Effects of Teamwork Training on Adverse Outcomes and Process of Care in Labor and Delivery

A Randomized Controlled Trial

Peter E. Nielsen, MD, Marlene B. Goldman, ScD, Susan Mann, MD, David E. Shapiro, PhD, Ronald G. Marcus, MB, BCh, Stephen D. Pratt, MD, Penny Greenberg, RN, Patricia McNamee, RN, MS, Mary Salisbury, RN, MSN, David J. Birnbach, MD, Paul A. Gluck, MD, Mark D. Pearlman, MD, Heidi King, MS, David N. Tornberg, MD, MPH, and Benjamin P. Sachs, MB, BS

OBJECTIVE: To evaluate the effect of teamwork training on the occurrence of adverse outcomes and process of care in labor and delivery.

METHODS: A cluster-randomized controlled trial was conducted at seven intervention and eight control hospitals. The intervention was a standardized teamwork training curriculum based on crew resource management that emphasized communication and team structure. The primary outcome was the proportion of deliveries at 20 weeks or more of gestation in which one or more adverse maternal or neonatal outcomes or both occurred (Adverse Outcome Index). Additional outcomes included 11 clinical process measures.

RESULTS: A total of 1,307 personnel were trained and 28,536 deliveries analyzed. At baseline, there were no differences in demographic or delivery characteristics between the groups. The mean Adverse Outcome Index prevalence was similar in the control and intervention groups, both at baseline and after implementation of teamwork training (9.4% versus 9.0% and 7.2% versus 8.3%, respectively). The intracluster correlation coefficient was 0.015, with a resultant wide confidence interval for the difference in mean Adverse Outcome Index between groups (–5.6% to 3.2%). One process measure, the time from the decision to perform an immediate cesarean delivery to the incision, differed significantly after team training (33.3 minutes versus 21.2 minutes, $P=.03$).

CONCLUSION: Training, as was conducted and implemented, did not transfer to a detectable impact in this study. The Adverse Outcome Index could be an important tool for comparing obstetric outcomes within and between institutions to help guide quality improvement.

CLINICAL TRIAL REGISTRATION: ClinicalTrials.gov, www.clinicaltrials.gov, NCT00381056

(Obstet Gynecol 2007;109:48–55)

LEVEL OF EVIDENCE: I

From the Department of Obstetrics and Gynecology, Madigan Army Medical Center, Tacoma, Washington; the Departments of Obstetrics and Gynecology and Anesthesia, Beth Israel Deaconess Medical Center, Boston, Massachusetts; the Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts; Miller School of Medicine, University of Miami, Miami, Florida; the Department of Obstetrics and Gynecology, University of Michigan, Ann Arbor, Michigan; TRICARE Management Activity, Falls Church, Virginia.

This study was funded through a subcontract with Dynamics Research Corporation under contract to the Armed Forces Institute of Pathology and Office of the Assistant Secretary of Defense (Health Affairs), TRICARE Management Activity Contract DAMD171-C-0052, the Risk Management Foundation of the Harvard Medical Institutions, and the Beth Israel Deaconess Medical Center Obstetrics/Gynecology Foundation. The sponsors participated in the design and conduct of the study, data collection, and preparation of the manuscript.

The views expressed in this article are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. The American College of Obstetricians and Gynecologists does not officially endorse the Adverse Outcome Index measure.

The authors thank Stephen W. Lagakos, PhD, Stuart J. Pocock, PhD, and Allan Donner, PhD, for their expert statistical consultation. The authors also thank the members of the MedTeams Labor and Delivery Consortium (see Appendix).

Corresponding author: Marlene B. Goldman, ScD, Department of Obstetrics and Gynecology, Dartmouth-Hitchcock Medical Center, One Medical Center Drive, Lebanon, NH 03756; e-mail: Marlene.B.Goldman@dartmouth.edu.

Financial Disclosures

Dr. Sachs's department received funding to collaborate with the Department of Defense and the American Research Institute to develop a new curriculum for team training that is now available to the public. Drs. Mann, Marcus, Pratt, and Sachs and Ms. Greenberg are developing quality assurance and team training programs with the Harvard Risk Management Foundation that will be provided through a limited liability company of which they are part owners. Ms. Salisbury is the founder and president of the Cedar Institute Inc, and in that capacity, contracts to deliver teamwork research subject matter expertise, curriculum development, and training to military and civilian organizations.

ISSN: 0029-7844/06

The Institute of Medicine (IOM) has issued two reports outlining a strategy to improve the quality of health care in the United States.[1,2] These reports speculated that team training and the implementation of team behaviors could reduce medical errors and improve patient safety. In the 1980s, the Department of Defense developed crew resource management training to improve the safety of air operations in the military.[3] Crew resource management was defined as "error management capability to detect, avoid, trap or mitigate the effects of human error and therefore prevent fatal accidents."[3] In the United States today, crew resource management has been instituted in all three branches of the military and in commercial aviation. The Department of Defense has had a long-standing interest in evaluating the concept of crew resource management as a teamwork tool to reduce human errors in medicine and has sponsored the development of a team training program in emergency medicine.[4,5]

From 2002 to 2004, we conducted a cluster-randomized clinical trial of crew resource management in obstetrics. Obstetrics was chosen because the labor and delivery environment requires intense, error-free vigilance and effective communication between many different clinical disciplines including obstetricians, midwives, nurses, anesthesiologists, and pediatricians. Secondly, the practice of obstetrics is being hurt by a liability insurance crisis in the United States. Therefore, it was felt that the obstetrics community would be open to a major change in clinical behavior.

The purpose of this study was to evaluate the effect of staff teamwork training on adverse outcomes and clinical processes in labor and delivery units, to develop tools for evaluating obstetric outcomes, and to address the methodologic challenges of conducting a cluster-randomized controlled clinical trial to assess the quality of care provided in the labor and delivery environment.

MATERIALS AND METHODS

We conducted a cluster-randomized controlled trial to evaluate the effectiveness of a teamwork training intervention in reducing adverse outcomes and improving the process of care in hospital labor and delivery units at 15 U.S. hospitals (Table 1). A balanced, masked randomization scheme at the hospital (cluster) level was implemented by the project biostatistician (D.E.S.), who used a permutation approach to assign seven hospitals to receive a teamwork-training curriculum and eight hospitals to a control arm, with balancing for hospital type (military, civilian with minimum external funding, or civilian externally funded) and annual number of deliveries.[6,7] At each step, a table of random numbers was used to simulate the toss of a coin. First, hospital identity (but not type) was masked by randomly assigning an alphanumeric label to each hospital in random order (eg, m1, m2 through m6 for the six military hospitals). All possible allocations of the hospitals to two arms balanced for hospital type and funding level were then generated. From among all allocations with a similar distribution of annual numbers of deliveries between the arms (as indicated by similar geometric means and standard deviations), one allocation was randomly chosen. Designation of one group as intervention and the other as control was then assigned randomly. The trial was not blinded, with personnel at each site aware of their assignment to either the intervention or control arm.

Table 1. Participating Hospitals and Mean Number of Annual Deliveries by Study Group

Study Site	Annual Deliveries
Intervention hospitals	3,543
Military	
National Naval Medical Center Bethesda	2,200
Madigan Army Medical Center	1,800
Naval Medical Center San Diego	3,600
Civilian	
Baptist Health Hospital of South Florida	3,900
William Beaumont Hospital Royal Oak	6,700
University of Alabama	3,000
University of Michigan	3,600
Control hospitals	3,798
Military	
Naval Hospital Camp Pendleton	1,680
Portsmouth Naval Medical Center	4,300
Tripler Army Medical Center	3,000
Civilian	
Baystate Medical Center	4,500
Johns Hopkins University	1,705
South Shore Hospital	4,000
Jackson Memorial Hospital, Miami	7,000
University of Vermont	2,200

All women with a pregnancy of 20–43 weeks of gestation who were admitted to one of the 15 hospitals between December 31, 2002, and March 31, 2004, were included. Data were collected during and immediately after delivery by labor and delivery staff under the supervision of centrally trained data coordinators. Data collection was divided into baseline (2 months before teamwork training) and post-implementation (5 months after the teamwork curriculum was adopted) periods. For continuity, data were collected at all hospitals during the mid-study training period of 4 months, but these data were not included in

the analysis. The coordinating hospital was the Beth Israel Deaconess Medical Center (Boston, MA), which served as a test-bed for the teamwork training intervention but was not randomized and did not contribute data to the analysis. Data management was conducted by Frontier Science and Technology Research Foundation (Amherst, NY). The study protocol was approved by the institutional review boards at the coordinating hospital, each participating hospital, the data management center, and the Office of Regulatory Compliance and Quality, U.S. Army Medical Research and Materiel Command (Fort Detrick, MD).

Investigators adapted a standardized teamwork training curriculum, called MedTeams Labor & Delivery Team Coordination Course, with principles based on crew resource management and a curriculum used in hospital emergency and obstetric departments.[4,5,8] Crew resource management attempts to capitalize on the ability of each crew (team) member to see, analyze, and react to the same situation in ways that reduce the potential for error. Clinical staff from the seven intervention hospitals attended a 3-day instructor training session comprising 4 hours of didactic lessons, video scenarios, and interactive training covering team structure and processes, planning and problem solving, communication, workload management, team skills, and implementation. Conflict resolution strategies were included to provide a means of enhancing team behavior. Teamwork training also included assistance with creation and structure of teams at each intervention hospital. Trainers returned to their respective hospitals to conduct onsite training sessions for staff members from obstetrics, anesthesiology, and nursing and to structure each unit into core work teams made up of those nurses, physicians, and staff in direct contact with patients and coordinating teams composed of immediate supervisors, clinical leaders, and unit resource personnel. In addition, a contingency team, a multidisciplinary group of experienced physicians and nurses drawn from practitioners that are on call during a 24-hour period, were trained to respond in a coordinated way to obstetric emergencies. The group was also empowered to draw on additional hospital-wide resources. In all, 1,307 labor and delivery room personnel were trained. All staff training occurred after the baseline data collection period.

Because there were no universally accepted, validated measures, we developed outcome and process measures to capture all major adverse events and timing delays that were considered preventable by the teamwork intervention. The derivation of the measures is described in detail in a related publication.[9]

The maternal outcomes recorded were maternal death, uterine rupture, unplanned admission to intensive care, unplanned return to the labor and delivery unit or to the operating room, blood transfusion, and presence of a third- or fourth-degree laceration during vaginal delivery (Table 2). Poor clinical outcomes assessed in the fetus or neonate were intrapartum death of a fetus weighing at least 500 g at 24 weeks of gestation or greater, neonatal death of a baby with a birth weight of 2,500 g or more within 7 days of birth, neonatal birth trauma (Erb's palsy or a vacuum or forceps injury noted within 24 hours of birth) in a fetus of at least 20 weeks of gestation, unplanned admission of a term neonate (birth weight 2,500 g or more and gestational age 37 weeks or more) to neonatal intensive care (NICU) within 24 hours of birth for 24 hours or more, and a 5-minute Apgar score less than 7 in a neonate with a birth weight of 2,500 g or more. Antepartum intrauterine fetal deaths were excluded. Multiple gestations counted as a single delivery; an adverse outcome was counted if any one of the fetuses or neonates had that outcome. Because individual adverse events are rare, an index outcome measure, called the Adverse Outcome Index, was developed to capture the proportion of all deliveries with at least one undesirable outcome and to serve as the primary response variable. The Adverse Outcome Index was defined as the number of patients with one or more adverse outcomes divided by the total number of deliveries. A second, weighted, index outcome measure, the Weighted Adverse Outcome Score, was developed with the assistance of the American College of Obstetricians and Gynecologists (ACOG) Committee on Patient Safety and Quality Assurance. The weights were developed using an iterative process with input from members of the ACOG Committee and the study's Scientific Oversight Commit-

Table 2. Clinical Maternal and Neonatal Measures and Assigned Weights

Index Measures	Weights
Maternal death	750
Intrapartum or neonatal death (more than 2,500 g)	400
Uterine rupture	100
Maternal admission to ICU	65
Birth trauma (Erb's palsy, vacuum or forceps injury)	60
Return to operating room or labor and delivery unit	40
Admission to NICU (more than 2,500 g for more than 24 h)	35
Apgar score less than 7 at 5 min	25
Blood transfusion	20
Third- or fourth-degree perineal tear	5

ICU, intensive care unit; NICU, neonatal intensive care unit.

tee. The Weighted Adverse Outcome Score was used to assess not only the occurrence of deliveries with poor outcomes but also the number and relative severity of the outcomes. It was defined as the total weighted score of each adverse outcome divided by the total number of deliveries. The weights assigned to each outcome in the Weighted Adverse Outcome Score were selected to capture the relative severity of the outcome and are shown in Table 2. The Weighted Adverse Outcome Score assigned to a delivery was the sum of the weights for each adverse outcome that occurred during that delivery, or zero if no adverse outcomes occurred. Lastly, we defined the severity of the types of adverse outcomes for those patients with one or more adverse outcomes as the Severity Index. The Severity Index was the total weighted score of each adverse outcome divided by the total number of deliveries with one or more adverse outcomes.

Time or process measures were developed to record length of stay or delay to action. The eleven measures were the time elapsed from the time the patient was registered in labor and delivery to the time that the patient assessment was initiated by a provider; the time elapsed from registration to the time that the maternal-fetal assessment was initiated; the time elapsed from the decision for an immediate cesarean delivery to the time of the initial incision of the cesarean delivery; the time elapsed from the time of identification of the need for group B streptococcus antibiotics to the time of initial administration; the time elapsed from the decision for an urgent cesarean delivery to the time of the initial incision of the cesarean delivery; the time elapsed from the time the patient registered for her scheduled induction of labor to the time of first introduction of the induction agent; the time elapsed between the request for regional anesthesia during labor and the arrival of an anesthesia clinician to administer regional anesthesia; the time elapsed from registration to delivery for a nullipara; the time elapsed from time of registration to delivery for a multipara; the time elapsed from the planned start time for an elective cesarean delivery to time of the initial incision of the cesarean delivery; and the time elapsed from the delivery of the last baby to the time that postpartum care was initiated for the mother. These measures were surrogates for teamwork because they captured situations where multiple caregivers needed to coordinate complex interpersonal tasks to deliver safe and timely care.

For a priori power calculations, we assumed an average of 1,000 deliveries per hospital during a post-implementation data collection period of 4 months and a control-group prevalence of the Adverse Outcome Index of 2.6–10%. The prevalence estimates were based on outcome data collected before the study at the coordinating hospital. Sample size and power calculations were performed taking into account the fact that clusters (hospitals), not individual women, would be randomized. Specifically, because observations on individuals in the same hospital tend to be correlated, the effective sample size in a cluster-randomized trial is less than the total number of individual participants, so that the sample size for an individual-randomized trial must be multiplied by an inflation factor called the design effect.[10,11] The design effect increases with larger average cluster sizes (larger numbers of deliveries, n) and greater within-hospital correlation (as measured by the intracluster correlation coefficient, ICC). Specifically, the design effect equals $1+([n-1]\times ICC)$.[10] The intracluster correlation coefficient ranges between zero (indicating no within-hospital correlation) and one (indicating perfect within-hospital correlation). However, if the average cluster size is large, even small values of the intracluster correlation coefficient (0.01 or less) can lead to a very large design effect, in which case large numbers of clusters are required to achieve high power and precise (narrow) confidence intervals for differences in the Adverse Outcome Index. Because no data were available to estimate the intracluster correlation coefficient of the Adverse Outcome Index for use in power calculations, we calculated the decreases in the Adverse Outcome Index that could be detected with at least 80% power for intracluster correlation coefficient values between 0.001 and 0.04, based on intracluster correlation coefficients estimated from hospital-level observational data on mortality and mode of delivery in the United Kingdom and with various numbers of hospitals per arm.[7] With six to nine hospitals per arm, the study would have 80% power ($\alpha=0.05$, two-tailed) to detect a 40% decrease in the Adverse Outcome Index if the control-group Adverse Outcome Index prevalence were 10% and the intracluster correlation coefficient was 0.01 or less, or if the control-group prevalence was 2.6% and the intracluster correlation coefficient was 0.001 or less.

Because the primary goal of the intervention was to reduce the overall frequency of adverse outcomes and improve clinical processes and the number of hospitals randomized was small, all analyses were conducted at the cluster (hospital) level according to a prespecified written analysis plan.[10] All analyses were by intention to treat. The baseline characteristics of the hospitals and the patient populations, and the hospital-specific values of the outcome measures (Adverse Outcome Index, Weighted Adverse Outcome Score, Severity Index) and process measures during the baseline and post-implementation periods were

summarized using group means in the implementation and control arms and compared between arms using cluster-level *t* tests.[7,10] The primary analyses performed to assess the effectiveness of the intervention were cluster-level analyses of covariance,[10,12,13] which compared the post-implementation mean outcome measure or process measure between arms with adjustment for the baseline mean outcome measure or process measure, respectively, in that arm as well as the following covariates: baseline prevalences of preterm delivery and low birth weight (less than 2,500 g), hospital type, and annual number of deliveries. In secondary analyses, the analyses of covariance were also conducted separately for preterm and term deliveries (gestational age less than 37 and 37 or more weeks, respectively) and cluster-level analyses of change in the Adverse Outcome Index from baseline to post-implementation on the log-odds scale were performed. Intracluster correlation coefficients were calculated for the outcome measures and each process measure using data from the baseline period for all 15 hospitals.[7,10,12] The intracluster correlation coefficients were used to calculate approximate 95% confidence intervals for the differences in these measures between arms with adjustment for the clustering.[10] Statistical significance was defined as a $P<.05$ (two-sided). Data analyses were performed using SAS 8.2 software (SAS Institute, Cary, NC).

RESULTS

Data collection was completed for 94.4% of deliveries at control hospitals and 95.9% of deliveries at intervention sites (Table 3). For both study arms, on average more than 44% of the deliveries were to first-time mothers and approximately one quarter were by cesarean delivery. In the control arm, a mean of 11.7% and in the intervention arm 13.3% of the deliveries occurred in pregnancies of less than 37 weeks of gestation. The mean prevalence of low birth weight was less than 10% and of multiple births was slightly greater than 2% in both arms. None of these baseline characteristics differed significantly between the control and intervention arms.

There were no statistically significant differences between the intervention and control arms for any individual or index measure at baseline. The mean Adverse Outcome Index (range) was 9.4 (6.5–16.6) in the control arm and 9.0 (5.9–14.7) in the intervention arm. The most common maternal outcome was a third- or fourth-degree perineal laceration after vaginal delivery: mean 5.0% (range 1.3–10.0) and 4.5% (range 3.1–5.4) of all deliveries in the control and intervention arms, respectively. The most common neonatal outcome was unplanned admission to the NICU: mean 4.5% (range 0–19.2) and 4.1% (range 0.2–10.0), respectively. These measures showed the greatest variability, with wide ranges across hospitals, and were more prevalent than all of the other outcome measures combined. The mean times elapsed and ranges for the baseline clinical process measures also showed wide variability. There were no statistically significant differences in the average elapsed times between the intervention and control groups (baseline data not shown).

No statistically significant differences between the arms were observed for the post-implementation outcome measures, even when any differences in baseline levels were controlled (Table 4). The adjusted post-implementation mean Adverse Outcome Index for the control arm was 7.2% and for the intervention

Table 3. Delivery and Demographic Characteristics by Study Group

Characteristic*	Control (n=8)	Intervention (n=7)
Total deliveries	14,336	14,200
Baseline data collection	3,779	3,894
Post-implementation data collection	10,557	10,306
Source of project funding		
Military (6 sites)	3,429	4,130
External funding (5 sites)	6,230	3,564
No external funding (4 sites)	4,677	6,506
Nullipara (%)	44.8 (40.2–54.1)	44.2 (39.3–48.8)
Cesarean deliveries (%)	24.8 (11.9–41.2)	28.2 (23.9–40.8)
History of prior cesarean delivery (%)	11.8 (5.6–19.4)	13.3 (11.5–16.7)
Gestational age less than 37 wk (%)	11.7 (5.6–18.7)	13.3 (9.6–27.4)
Birth weight less than 2,500 g (%)	8.6 (3.9–14.6)	9.7 (5.2–24.7)
Multiple births (%)	2.2 (0.8–3.4)	2.5 (1.0–3.5)

Data are expressed as number or mean (range).
* None of these baseline characteristics differed significantly between groups.

Table 4. Comparison of the Mean Adverse Outcome Index, Weighted Adverse Outcome Score, Severity Index, and Process Measures Between Control and Intervention Groups for the Post-Implementation Data Collection Period, With Adjustment for the Baseline Mean

Measure	Adjusted Mean*		Intracluster Correlation Coefficient	Approximate 95% Confidence Interval for the Difference Between Groups†
	Control Group	Intervention Group		
Adverse Outcome Index (%)	7.2	8.3	0.015	–5.6 to 3.2
Weighted Adverse Outcome Score	2.3	2.7	0.008	–3.4 to 1.4
Severity Index	30.6	31.9	0.017	–23.0 to 7.0
Process measures (time elapsed)‡				
Registration to provider assessment	1.0	1.1	0.268	–0.8 to 1.4
Registration to maternal-fetal assessment (min)	14.9	17.8	0.031	–17.5 to 17.6
Scheduled registration to induction	3.3	3.3	0.028	–2.7 to 0.9
GBS antibiotic order to first dose (min)	42.5	42.9	0.015	–22.3 to 14.6
Epidural request to initiation (min)	33.1	32.5	0.036	–9.6 to 11.1
Scheduled cesarean delivery start time to incision	2.0	2.0	0.203	–1.6 to 1.4
Immediate cesarean delivery decision to incision (min)	33.3§	21.2§	0.039	–36.9 to –0.7
Urgent cesarean section decision to incision (min)	65.8	77.0	0.034	–43.7 to 17.7
Registration to delivery – nullipara	14.4	13.8	0.042	–5.2 to 1.1
Registration to delivery – multipara	8.1	8.3	0.021	–3.4 to 0.9
Delivery to end of care in labor and delivery	3.4	3.3	0.141	–2.1 to 1.1

GBS, group B streptococcus.
* Adjusted for baseline levels using cluster-level analysis of covariance.
† Calculated using the intracluster correlation coefficient to adjust for clustering, as described by Donner and Klar.[10]
‡ In hours, except where otherwise indicated.
§ P=.03.

arm was 8.3% (P=.30). The range of the Adverse Outcome Index across the sites was 4.1–16.5% at the control hospitals and 4.7–12.6% at the intervention hospitals. The intracluster correlation coefficient was 0.015, resulting in a wide confidence interval (CI) for the difference between arms. When third- or fourth-degree vaginal lacerations, considered a minor outcome, were removed from the Adverse Outcome Index, the adjusted mean prevalence of adverse outcomes was 5.2% in the control arm and 4.8% in the intervention arm (P=.64; intracluster correlation coefficient 0.034, 95% CI for difference in means –8.0% to 2.5%). We repeated the analysis excluding both third- or fourth-degree vaginal lacerations and unplanned admission to the NICU because these two measures demonstrated the greatest variability. The resulting adjusted mean Adverse Outcome Index was 2.2% in the control arm and 2.3% in the intervention arm (P=.77; intracluster correlation coefficient 0.003, 95% CI for difference in means –1.2% to 1.4%). In analyses of change from baseline to post-implementation on the log-odds scale, there were no statistically significant differences between arms for any of these three versions of the Adverse Outcome Index (odds ratios 1.2, 1.1, and 1.1, and P=.18, .83, and .64, respectively). The adjusted mean Weighted Adverse Outcome Score was 2.3 and 2.7 for the control and intervention arms, respectively (P=.50). Among deliveries with one or more adverse outcomes, the adjusted mean Severity Index scores were 30.6 and 31.9 for the control and intervention arms, respectively (P=.80).

The length of time elapsed for the 11 process measures capturing the flow of activities related to optimal labor and delivery care was compared during the post-implementation period, controlling for the observed baseline values (Table 4). Only the mean time elapsed between the decision to perform an emergency cesarean delivery and the time of the incision differed significantly between the arms (33.3 minutes for the control arm compared with 21.2 minutes for the intervention arm, P=.03). The intracluster correlation coefficients for the process measures ranged from 0.015 to 0.268, leading to wide confidence intervals for the differences between arms.

Twelve percent of deliveries in each arm were in pregnancies of less than 37 weeks of gestation (1,750 deliveries in the control arm and 1,758 in the intervention arm). The adjusted mean post-implementation index outcome measures, excluding third- or fourth-degree lacerations and unplanned NICU admissions, for term deliveries of 37 weeks or more were 1.8% in the control arm and 2.1% in the intervention arm (P=.49); and for premature deliver-

ies (defined as those less 37 weeks of gestation) were 4.5% and 4.8% (P=.82).

DISCUSSION

We observed no statistically significant differences between the intervention and control groups for the Adverse Outcome Index, weighted index, any of the individual outcomes, or 10 of the 11 process measures. The results did not change after adjustment for the baseline values of the measures, the rate of preterm delivery or low birth weight, and the two factors used to balance the randomization, or when analyses were conducted separately for deliveries at less than 37 or at 37 weeks or more of gestation.

The time from the decision to the incision for an immediate cesarean delivery was significantly shorter in the intervention group (P=.03). This result must be interpreted in light of the fact that we conducted a number of statistical significance tests, which could give rise to a chance observation being accorded statistical significance. However, performing an immediate cesarean delivery 12 minutes sooner (as in the intervention arm) could have significant patient safety implications. The improvement may have resulted from the use of a contingency team. As part of the teamwork structure, all hospitals developed contingency teams consisting of the obstetric and anesthesia attending physicians, chief resident, labor nurse, and surgical scrub nurse. The contingency team, one of the changes in culture that is the easiest and fasted to implement, may be viewed as comparable to a code or disaster drill team and may be an effective organizational change in the labor and delivery environment.

There are a number of possible explanations for why this study of teamwork intervention did not demonstrate a significant impact on study defined measures. Foremost, of course, the training may simply not have been effective. However, a number of other explanations need to be considered. For logistical reasons, we had only 120 days for training and implementation and 150 days to observe the outcome and process measures post-implementation. Teamwork that results in a detectable impact may require more than a 4-hour training session and more than 4 months to practice behaviors regularly. Many other variables may need to be considered to effect lasting change. Subsequent to the completion of the cluster-randomized trial, our experience after implementation of these changes at the coordinating institution, Beth Israel Deaconess Medical Center, has indicated that it may take 9–12 months before a significant decline in the Adverse Outcome Index occurs.

Compared with the baseline Adverse Outcome Index, the post-intervention Adverse Outcome Index decreased in both the control and intervention groups (21% and 10%, respectively), although the magnitude of the decrease did not differ significantly between the two groups. Similar declines were seen for those high-risk patients delivered at less than 37 weeks of gestation. The decline in the Adverse Outcome Index in both groups could have reduced our ability to observe a positive intervention effect. Possible explanations for this finding include incomplete ascertainment of outcomes, the influence of collecting data (ie, a Hawthorne effect), other quality control initiatives and procedural changes made at the institutions to improve care, or random variation.

Because there are no national measures of error in obstetrics, we developed the Adverse Outcome Index and used process measures as surrogates for teamwork behaviors. It is possible that the measures chosen do not capture teamwork behavior or medical errors in obstetrics. Published data suggest that crew resource management principles in medicine improve the individual's attitude that the training will reduce errors in their practice.[14,15] However, the precise impact on improvements in safety is uncertain, even in aviation.[16] Proving that crew resource management principles improve safety will be challenging in medicine without the development of better evaluation tools. Since this study, teamwork competencies (knowledge, skills, and attitudes) identified from various high-risk industries have been defined as an approach to measure and assess team performance within health care, specifically within the professional education of physicians and nurses.[17]

Power to detect important intervention effects may have been lacking. At the time we planned the study, we determined that, with the anticipated Adverse Outcome Index prevalence and number of hospitals, the intracluster correlation coefficient would need to be 0.01 or smaller for the study to have good statistical power to detect a 40% decrease in the Adverse Outcome Index, but no data were available to estimate the intracluster correlation coefficient for our outcome measures and setting. The intracluster correlation coefficient calculated using the baseline data from the study was 0.015, rather than 0.01, resulting in more variability than expected and 95% confidence limits for the difference in the Adverse Outcome Index between the intervention and control groups that are consistent with a result between a large positive effect and a large negative effect. With knowledge of the intracluster correlation coefficient and Adverse Outcome Index observed in this study,

future cluster-randomized trials will require a minimum of 11–13 hospitals per arm to achieve 80% power to detect a 40% reduction. Of note, a recently published cluster-randomized controlled trial to evaluate medical emergency teams also failed to show a difference.[18] Our studies were very similar in design and both, in the end, included too few hospitals.

It is interesting to note that approximately 8% of deliveries in each arm resulted in at least one adverse outcome. A surprising finding is the wide range of the Adverse Outcome Index across the hospitals in both groups, from a low of 4% to a high of 16.5%. These results demonstrate new information about the prevalence of adverse clinical outcomes in obstetrics using a standard set of measures. The outcome measures and Adverse Outcome Index may be useful tools for comparing obstetric outcomes within and between institutions and to help guide quality improvement processes. In addition, the application of the Weighted Adverse Outcome Score provides an assessment of how severe the adverse events are, not just how frequently they occur. In this study we highlight the need for the implementation and evaluation of teamwork training programs in obstetrics and suggest a set of uniformly defined and collected outcome and process measures that will provide a foundation for future trials as we search for effective strategies to improve the safety of obstetric care.

REFERENCES

1. Kohn LT, Corrigan JM, Donaldson MS, editors; Institute of Medicine. To err is human: building a safer health system. Washington, DC: National Academy Press; 2000.
2. Committee on Quality Health Care in America. Institute of Medicine. Crossing the quality chasm: a new health system for the 21st century. Washington, DC: National Academy Press; 2001.
3. Helmreich RL, Merritt AC, Wilhelm JA. The evolution of crew resource management in training in commercial aviation. Int J Aviat Psychol 1999;9:19–32.
4. Morey JC, Simon R, Jay GD, Wears RL, Salisbury M, Dukes KA, et al. Error reduction and performance improvement in the emergency department through formal teamwork training: evaluation results of the MedTeams project. Health Serv Res 2002;37:1553–81.
5. Locke A, Evangelista J, Langford V, Morey J, Risser D, Simon R, et al. MedTeams Emergency Team Coordination Course. Andover (MA): Dynamics Research Corporation; 2001.
6. Moulton LH. Covariate-based constrained randomization of group-randomized trials. Clin Trials 2004;1:297–305.
7. Ukoumunne OC, Gulliford MC, Chinn S, Sterne JA, Burney PG. Methods for evaluating area-wide and organization-based interventions in health and health care: a systematic review. Health Technol Assess 1999;3:iii–92.
8. Sachs BP. A 38-year-old woman with fetal loss and hysterectomy. JAMA 2005;294:833–40.
9. Mann S, Pratt S, Gluck P, Nielsen P, Risser D, Greenberg P, et al. Assessing quality in obstetrical care: development of standardized measures. Jt Comm J Qual Patient Saf 2006;32:497–505.
10. Donner A, Klar N. Design and analysis of cluster randomized trials in health research. London: Hodder Arnold; 2000.
11. Campbell MK, Elbourne DR, Altman DG; CONSORT Group. CONSORT statement: extension to cluster randomized trials. BMJ 2004;328:702–8.
12. Ukoumunne OC, Thompson SG. Analysis of cluster randomized trials with repeated cross-sectional binary measurements. Stat Med 2001;20:417–33.
13. Fleiss JL. Statistical methods for rates and proportions. 2nd ed. New York (NY): John Wiley & Sons; 1981.
14. Barrett J, Gifford C, Morey J, Risser D, Salisbury M. Enhancing patient safety through teamwork training. J Healthcare Risk Manag 2001;21:57–65.
15. Grogan EL, Stiles RA, France DJ, Speroff T, Morris JA, Nixon B, et al. The impact of aviation-based teamwork training on the attitudes of health-care professionals. J Am Coll Surg 2004;199:843–8.
16. Salas E, Burke CS, Bowers CA, Wilson KA. Team training in the skies: does crew resource management (CRM) training work? Hum Factors 2001;43:641–74.
17. Baker D, Salas E, King H, Battles J, Barach P. The role of teamwork in the professional education of physicians: current status and assessment recommendations. Jt Comm J Qual Patient Saf 2005;31:185–202.
18. Hillman K, Chen J, Cretikos M, Bellomo R, Brown D, Doig G, et al; for the MERIT study investigators. Introduction of the medical emergency team (MET) system: a cluster-randomised controlled trial. Lancet 2005;356:2091–7.

APPENDIX

Members of the MedTeams Labor and Delivery Consortium

Lauren Bales, MD, Naval Medical Center Camp Pendleton; David J. Birnbach, MD, Jackson Memorial Hospital, Miami; Ronald Burkman, MD, Baystate Medical Center; Cynthia Brumfield, MD, University of Alabama at Birmingham Hospital; Peter Cherouny, MD, University of Vermont-Fletcher Allen Health Care; Jack Cooley, MD, National Naval Medical Center; Harold Fox, MD, Johns Hopkins Medical Center; Elizabeth Golladay, MD, Tripler Army Medical Center; Lynn Leventis, MD, Naval Medical Center of San Diego; Robert Lorenz, MD, William Beaumont Hospital; William Lucky, MD, Baptist Hospital of Miami; Patrick Nugent, MD, South Shore Hospital; Spike Lipschitz, MD, South Shore Hospital; Chris Stolle, MD, Naval Medical Center of Portsmouth; Cosmas van De Ven, MD, University of Michigan Medical Center; Frank Witter, MD, Johns Hopkins Medical Center.

The trial's Scientific Oversight Committee comprised the study authors, members of the consortium, and Stanley Zinberg, MD, ACOG, Lucian Leape, MD, Harvard School of Public Health, Molly McCarthy and Karen Peddicord, AWHONN, and Tom Dimino, MD, Baptist Health Hospital South Florida.