9.1 %</td>
<td>60.4 %</td>
</tr>
<tr>
<td>SF-36 Vitality</td>
<td>16.4 %</td>
<td>8.5 %</td>
<td>75.1 %</td>
</tr>
<tr>
<td>SF-36 Social Functioning</td>
<td>13.1 %</td>
<td>20.2 %</td>
<td>66.7 %</td>
</tr>
<tr>
<td>SF-36 Role Emotional Functioning</td>
<td>18.7 %</td>
<td>49.1 %</td>
<td>32.2 %</td>
</tr>
<tr>
<td>SF-36 Mental Health</td>
<td>25.3 %</td>
<td>13.9 %</td>
<td>60.9 %</td>
</tr>
<tr>
<td>SF-36 Physical Component Score</td>
<td>12.2 %</td>
<td>0 %</td>
<td>87.8 %</td>
</tr>
<tr>
<td>SF-36 Mental Component Score</td>
<td>40.5 %</td>
<td>0 %</td>
<td>59.5 %</td>
</tr>
</tbody>
</table>

Appendix 5.4.1 contains these data in significantly more detail, outlining the number, the mean and the standard deviation within each category for both the baseline and the 6-month follow-up data. Appendix 5.4.1 demonstrates that in all WOMAC and SF-36 domains the baseline scores for those who do not improve are substantially better than in those who did show improvement. This implies that those patients still had a fairly high level of functional ability, thus they had less functional ability to gain from the surgery. Given the risks associated with surgery, this group had potentially less to gain and more to lose by receiving surgery, and may in fact have received the surgery too early.

Regression Analyses

There are a large number of stepwise regression analyses, as there are 13 outcomes (3 WOMAC domains, eight SF-36 domains and two SF-36 summary scores). In addition, regression analyses were also completed for the Lifestyle Score (the LS score is
described below). Fourteen analyses were therefore run for the entire sample (combining hips and knees) and for the hip patients and the knee patients separately. When analyzing the entire sample, a variable identifying hips versus knees was added as a covariate. The models are contained in Appendix 5.5.1, along with a one-page summary of all 14 models.

All regression models were highly significant (p < .001 in all cases). The r-square values ranged from a low of .119 for the role physical domain of the SF-36 to a high of .467 for the role emotional domain of the SF-36. Several points of interest can be identified. First, it is clear from the direction of the relationship that those with poorer pre-surgical scores achieve a greater improvement in their post-surgical scores. As mentioned previously, high pre-surgical WOMAC scores reflect poor functioning while high pre-surgical SF-36 scores reflect better functioning. However, since the regression analyses would be easier to interpret if positive difference scores represent positive outcomes, the difference scores for the SF-36 were calculated by subtracting the pre-surgical score from the post-surgical score, while the reverse was done for the WOMAC scores.

The models also identify the larger improvements in the hip patients as compared to the knee patients (hip was coded as 0, knee was coded as 1). Those who were working had more improvement in 5 domains of the SF-36. Comorbid disease was commonly associated with less improvement, particularly a history of cancer, history of heart disease (angina and/or myocardial infarction) and the variable coded as ‘other significant comorbidities’, which included a number of low-frequency but serious health problems such as liver disease, chronic renal failure and history of aortic abdominal aneurysm. The presence of rheumatoid arthritis was associated with poorer outcomes, as was a higher ASA risk score (American Society of Anesthesiology score where 1 = no systemic disease and 5 = a moribund patient with little chance of survival). The assignment of the ASA risk score relies heavily on the presence and severity of comorbid conditions, so this finding is consistent with the finding of less improvement in those with comorbid conditions.
The stepwise regression models for the subset of total hip replacement patients (Appendix 5.5.2) and for the subset of total knee replacement patients (Appendix 5.5.3) are similar to the models for the two groups combined. For the THR patients, all the regression models were highly significant (p < .001) with the exception of the model for the SF-36 role physical, which was significant at p < .05. The r-square values ranged from a low of .07 in the SF-36 role physical domain, to a high of .58 for the SF-36 role emotional domain. As in the overall sample, those with poorer pre-surgical scores have higher post-surgical scores. Those with poorer pre-surgical mobility had better outcomes. High ASA risk scores, higher body mass index and a number of comorbidities were associated with less improvement.

For the TKR patients, all models were highly significant, with (p ≤ .001), with r-square values ranging from a low of .16 (SF-36 Physical Component Score) to a high of .51 (SF-36 Mental Component Score and the SF-36 mental health domain score). Comorbid conditions were once again the most common variables associated with a poorer outcome, particularly a history of cancer, depression, rheumatoid arthritis, myocardial infarction and the miscellaneous comorbid conditions captured under ‘other significant comorbidities’. Two factors that were associated with a positive outcome in a number of domains included current employment and higher education levels.

**Time to Surgery**

Time to surgery was very heterogeneous across surgeons, with median waiting times varying from 6.14 months to 20.0 months depending on the surgeon (Table 5). The log-rank tests for the equality of waiting times across surgeons yields a p-value less than 0.0001, indicating highly significant differences in waiting time by surgeon.

The Cox proportional hazards model was employed to examine the effect of several factors on time to surgery. Each factor was modeled individually since some of the factors were highly correlated and we were interested in the effects of the individual factors rather than developing a predictive model. Due to the heterogeneity of waiting times (Table 5), all models were stratified by surgeon (1 – 5) in the analysis.
Appendix 5.6.1 demonstrates that each of the 3 WOMAC scores and the SF-36 physical functioning domain score have a significant impact on waiting times. In particular, patients with a greater physical impairment (high WOMAC scores or low SF-36 physical functioning scores) had an accelerated time to surgery compared to patients with less physical impairment. This suggests that the surgeons were aware of their disability level and may have stratified their lists on the basis of this knowledge. The trend is similar in the non-physical domains of the SF-36, but the significance is not as clear. A hazard ratio of >1 is associated with an accelerated time to surgery. The scores are scaled such that the hazard rates are associated with an increase in one point for the WOMAC scores and 10 points for the SF-36 scores. The scaling is arbitrary and does not affect the final p-values. The effect of the demographic factors on waiting times is also provided in Appendix 5.6.1. No factors are significant at p=0.05, although the difference between hip and knee patients approaches significance (p = 0.08), as does the level of education (p = 0.09).

### Table 5: Median Waiting Times by Surgeon

<table>
<thead>
<tr>
<th>Surgeon</th>
<th>Total Patients on List</th>
<th>Still Waiting at end of Study (Censored)</th>
<th>Had Surgery While Followed</th>
<th>Median Time to Surgery (Months)</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower</td>
</tr>
<tr>
<td>1</td>
<td>45</td>
<td>8</td>
<td>17.8%</td>
<td>37</td>
<td>7.359</td>
</tr>
<tr>
<td>2</td>
<td>85</td>
<td>26</td>
<td>30.6%</td>
<td>59</td>
<td>7.819</td>
</tr>
<tr>
<td>3</td>
<td>232</td>
<td>123</td>
<td>53.0%</td>
<td>109</td>
<td>19.679</td>
</tr>
<tr>
<td>4</td>
<td>106</td>
<td>64</td>
<td>60.4%</td>
<td>42</td>
<td>25.988</td>
</tr>
<tr>
<td>5</td>
<td>168</td>
<td>48</td>
<td>28.6%</td>
<td>120</td>
<td>6.144</td>
</tr>
<tr>
<td>Overall</td>
<td>636</td>
<td>269</td>
<td>42.3%</td>
<td>367</td>
<td>9.331</td>
</tr>
</tbody>
</table>

Disability level and hip/knee patients approaches significance (p = 0.08)
The final table in Appendix 5.6.1 presents the effect of support at home, a number of comorbid conditions and ASA risk on the waiting times. Since these variables were only collected for patients who had surgery, there are no censored patients in these models. The number of events is similar since data were present for patients who had surgery. Presence of osteoarthritis in the operative joint is associated with an accelerated time to surgery (hazard ratio=1.6 p=0.016). In addition, history of stroke, presence of a caregiver in the home and presence of diabetes (insulin-dependent) are all of borderline significance, although their significance is questionable in light of the number of comparisons being made.

This analysis provides strong evidence that patients with worse physical functioning, as measured by the 3 WOMAC scores and the physical functioning domain of the SF-36, received surgery sooner than those with less pain and impairment. For the most part there is little indication that the demographic or comorbid factors played an important role in the waiting times.

**Pre-Surgical Changes While Followed on Waiting List**

**Baseline and Pre-Surgical Scores**

Prior to analyzing the change in patient pre-surgical scores, it is first necessary to describe the distributions of both the initial and immediate pre-surgical scores. Pre-surgical waiting list data are available for 636 patients. Appendix 5.7.1 describes the first score collected for these patients. Of the 636 patients, 122 were already on the waiting list at the time of study initiation. For this group the first measure is not a measure of the patient functioning at the time they were entered onto the waiting list, but rather it is a measure of their state at the time they were entered in to this study. The median time from entry on the waiting to enrollment into this study was approximately 3 months overall and 6 months for the patients already on the waiting list at the start of the study.

The initial scores are also presented by surgery type (hip versus knee) and surgeon (1 – 5) in Appendix 5.7.1. In order to collectively test if the means of these initial scores were significantly different across surgery type, we employed the multivariate Wilks’ Lambda test. This test was performed using a one way layout with the three WOMAC scores and
the eight SF-36 domain scores as the dependent variables, with surgery type as the independent variable. The component scores (PCS and MCS) were excluded, since they are simply linear combinations of the domain scores already included in the model. The Wilks’ Lambda resulted in a p-value of 0.0004 under the assumption of multivariate normality. This provides evidence that there is a statistical (although perhaps not clinically) significant difference in the health state of the two groups as measured by the first collected QOL scores. A univariate analysis of the initial scores shows that the QOL scores are significantly worse for the WOMAC physical functioning, SF-36 bodily pain and SF-36 vitality scores (p=0.0012, p=0.035 and p=0.025 respectively) in the hip patients when compared to the knee patients. The multivariate approach was repeated to test if there was a difference in the initial scores across surgeon. The Wilks’ Lambda resulted in a p-value of 0.0041, indicating that there was a difference in the initial QOL scores across surgeons. There was, however, no clear indication that those patients experiencing greater pain and disability were assigned to the surgeons with shorter waiting lists, with the exception of Surgeon 2 who had patients with higher WOMAC scores (greater disability) and lower SF-36 scores (greater disability) than the other surgeons at the time the initial assessment.

Appendix 5.7.2 contains the final pre-surgical scores for the subset of patients who had surgery. Generally these pre-surgical scores are within three months of surgery. The scores are also presented by surgery type (hip versus knee) and surgeon (1 – 5) respectively. The multivariate test for a difference in surgery type results in a Wilks’ Lambda p-value of 0.0054. As in the first assessment, the hip patients had worse QOL scores prior to undergoing surgery than did the knee patients. Testing for an overall surgeon effect prior to surgery was not significant (p = 0.095), with roughly equivalent levels of disability on both the WOMAC and the SF-36 domains. Although Surgeon 2 still had patients with somewhat lower SF-36 domain scores than the other four surgeons, the patients of three of the other four experienced more decline while waiting.
**Change Scores**

For 284 (44.7%) patients, there was only one pre-surgical assessment. As a result, these patients could not contribute to the waiting list portion of the study. One hundred and eighty six (29.3%) patients had two assessments, 67 (10.5%) had three assessments, 37 (5.8%) had four and 62 (9.7%) had five or more.

The change scores represent the change from the first observed to the last observed pre-surgical measure. As mentioned earlier, 122 (34.7%) of the 352 evaluable patients were already on the waiting list at the time they were entered into the study, and 144 (40.9%) patients did not have surgery during the duration of the study. Thus, the change from first to last observed pre-surgical measure should not be interpreted as an estimate of the change while on the waiting list, but rather the change that we observed while following the patients pre-surgically.

The average time from the first to the last observed pre-surgical visit was 7.7 months. Appendix 5.7.3 contains the average change scores during this period for each of the WOMAC and SF-36 scores. A statistically significant worsening is evident in all three WOMAC scores and all SF-36 domain and summary scores. The most significant deterioration can be seen in the SF-36 domains of role physical functioning (mean change of −6.061, SD 32.44), role emotional functioning (mean change of −8.879, SD 45.96) and social functioning (mean change of −6.108, SD 25.14). These are the only three domains that had a change large enough to be considered socially and clinically relevant by the developers of the SF-36\(^{23}\).

Deterioration was much more pronounced for the hip patients than for the knee patients. For the hip patients, role physical functioning mean change was −11.450 (SD 33.72), role emotional functioning mean change was −11.959 (SD 45.30) and social functioning mean change was −8.874 (SD 23.23). In addition, the physical functioning domain of the SF-36 also deteriorated significantly (mean change −7.123, SD 20.87). For the knee patients, role physical functioning mean change was −2.866 (SD 31.30), role emotional functioning mean change was −7.045 (SD 46.35) and social functioning mean change was −4.468 (SD 26.12). The hip replacement patients experienced a greater mean
deterioration than the knee replacement patients in all scores other than the SF-36 general health perceptions domain, where the change was very similar for the two groups (mean change of –3.250 for THR and -3.508 for TKR patients).

The three domains that appeared to change the most are the role physical, the role emotional and the social functioning domains of the SF-36. Because these domains measure a person’s ability to perform their day-to-day activities, both physically and mentally, and their ability to carry on normal social activities, we examined these three domains as a single entity that we termed the Lifestyle (LS) domain. The other reason for focusing on these three domains is that total hip and total knee replacements are, to a certain extent, lifestyle surgeries. All of those on the list had significant levels of pain, stiffness and problems with physical functioning on the WOMAC measure. The domains contributing to our LS score are those that continue to deteriorate, suggesting that perhaps these factors should be measured along with functional level and pain, and used to prioritize those on the waiting list. Using a principal components analysis, we identified that the first principle component of these domains weighted each of the three factors near equally and explained most of the variation, therefore the average of the three scores was calculated to obtain the LS score. Mean change in the LS score was –6.629 (SD 25.36) for the entire sample, -10.523 (SD 24.52) for the THR patients and –4.321 (SD 25.61) for the TKR patients.

In order to collectively test if the means of these change scores were significantly different in the two treatment groups, we employed the multivariate Wilks’ Lambda test as described previously. The Wilks’ Lambda resulted in a p-value of 0.29 under the assumption of multivariate normality. Although the multivariate test did not reveal an overall significant effect, univariate tests indicated a significant difference between the hip and knee patients for the physical functioning domain, the role physical domain and the physical component score at p=0.0021, 0.017 and 0.012 respectively.

Appendix 5.7.3 also contains the change scores by surgeon. The same multivariate strategy was applied to detect an overall surgeon effect. In this case, the Wilks’ Lambda resulted in a p-value of 0.22.
Appendix 5.7.4 contains information relating to the average change per month. This value is calculated by dividing the difference between the first and last measure by the duration (in months) between these measures. While the change per month trends are similar to the absolute change trends, the significance is not as striking and the average change for some scores is no longer significantly different from zero.

Appendix 5.7.4 also presents the average monthly changes by surgery type and surgeon. The multivariate Wilks’ Lambda for overall surgery type effect yielded a borderline p-value of 0.053 while the same statistic yielded a p-value of 0.020 for the surgeon effect. This provides an indication that the hip patients may deteriorate at a steeper rate than the knee patients, and that the rate of deterioration may differ across surgeons. These results must be interpreted with caution, as the Wilks’ Lambda test statistic requires that the dependent variables come from a multivariate normal distribution. Upon investigation it was seen that while the individual slope scores have a roughly normal symmetric shape, there is moderate excess in the tails of the distributions. Furthermore, the significance of these borderline p-values is quite questionable when one considers the number of comparisons being made in this study.

**Graphical Representation of Changes on the Waiting List**

Figure 8a – 8n plot each of the WOMAC and SF-36 domain and summary scores over time for patients while they were waiting for surgery. The immediate pre-surgical measures are denoted by a "+" symbol. These measures are the last taken prior to surgery for those patients who had surgery. Surgery was normally performed within three months of these immediate pre-surgical measures. Individual patients may have repeated measures on these graphs. As described earlier, 352 of the 636 patients with pre-surgical measures had more than one assessment. The average number of assessments was 2.2 with a range of one to nine.

The robust LOWESS (locally weighted scatterplot smoother) smoother\textsuperscript{39,40} was used to display the trends of the data over time. The solid line describes the centre of the data points (waiting list scores) at each time. The dotted line represents a smooth of the 6-month post-surgical scores for those patients followed at the given times. For the SF-36
scores, we also include a smooth of the Canadian normative scores matched by age and gender for the patients followed at a given time\textsuperscript{33}.

Through these figures, several consistent and interesting characteristics of the data emerge. For most of the scores there is little, if any, change in the scores over time. This indicates that patients who have been on the waiting list for a long time are not noticeably worse those patients who have been on for a short time. We know that individual patients do deteriorate somewhat over time, so we are probably seeing the effect of the worst patients being treated and removed from the list. In the time on waiting list analysis described earlier, we do in fact show that patients with worse initial scores are treated earlier.

The difference between the dotted and solid lines describes the cost of waiting in terms decreased functioning and quality of life. The solid line represents the local centre of the scores while patients were waiting. These same patients reach a QOL score represented by the dotted line following surgery. It can be seen that there is a substantial post-surgical improvement in the scores for the patients who have been on the waiting list for less than two years. However, this improvement substantially decreases for patients who have been on the list for a longer time. In fact, the post- surgical and waiting list curves actually cross at approximately three years for many of the outcomes, indicating that the few patients on the list for greater than three years may actually get worse after surgery. Since there are only five patients on the list after three years who had surgery, the crossing of the curves after three years may be an artifact of this sample and not reflective of the general population. However, it does appear that in this sample, patients who received greater benefit from surgery tended to be done earlier than those who benefited less.

**Waiting List Algorithm**

Although 658 evaluable patients were enrolled in the study, only 389 had surgery during the study period and only 250 (38\%) had six-month follow-up PCS and MCS data. An additional 17 patients were lost because their PCS and/or MCS score could not be calculated either pre-surgically or post-surgically due to missing data. Since we compared the average waiting list areas (HRQOL cost) under various re-ordering scenarios to the actual observed areas, we needed both PCS values and both MCS values.
for each patient in the analysis. Thus, this analysis was only performed on the 233 patients who had surgery and had complete pre-surgical and six-month follow-up data on the PCS and the MCS. We acknowledge that the limited proportion of data available at the time that the database was locked for analysis is a significant limitation of this study. A further year or two of follow up would substantially increase the proportion of available data and allow for much more reliable results.

The cost of waiting in terms of the PCS and MCS area under various re-ordering schemes is presented in Table 6. All comparisons are based on the same 233 patients described above. Each of the compared schemes simply reallocated the 233 patients to different waiting times on the basis of the criteria used (e.g. maximum expected change in PCS, first come first served, etc.). As described previously (methods), the waiting time was calculated by subtracting the date a patient was entered on the waiting list from their surgery date. Although a given patient’s waiting time differed under the various schemes, the overall mean waiting time of these 233 patients was maintained at 10.6 months. The absolute area (HRQOL cost) was not particularly useful in its own right, but was used as a benchmark to compare the various schemes. If the study duration had been longer, the average area would likely increase since the analysis would include more of the patients with longer pre-surgical waits.

It can be seen that the average PCS cost actually observed across the sample of 233 patients was 97.5 units. This can be thought of as PCS points*month, since it represents the difference between the pre-surgical and post-surgical PCS scores multiplied by the number of months of waiting for surgery. Similarly, the average observed MCS cost was 24.5 points*months. The PCS cost (physical cost of waiting) was substantially higher than the MCS cost (mental cost of waiting), indicating that surgery had a larger impact on the PCS than on the MCS. This is not surprising since the surgery addresses physical rather then mental impairment.
Table 6: Cost of waiting for surgery in terms of PCS and MCS point months (HRQOL cost)

<table>
<thead>
<tr>
<th>Re-ordering Criteria</th>
<th>PCS Area/cost (% change from as done)</th>
<th>MCS Area/cost (% change from as done)</th>
</tr>
</thead>
<tbody>
<tr>
<td>As actually done by surgeons</td>
<td>97.5 ( 0.0%)</td>
<td>24.5 ( 0.0%)</td>
</tr>
<tr>
<td>In order entered on waiting list (first come first serve)</td>
<td>104.2 ( 6.9%)</td>
<td>28.0 ( 14.3%)</td>
</tr>
<tr>
<td>Maximum expected change in PCS (within surgeon)</td>
<td>82.9 (-15.0%)</td>
<td>28.9 ( 18.0%)</td>
</tr>
<tr>
<td>Maximum expected change in PCS (across surgeons)</td>
<td>82.0 (-15.9%)</td>
<td>29.7 ( 21.2%)</td>
</tr>
<tr>
<td>Maximum expected change in MCS (within surgeon)</td>
<td>116.8 ( 19.8%)</td>
<td>-8.6 (-135.0%)</td>
</tr>
<tr>
<td>Maximum expected change in MCS (across surgeons)</td>
<td>117.7 ( 20.1%)</td>
<td>-9.3 (-138.0%)</td>
</tr>
<tr>
<td>Minimal possible cost (within surgeon)</td>
<td>52.8 (-45.8%)</td>
<td>-30.2 (-223.3%)</td>
</tr>
<tr>
<td>Maximal possible cost (within surgeon)</td>
<td>153.9 ( 57.8%)</td>
<td>82.5 ( 236.7%)</td>
</tr>
<tr>
<td>Minimal possible cost (across surgeon)</td>
<td>50.7 (-48.0%)</td>
<td>-32.2 (-231.4%)</td>
</tr>
<tr>
<td>1st percentile of random orderings</td>
<td>92.5 (-5.1%)</td>
<td>15.3 (-37.6%)</td>
</tr>
<tr>
<td>5th percentile of random orderings</td>
<td>94.8 (-2.8%)</td>
<td>19.2 (-21.6%)</td>
</tr>
<tr>
<td>Median of random orderings</td>
<td>101.4 ( 4.0%)</td>
<td>26.3 ( 7.3%)</td>
</tr>
<tr>
<td>95th percentile of random orderings</td>
<td>107.2 ( 9.9%)</td>
<td>33.8 ( 38.0%)</td>
</tr>
<tr>
<td>99th percentile of random orderings</td>
<td>109.8 ( 12.6%)</td>
<td>36.8 ( 50.2%)</td>
</tr>
<tr>
<td>Mean of random orderings</td>
<td>101.2 ( 3.8%)</td>
<td>26.3 ( 7.3%)</td>
</tr>
<tr>
<td>Standard Deviation of random orderings</td>
<td>3.8</td>
<td>4.5</td>
</tr>
</tbody>
</table>

When patients were re-ordered within surgeon categories, the possible range of the average waiting costs ranged from 52.8 to 153.9 for the PCS, and from -30.2 to 82.5 for the MCS. The midpoints of these ranges are 103.4 and 26.2 respectively. The minimal possible cost was estimated by prioritizing patients according to the change score that was actually observed. The maximal possible cost was estimated by reversing this prioritization. While these ranges suggest that there may be room for improvement, it should be noted that the optimal values defined by these ranges could never be achieved in practice, since surgeons could not know the precise effect of surgery on patient HRQOL before the patients were treated.

Optimal re-ordering across surgeons is also presented. The optimal average areas of 50.7 and –32.2 for the PCS and MCS respectively provide only a minimal improvement from the within surgeon prioritization.

If these patients had been treated on a first come first serve basis according to the date they were entered onto the waiting list, the average PCS and MCS costs would have been 104.2 and 28.0 respectively. Since the observed ‘as treated’ HRQOL costs (PCS 97.5, MCS 24.5) were smaller than the ‘first come first served’ costs (PCS 104.2, MCS 28.0),
one may argue that the surgeons are already prioritizing patients according to some criteria other than simply length of time on the waiting list. However, the statistical as well as clinical significance of this difference is difficult to assess.

Undergoing a hip or knee replacement has a large impact on the physical aspects of patient health, and therefore prioritizing patients according to their expected gain in PCS seems reasonable. Earlier in this report, linear regression models for the change in score from the pre-surgical to the post-surgical assessment were presented. It was found that pre-surgical PCS, BMI, ASA risk level, presence of RA in the operative joint and presence of other significant comorbidities were significant predictors of the change in PCS. The actual regression equation was as follows:

\[ 34.9 - 0.433 \times \text{pre}_\text{pcs} - 0.202 \times \text{bmi} - 2.70 \times \text{asarisk} - 6.50 \times \text{raopjoin} - 3.80 \times \text{othcomor} \]

where ‘pre_pcs’ is the pre-surgical PCS, ‘bmi’ is the body mass index, ‘asarisk’ is the ASA risk of 1 to 5 and ‘raopjoin’ and ‘othcomor’ are dummy variables that take on the value of one when true and zero otherwise.

Using this model, the greatest expected improvement in PCS would occur for patients with low pre-surgical scores in the absence of other comorbidities. On the other hand, patients with relatively high pre-surgical PCS and the presence of comorbidities may achieve much less benefit in terms of PCS. To illustrate this point, a patient with a pre-surgical PCS of 25.3, a BMI of 29.55, an ASA risk level of one, no RA in the operative joint and no other significant comorbidities would have an expected post surgery increase in their PCS of 15.3. (95% prediction interval of –2.03 to 32.53). Another patient with the same pre-surgical PCS and BMI, but with an ASA risk of three, presence of RA in the operative joint and other significant comorbidities would have an expected post surgery change of -0.44 (95% prediction interval of -18.47 to 17.59). Unfortunately, the model predicting change in PCS has a modest coefficient of determination (R\(^2\)) of 0.18, indicating that the predictors can only explain 18% of the variability in the change score. It should be noted that the model identified to predict change in MCS was much more predictive with an R\(^2\) of 0.45.
Using the above model, we also substituted the first PCS (the first assessment obtained from each patient) for the pre-surgical PCS (the last one prior to surgery) and again reordered the patients within surgeon. This re-ordering resulted in a reduction of 14.6 in the PCS cost, from 97.5 to 82.9. However, it also resulted in a 4.5 points*month increase in the MCS cost from the value achieved by the surgeons, suggesting that although there may be additional gain in PCS, it may be at the cost of mental health.

When we focused on prioritizing patients according to the expected change in the MCS rather than the PCS, we achieved a reduction of 33.1 points*months in the MCS cost (from 24.5 to -8.6). The negative area in terms of MCS cost is achievable since the average positive effect of surgery on MCS is quite small and many patients were observed to have a post-surgery decrease in MCS. Unfortunately, when the list is reordered by prioritizing patients according to their expected change in MCS, the resulting PCS change is 19.3 points*months higher than the value actually achieved by the surgeons. This indicates that there is a higher PCS cost when looking at gains in MCS, as compared to the MCS cost when looking at gains in PCS. This suggests that physical parameters are more appropriate than mental parameters for use in evaluation.

The prioritization of patients by expected change in PCS and MCS was repeated without restricting patents to the surgery slots available for their surgeon. This across-surgeon re-ordering reduced HRQOL costs by only a negligible amount beyond what had been achieved with within-surgeon re-ordering. Thus, this sample provides no evidence that a shared waiting list would decrease the cost of waiting conditioned on fixed surgery times.

In order to further describe the distribution of the average waiting costs for this sample, we ran 1000 simulations which randomly prioritized patients. The results of the simulations provide an estimation of the distribution of the PCS and MCS costs on the basis of the observed patients and surgery times, under the null hypothesis that no criteria related to outcome was used to prioritize the patients. The random PCS cost and MCS cost distributions were observed to be approximately normal. The mean and standard deviations are 101.2 (3.8) and 26.3 (4.5) for the PCS and MCS respectively. Medians, 1st, 5th, 95th and 99th percentiles are also presented in Table 6. It may be seen that for both the PCS and MCS, the HRQOL cost as observed for the surgeons (97.5 and 24.5) is
less than the mean cost observed under random re-ordering, but greater than the 5th percentile. The cost in terms of PCS when re-ordered by expected change in PCS, and the cost in terms of MCS when re-ordered by expected change in MCS are both well below the 1st percentile under random reordering.

Since our predictive models and HRQOL cost evaluation used the same limited sample, it is not surprising that we are able to reduce the PCS cost if we prioritize patients by their expected change in PCS, and reduce the MCS cost by prioritizing by expected change in MCS. It is interesting that prioritizing by the physical domain results in an increase in the cost of the mental domain and prioritizing by the mental domain results in an increased cost in the physical domain. This may indicate that the surgeons participating in this study consider factors related to both the physical and mental domain when prioritizing patients.

**Discussion and Conclusions**

This longitudinal study has identified that there is a significant waiting time for surgery, during which the patients continue to experience a decline in both functional ability and HRQOL. It has also shown that the majority of patients experience a significant improvement in function and HRQOL following surgery. Thus, it is appropriate to identify a mechanism for equitable allocation of surgery.

WOMAC and SF-36 domain scores continue to deteriorate while on the waiting list, even though patients are already very impaired compared to age matched controls. Patients with hip osteoarthritis experience a greater decline in their scores while on the waiting list than do those with knee osteoarthritis. This is in keeping with clinical experience. It is more difficult to compensate for a stiff or arthritic hip than for an arthritic knee, hence the hip patients experience a higher degree of impairment in daily activities such as sitting, walking and standing.

There was overall significant improvement following surgery for the entire sample as well as within the THR and the TKR sample. However, not all patients improved. In general, those with higher levels of initial disability experienced a greater improvement
in their scores following surgery, but did not reach as high a level as those who were less disabled initially. Factors associated with less improvement included higher pre-operative scores, a number of comorbid conditions, higher ASA risk and elevated body mass index. These are important findings, and suggest that both physician and patient expectations need to be tempered under these circumstances. These results also suggest that there may be some patients who receive surgery too early to receive any significant benefit as compared to risk.

Greater physical impairment was associated with shorter waiting times, suggesting that the surgeons in this study were aware of the relative disability of patients on their waiting lists. Those awaiting a TKR replacement waited somewhat longer than those awaiting a THR. Those with no caregiver in the home also received surgery more quickly, possibly in order to preserve their independence. The difference in initial quality of life scores across surgeons likely relates to the referral and practice patterns of the involved surgeons.

While there may be a perception that waiting lists are managed on a ‘first come first served’ basis, this did not appear to be the case in Kingston. Each of the surgeons had a method for stratifying their lists. All used a subjective estimate of the severity of disease and the patient’s level of disability as the criteria for priority of treatment. This priority assessment was usually done at the time of initial consultation when the patient was placed on the surgical waiting list, and was based on such factors as severity of night pain, use of narcotic medications, whether the patient had support at home, and whether the patient was house or wheelchair confined. This priority assessment was subjective, ad hoc and potentially biased, but the results of the study show that it was effective since more severely disabled patients had shorter waiting times than their less disabled counterparts.

Attempts to re-order the list on the basis of maximum expected improvement in the SF-36 PCS and MCS indicated that while some small gains could be made under optimal circumstances, the surgeons performed better than a ‘first come first served’ and better than a random allocation. This suggests that they are already effectively stratifying their waiting lists and prioritizing on the basis of disability and functional status. Re-ordering
across surgeons in order to test the potential effect of a shared waiting list produced no
evidence that sharing the waiting list would decrease the impact of the burden of waiting
on HRQOL.

Our study suggests that there may be two separate populations requiring total joint
replacement surgery, those whose expectations should be almost entirely focused on the
relief of pain, and those who should reasonably seek a significant improvement in both
pain and daily function. The former group will see less improvement in daily function
due to significant comorbidity. This conclusion is based on the regression models of
changes in functional status and HRQOL between the pre-surgical and six-month follow-
up assessment, where a number of comorbid conditions were consistently associated with
poorer outcomes.

The identification of the SF-36-based Lifestyle Domain, on the basis of the three domain
scores that showed the most significant decline while patients remained on the waiting
list, suggests that a questionnaire might be developed to assist in the prioritization of
surgery. Lifestyle factors could be utilized in addition to pain, stiffness and functional
ability to determine the optimal time to intervene, in those patients who should be
expected to achieve a significant improvement in functional ability. Thus, it might be
appropriate to determine patient expectations, discuss them in light of the presence or
absence of significant comorbidity, and use two different sets of criteria for ordering the
two groups of patients.

The most significant limitation of this study is the limited proportion of data available at
the time that the database was locked for analysis. Longer follow-up on the enrolled
patients would substantially increase the proportion of available data and allow for much
more reliable and generalizable results. It should also be noted that this sample consists
only of patients from five surgeons. Given the high level of heterogeneity in waiting
times and waiting list composition found across surgeons, this limited sample is not
guaranteed to be reflective of the general Ontario patient population.

This report therefore supports the findings of a number of other researchers, who
identified significant gains in quality of life and functional status following total joint
replacement\textsuperscript{3,8,12-15,19-22}. It also finds that patients continued to decline while on the
waiting list, but that surgery appears to be allocated in an equitable manner, based at least in part on the surgeons’ subjective estimate of disease severity and patient disability.
References


26. McHorney CA, Ware JE & Raczek AB. The MOS 36-Item Short Form Health Survey (SF-36): II Psychometric and clinical tests of validity in measuring physical and mental constructs. Medical Care, 1993;31:247-263.


