Chapter 3. Results

Search Results in the General Population

The details of the paper identification are outlined in Figure 3. The electronic search identified 6,790 possible references. After review of the titles and abstracts, 260 of these references were felt to possibly meet inclusion criteria. The search of the references used in the previous systematic reviews resulted in the identification of 428 possible references. After review of the title and abstracts, 263 references were felt to possibly meet inclusion criteria. Because of overlap, these two sources provided 477 references to be pulled for review of the entire paper.

In addition to the papers identified through the electronic search and the review of previous reviews, two additional references were suggested by the Technical Expert Panel. Further, in the process of reviewing the references, the abstractors identified 47 additional references from the reference lists of the papers. (These were primarily related to the study under review). In total 526 references were identified for full review.

From the references identified for full review, six could not be obtained. These were references that were not available from any library through interlibrary loan. One was a thesis, two were conference proceedings, and three were from journals that could not be obtained. Of the 520 remaining studies that underwent full review, 47 studies were identified that met inclusion criteria. Eighteen studies had important information found in multiple references and one reference contained two studies. In sum, 87 references were identified for inclusion in the study.

Figure 3 also shows the reasons for exclusion of the excluded references. Exclusion criteria were considered in the order presented so that in general a reference that was excluded for a reason lower in the list was felt to likely meet the criteria higher in the list. For example, references excluded because the last measure of physical activity was less than three months after the end of the intervention were felt to have met the criteria above that measure in the inclusion/exclusion list (e.g., ≥75 subjects).

Of the 433 excluded references, 40 did not contain behavioral or policy intervention to increase physical activity. One common study of this type was trials in which the outcome of interest was actually the effect of exercise and the control group was told not to change their physical activity. Any study in which the control group was told not to exercise was thus excluded. Forty-four of the references were excluded because they did not contain a concurrent control group. Insufficient study size (<75 subjects total enrolled) was the reason for exclusion of 75 references. The largest reason for exclusion was a lack of an outcome measure three months or more after the end of the intervention. Two hundred nineteen references, which were about half of those identified for full review, were excluded for this reason. This may be a small overestimate of the percent of the references that otherwise met criteria since attempts were not made to certify with a complete review that the other criteria were met when a clear exclusion was identified. Thus, for example, if the nature of the outcome was unclear (whether it was a physical activity outcome or not) but it was clear that there was inadequate followup time, the reference was excluded and no further attempt was made to adjudicate whether the outcome was a physical activity outcome. Because one exclusion is occasionally more obvious than another, it is possible that a portion of the studies that appeared to meet all criteria applied before the length of followup criterion may have met other exclusion criteria if they underwent complete review.

Study Characteristics: General Population

Populations studied. The 47 studies identified addressed a variety of populations. Adults were studied exclusively in 41 of the studies, four exclusively studied children, and two included both. Of the studies of adults, eight included only women, whereas two included only men. In all but two of the studies where race was reported, white subjects were in the majority. Of the remaining two, one studied an exclusively black population and the other a population that was 50 percent black and 50 percent Hispanic (with the race of the Hispanic subjects not stated). The setting of recruitment also varied across studies with 16 from a healthcare setting, 12 from community, six in school, two from a government agency, eight from a worksite, two from an exercise center, and one from both the community and worksite.

Study characteristics. By the inclusion criteria, all of the studies had a concurrent comparison group. The intervention and comparison groups could be either randomly assigned or use some other method of assignment. Further, the assignment could be done either on an individual level or a group (e.g., clinic or school) level. Within the 47 studies, five were assigned non-randomly on the group level (though two of these were analyzed as if randomized at the individual level), five non-randomly at the individual level, 14 randomly at the group level, and 23 randomly at the individual level.

Intervention characteristics. Within the 47 studies there were 72 interventions examined (exclusive of the comparison or control intervention(s)). Thirty studies examined one intervention, 11 examined two, 4 examined three, and two examined four. A complete description of all of the interventions is given in Appendix D. Twenty-two of the studies delivered a physical activity intervention to the control group as well as the intervention group. These interventions are also described in Appendix D. Control interventions not designed to increase physical activity are excluded from Appendix D.

There was a great deal of diversity within the interventions and across studies (Table 1). Across the studies, the intervention occurred in nine different settings and some interventions occurred in more than one setting. The most common intervention setting was a health care facility, which was used in nearly one-third of the studies. The next most common sites were worksites (28 percent) and community (26 percent), with home and school each accounting for about 15 percent of the studies.

Many of the interventions were aimed at other behaviors in addition to physical activity. Slightly over half of the studies (25, or 53 percent) included an intervention aimed at diet and/or smoking in addition to the physical activity intervention.

Where the type or mode of physical activity that was targeted by the intervention was stated, the studies were rather uniform. All that specified a type of physical activity specified a type of aerobic activity, but 58 percent of the interventions did not specify the activity mode and 49 percent of the studies did not specify activity mode for any intervention. Where the physical activity intensity was noted, it also was rather uniform, with moderate intensity most common. However, over two-thirds of the interventions did not specify intensity and 60 percent of studies did not specify intensity.

The interventions and studies also differed as to whether there was any in-person contact. Three-fourths of the interventions and studies did include some sort of in-person contact, but that leaves a sizeable minority in which the only contact with the subjects was by mail and, occasionally, telephone.

Half of the interventions (50 percent) and 43 percent of studies were tailored to the individual subject in some way. Those means of tailoring are shown in Table 1. Nearly a quarter of the interventions were tailored to a "Stage of Change." Other means of tailoring that were used in more than five percent of the interventions included tailoring to an individual's risk factor status, fitness level or exercise preference, or individualized counseling.

A wide range of behavioral intervention components (which are often also theoretical constructs) were used. Some were commonly used—over two-thirds of the interventions (67 percent) employed 'education on the benefits of exercise' and a similar amount (46 percent) provided written and/or verbal feedback and/or encouragement. Yet there was also a great deal of diversity. There were 11 behavioral intervention components that were used in 13 to 43 percent of the interventions, and there were 14 behavioral intervention components that were employed in four or fewer interventions. Nearly 20 percent of the interventions did not specify any behavioral intervention components.

Like the other aspects of the interventions, there was diversity in whether the study authors elucidated a theory underlying the intervention tested. For half of the interventions and studies (51 and 49 percent, respectively) no theory was discussed as the basis of the intervention. For 11 interventions two theories were said to underlie the intervention, and for 24 of them one theory was said to underlie the intervention. There were a variety of theories used. The most common theory was the Transtheoretical or Stages of Change model that was said to underlie about a third of the interventions (29 percent). No other theory accounted for more than ten percent of the interventions where theory was reported.

Even in the most fundamental aspect of the overall intervention intensity, the studies differed widely. The intensity of the most intensive intervention in each study is shown in Table 2. The number of contacts with the study subjects over the course of the intervention varied quite widely from just one to over 200. Further, the length of the intervention varied from a single encounter to seven years. One-quarter of the interventions went on for over six months.

We combined the type of contact, frequency of contact, and length of the intervention to classify the studies into an ordinal intensity scale. Studies in which there was no in-person contact were scored as "1". If there was in-person contact, but less than a total of eight times, and the study was less than two years long, it was scored as a "2". Studies that had ten or more in-person contacts and/or were large community trials that had a number of environmental and media changes and lasted five to seven years (such as Minnesota Heart Health Program,[49] Pawtucket,[50] and UK Heart Disease Prevention Project[51]) were scored as "3". The remaining studies, one of which met four times weekly for four months and three of which had in-person contact three to five times weekly from one to three years, were scored as a "4". Using this scoring system, four studies were scored in the highest category, ten in the lowest, and the remainder closely split in the middle categories. It should be noted that the decision as to where to place large community trials in such a scale is somewhat arbitrary.

Outcomes examined. A range of different physical activity outcomes was found in the included studies, and many studies included more than one. Twenty-four studies had one physical activity outcome, eight studies had two physical activity outcomes, 11 had three, one had five, two had six, and one had nine, for a total of 99 individual outcomes. No specific outcome was used as the primary outcome across studies. Further, what may have been considered the primary outcome domain in one study (such as a measure of leisure time activity) may have been a secondary domain in another study (where the primary outcome could have been overall activity).

The diversity of outcomes presents a significant challenge in comparing the results of different trials. Two possible conditions may exist: 1) the different outcome measures may be measures of the same underlying physical activity domain assessed in different ways (e.g., leisure activity measured by self-report and accelerometry); or 2) the measures, although both measures of physical activity, may be measuring different underlying domains (e.g., self-report of vigorous activity and self-report of total activity).

There are a number of examples in this literature of different outcome measures that are assessing the same underlying domain. For example, a number of measures attempt to assess the total activity an individual performs in a day. This underlying domain may be assessed with a log of all activities, an objective measure (e.g. accelerometer), or a survey of activities for a recent period of time. Each of these methods of measurement may be more or less valid and reliable but they all reflect measurement of total activity. An intervention that actually increases total activity would be expected to have a similar effect on all three of the measures. Therefore, it may be reasonable to compare these outcomes that have been converted to a standardized metric such as an effect size.

There are also many examples of different outcome measures that are assessing different underlying levels of intensity within one domain (vigorous versus total leisure time activity) or differing domains (leisure time activity versus household chores). One could imagine that an intervention could have one effect on total leisure time physical activity and a different effect on vigorous leisure time physical activity. For example, the CATCH trial[52,53] sought to increase the physical activity of school children. They found that children who underwent the CATCH intervention had a statistically significant *increase* in vigorous leisure time physical activity and a statistically significant *decrease* in total leisure time physical activity. If we were comparing two distinct studies in children, one of which reported a decrease in total leisure time physical activity and one that reported an increase in vigorous leisure time physical activity and we compared the reported effects of the two studies, we would conclude that one was harmful (as it had a statistically significant negative result) and one was beneficial (as it had a statistically significant positive result). In truth, both occurred in the same study, and interpretation of these findings is complicated. This example is intended to point out the caution required in comparing results that assess different underlying domains or differing intensities within the same domain.

It would be optimal if there were a common measured domain across the studies included in the review to facilitate comparison of the effects of the different interventions. We grouped the outcomes in two ways to attempt to assess the effects of interventions. Because guidelines have targets for both moderate and vigorous activity[2] we first classified outcomes as measures of vigorous, moderate, or total activities. Measures of exercise sessions, fitness activities, fitness and vigorous activities were grouped as “vigorous activities.” Measures of walking activities, other moderate activities and leisure activities were grouped as “moderate activities.” Finally measures of daily activities and total activities were grouped as “total activities.” Measures that did not fit these categories were classified as “other.” Of the 99 outcome measures in the studies, 23 (23 percent) were classified as "total activities," 50 (51 percent) were classified as "vigorous activities," 25 (25 percent) were classified as moderate activities, and 1 (1 percent) was classified as "other." Of the 47 studies 20 (43 percent) contained a measure of "total activity," 28 (60 percent) contained a measure of "vigorous activity," and 18 (38 percent) contained a measure of "moderate activity." Because, each of these are collections of measures, when presented in the results they will be referred to as "group." For example, the "moderate activities" will be referred to as "moderate activities group" so it is clear that it is not a measure of total moderate activities.

As discussed above, one potential problem with the above categorization is that to the extent that some of the measures assess only a portion of the domain, it is possible that changes could be seen in the measures that do not in actuality reflect changes within the complete domain. For example, it is possible that individuals in a walking program could substitute the activity in the program for physical activities they would otherwise do. One might then see an increase in walking but in reality there is no change in overall moderate activity. There is little literature on this point, although observations in this literature review such as the differences seen in CATCH in which the vigorous activity promoted in CATCH substituted for other activity to result in a net decrease in total activity (see above) and the study of Goran et. al. in which elderly subjects in exercise training reduced their activity in the rest of the day for no net change in activity suggest this is certainly possible.[54] We therefore attempted to create distinct domains of physical activity outcomes. Some of these are subsets of other domains (e.g., walking activity is a portion of total moderate activity). For example if two studies each attempted to increase walking but one measured walking as an outcome and one measured total moderate activity as an outcome, differences could result either from differences in the interventions or because the interventions affect walking but not total moderate activity. This issue would not exist if they both measured walking or both measured total moderate activity and further underscores caution in interpretation of results.

The domains examined are shown in Table 3. We do not claim these are unique domains. Determining whether they are unique would require empirical testing. However, they provide an attempt to classify the outcomes of the studies in these reviews. Unfortunately, no one outcome domain was measured by more that 40 percent of the studies, so it was not possible to select one domain to examine across all of the studies. This diversity of domains should be kept in mind, however, when interpreting the overall results.

An attempt was made to use all of the existing information in the studies to create a measure of overall energy expenditure but this failed (see Methods). We therefore elected to include all of the physical activity outcomes reported in the results that follow. The complete list of outcome measures can be found within the main evidence tables (Appendix E). As the results contain a variety of outcomes, caution must be used in comparing the effects across studies as differences may result from differences in the outcomes assessed rather than differences in the intervention effects.

Followup time. There was a significant range in time between the end of the intervention and the final outcome measurement ranging from three months to 11 years (Figure 4). Most studies did not report multiple followup times, so it was not possible to pick a followup interval that was comparable across studies. The distribution of followup times is little different when one examines the first followup greater than or equal to three months following the end of the intervention. The followup point used is stated in each of the analyses that follow.

Assessment of Outcomes

Two criteria for a positive effect of the intervention on outcome were used throughout the results: effect size and statistically significant positive effect. Each has its strengths and weaknesses. Used together these two criteria give a fuller picture of the results of the interventions.

Effect size. Effect sizes (e.g. standardized mean differences) are frequently used in pooling studies so that the results of studies that use different measures for the same outcome can be examined together. They have great strength because whatever the outcome is, if sufficient information is provided, an effect size can be calculated for it. This then allows that outcome to be compared to the same outcome from a different study measured in a different way (and also converted to an effect size). Thus, for this review it is possible to use effect size to get a sense of the effects of diverse outcomes without needing to understand exactly the metric employed in the study.

The ability to convert the effects of the various studies to effect sizes, however, comes at a price. Because results of studies are on the same metric, it is tempting to make comparisons between studies that should not be made. As discussed above, the outcomes of these studies may be of different domains of physical activity or differing intensities within a single domain. An effect size in a measure in one domain may or may not be analogous to an effect size in another domain. Although we think it is useful to examine the range of effect sizes in the included studies, any assessment of the actual effectiveness of an individual study requires a closer examination of the specific outcome measured. This information is provided in the evidence tables.

One additional weakness of effect size as a measure of outcome is that it cannot be calculated for all of the outcomes and in some circumstances when it can be measured, the results are known to be biased (usually downward). We were unable to calculate an effect size for 13 (28 percent) of the included studies. In the presentation of the effect size results, effects that could not be calculated are noted. It should be noted that the inability to calculate some effect sizes may artificially inflate the overall results reflected by effect sizes because the manner of reporting results in statistically insignificant studies tends to be less detailed, leading to inadequate data for effect size calculation. For example, for statistically significant studies, a p-value is generally either reported or is stated to be ‘less than 0.05’, which is part of the information needed to calculate an effect size. However, in statistically insignificant studies, the p-value may just be reported as ‘NS’ for not-significant. Reasons that effect size calculation was not possible for individual studies included no available variance estimates, no significance levels, insufficient information about number analyzed, or missing correlation information in multinomial models. Specifics on data needed for calculations of the effect sizes are provided in the Methods section.

There are no criteria that could classify effect sizes as small, moderate, or large that would make sense across all studies. Some relatively small effects may have a large impact if applied across a large population. However, for the purpose of ready comparison here we provide reference lines in the graph for effect sizes of .2, .5, and .8. If one considers the mean of the treated group as a percentile ranking of the control group, these guidelines correspond to a percentile ranking of 58, 69, and 79 respectively.[55] In the text that follows, these will be referred to as small, medium, and large effects with the caveat that small effects may in reality have large impacts in a population and the reader should examine the details of the measures and effects in the evidence tables.

Statistical significance. We also examine whether interventions have a statistically significant effect. The advantage of this metric is that, unlike the effect size metric, it supports whether changes seen are real or reflect random chance. However, examining whether an intervention has a statistically significant positive effect may underestimate the effect of the interventions because the study may not have been sufficiently powered to detect a meaningful effect. This issue can be overcome by pooling similar studies to provide greater power. After examining the diversity of populations, interventions, and outcomes it was decided that formal pooling of the effects from the studies to increase statistical power was not appropriate.

Level of assessment. Within the 47 studies there were 72 interventions examined and 99 outcomes. Six outcomes were reported by subgroup only. A total of 166 outcomes for interventions were examined. As discussed above, it was not possible to establish one "best" outcome to examine from each study. Further, there is benefit to examining multiple interventions within studies independently because a specific intervention within a study may have been effective, and this level of evaluation will allow for examination of intervention components that are effective versus ineffective. Finally, on the study level we are able to see the overall effect of the study as a whole.

Outcome level examination. The effect of the intervention on each unique outcome of the included studies is reported. Again many studies examined multiple unique outcomes. Wherever possible the results for the whole intervention and control groups were used. In a few studies results were reported by subgroup only. In these cases the subgroup analyses were used. All of the effect sizes that could be calculated are reported in the evidence tables and are used in the graphs of effect size on the outcome level.

An effect was considered a statistically significant positive outcome if a statistical test was performed that demonstrated that the intervention group had greater physical activity (however measured) than the control with a significance of $p<.05$. Where sufficient data were presented to perform a statistical test but the statistical test was not reported in the paper, that testing was done as part of the review and if $p<.05$ the outcome was reported as statistically significant. Where 95 percent confidence intervals were reported, an outcome was reported as statistically significant if the intervals were non-overlapping.

Intervention level examination. For each intervention there could be several outcomes reported. To report an effect size for an intervention it would be therefore necessary to calculate one effect size out of multiple effect sizes and do it consistently across studies. For studies that had only one intervention tested, the intervention level would be the same as the study level (see below). Although a mean of multiple effects may appear appealing as a means of calculating the effect of an intervention that had multiple outcomes, the fact that the number of effects presented is arbitrary may result in penalizing studies that more thoroughly report the results. This would occur if authors prejudiciously fail to report results of lesser effect over those of greater effect. Therefore, we assumed that authors may report the outcomes that show the greatest effect and used the largest effect to give the best comparison across interventions and studies. This may bias the effect seen for the individual interventions and studies upward for the true effect but allows a greater degree of comparability across interventions.

An effect was considered a statistically significant positive intervention if any one of the outcomes examined within the intervention was statistically significant. The intention here is to convey a level of positivity of the results, not to perform a statistical test. Significance was not corrected for multiple tests so classifying an intervention as a statistically significant positive effect does not necessarily mean that the intervention was indeed significant at the .05 level.

Study level examination. When there were multiple interventions used in a study it was necessary to calculate one effect across the interventions to be able to report an overall effect of the study. The same reasoning was used to combine interventions as was used to combine outcomes within interventions (see above). Therefore, in combining the effects of studies with multiple interventions, we chose the largest effect to report as the study effect. Again, this may bias the effects on the study level upward but eliminates the role of number of outcomes reported on effect size.

A study was considered a statistically significant positive intervention if any one of the outcomes examined within the study was statistically significant. The intention here is to convey a level of positivity of the results, not to perform a statistical test. Significance was not corrected for multiple tests, so classifying an intervention as a statistically significant positive effect does not necessarily mean that the intervention was indeed significant at the .05 level.

Overall Effect

The overall effect sizes at the outcome, intervention, and study level are shown in Figure 5. There were 102 outcomes and 50 interventions within the 34 studies for which effect sizes could be calculated. Of the 102 outcomes, 7.8 percent (eight) had an effect size greater than .8, and 2.9 percent (three) had an effect size between .5 and .8. An additional 32.4 percent of the 102 outcomes (33 outcomes) had an effect size that exceeded our criteria for a small positive effect of .2. Of the 50 interventions for which we could calculate an effect size, 10 percent (five) had an effect size greater than .8 and 4 percent (two) had an effect size between .5 and .8. An additional 44 percent (22 interventions) had an effect size that exceeded our criteria for a small positive effect of .2. Finally, on the study level, 5.9 percent (two) of studies had an effect size greater than .8, and 5.9 percent (two) had an effect size between .5 and .8. An additional 47.1 percent (16 studies) had an effect size that exceeded our criteria for a small positive effect of .2. Overall, 58.8 percent of studies had an effect size that exceeded our guideline of small (.2).

There were only two studies exclusively of children for which an effect size could be calculated.[53,56] The overall effect size of these studies was similar to those of the other studies (.597 and .145). Arguments could be made either way as to whether it is reasonable to include studies of children with those of adults. We elected not to exclude these studies from the other analysis that follows. This decision has no effect on the conclusions derived from the results.

Approximately one-fourth of the outcomes reached statistical significance (see Table 4). Nearly a third of the interventions overall had at least one outcome that was significant at the .05 level. Nearly half of the studies (44.7 percent) had at least one outcome that was statistically significantly positive. Again, this is not corrected for multiple tests within studies.

Effect by outcome group. Because of the number of different outcomes examined in these studies it is possible to examine the range of effect sizes and percent statistically significant for the different outcome groups. One issue with examining whether the effects observed varied by outcome group is that some outcomes occur multiple times within an individual study because results may be reported for multiple interventions and subgroups. If this is not accounted for, the effect would be to overweight these outcomes in the examination of the distribution of effect sizes and statistical significance. Therefore, in examining the effect by outcome group, it is necessary to assign one effect to each of the 99 individual outcomes examined. To assign one effect to each outcome we used the greatest effect size observed for that outcome (if an effect size could be calculated) and if any of the observations for that outcome were statistically significant, the outcome was considered statistically significant. For example, if the outcome "walking sessions per week" was reported for two interventions in a study with an effect size of 0.1 and 0.2, we assigned the effect size of 0.2 to the outcome. Similarly, if the effect of "walking sessions per week" was statistically significant for the effect size of 0.2, we classified the outcome of "walking sessions per week" as statistically significant. The effect size for all outcomes by the outcome group is shown in Figure 6. The percent of each outcome group that was statistically significant is shown in Table 5.

Within the outcome groups, only the moderate activity group and the vigorous activity group had any outcomes that exceeded our guide of a large outcome of .8 (two moderate and one vigorous). Approximately 60 percent of moderate activity outcomes had an effect size greater than our guide of .2, whereas approximately 40 percent of the vigorous activity outcomes and total activity outcomes exceeded that threshold. A greater percentage of moderate activity outcomes was statistically significant compared to total activity outcomes (48 percent versus 13 percent; p=.008). The percentage of vigorous activity outcomes that was statistically significant fell between the other two outcome groupings, (28 percent) but was not statistically significantly different from either the moderate or total outcome groups.

Description of the Specific Effects

A more full understanding of the effects seen in this literature may be obtained by closer examination of the individual studies. For that reason, we describe in greater detail the interventions and results of those trials that meet the traditional measure of success and are statistically significant. For ease of understanding, they are discussed by the setting in which the intervention took place.

Health care. Bull et al. examined whether brief advice from a family practitioner combined with a mailed pamphlet would increase sedentary patients' physical activity.[57] Seven hundred sixty-three sedentary subjects were allocated to a control group, advice plus a standard pamphlet mailing or advice plus a tailored pamphlet mailing based upon the day of the week they attended the clinic. They found that six months after the intervention the percent of patients who were "now active," defined as any walking or exercise in the previous two weeks was greater in the combined intervention groups than the control group (38 percent vs. 30 percent; p not reported but stated to be significant). The difference at 12 months of followup was smaller and non-significant (36 percent vs. 31 percent). There were no statistically significant differences between the control and intervention groups in the number of exercise sessions in the previous two weeks at six months or 12 months, and there were no statistically significant differences between the two intervention groups.

In the "Change of Heart Study" Steptoe et al. examined the effect of behavioral counseling on coronary heart disease risk factors including exercise.[58] Eight hundred eighty-three men and women with one or more modifiable risk factors attending a general medical practice were given either routine counseling or behaviorally oriented counseling depending upon the clinic they attended. Behavioral counseling subjects received two or three counseling sessions depending upon their number of risk factors. They found that approximately eight months after the end of the intervention the intervention group had increased the average number of exercise sessions in the previous four weeks from 5.56 to 8.2, whereas the control group had decreased slightly from 4.82 to 4.3. This change in number of exercise sessions in the intervention group compared to the control group was statistically significant.

Halbert et al. examined the effect of physical activity advice given by an exercise specialist during three general practitioner appointments versus no advice.[59] Two hundred ninety-nine subjects over age 60 were randomly selected from two general practices in Adelaide, Australia. Approximately six months after the end of the intervention, intervention subjects were exercising more than control subjects on three of five measures of physical activity: walking sessions per week (median 3 vs. 2; p<.05), vigorous exercise sessions per week (median 2 vs. 0; p<.05), and minutes of vigorous exercise per session (median 20 vs. 0; p<.05). There were no significant differences in the minutes of walking per session or in energy expenditure as measured with an accelerometer although the latter was done only on a subset of 59 individuals so power may have been an issue (no data is presented to allow evaluation of power).

Kerse et al. examined the effect of educating general practitioners in health promotion (including increasing physical activity) for elderly people.[60] Forty-two Australian general practitioners were randomly assigned to either an education group or a control group and 267 of their patients were randomly selected from their practices. Approximately nine months after the physicians completed their educational program, the patients of intervention physicians were performing more physical activity on one of three continuous self-reported measures. The results are reported as net differences in the physical activity changes between treatment and control participants: minutes walking in the previous fortnight (88 minutes more in treatment than control participants; p=.032); as well as two of three categorical self-report measures: walking minutes per day on a five-point likert scale (.34, p = .005), walking minutes over previous fortnight on a three-point scale (.27, p = .025). There were no statistically significant differences in minutes walking per day as a continuous (8.4 minutes p = .059), total activity as a continuous measure (148 minutes, p = .34), or total activity total in the last fortnight on a five point scale (.23, p = .30).

Green et al. examined the effect on 316 primary care patients of three session of telephone based motivational counseling.[61] Using intention to treat analysis, intervention subjects were exercising more than controls approximately three months after the intervention as assessed using the Patient-centered Assessment and Counseling for Exercise (PACE) score, which is a self-report measure of both stage of change and level of exercise (5.37 vs. 4.98; p=.049). However, the change in PACE score from baseline was not statistically different between the two groups (.426 vs. .102; p=.145).

Stevens et al. examined the effect within 363 inactive subjects selected at random from 714 subjects recruited from two London general medical practices, of meeting with an exercise development officer followed by a personalized ten-week exercise program.[62] Eight months after the intervention participants reported more sessions for moderate physical activity (5.09 vs. 3.64

control; p<.05), more sessions of vigorous activity (.86 vs. .78 control; p<.05), and more overall episodes of physical activity (5.95 vs. 4.43 control; p<.05) in the four weeks prior to the end of followup.

Community. Gillett et. al. randomly assigned 182 sedentary obese 60-70 year old women recruited from newspaper ads to fitness education, fitness education with aerobic training, or a control group.[63] They examined fitness three months and six months following the intervention and found that overall the aerobic training group had a better VO2 max than the other two groups (at six months average VO2 max increased in aerobic training group 14.9 percent vs. education 1.8 percent vs. control -1.0 percent; overall group effect p<.001). However, the aerobic training group reported exercising fewer days per week at six months than the education group (2.3 vs. 3.3; p<.01). Results for the control group were not reported.

Pereira et al. reported on the exercise status of subjects ten years after the conclusion of a randomized controlled trial of a walking program to examine the effect of walking on bone mass.[64] Two hundred twenty-nine female postmenopausal subjects were randomly assigned to a control group (instructions to control group participants were not described) or a walking program consisting of 16 organized group walking sessions over eight weeks followed by either group walking sessions or walking on their own for the duration of the clinical trial (1982-1985). Ten years after the conclusion of the clinical trial, walking program subjects reported more weekly kilocalories (kcal) expenditure for total usual walking (median 1,344 vs. 924 control; p=.01) and more weekly kcal expenditure for usual walking for exercise (median 1,008 vs. 302; p=.01). There were no statistically significant differences in weekly kcal for sport and recreation, weekly kcal for past year exercise, Paffenbarger sport and recreation index, or Paffenbarger sport and recreation index with walking excluded.

School. Burke et al. examined the effect of a physical activity and nutrition program during two ten-week terms for 800 11-year olds in Australia.[65] Schools were randomly allocated to a physical activity program consisting of classroom lessons and fitness sessions (six standard intervention schools), the fitness program combined with an education enrichment program for high-risk children (seven enriched program schools) or no program (five control schools). Results were reported by gender and risk group. The results were reported graphically and significance was determined by non-overlapping confidence interval bars. Six months after the intervention six of the intervention groups had statistically significant improvements in fitness as measured by change from baseline in shuttle run time (measured in minutes) as compared to the comparable group at the control schools. Three of the groups that improved more than the comparable control schools were at the standard intervention schools (low-risk girls 9.5 vs. 1; high-risk girls 8 vs. 4; low-risk boys 8 vs. 5) and three in the enriched program schools (low-risk girls 9.25 vs. 1; high-risk girls 9.75 vs. 4; low-risk boys 10 vs. 5). In a second measure of fitness, time in minutes of a 1.6 km run, only high-risk boys at the enrichment schools had what appeared to be a borderline significant improvement (-1.1 min vs. -.4).

Dale et al. examined the effect of a "conceptual physical education" program for ninth grade students at one high school.[66] They were compared to students who moved to the school after the program started. They analyzed male and female students separately and two cohorts separately (the program was done in two subsequent years). Two, three, and four years after the intervention they assessed the percent of individuals who reported doing moderate activity five or more days per week and vigorous activity three or more days per week. From the 24 comparisons (two genders, two levels of exercise, three points in time, two cohorts) they found two statistically significant differences. A larger percentage of men in one intervention cohort were doing moderate exercise compared to the control three years after the intervention (34 percent vs. 13 percent; p=.04 without correction for multiple comparisons) and a larger percentage of men in a different cohort were doing vigorous exercise four years after the intervention (65 percent vs. 29 percent; p=.01 without correction for multiple comparisons). There were no statistically significant differences in the other 22 comparisons.

Howard et. al. examined the effectiveness of a cardiovascular risk reduction program for children in grades four through six.[56] The study was conducted at one private parochial school. One class in each of the fourth through sixth grades was given the intervention and the other class within each grade served as the control group. The intervention included five sessions including "physiology of the heart, smoking, hypertension, diet, and physical activity" developed from materials from the American Heart Association. One year after the end of the intervention, the intervention group was exercising fewer times per week (for at least 30 minutes per time) than the control group (5.89 versus 10.4; p=ns) although the difference was not statistically significant. Further, there were no statistically significant differences in fitness between the intervention and control groups at followup as measured by the Canadian Aerobic Fitness test (4.17 intervention group versus 4.08 in control; p=ns). Yet the intervention group reported that a greater percentage of their exercise was running compared to the control group (68.7 percent versus 38.3 percent; p<.05).

Nader et al. reported on the post-intervention findings of the CATCH trial, which was a three-year cardiovascular health promotion program given to students in third through fifth grades at 56 randomly assigned schools in four states.[52] Outcomes were compared to students from 40 control schools. One year after the end of the intervention, intervention students reported doing more minutes of vigorous physical activity per day than control students (53.2 vs. 42.2 control; p=.001) but control students reported doing more total physical activity minutes per day (164.5 vs. 172.1 control ; p=.04). [Note: the text of the paper states that the direction of the total physical activity effect favored intervention students, but the table presented showed the opposite. The authors confirmed with us that the table is correct, which is the data presented here.] (Personal communication, Henry Feldman.) At three years following the intervention, intervention students still reported more vigorous physical activity minutes per day (30.2 vs. 22.1 control; p=.001), but the differences in total physical activity minutes were no longer significant (121.1 vs. 125.4 control; p=.59).

Worksite. O'Loughlin et al. examined the effect of workplace-based health screening on employees of eight elementary schools compared to eight matched comparison schools.[67] Screening was done for all the subjects at the school during one day in February. Two hundred nine subjects completed baseline information at the intervention schools and 177 at the control schools. Four months after the health screening, intervention subjects reported a greater increase in leisure time exercise behavior score (sessions per week x intensity weight per session) (4.6 vs. –0.4 p=.05).

Gemson et al. examined the effect amongst 161 financial services workers of a worksite based computerized health risk appraisal with counseling compared to just the computerized health risk appraisal.[68] Subjects were randomly allocated to get the intervention. Of the 56 percent of the subjects who followed up at six months, the intervention group reported a greater increase in episodes of physical activity per week (.33 vs. -.13; p<.05).

Lombard et al. examined the effect of telephone prompting to increase physical activity in 135 subjects recruited from faculty and staff at a southeastern university.[69] Four different telephone prompts were examined: high frequency low structure (HF/LS), high frequency high structure (HF/HS), low frequency high structure (LF/HS) and low frequency low structure (LF/LS). At three months after the end of the intervention, a larger percentage of each of the intervention groups except low frequency high structure were walking at least one day per week for 20 minutes (63 percent HF/LS vs. 63 percent HF/HS vs. 26 percent LF/LS vs. 22 percent LF/HS vs. 3.7 percent control; significance not reported but by t-test using data in paper p<.05 for all but LF/HS compared to control). Similar results were seen for percent of subjects meeting Centers for Disease Control/ American College of Sports Medicine (CDC/ACSM) criteria (52 percent HF/LS vs. 41 percent HF/HS vs. 11 percent LF/LS vs. 15 percent LF/HS vs. 3.7 percent control; significance not reported but by t-test using data in paper p<.05 for HF/LS and HF/HS compared to control).

Mutrie et al. examined the effect of distributing a "walk in to work out" pack (consisting of interactive materials based on the transtheoretical model of behavior change, local information about distance and routes, and safety information) to 145 employees randomly selected from 295 employees from three work places who expressed an interest in walking or cycling to work.[70] Six months after the intervention they found that the intervention subjects spent nearly twice as much time walking to work than did the controls (1.93 average relative increase in time compared to controls with 95 percent confidence intervals 1.06 to 3.52). There was no difference in time cycling to work (data not reported).

Linenger et al. examined the effect of multiple environmental interventions undertaken at the San Diego naval air station including new recreational facilities, paths, events, and equipment.[71] The fitness of residents was compared to a comparison naval air station and a random sample from the Navy as a whole. One year after the intervention, those in the intervention community had a greater average improvement in 1.5 mile run (intervention group 12.6 minutes baseline, 12.3 minutes one year; comparison groups 12.3 and 12.1 baseline, 12.2 and 12.2 one year; p<.01 time by group interaction). However, energy expenditures in kcals per week did not differ between the groups.

Other. Perkiö-Mäkelä examined the effect of 2.5 months of aerobic training and lectures on work issues on 62 female dairy farmers aged 25-45 with moderate musculoskeletal symptoms randomly selected from a group of 126.[72] At one-year followup, intervention subjects reported more leisure time physical activity than the control group (≥2 times per week 34 percent intervention vs. 20 percent control; p=.003). At three years the control group was more active, but the difference between groups was not statistically significant.

Marcus et. al. examined whether a computerized report which gave motivation feedback, comparative feedback, and progress feedback increased physical activity more than standard self-help materials in 194 healthy adults recruited through newspaper advertisements.[73] Materials were provided at baseline, one, three, and six months. Six months after the last intervention materials more intervention subjects met CDC/ACSM criteria for physical activity (42 percent vs. 25 percent; p<.05) but there were no statistically significant differences in minutes walked per week (187 vs. 133; p=.1).

Belisle et al. did two studies among registrants to exercise groups at the University of Montreal sports center.[74] Intervention participants received a special health education program designed to increase awareness of obstacles to exercise and develop appropriate techniques for coping with them in addition to the structured exercise program given to both intervention and control participants. Intervention subjects reported more exercise sessions per week in the three-month period following the intervention in both studies (4.2 vs. 3.68; p<.01 in study one and 4.24 vs. 2.68; p<.01 in study two).

Moderators of Effect

Moderators of the effect of a physical activity intervention are those characteristics of the subject or environment that alter the effect of the intervention in that subject or environment (i.e., the effect of the intervention is modified by the moderator). For example, if the same intervention had a different effect in men than women, then gender could be viewed as a moderator of the effect of the intervention. Similarly, if the same intervention had a different effect when undertaken in a workplace rather than a community center, the setting could be viewed as moderating the effect. Most characteristics of individuals that may affect how they respond to a physical activity intervention may be thought of as moderators.

The original hope was that sufficient studies would have similar interventions and comparable outcomes so that the effects of moderators across pooled studies could be studied (e.g., examine the effect of age within a group of pooled studies using meta-regression techniques). Unfortunately, as discussed above, the literature proved to be too heterogeneous and the outcomes could not be pooled. Therefore we restrict our examination of moderators to those with sufficient numbers that they can be examined across studies without pooling and those that were explicitly tested within studies. Within studies a moderator was considered to be explicitly examined if there was an explicit comparison of the effect of the intervention in two or more subgroups or if there was a multinomial model (e.g., ANOVA or multiple regression) that tested an interaction between a moderator and the intervention. We did not attempt to deduce a moderator effect when subgroups were reported but not compared or when multinomial models were presented but interactions were not tested.

Effect of Intervention Setting

An examination of the effect of intervention setting on outcomes is presented in Figures 7, 8, and 9 and Table 6. Examination of the effect size results at the study level is perhaps most informative as the distribution is not affected by the fact that some studies examined multiple outcomes or had multiple interventions. Two of the studies had effects greater than our guideline of .8; one in the community setting and one in the work setting. One school-based study had a moderate effect as did one study at a government agency. The numbers are too small to draw any conclusions about the percent of studies in each setting that had an effect size of at least .2. By setting, the criterion of an effect size of 0.2 or greater was met by: healthcare 54 percent (seven), home 33 percent (two), community 67 percent (six), school 50 percent (one), work 57 percent (four), government 100 percent (one), and other settings 75 percent (three). The range for statistically significant interventions has a similar magnitude ranging from a low of 28.6 percent for community-based interventions to a high of 100 percent for the government institution based intervention. However, the numbers are again small overall so it would be wrong to attempt to draw any firm conclusions. The differences seen could be random statistical variation.

Moderators within Studies

Although a large number of the studies examined measured baseline characteristics of the study subjects, only one used this information to examine whether baseline characteristics moderated the effect of the intervention to increase physical activity. Steptoe et al. investigated the effect of brief behavioral counseling in primary care.[75] They found a significant interaction effect between the intervention and measures of support from family and friends, having a partner who exercised, perceived greater benefits from exercise, and perceived lower barriers to exercise. This suggested that these factors were moderating the effect of the intervention. That is, subjects who possessed these characteristics were more likely to respond to the intervention than other subjects. They did not see an effect for a measure of stages of change. None of the studies reported explicitly testing the effects between subgroups of the population.

Mediators of Effect

Mediators of the effect of physical activity intervention are constructs that are hypothesized by the interventionist to fall in the causal pathway between the intervention components and behavior. For example, one reason individuals may not exercise is because they perceive barriers to exercising. If one intervenes to reduce those perceived barriers, subjects may then exercise more. A change in perceived barriers to exercising would then be considered a mediator of physical activity change. Support for the possibility that a factor is a mediator of an intervention is provided if the intervention is found to have a positive effect on the mediator. Further support is provided if changes in the mediator with the intervention are associated with changes in physical activity.

Like the examination of moderators, it was not possible to pool studies to examine the effect of mediators of physical activity because of the heterogeneity of the studies and the small number of these studies that examined mediators. We therefore examined the effect of mediators as described within studies.

Effect of intervention on hypothesized mediators. Eleven studies hypothesized mediators (Table 7). All 11 of them intervened on at least one of the hypothesized mediators. Nine of the studies measured the effect of the interventions on the hypothesized mediator although two[76,77] did not report any of the results.

The only statistically significant changes in mediators reported were for ‘greater intention to exercise,’ and this was reported in one study. In the other studies that reported results, there was either no effect or a nonsignificant change in mediators resultant to the physical activity interventions.

Effect of hypothesized mediators on physical activity. Only one study examined whether a hypothesized mediator affected the physical activity outcome.[78] They examined whether including the mediator in the overall model of the effect of the intervention on outcome would change the intervention effect. They found that including partner support and self-efficacy in the model attenuated the effect of the intervention seen. They therefore concluded that the effect may be acting through the hypothesized mediators.

Effect of Intervention Type on Outcome

Because of the heterogeneity of populations and outcomes it was not possible to closely examine the possible different effects of different types of interventions. We felt it was possible, however, to attempt to look at a gross level of how intensive the intervention was and whether that predicted the outcome of the intervention. The effect size seen in the studies by the intensity of the intervention is shown in Figure 10. The percent of the studies that were statistically significantly positive by intensity of the intervention is shown in Table 8. The intensity measure is described in greater detail above. Within this data there is not a clear effect of intensity of the intervention on the magnitude of the effect size. Seventy-one percent of the lowest intensity interventions had an effect size greater than .2, compared to 57 percent of the level two intensity studies, 45 percent of the level three intensity studies, and 50 percent of the highest intensity studies. Yet, the two studies with effect sizes greater than .8 were studies of intensity level three and four and none of the studies in intensity level one had an effect size larger than our guideline for a small effect size of .5. All four of the most intensive studies were statistically significant, but no clear trend of statistical significance was seen in the other intensity levels ranging from 33 percent statistically significant for level three studies to 44 percent significant for level two studies. Again, sample sizes are small so it is not possible to conclude whether there is an effect of intervention type on intervention success.

One possible limitation to physical activity is lack of access of places to exercise. Some interventions specifically address this issue by providing greater access for subjects to places to be physically active such as exercise facilities or parks. Figure 11 shows the effect size of studies that addressed access compared to those that did not. Of the nine studies that addressed accessibility, seven (78 percent) had an effect size greater than our guideline of .2, as opposed to 13 (52 percent) of those that did not address accessibility. This difference was not statistically significant. There were the same number of studies with moderate and large effects within the studies that addressed accessibility and those that did not (one in each category each).

Another way in which the studies differed was whether they addressed other health issues beyond physical activity. We examine specifically whether the studies combined the physical activity intervention with diet or smoking cessation interventions. Figure 12 shows the effect size of studies that had interventions that included smoking cessation and/or diet interventions and those that did not. Again, the numbers are small but there does not, in this set of studies, appear to be a notable difference. Twelve of the 19 studies (63 percent) that did not include a smoking cessation or diet intervention had an effect size that exceeded our guideline of .2, compared to eight of the 15 (53 percent) of the studies that did not have those components.

The other major variation in intervention type that we set out to test was whether theoretically-based interventions were more effective than those that do not explicitly use theory. In this review we accepted the authors' statements about whether an intervention was designed with the use of theory and which theory was used. The effect size by whether an intervention was theoretically based is shown in Figure 13. The four studies with the largest effect sizes were not based upon theory. Overall, 12 of the 18 studies that did not use theory (67 percent) had an effect size greater than our guideline of .2, whereas eight of the 16 studies (50 percent) that used theory exceed that criterion. Theory based interventions were less likely to be statistically significant when examined on the level of outcomes and interventions (Table 9). On the level of outcomes, 12 percent of theory based interventions were statistically significant compared to 28 percent of those that did not explicitly use theory (p=.02). Similarly, on the intervention level 19 percent of theoretically-based interventions were statistically significant compared to 44 percent of those that did not use theory (p=.02). At the study level, a similar pattern was seen (57 percent versus 33 percent; p=.110) but it was not statistically significant. Although there are no clear differences in effect size between the theoretically based interventions and others, the results for statistical significance do not support that theoretically based interventions are more effective.

One point to keep in mind in examining these factors is that they most likely are not completely independent. That is, studies that use theory may have other characteristics in common that may also influence the results. Table 10 shows the relationship between the use of theory and intensity level. Although not statistically significant, there was a suggested trend in more intensive interventions to be less likely to be theory based. Hence, if more intensive interventions have greater effect, as was suggested by our non-significant finding that none of the lowest intensity interventions had an effect size greater .2, the apparent negative effect of theory may be a result of the intensity of the interventions rather than the effect of theory. The relatively small number of studies in the review does not allow further exploration of these questions as the number of studies in any category becomes quite small.

Time to Followup

One potentially confounding issue in this literature is the inconsistent length of followup. We used a criterion that studies must report followup data three months or more from the end of the intervention. Yet as was clear in the Study Characteristics discussion above and shown in Figure 4, there is a significant range of followup intervals. As one might expect that the effect may decrease over time, part of the difference in effects between studies may be related to the length of followup after the end of the intervention.

We examined the percent of studies that were statistically significant by the length of time to the first followup greater than or equal to three months after the intervention (Table 11). There is no clear effect of followup time on whether interventions were statistically significant ranging from 46 percent significant for those with the shortest time to 41 percent significant for those with the longest. Similarly, there is no clear pattern in effect size by the length of time between the end of the intervention and followup (Figure 14). There was no clear trend in the percent of studies with a small effect or greater, nor in those with a moderate or large effect.

The lack of clear effect of followup time seen looking across could be related to other differences between the studies other than length of time A more direct test which controls for this possibility is to examine the change in effect across time within studies. Unfortunately, few of the studies have measures of outcome at points in time short of the final outcome. We abstracted the effect of the interventions at the end of the intervention, a point greater than or equal to three months after the intervention and at the last followup point. Only 17 of the studies in the review provided sufficient data on effect size at more than one of these points in time.

The effect size by time from the end of the intervention for those 17 studies with more than one measurement is shown in Figure 15. Three quarters (73 percent) had an overall decrease in effect size over time with the average decrease in effect size per month of followup of .03 (range decrease of .14 per month to increase of .04 per month). Because of this effect, the length of followup should be taken into account when judging the individual effects in the evidence table.

Size of Study

The relationship between effect size, statistical significance, and the number of subjects analyzed is shown in Figure 16. A couple of observations may be made from this figure.

Although it is difficult to get a good measure of power of each of the studies because of insufficient variance estimates of outcomes, the measure of sample size is a reasonable surrogate looking across studies all examining a physical activity outcome or outcomes. The figure fails to show a clear positive relationship between sample size and statistical significance. Of the eight studies that analyzed over 700 subjects for which effect sizes could be calculated, only two were statistically significant compared with 44 percent of the studies overall Although there may well be interventions within this review that would have shown a statistically significant effect if they had had greater power, overall lack of sample size does not appear to be a major determining factor driving the differences seen in statistical significance between the studies.

It is difficult to assess publication bias with the diverse set of interventions and populations examined in this review. With a consistent literature, one expects to see a relationship between the distribution of effect sizes related to the size of the study (narrower at larger sample sizes) with the mean at each size the same. This sort of observation could be confounded by differences in the studies and populations at different study sizes. Nonetheless, it does appear that the smaller studies have larger effect sizes on average, suggesting that there may well be smaller negative studies missing from the literature.

Other Outcomes: Potential Harm

The entire purpose of the interventions examined in this review is to increase the physical activity of individuals to reduce their risk of adverse health outcomes. However, it is at least in theory possible that there could be adverse health outcomes associated with the interventions themselves. It is conceivable that this could be an important factor in the overall health impact of these interventions. In a most extreme example, it would not take too many elderly subjects falling and breaking a hip as a result of a fall in a walking program to outweigh the overall health benefits to the group. Yet, only one study examined potential harm of these interventions and that was simply a statement that no injuries were reported by study subjects.[79]

Measure Quality

Four means of measuring physical activity outcomes were used in the included studies: diary/log, patient recall on survey, accelerometry, and physiologic fitness measure. Of the 166 individual outcome results, 115 were obtained by survey, 16 by diary or log, two by accelerometer, and 33 were fitness measures. One concern about this literature is that the subjective measures may be more prone to bias. As these are unblinded studies one might expect that this may increase the effect size as individuals in the intervention groups report greater exercise than actually performed. More subjective measures could also act to decrease effect size by introducing random noise that may decrease the differences between groups. We examined the effect size by the type of measure (Figure 17). Again, each of the 99 individual measures was included just once in the analysis. There was little difference in effect size regardless of how physical activity was measured. Fifty three percent of survey measures had an effect size less than .2 compared to 57 percent of diary/log and fitness measures. There was only one measure using an accelerometer and it also had an effect size less than .2.

Study Quality

Two measures of quality were used. The first was an adaptation of the quality measure from the Guide to Community Preventive Services.[46] The advantage of this measure is that it has been successfully applied to the physical activity literature previously. Further, the specific criteria within the measure may be applied to all of the studies in the review. One shortcoming of the measure is that it was not designed to evaluate randomized controlled trials and therefore omits criteria that may be important in evaluating these sorts of studies, specifically those criteria relating to randomization and blinding. As 49 percent of our studies had some sort of random assignment at the individual level, we elected to use an additional quality measure to examine those criteria related specifically to randomized trials.

Measure derived from the community guide. Eighteen criteria of study quality were examined. The results are shown in Table 12. On average, the studies met under half of the quality criteria (average 7.5) but there was a wide range from a low of three criteria met to a high of 16. There was considerable variation within the criteria in the percent of the studies that met the criteria (Table 13). For example, all of the studies met the criteria, "conducting statistical testing when appropriate," whereas only one study met the criteria "controlling for differential exposure to the intervention." There is a subjective element to the assessment of whether a study met a quality criterion and so these results cannot be viewed as definitive. However, it is still of note that the abstractors felt that important details of the study populations and the interventions were lacking.

The quality measure was not specifically designed to be used as a scale. Different criteria within the scale may have different weights if one were to consider aggregating them into one overall measure of study quality. This would be a particularly important issue if one were to attempt to compare individual studies where weighting differences of individual criteria may shift one study relative to another. Yet, it is probable that the overall pattern of study results from a simple aggregation will still have a relatively good reflection of the true underlying quality of the studies. That is, studies that meet few criteria likely are poorer than those studies that meet many which are likely to be poorer than those that meet most, even if the relationship is less than exact. We therefore feel it is reasonable to examine the count of criteria used against the effect size.

Figure 18 plots effect size against the number of quality criteria met. It does appear as though the poorest studies have on average a greater effect than better studies, as only two of the ten studies (20 percent) that met six or fewer quality criteria had an effect size less than our guideline of .2, versus 12 of the 24 studies (50 percent) that met seven or more quality criteria.

One important potential quality issue with this literature is the percent of participants who were available at followup. One could hypothesize that subjects who were less adherent to the exercise recommendations may also be more likely to be lost at followup. This loss to followup may then adversely affect the study in one of two ways. If all subjects for whom there is no followup data are treated as non-responders to the intervention (as was done in some trials) the results may be biased downward. If those lost to followup are ignored in the analysis, then the intervention will likely appear to have more of an effect than it really does. Clearly, the greater the percent of subjects who are lost at followup, the greater the potential issue. Figure 19 examines the relationship in this literature between the percent of the enrolled population who are assessed at followup and the effect size. The hypothesized negative relationship (larger percentage of sample retained at followup causing smaller effect size) is not clearly borne out in this data. However, it also may be confounded by other factors that affect both quality and effect size. Therefore as a quality measure we used an arbitrary cut-off of 80 percent of subjects completing the trial as our criterion for good quality of study followup. Only 19 of the 47 studies (40 percent) met this criterion, suggesting that if followup is important, it is an issue with this literature.

Randomized controlled trial measure. To examine the additional quality dimensions specific to randomized controlled trials we used the scale of Chalmers et. al.[80] This scale was used because it contains specifically those elements that are relevant to randomized controlled trials that are missing in the community guide scale. Specifically, it examines the randomization of subjects, how withdrawals are dealt with, and blinding. The quality of the studies that randomized individual subjects is shown in Table 14.

Studies were given a rating of “one” for the method of treatment assignment if no details of randomization were given. Eighteen of the 23 studies that randomized individuals fell into this category. Three studies that provided information on randomization used methods rated as intermediate quality (such as opaque envelopes). Only one study described a randomization scheme that met the highest quality criteria on the scale.

The studies did little better on the measure of control of selection bias after treatment assignment. This measure assesses how withdrawals are treated in the analysis and what percent of the subjects were withdrawals. Studies in which more than 15 percent of the subjects randomized withdrew are given a rating of “zero.” Eighteen of the 23 studies rated “zero” on this scale. Studies get the highest rating in this measure if the results are analyzed both as treatment assigned and treatment given and the withdrawals are further examined. None of the studies met this criterion. One study received a rating of “two” because withdrawals were examined and results were analyzed by original treatment assignment (but not treatment received).

The final criterion in the scale is blinding of participants and investigators. Uniformly the studies received a “one” for this criterion. In most cases this is due to the fact that these studies do not lend themselves well to blinding. Given the nature of the intervention, it is clear to both the subjects and the investigators what treatment has been assigned. Further, as the results are usually obtained from reports of the subjects, the measure of outcome is generally obtained from an unblinded observer (i.e., the subject). However, four of the studies used outcome measures that could possibly have been blinded, such as stress testing in which the individual conducting and reading the test could have been blinded to treatment assignment, but blinding was not used in any of these studies either. Hence, these studies also received a “one” for this criterion.

Figure 20 shows the effect size of individual studies by the rating of the study on the quality scale. The possible scale range is 0-9 with large numbers representing better quality. Most studies received a rating of “two,” the highest rating amongst these studies was “five,” which was obtained by only one study. There is no clear pattern between study quality and effect size.

Search Results for Cancer Survivors

The details of the process to identify eligible exercise intervention studies conducted in cancer survivors are outlined in Figure 21. There were two MEDLINE® searches undertaken, one in July, the second in September. References from identified papers from the earlier search and references from a recent review[7] were also included. This resulted in a total of 128 papers in an EndNote® file.

This EndNote® file was reviewed twice by a project staff member with content area expertise to ascertain whether the papers needed to be pulled for full review. Of the 128 papers, 77 were identified as not being exercise interventions on the basis of study title and or abstract contents and were not obtained. The remaining 51 papers were obtained and fully reviewed. Two additional papers were identified in reviewing these 51 papers. In the process of peer review, an additional 14 papers were identified. Of these, one was eligible for inclusion, one was an exercise intervention with no concurrent comparison group, and the remaining 12 were review papers.

Figure 21 also shows the reasons for excluding 26 papers from the review. The most common exclusion criterion was the lack of a concurrent comparison group (14 papers). There were also a small number of papers that were not interventions on cancer patients (four papers) and five that were reports of baseline data and design of studies currently underway. The final number of papers included was 29. These papers described 24 unique studies.

Study Characteristics

Populations studied. Table 15 includes a description of populations studied and the interventions employed. Of the 24 studies included in the review, 54 percent conducted interventions during active cancer treatment. The sample sizes were often small, with a range of four to 101 per group and a mean of 22 in the control groups and 23 in the treatment groups. The most common diagnosis included in the studies was breast cancer, with 83 percent of the studies reporting inclusion of breast cancer survivors. After breast, the two other most common diagnoses were lung cancer and sarcoma. All included studies had concurrent comparison groups, 83 percent of them were randomized controlled trials. The PEACE framework suggested by Courneya and Friedenreich[13] was described in detail in the introduction and the percentage of studies that fall within each of the post-diagnosis PEACE framework categories is also provided in Table 15. The majority of the studies focus on the time period during or immediately following active cancer therapy, as evidenced by 50 percent of studies in the coping category and 42 percent in the rehabilitation category.

Dropout rates ranged from 0 to 25 percent with a mean of 10.8 percent. Dropout rates tended to be higher in studies that focused on those who had completed treatment (11.5 percent average dropout rate) than patients currently undergoing treatment (10.3 percent average dropout rate).

Intervention Characteristics

The majority (79 percent) of the interventions were exercise only interventions, the remaining 21 percent including some dietary, psychological counseling, or other intervention elements. The interventions tended to be relatively short, compared with those described in the other section of this evidence report. The majority of the interventions were between five weeks and three months long, with no followup after the end of the intervention. The longest intervention was 26 weeks. The vast majority (88 percent) focused on aerobic activity, 83 percent prescribed moderate to vigorous intensity activity, 88 percent prescribed physical activity three or more times per week. Fifty-eight percent of the interventions prescribed physical activity of less than 40 minutes per session, though 29 percent never specified a length of exercise session.

Of the 24 studies reviewed, 75 percent involved pre-planned exercise sessions, usually supervised, in an exercise or physical therapy facility, with the equipment and supervision provided at no cost to participants. These 18 studies cannot be evaluated with regard to the ability to change exercise behavior. By contrast, six studies (25 percent) intervened to change exercise behavior, did not tell the control group to stop exercising, and assessed whether the intervention resulted in behavior change (or some surrogate for behavior change). Based on these characteristics, these studies could be considered behavioral interventions. Further, there was one additional intervention in which an exercise prescription was given, but the program was done entirely independently, in the home, with no supervision.[81] We have identified studies as being either 'behavioral interventions' or 'pre-planned exercise' studies in each of the outcomes tables (Appendix F).

As required by the inclusion criteria, each of the 24 studies had a comparison group. The majority (17 studies) had two groups and the comparison group was a control group, in which no exercise or other treatment was prescribed. The only study to provide an intervention for non-exercising controls was the Group-Hope trial,[82, 83] in which non-exercise and exercise group participants were offered a group psychotherapy intervention.

There were seven studies with more than two groups. Segar et al. included an exercise only group, an exercise and behavior modification group, and a control group.[84] Cunningham et al. included a control group and two intervention groups.[85] One of the intervention groups received physical therapy three times weekly, the other received physical therapy five times weekly. Burnham and Wilcox et al. included a control group, a low intensity and a moderate intensity exercise group.[86] Segal et al. 2001 also included two intervention groups and a control group:[87] Intervention Group 1 had a home based self-directed exercise prescription, while Intervention Group 2 performed supervised exercise. Djuric et al. included four groups: a control group, a Weight Watchers only group, an individualized weight loss plan group, and a group that received a combination of the Weight Watchers and individualized weight loss plans.[81] MacVicar and Winningham[88] and Winningham and MacVicar[89] both included three groups: an aerobic exercise group, a placebo group that received equal attention but performed flexibility exercise, and a control group. In our outcomes tables (Appendix F), we have presented the placebo group as an exercise intervention group, because they did receive an exercise intervention (stretching), just not the same exercise intervention as the aerobic exercise group.

The loss to followup from these studies was relatively minimal, with an average of 10.8 percent overall, with a slightly lower dropout rate in studies of patients during treatment. These dropout rates should be viewed in context of the percent of cancer survivors approached regarding study participation who agree to participate or even to be screened for eligibility. The seven studies that provided data regarding the percentage of cancer survivors approached who agreed to participate or to at least be screened for study eligibility reported values of 28, 30.6, 32.5, 43, 68, 75, and 81 percent, with a mean of 51 percent.[83, 87, 90-94]

In addition to identifying the timing of the interventions with regard to whether they took place during or after treatment, each of the 24 studies has been placed into a category according to the PEACE framework proposed by Courneya and Friedenreich[13] described in the introduction section. The evidence tables (Appendix E) and the outcomes tables (Appendix F) identify whether the interventions focused on buffering (one study), coping (13 studies), rehabilitation (ten studies), health promotion (five studies), survival (one study), or palliation (zero studies). Further, five studies were found in multiple PEACE framework categories.

Outcomes examined. We grouped outcomes from the 24 studies into 16 categories and present these in Table 16, along with the number of studies that examine each of these categories or subcategories and the number of measurement tools that were used to examine a given construct. The measurement tools used to assess cardiorespiratory fitness, strength, flexibility, and body size, as well as all self-reported outcomes are described in Table 17. The two most common outcomes examined were cardiovascular fitness and fatigue or tiredness, which were examined in 12 of the 24 studies. Depression, anxiety, and quality of life were also commonly examined (ten studies), as well as body weight or body mass index (BMI) (eight studies).-

Outcome level examination. There are no subgroup analyses reported; only comparisons between treatment group(s) and control group were considered. There was one study for which there was a tremendous diffusion of effect, for which results were presented by exercise level, like a cohort analysis.[95, 96] At the level of outcome type there were three methods used to assess intervention effects. We calculated effect sizes, examined whether results were statistically significant, and assessed whether results were in the hypothesized direction, regardless of statistical results.

An attempt was made to examine effect size at post test only. If means and standard deviations were available for both groups at post intervention testing, an effect size was calculated. These post intervention effect sizes are more useful in studies with no baseline differences between groups, which is more likely in the larger randomized controlled trials than smaller and non-randomized controlled trials. Comments on between group differences at baseline have been included in the text in order to guide interpretation of effect sizes. Because of the potential for overestimating effects if there were pre-intervention between group differences, effect sizes are commented on individually within each outcome type rather than considering an effect size over a given value to be 'large' or 'small.'

If insufficient data were available to calculate an effect size, the p-values for the outcomes were provided in the outcomes tables (Appendix F). An effect was considered to be statistically significantly positive if a statistical test was performed that demonstrated that the intervention group had greater improvement in the outcome of interest than the control group with an alpha value of 0.05. An effect was considered positive if the results were in the hypothesized direction but not statistically significant. This criterion was included because there is no clinically important threshold known for the wide variety of outcomes reported.

As discussed in the results from the non-cancer population, if all of the data were reported, the probability of a positive effect would be 50 percent if there was actually no effect of the intervention. Hence, a rate of 50 percent positive outcomes would be evidence of no effect of the interventions. The results here may be skewed below 50 percent where there is no effect because some of the studies that had small positive effects may have reported 'no effect.'

Intervention level examination. The number of studies with positive and statistically significant effects, as well as the mean and range of calculable effect sizes are provided in Table 18 for each of the 16 outcome categories. The criterion for considering an intervention positive was if one or more of the outcomes in a given outcome category was positive. An effect was considered to be statistically significantly positive if any one of the outcomes examined within a category was statistically significant. The intention here is to convey a level of positivity of results, not to perform a statistical test. Significance was not corrected for multiple tests. The effect sizes reported are a comparison of between group means at post-intervention only, given that pre-post correlations for all 16 outcome categories was not available. The mean effect sizes are not corrected for sample size.

Overall Effect

The overall effect of interventions on all studies, within 16 outcome categories, is provided in Table 18. Categories with 100 percent positive findings include strength, flexibility, fatigue/tiredness, confusion, difficulty sleeping, self-esteem, psychosocial outcomes, body size (goal to reduce) vigor/vitality, immune parameters, and mental health quality of life.

The percent of studies reporting statistically significant results within the 16 categories ranged from zero percent for confusion and body size (goal to gain or avoid muscle loss) to 100 percent for flexibility and difficulty sleeping. There were eight categories with 75 percent of studies reporting at least one statistically significant finding: cardiorespiratory fitness, flexibility, fatigue/tiredness, quality of life, difficulty sleeping, psychosocial outcomes, physiologic outcomes, and immune parameters.

Mean effect sizes ranged from –0.055 for immune parameters to 2.93 for physical activity behavior. Outcome categories with effect sizes of 0.20 or greater include physical activity behavior, cardiorespiratory fitness, flexibility, fatigue/tiredness, body image/dissatisfaction, quality of life, confusion, vigor/vitality, symptoms/side effects, depression, anxiety, and the combined multiple constructs section of mental/emotional/psychological well-being.

The categories vary with regard to the appropriateness of combining results, thus, results are also presented for each outcome category in a later section.

Effect by Timing: During Versus Post Treatment

We further examined whether the results of studies would be more likely to be positive during versus post active cancer treatment. Table 19 presents the results again, divided by timing of intervention: during versus post treatment. For many categories, there are too few studies to compare the results across this timing variable. Exceptions include the negative effect size for immune parameter changes post-treatment versus the positive effect size during treatment and the larger positive effect sizes during treatment for quality of life, self-esteem, psychosocial outcomes, physiological outcomes, anger/hostility, and the multiple constructs section of the mental/emotional/psychological well-being category.

Effect by Outcome Category

Tables F-1 to F-16 in Appendix F provide descriptions of the studies as well as outcomes in each of the 16 categories defined earlier.

Physical activity behavior (Table F-1). There were six studies that could be considered behavioral interventions, defined as interventions designed to examine whether cancer survivors would adhere to an exercise prescription on their own, and in which the control group was not asked to stop exercising or avoid starting exercise. One of these studies[81] was a weight loss intervention that included an exercise component and no physical activity behavior data was provided. Therefore, this study is not included in Table F-1. One of these also included a pre-planned, supervised exercise group[97] and reported changes in physical fitness that can be used as a surrogate for physical activity behavior. All five behavioral interventions that reported a physical activity behavior outcome reported statistically significant increases in at least one physical activity behavior variable (or surrogate) as a result of the intervention. Only one study provided adequate information to calculate an effect size for post intervention between groups differences. The large effect size (2.93) is mostly due to pre-intervention between group differences in this small randomized controlled trial, though the Mann-Whitney U test for intervention effect on a categorical exercise level scale did have a p-value < 0.01.

Three of the behavioral studies came from one research group and the intervention for these three studies was identical.[84, 95, 96, 98, 99] In one of these studies,[84, 96] there was significant cross-over after randomization: Fifty percent of the usual care group was exercising at levels as high or higher than the level prescribed for the treatment group, one-third of the treatment group failed to do any exercise. The investigators analyzed the data as an observational cohort.

One of the 24 reviewed studies examined the psychosocial mediators of adherence to the exercise intervention[83, 94] and reported that past exercise and female gender were associated with exercise during the study, regardless of experimental condition.

Physical fitness: cardiovascular, strength, and flexibility (Table F-2). All 13 studies that examined the efficacy of physical activity interventions to increase one or more aspect of physical fitness reported fitness improvements. We were able to calculate effect size for cardiovascular fitness from seven studies with a range of effect sizes of 0 to 1.242, though this largest effect size is strongly influenced by large baseline differences. The range of post intervention effect sizes for cardiovascular fitness in the four studies with minimal baseline between group differences was 0.319 to 0.950, indicating a consistently positive effect on cardiovascular fitness. There were two studies that reported results regarding flexibility, both provided sufficient data to calculate post intervention effect sizes (0.024 and 0.666); neither study had between group differences in flexibility at baseline. Two studies reported results regarding muscular strength, both reported improvements, though only one of them reported a statistically significant improvement; neither provided sufficient data to calculate post intervention effect size.

Fatigue/tiredness (Table F-3). There were 12 studies that examined whether an exercise intervention would positively alter symptoms of fatigue or tiredness in cancer survivors. Of these, six were conducted post-treatment and six during active cancer therapy. All but one of the 12 studies reported a positive effect,[93] though the size of the effect varied. During active treatment, exercise interventions positively affected fatigue or tiredness in all six studies, with three reporting statistically significant improvements. Sufficient data was provided for the post intervention effect size calculation in one of the five studies conducted during active cancer treatment (Effect Size 0.130), this study reported no between group baseline differences for fatigue.[91]

Of the six studies conducted with survivors post treatment, five reported improvements in fatigue, though the magnitude of the improvement varied. Sufficient data were provided for post intervention effect size calculation in three of these studies and the effect size ranged from 0.031 to 0.645. The study with the largest effect size,[86] prescribed the lowest intensity exercise. None of these three studies reported baseline differences that would make these effect sizes an over estimation. In fact, in one study,[90] baseline differences make the effect size of 0.063 an underestimation of the intervention results. The p-value for the ANCOVA analysis of fatigue effects in this study was 0.006. Further, the smallest effect size of 0.031 was calculated for consistency with the other effect sizes in this report, which were all calculated with post intervention values only, given missing data on pre-post correlations for outcomes. However, in this study,[82] the authors report the post intervention minus pre intervention effect size[55] to be 0.28, much larger than what is observed using only post-test data.

As shown in Tables 19 and F-3, eight different instruments were used to assess the effect of exercise interventions on fatigue. Studies that used the Piper or Functional Assessment of Cancer Therapy (FACT) fatigue scales all reported statistically significant improvements in fatigue as a result of exercise participation.

Body image/dissatisfaction (Table F-4). There were four studies that reported outcomes related to body image or body dissatisfaction. All four studies included in Table F-4 included breast cancer patients and were conducted post treatment. One of these studies reported positive, statistically significant improvement in body image and dissatisfaction after a moderate intensity aerobic exercise intervention, 10-45 minutes per session, four to five days per week.[94] Adequate data were provided to calculate an effect size from one study.[99] For this study, the effect sizes for body image, measured on two separate scales, were 0.301 and 0.318. However, these effect sizes were mostly driven by between group differences at baseline and may reflect an overestimation of the impact of the exercise intervention on body image. Both of these positive studies[94, 99] were conducted in breast cancer survivors exclusively. The two non-positive studies included a variety of cancer diagnoses and the intensity and frequency of prescribed exercise was lower.

Quality of life (Table F-5). Ten studies examined the effects of exercise on quality of life (QOL) in cancer survivors. There were eight unique QOL instruments used in these ten studies (see Table 17). Six of these studies were conducted post-treatment and four during active cancer therapy. Three of the four studies conducted during active treatment reported statistically significant improvements in at least one measure of quality of life. Effect sizes of 0.168 and 1.155 were calculated for two of these studies;[95, 99] both of these studies observed between group differences at baseline that make these effect sizes likely underestimates. Health-related quality of life was consistently reported to be improved as a result of exercise interventions conducted during active cancer treatment. Three of the exercise prescriptions in these active treatment studies focused on aerobic activity of moderate to vigorous intensity, four to six days per week, with exercise sessions of ten to 45 minutes in duration. One of the studies focused exclusively on strength training, three times weekly.[91]

In the six studies conducted post cancer therapy, five reported statistically significant improvements in at least one measure of quality of life. Post intervention effect sizes ranging from zero to 1.689 were calculated for results of three studies. The two studies with smaller effect sizes[82, 83] observed baseline differences that likely make some of the post test effect sizes underestimates. Courneya et al.[82] reports effect sizes of 0.18 for physical well-being (compared to 0.02 in Table F-5), and 0.03 for functional well-being (compared to 0.049 in Table F-5). Our calculations are based on data provided in the paper, while the effect sizes reported in the publication are based on pre-intervention values, post-intervention values, and the correlation between pre and post values. There was consistency in the positive direction if not the magnitude of changes observed. The exercise intervention in the post-treatment interventions all focused on aerobic activity, with intensity (where reported) ranging from 25-40 percent of heart rate reserve to 70-75 percent of maximal oxygen consumption, frequency of one to five days per week, and duration of 14 to 60 minutes per session.

Confusion (Table F-6). Two studies examined the effect of aerobic exercise on measures of confusion, one during and one post cancer treatment. Both reported small improvements in confusion as a result of an exercise intervention, though neither result was statistically significant. An effect size of 0.402 was calculated for the post treatment intervention in breast cancer survivors. Both studies prescribed aerobic activity of moderate to vigorous intensity, three days weekly, from 14 to 32 minutes per session.

Difficulty sleeping (Table F-7). The two studies that examined the effect of aerobic exercise on difficulty sleeping post treatment for breast cancer both reported statistically significant improvements after a program of moderate intensity aerobic activity four to five days a week for ten to 45 minutes per session. Insufficient data were provided to determine effect size for either study.

Self-esteem (Table F-8). Three studies examined whether moderate to vigorous intensity aerobic exercise three or more days per week would improve self-esteem in post treatment breast cancer survivors. All three observed improvements in exercise participants, though only one reported a statistically significant difference between groups. Differences between this study and the other two include higher intensity of exercise prescribed (70-75 percent of aerobic capacity versus 60 percent of age predicted maximal heart rate) and a longer intervention (15 weeks versus ten weeks). Effect sizes of 0.044 and 0.154 were calculated for two of the studies, and baseline differences indicate that the smaller of these two values is likely an underestimate.

Psychosocial outcomes (Table F-9). There were six studies that examined a variety of psychosocial outcomes, including activities in the community, activities in the home, change of lifestyle, participation in patient organizations, satisfaction about information provided as a patient, sick leave, work status, cognitive functioning, communication with clinic staff, information problems, happiness, social/family well-being, role limitations, social functioning, hope, and power. Only one of these studies neglected to show any positive or statistically significant effect of exercise on psychosocial outcomes. Multiple outcomes were measured in each of these studies, resulting in multiple comparisons within each study. There were 14 separate measurement tools used for a variety of constructs measured (see Table 17). Effect size was calculated for the effects of aerobic exercise on post treatment breast cancer survivors for happiness (ES = 0.302), social/family well-being (two ES calculations: 0.005 and 0.113), satisfaction with life (ES = 0.028), and spiritual well-being (ES = 0.00). Further, effect sizes of 0.280 and 0.612 were calculated for the buffering effects of pre-lung cancer surgery exercise effects on hope and power, respectively.

Physiologic outcomes (Table F-10). The three of the four studies that examined physiologic outcomes focused on the active cancer treatment time frame. Fairey et al.[100] examined the effects of aerobic exercise on insulin, glucose, and insulin-like-growth factor (IGF) variables (IGF-1, IGF-2, and two IGF binding proteins: IGFBP-1, IGFBP-3) in post treatment breast cancer survivors. Changes in the hypothesized direction were reported for IGF-1, IGFBP-1, IGFBP-3, and the IGF-

1:IGFBP-3 molar ratio, with effect sizes of 0.414, 0.025, 0.425, and 0.657, respectively. Other reported variables either did not change or changed in the opposite direction of what was hypothesized. Cunningham et al. examined the effect of three or five times weekly physical therapy exercises on muscle mass loss in acute leukemia bone marrow transplant receipients.[85] Results indicated a muscle sparing effect of exercise that was mostly too small to be detected statistically. Dimeo et al. also examined effects of exercise on physiologic parameters during bone marrow transplant, though the exercise prescription was aerobic activity in this study.[101] Effect sizes of 0.00 to 0.528 are reported in Table F-10 for the physiologic parameters assessed in Dimeo et al.,[101] indicating no harm of exercise for any these variables and significant improvement for a subset of physiologic outcomes, particularly number of in-hospital days (ES = 0.528). Dimeo et al. also examined the effects of high intensity walking post-hospital discharge for bone marrow transplant on cardiac function and hemoglobin.[102] An effect size of 0.822 for hemoglobin was calculated. The authors report a p-value of 0.04 for between group differences in hemoglobin after seven weeks of exercise training, though the actual values were 13.0 versus 12.0 g/dL in the treatment and control groups, respectively. Segal et al. examined whether resistance training in men undergoing androgen deprivation therapy for prostate cancer would result in increased PSA or testosterone levels.[91] The non-significant changes in both groups were reported by the authors to indicate the safety of resistance exercise for this population.

Body size (Table F-11). Ten studies examined the effect of exercise on some measure of body size. In Table F-11, these have been divided into two subsets according to whether the goal was to decrease body weight or body fat versus a goal of maintaining muscle mass, avoiding cachexia, or avoiding arm volume increases.

Of the six studies that examined whether exercise could decrease weight or body fat, alone or in combination with diet changes, four reported significant reductions in at least one body size related variable in the treatment group(s) when compared to changes in the control group(s). The only study to report a significant decrease in body weight in the treatment compared to the control group included a strong dietary intervention component.[81] Effect sizes for these studies, where calculable, ranged from 0.015 for body weight in Courneya et al.[90] to 0.636 for body weight in Burnham and Wilcox et al.[86] The large effect size from Burnham and Wilcox is mostly a reflection of large between group differences at baseline. In general, body size changes were small in all exercise interventions that stated goals of decreasing fat or weight, for studies conducted during as well as post treatment.

There were three studies that examined the effects of exercise on cancer survivors during either bone marrow transplant or androgen deprivation therapy that included a measure of body size. Cunningham et al.[85] and Dimeo et al.[102] examined whether physical therapy exercises[85] or aerobic activity[102] would prevent muscle wasting during bone marrow transplant. Both studies reported no muscle mass or body mass index change resultant to exercise training. Segal et al. examined whether strength training during androgen deprivation therapy would prevent muscle mass loss and reported no significant differences between groups after 12 weeks of strength training three times weekly at a relatively high intensity.[91]

Finally, one pilot study was conducted to assess the safety of upper body aerobic and resistance exercise in breast cancer survivors with lymphedema.[103] This pilot study reported no changes in arm volume. The effect sizes for both measures of arm volume were 1.642 and 1.262 mostly reflect between group differences at baseline. The arm volumes of control group participants were larger at baseline and stayed larger through out the eight-week intervention.

Pain (Table F-12). There were three studies that examined changes in self-reports of pain after an exercise intervention. None of these studies provided adequate information for calculation of effect sizes. One of these reported a statistically significant improvement in self-reported pain among post treatment survivors with a variety of cancer diagnoses after four weeks of low intensity aerobic activity and strength training at a frequency of one time weekly.

Vigor/vitality (Table F-13). Of the six studies that examined changes in vigor or vitality, five reported some positive effect. Effect sizes of 0.434 and 1.265 were calculated from one study conducted post-treatment and one study conducted during treatment, respectively. Neither study reported baseline differences between groups for vigor. Both of the studies conducted post-treatment showed improvements in vigor (ES = 1.265 for one and p = 0.023 for the other) and both prescribed aerobic exercise three days a week at moderate intensity, ranging from 14 to 60 minutes per session.

Three of the four studies conducted during active cancer treatment reported improvements (ES = 0.434 and p-values of 0.023, 0.00, and 'mean changes showed an increase in treatment group'). Of the four studies conducted during treatment three focused exclusively on breast cancer patients. The longest of these (a 26-week intervention) showed no effect on vitality. In contrast, a six-week intervention in breast cancer patients during treatment showed a significant improvement in vigor with a very similar exercise prescription (moderate intensity aerobic activity four to six days a week, 20-30 minutes per session).

Symptoms/side effects (Table F-14). There have been five studies that have examined whether exercise during or post treatment might improve patients' experiences of cancer treatment related symptoms and side effects. One was conducted post treatment and showed no effect of exercise training. Two of the four interventions that took place during treatment were specifically conducted in bone marrow transplant patients. Effect sizes for these two studies were 0.547 for somatization, 0.507 for diarrhea, 0.225 for severity of infection, -0.130 for mucositis, and 0.849 for severity of pain, indicating that exercise resulted in positive changes in most symptoms/side effects assessed in bone marrow transplant patients. The exercise intervention in both of these studies was 15 minutes of aerobic exercise seven days weekly, at 50 percent of the heart rate reserve. The other two interventions that took place during treatment focused on breast cancer patients, one of these showed significant improvements in vomiting and nausea resultant to ten weeks of moderate intensity aerobic activity three days a week.

Immune parameters (Table F-15). Of the four studies conducted to assess the effect of exercise on immune parameters, three took place during treatment. All three studies conducted during treatment showed statistically significant improvements in a variety of immune parameters, including T-cells, lymphocytes, white blood cells, and natural killer cell activity.[101, 104, 105] Insufficient data were provided to calculate effect sizes for two of the studies. For an intervention among bone marrow transplant recipients by Dimeo et al.,[101] effect sizes of 0.643 and 0.442 were calculated for duration of neutropenia and thrombopenia, respectively. All three interventions conducted during active treatment prescribed moderate intensity aerobic activity on three to seven days weekly for 15 to 90 minutes per session. For the only study conducted post treatment,[103] effect sizes were 1.047 and 0.636, for natural killer cell cytotoxicity (E:T20:1 and E:T40:1 respectively), indicating improved cytotoxicity. Effect sizes for lymphocytes, neutrophils, natural killer cells, t-cells, and total leukocytes were below zero and ranged from –0.417 to –7.99, suggesting less favorable values for immune parameters in exercise participants compared to controls at post-testing. There were differences between groups at baseline for the measures of cytotoxity (which had positive effect sizes), but not for the other immune parameters. None of the effects in Nieman et al. were statistically significant.[103]

Mental/emotional/psychological well-being (Table F-16). The effect of physical activity has been assessed on a variety of mental health related parameters in cancer survivors. We review here the results related to anxiety, depression, anger or hostility, general mental health, and a comment on the remaining studies that have examined a broad variety of other mental health constructs. Overall, the majority of the nine studies conducted post treatment had at least one significant improvement in a parameter related to mental and emotional health. Half of the four studies conducted during treatment had at least one significant improvement in a parameter related to mental and emotional health.

Anxiety. Of the ten studies that have been conducted to assess the impact of exercise training on anxiety in cancer survivors, five were conducted post treatment. Of these post-treatment studies, three report statistically significant improvements in anxiety after exercise training. Effect sizes of 0.00 and 0.901 were calculated for two of the studies.[82, 86] Of the five studies conducted during treatment, four reported improvements in anxiety with exercise training and three reported statistically significant improvements. An effect size of 0.278 for anxiety, as a result of 15 minutes of daily aerobic exercise at 50 percent of heart rate reserve in an intervention conducted with bone marrow transplant patients.[106]

Depression. Of the ten studies that assessed effects of exercise training on depression during or post cancer treatment, five were conducted post treatment. Of these five post treatment studies, two report statistically significant improvements in depression after exercise training. Effect sizes of 0.005 and 1.279 were calculated for two studies that prescribed aerobic activity of moderate to vigorous intensity 14-32 minutes per session, three to five days weekly; both studies were small, neither of these effects were statistically significant as reported by the authors.[82, 86] All five studies conducted during treatment reported some improvement in depressive symptoms, though the magnitude of these improvements was often small and were only statistically significant in two studies. Effect sizes of 0.079 and 0.263 were calculated for the Profile of Mood States and Symptom Check List (SCL-90-R) assessments of depression in the only study that reported sufficient data to calculate an effect size.[106]

Anger/hostility. There were three studies that assessed changes in anger or hostility resultant to exercise training. Two of these studies were conducted during treatment, one post treatment. Both of the interventions conducted during treatment reported small improvements. In one study conducted in bone marrow transplant receipients,[106] effect sizes of 0.063 and 0.266 were calculated for the anger and/or hostility scales from the Profile of Mood States or the Symptom Check List (SCL-90-R) respectively. The study conducted post treatment showed a slightly negative effect size of –0.114 after 10 weeks of aerobic exercise. This negative effect size is mostly reflective of baseline differences between groups. The authors reported no statistically significant difference between groups for anger based on repeated measures ANOVA.[86]

Mental health. Two studies included a measure that was called 'mental health quality of life.' Both studies were conducted exclusively in breast cancer survivors, one during treatment, one post treatment. The post treatment intervention showed a significant increase;[107] the other did not.[87] The exercise intervention in the study with the significant increase included moderate intensity aerobic exercise and resistance training for 30 to 60 minutes three times a week and lasted eight weeks.

Other constructs related to mental health. Other constructs assessed included global psychologic distress, total mood disturbance, avoidance, fatalistic, fighting spirit, hopelessness, emotional well-being, trial outcome index score (a well-being score for breast cancer survivors), impact of medical illness on subject, and psychologic distress. These constructs were studied in survivors with a variety of cancer diagnoses, during treatment (three studies), and post treatment (three studies). There was no obvious pattern of findings to report.

Study Quality

Tables 20 and 21 provide information about study quality variables abstracted from each of the included studies. Of the 24 studies described, only two described the sample as to cancer diagnoses and treatment course, race/ethnicity, gender, and sociodemographic variables. The rest either neglected to provide these variables at baseline and/or provided some of these variables for those who completed the study only. This makes it difficult to determine who was recruited versus who was able to complete the study.

There were seven studies that described the exercise intervention with inclusion of exercise modality, intensity, frequency, duration per session, and progression of these variables throughout the intervention in a manner that would allow others to repeat what they had done. Only four studies failed to include information about the reliability and or validity of the measured outcomes of interests.

Of the 24 studies reviewed, only one did no statistical testing. In this very small feasibility study, the pre and post intervention mean values for oxygen uptake, mood disturbances, and Profile of Mood States scores were compared in exercising breast cancer patients (n=6) versus non-exercising breast cancer patients (n=4) and healthy non-exercising controls (n=6). The authors made qualitative statements regarding the results, with no statistical testing provided.

None of the included studies examined or controlled for differential exposure to the intervention in assessing treatment effects.

Each of the studies included measures repeated at least two time points (pre and post intervention). Approximately half of the studies (12) conducted analyses that were appropriate for repeated measures, such as ANCOVA or repeated measures regression analysis. The rest of the studies conducted tests that did not account for within person correlations between repeated measures.

Only two of the included studies adequately reported baseline values for all participants, including sociodemographic variables, race/ethnicity, age, gender, and cancer diagnosis and treatment. Half of the studies reported baseline values for all participants who started the study, as opposed to all participants who finished the study. The other half did not. This could introduce bias into the results if those who do not finish the study are in some way different from those who do. Of the 12 studies that did not report baseline values for all participants who started the study, two reported that the dropouts were lost prior to baseline measures.

Re-examination of outcomes in Tables F-1 to F-16 in Appendix F according to study quality variables in Table 20 revealed no obvious pattern of differences. For example, examination of results of studies that reported 80 percent of participants finishing the study compared to studies that suffered greater loss to followup did not reveal any pattern of differences in results. On the other hand, there are several examples in this literature of larger studies that met more of the 11 study quality criteria that differed from smaller, less well-conducted studies. First, Berglund et al.[92] included 30 participants per group and lost more than 20 percent of the sample, and did not report reliability or validity of measures. The results from this study often stand in sharp contrast to another paper from the same author[93] in which these study quality deficits were corrected and the sample size was larger (98 in the intervention group, 101 in the comparison group). Throughout the outcomes tables (Appendix F), the results of these two studies differ from one another, despite similarities of intervention. The larger, later study[93] is the better quality study and results from this study may be more likely to be accurate as a result. In addition, Mock et al.[99] can be compared to Winningham et al.[89, 108] for the specific outcome of nausea. Winningham's study was larger and met more of the study quality criteria evaluated for this report and reported a statistically significant improvement in nausea, despite very similar interventions in the two studies.

Adverse Events Issues

Of the 24 reviewed studies, 11 commented on the presence or absence of adverse events. In ten of the studies that commented on adverse events, the comments indicate that no harm was observed as a result of exercise during or after cancer treatment.[85-87, 91, 95, 101-104, 107] The exception was Courneya et al.[90] in which overall rates of adverse events were similar between groups of breast cancer survivors, but the rate of lymphedema in the exercise group was higher. The authors note that two of the three participants who developed lymphedema had had axillary irradiation, a strong risk factor for lymphedema. The authors commented that it was not clear whether the onset of lymphedema was due to the exercise. Further, a group of researchers in Edmonton, Alberta, Canada,[107] conducted a pilot study specifically to examine the safety of upper body exercise in breast cancer survivors with lymphedema and reported no increases in arm volume in the treatment as compared to the control group.

There were several studies that commented on issues related to the potential for harm from exercise in cancer survivors. For example, Nieman et al.[103] notes that there is evidence from animal studies that high intensity high volume physical activity in cancer patients can increase the spread of the disease.[109-111] The results of the reviewed studies do not allow for evaluation of this possibility in human subjects, but this animal data cannot be ignored in considering the appropriate exercise prescription for human cancer survivors.

Mock et al.[95] commented that self-reported data collection of worsening of side effects leaves open the possibility that survivors with more extreme side effects brought on by exercising may not have felt well enough to complete data collection at the end of the study. MacVicar and Winningham[112] noted that there are conditions during cancer treatment and recovery that preclude any physical activity, including chest pain, irregular pulse, acute vomiting, blurred vision, sudden onset dyspnea, bleeding, and extreme immune-compromised states. A balance of harm and benefit needs to be considered when prescribing activity for cancer survivors.

Caveats: Measurement Limitations, Quality of the Literature

Cancer Survivors. All but four of the studies reported some information on the reliability and validity of outcomes measured. However, the broad variety of measurement tools used for each construct (see Table 17) makes it difficult to compare results across studies. Combine this limitation with the broad variety of timing with regard to treatment and populations and the potential for quantitative analysis all but disappears.

All 24 included studies used convenience samples, as would be expected in this patient population. The source of patients recruited into these studies and flow of subjects from recruitment to study end was generally not well explicated. There are a few notable exceptions. Courneya et al. started with the Alberta Cancer Registry and provides a flowchart of recruitment, intervention, and measurement subject participation.[90] Segal et al. provide a similar flow chart, but does not indicate how or where the original 378 patients were found.[87]

Further, none of the interventions had any followup beyond the end of the intervention to assess whether the physical activity behaviors or the other outcomes of interest were maintained. The sample sizes of the included studies was relatively small (average control group = 22, average treatment group = 23 subjects) and study quality requires improvement in the future, given that of the 24 studies reviewed, half the studies (or more) failed to meet six of 11 study quality criteria applied.