HHS Public Access

Author manuscript

*J Am Geriatr Soc*. Author manuscript; available in PMC 2016 December 14.

Published in final edited form as:

*J Am Geriatr Soc*. 2015 December ; 63(12): 2463–2471. doi:10.1111/jgs.13793.

# The Successful Aging after Elective Surgery (SAGES) Study: Cohort Description and Data Quality Procedures

**Eva M. Schmitt, Ph.D.**[#a], **Jane S. Saczynski, Ph.D**[#b], **Cyrus M. Kosar, M.A**[#a], **Richard N. Jones, Sc.D**[c], **David C. Alsop, Ph.D**[d], **Tamara G. Fong, M.D., Ph.D**[a,f], **Eran Metzger, M.D**[g,h], **Zara Cooper, M.D., MSc**[i], **Edward R. Marcantonio, M.D., S.M**[#e], **Thomas Travison, Ph.D**[#a], **Sharon K. Inouye, M.D., M.P.H**[#a,e], and **the SAGES Study Group**

[a]Aging Brain Center, Institute for Aging Research, Hebrew SeniorLife, Boston, MA

[b]University of Massachusetts Medical Center and Meyers Primary Care Institute, Worcester, MA

[c]Department of Psychiatry and Human Behavior, Warren Alpert Medical School, Brown University, Providence, RI

[d]Department of Radiology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA

[e]Department of Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA

[f]Department of Neurology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA

[g]Department of Psychiatry, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA

[h]Department of Psychiatry, Hebrew SeniorLife, Boston, MA

[i]Department of Surgery, Brigham and Women's Hospital, Boston, MA

[#] These authors contributed equally to this work.

## Abstract

**Background/Objectives—**Delirium is the most common complication of major elective surgery in older patients. The Successful Aging after Elective Surgery (SAGES) study was designed to examine novel risk factors and long-term outcomes associated with delirium. This report describes the cohort, quality assurance procedures, and results.

**Design—**Long-term prospective cohort study.

**Setting—**Three academic medical centers.

**Participants—**A total of 566 patients age 70 and older without recognized dementia scheduled for elective major surgery.

**Measurements—**Participants were assessed preoperatively, daily during hospitalization, and at variable monthly intervals for up to 36 months post-discharge. Delirium was assessed in hospital by trained study staff. Study outcomes included cognitive and physical function. Novel risk factors

Internal Advisory Board: Gary Gottlieb, MD, MBA (BWH, MGH, HMS), John Orav, PhD (BWH, HMS), Reisa Sperling, MD, MMSc (BWH, HMS).

**Conflict of Interest:** Dr. Alsop receives institutional royalties from GE Healthcare for patents related to the arterial spin labeling MRI technique used for this study. None of the other authors report any conflicts of interest. All other co-authors fully disclose they have no financial interests, activities, relationships and affiliations. The co-authors also declare they have no potential conflicts in the three years prior to submission of this manuscript.

for delirium were assessed including genetic and plasma biomarkers, neuroimaging markers, and

| Elements of Financial/Personal Conflicts | Eva M. Schmitt* | | Jane S. Saczynski | | Cyrus M. Kosar | | Richard N. Jones | |
|---|---|---|---|---|---|---|---|---|
| | Yes | No | Yes | No | Yes | No | Yes | No |
| Employment or Affiliation | | X | | X | | X | | X |
| Grants/Funds | | X | | X | | X | | X |
| Honoraria | | X | | X | | X | | X |
| Speaker Forum | | X | | X | | X | | X |
| Consultant | | X | | X | | X | | X |
| Stocks | | X | | X | | X | | X |
| Royalties | | X | | X | | X | | X |
| Expert Testimony | | X | | X | | X | | X |
| Board Member | | X | | X | | X | | X |
| Patents | | X | | X | | X | | X |
| Personal Relationship | | X | | X | | X | | X |

| Elements of Financial/Personal Conflicts | David C. Alsop | | Tamara G. Fong | | Eran Metzger | | Zara Cooper | |
|---|---|---|---|---|---|---|---|---|
| | Yes | No | Yes | No | Yes | No | Yes | No |
| Employment or Affiliation | | X | | X | | X | | X |
| Grants/Funds | | X | | X | | X | | X |
| Honoraria | | X | | X | | X | | X |
| Speaker Forum | | X | | X | | X | | X |
| Consultant | | X | | X | | X | | X |
| Stocks | | X | | X | | X | | X |
| Royalties | X | | | X | | X | | X |
| Expert Testimony | | X | | X | | X | | X |
| Board Member | | X | | X | | X | | X |
| Patents | | X | | X | | X | | X |
| Personal Relationship | | X | | X | | X | | X |

| Elements of Financial/Personal Conflicts | Edward R. Marcantonio | | Thomas Travison | | Sharon K. Inouye | | | |
|---|---|---|---|---|---|---|---|---|
| | Yes | No | Yes | No | Yes | No | | |
| Employment or Affiliation | | X | | X | | X | | |
| Grants/Funds | | X | | X | | X | | |
| Honoraria | | X | | X | | X | | |
| Speaker Forum | | X | | X | | X | | |
| Consultant | | X | | X | | X | | |
| Stocks | | X | | X | | X | | |
| Royalties | | X | | X | | X | | |
| Expert Testimony | | X | | X | | X | | |
| Board Member | | X | | X | | X | | |
| Patents | | X | | X | | X | | |
| Personal Relationship | | X | | X | | X | | |

* Corresponding Author

cognitive reserve markers. Interrater reliability (kappa and weighted kappa) was assessed for key variables in 119 of the patient interviews.

**Results—**Participants were an average of 77 years old and 58% were female. The majority of patients (81%) were undergoing orthopedic surgery and 24% developed delirium post-operatively. Over 95% of eligible patients were followed for 18 months. There was >99% capture of key study outcomes (cognitive and functional status) at every study interview and interrater reliability was high (weighted kappas for delirium = 0.92 and for overall cognitive and functional outcomes = 0.94 -1.0). Completion rates for plasma biomarkers (4 timepoints) were 95%-99% and for neuroimaging (one year follow-up) was 86%.

**Conclusion—**The SAGES study will contribute to the understanding of novel risk factors, pathophysiology and long-term outcomes of delirium. This manuscript describes the cohort and data quality procedures, and will serve as a reference source for future studies based on SAGES.

### Keywords

Delirium; Surgical Outcome; Data Quality; Longitudinal Study

## INTRODUCTION

Delirium, an acute change in attention and cognition, is the leading postoperative complication in older persons, affecting up to half of older patients following surgery. The development of delirium is associated with increased mortality, prolonged functional and cognitive impairment and substantial health care costs, estimated at over $164 billion per year in the U.S.[1-8] Delirium is an important patient safety issue and is increasingly considered as a healthcare quality indicator in older hospitalized patients.[9, 10] Moreover, delirium is likely to be a major contributor to the 'post-hospital syndrome', characterized by prolonged functional impairment following hospitalization and often unrelated to the admitting diagnosis.[11] The multifactorial contributors to delirium are often preventable, underscoring the importance of interdisciplinary prospective studies to further elucidate the risk factors, mechanisms, and outcomes of delirium.[12]

The Successful Aging after Elective Surgery (SAGES) study is an innovative, 5-year, prospective program project, funded by the National Institute on Aging, aimed to elucidate novel risk factors (including biomarkers, neuroimaging, and reserve markers) and to examine the contribution of delirium to long-term cognitive and functional decline.[13] State-of-the-art measurement of delirium and its risk factors and outcomes utilizing quality data collection and management procedures makes the SAGES project uniquely poised to contribute to our understanding of the causes and consequences of delirium.

During the early stages of the study, we published the design and methods[13] along with several methodological reports validating our key exposure and outcome variables.[14-18] We have now completed enrollment and follow-up to 18 months. The purpose of this manuscript is to provide the baseline characteristics of the SAGES cohort, to describe the procedures used to optimize data collection, and to lay the foundation for future work.

# METHODS

## Study Population

The SAGES study is an ongoing prospective cohort study of older adults undergoing elective major non-cardiac surgery. The study design and methods have been described previously.[13] Briefly, eligible participants were age 70 years and older, English speaking, scheduled to undergo elective surgery at one of two Harvard-affiliated academic medical centers and with an anticipated length of stay of at least 3 days. Eligible surgical procedures were: total hip or knee replacement, lumbar, cervical, or sacral laminectomy, lower extremity arterial bypass surgery, open abdominal aortic aneurysm repair, and colectomy. Exclusion criteria included evidence of dementia, delirium, hospitalization within 3 months, terminal condition, legal blindness, severe deafness, history of schizophrenia or psychosis, and history of alcohol abuse. A total of 566 patients were enrolled between June 18, 2010 and August 8, 2013. Written informed consent was obtained from all participants according to procedures approved by the institutional review boards of Beth Israel Deaconess Medical Center and Brigham and Women's Hospital, the two study hospitals, and Hebrew SeniorLife, the study coordinating center, all located in Boston, Massachusetts.

## Recruitment and Follow-up

Participant identification, eligibility screening, informed consent, baseline interviews, phlebotomy, and imaging studies were performed on average 2 weeks (mean = 13 ± 15 days) prior to the index surgery. Eligibility was first determined using medical record review, and confirmed by the SAGES team, with formal adjudication by physician investigators if required. Patients were interviewed daily for delirium during the index hospitalization and followed at intervals for at least 18 months and up to 36 months after discharge.

Of 1,052 patients screened for eligibility, 318 (30%) declined to be interviewed. Thus, 734 patients were assessed for eligibility, of which 163 were ineligible and 5 were eligible but refused participation resulting in a total of 566 patients enrolled (**Figure 1**). The response rate (the percent of the estimated number eligible who were enrolled) was 70% (Appendix A) which is comparable to that observed in many important observational studies.[19] Patients who refused eligibility screening were more likely to be female and less likely to be undergoing orthopedic surgery than patients who agreed to be screened (Appendix Table 1).

## Data Collection Protocol

**Baseline Interviews—**A 90-minute baseline interview, conducted before the index surgery in participants' homes, included detailed data collection on health and functioning. Specific assessments have been described previously.[13] Key outcomes and other study variables are reviewed below.

**Daily In-hospital Delirium Assessment—**Delirium was assessed by trained research assessors once daily during the day shift. The Confusion Assessment Method (CAM)[20] was rated based on information from patient interviews including a brief cognitive screen (orientation, short-term recall, sustained attention) described previously,[13] the Delirium

Symptom Interview (DSI),[21] and information related to acute changes in mental status noted by nurses or family members. The CAM is a standardized method for delirium identification with high interrater reliability and sensitivity (94%) and specificity (89%) when compared to clinical expert or consensus-based diagnoses of delirium.[22] Delirium severity was assessed using the CAM-S,[15] a validated severity measure based on the CAM that demonstrates strong psychometric properties and strong associations with important clinical outcomes.

**Index Hospitalization Medical Record Review—**After discharge from the index hospitalization, medical records were reviewed by study clinicians to collect information on surgical procedure, anesthesia type and duration, abnormal laboratory results, baseline diagnoses, development of delirium, precipitating factors for delirium (e.g., medications, iatrogenic events, or catheters), postoperative complications and death. Chart abstraction data were randomly checked for illogical values and against data collected as part of the screening process (e.g., surgery type).

A standardized chart review method for identification of delirium,[14] was used to supplement the CAM interviews. The chart-based delirium diagnosis abstracted information on acute changes in mental status, time and duration of such episodes, evidence of agitation and reversibility or improvement of the acute confusion from in-patient and preoperative notes, discharge summaries, and outpatient visit notes. All chart-based delirium cases were adjudicated by a delirium expert panel and discrepancies were resolved during consensus conferences.

**Follow-up Interviews—**Home-based interviews were conducted at 1 and 2 months after discharge and then every 6 months up to 36 months. Interviews included assessments of cognitive function, using a neuropsychological test battery (detailed below), and physical function, using the Activities of Daily Living (ADL), Instrumental Activities of Daily Living (IADL) and Short Form 12 (SF-12). Brief phone interviews were conducted with participants at 4, 9, 15, 21 and 27 months after the index hospitalization. Telephone follow-up interviews included a cognitive screen and delirium rating, ADL and IADL assessment, and healthcare utilization since the last interview.

**Subsequent Hospitalizations—**Information on rehospitalizations was obtained during follow-up interviews with patients and family members. Charts for rehospitalizations were abstracted for delirium and intercurrent illnesses.

## Key Study Variables

**Delirium—**Delirium incidence was defined as presence of delirium according to the CAM criteria during one or more days of hospital interviews or by the chart review method at any time during the hospitalization. Delirium severity was measured using the CAM-S severity score (0-19, 19 most severe) scored from the daily 10-item CAM assessments.

**Cognitive and Physical Function—**Cognitive function was assessed by a neuropsychological test battery, described previously.[13] Global cognitive function was assessed using a composite variable created for the study, the General Cognitive Performance (GCP) composite (scaled, 0-100) which synthesized the neuropsychological

test results (general population mean = 50, standard deviation =10), and has been demonstrated to be highly sensitive to cognitive change.[16] Physical functioning was assessed by the ADL[23] and IADL,[24] and the physical function subscore of the Short Form Health Survey (SF-12).[25] A composite physical functioning variable based on these 3 measures was created for the SAGES study, and demonstrated to have strong predictive criterion validity with higher scores on the composite associated with lower risk of discharge to a rehabilitation facility and shorter hospital stays.[17] These outcomes were assessed at baseline, 1 month, 2 months, 6 months and every 6 months thereafter.

**Other Study Measures—**Race and ethnicity were self-reported by participants. Depression was assessed using the short form of the Geriatric Depression Scale (GDS-15) at baseline, 6 months, and every 6 months thereafter.[26] Scores range from 0 to 15, with higher scores reflecting higher severity of depressive symptoms. From the baseline assessment, vision impairment was defined as corrected binocular near vision worse than 20/70 on the Jaeger vision test[27] and hearing impairment was defined as hearing six or fewer of 12 numbers out of both ears on the Whisper test.[28] The hospital medical record was reviewed to obtain preoperative comorbidity burden measured by the Charlson Comorbidity Index,[29] and immediate postoperative severity of illness measured by the Acute Physiology and Chronic Health Evaluation II (APACHE II).[30] Participants' education, occupation, income, participation in cognitively stimulating activities at various points in the life course (18 years, 40 years, etc.) and parents' education, were collected from participants to operationalize measures of cognitive reserve.

Blood was collected at baseline (prior to surgery), immediately after surgery in the post-anesthesia care unit, on postoperative day 2, and 1-month after surgery. Blood was collected in heparinized (grey top) tubes, placed immediately in ice and transported to the Clinical Research Center, where it was processed within 4 hours of collection. Tubes underwent low speed centrifugation to separate out cellular and plasma components. The cellular component from the baseline time point was used for DNA extraction via the whole blood method.[31] Plasma was aliquoted into 15 tubes (baseline) or 10 tubes (follow-up) and stored at −80°C to create a SAGES biorepository.

Approximately 25% (n = 146) of enrolled participants elected to undergo brain Magnetic Resonance Imaging (MRI) prior to surgery and 1 year later. Participants who agreed to MRI imaging did not differ significantly from those who did not agree (Appendix Table 2). Exclusion criteria for the nested cohort MRI study included contraindications to 3 Tesla MRI, such as pacemakers, stents, and implants. Participants were imaged on a 3T HDxt MRI (General Electric Healthcare, Waukesha, WI) scanner using a standard eight-channel head coil. The MRI protocol included: Coronal T1, Fluid Attenuated Inversion Recovery (FLAIR), High Resolution 3D T2-weighted imaging, Arterial Spin Labeling (ASL) perfusion, and Diffusion Tensor Imaging (DTI).

## Interrater Reliability Assessments

A total of 119 interrater assessments of the in-person interviews were conducted semi-annually. During interrater sessions, two interviewers observed the patient simultaneously

but rated the responses independently and were blinded to each other's ratings. All key study variables underwent interrater reliability assessments, including the neuropsychological exam, functional measures (ADL, IADL, SF-12), and delirium. Interrater assessments of medical record abstractions were conducted on a 10% subsample (approximately 60 participants), which involved abstraction of the medical record by two independent raters. For both interview and medical record data, the data stored for future analyses came from one rater that was preselected before any comparison of results.

## Data Management and Quality

Data quality procedures were implemented to address four dimensions of quality: completeness (lack of missing data); accuracy (freedom from error); consistency (stability of definitions across databases); and timeliness (sensitivity to temporal change), and were applied to all data collection and management activities.[32, 33] To optimize accuracy and timeliness, interview staff underwent training including didactics, post-tests, standardization procedures, practice with volunteers, and shadowing of experienced personnel. Completeness and consistency of data capture was assured through database programming and continual cross-checks. Study organizational procedures, including ‘Cores’ responsible for each specific element of data collection (e.g., patient interviews, MRI, phlebotomy)[13] and team meetings of all Cores resulted in consistent data capture and quality across all measures. The procedures for each data quality dimension across each study activity are described in **Table 1.** Activities at several stages of data collection and management are described below.

**Participant interviews and medical record abstraction—**Reports detailing upcoming interviews were produced weekly. Data collection forms had unique identifiers to prevent misassignment. Submitted forms were checked by independent interviewers to ensure completeness and correct assignment. The data collection team met weekly with study investigators to discuss coding and scoring questions; decisions were recorded in a field operations manual. Interrater reliability assessment was performed semiannually (**Table 1**).

**Blood collection—**Phlebotomy tubes were coded with unique identifiers and tracked using the study data management system. Data were regularly cross-checked for alignment with expected dates, and consistent recording of volumes and processing times.

**Neuroimaging—**Neuroimaging scans were coded with unique identifiers maintained in a separate database and continually cross-checked with the main SAGES database to ensure consistency. Visual assessment of scan quality was conducted in real time and problems were carefully documented.

**Data management—**Double data entry was performed for paper-based data collection. A codebook with a formal definition of derived variables, description of missing records, chronological records of changes to definitions and coding, and a catalog of ongoing and published analyses relying on each derived metric was regularly updated. A “frozen”

compendium of all validated data was created semiannually, and an audit trail of database changes was maintained.

Electronic data were captured with REDCap. Real-time warning messages were triggered when data elements on interviews were left empty. Missing data were queried and reported to staff at weekly meetings, ensuring completeness of the data. Data accuracy was achieved through field validation. Embedded rules, including acceptable values or ranges, prevented erroneous entries.

### Statistical Analysis

To describe the cohort, standard descriptive statistics were utilized, including means, standard deviations, proportions and percentages. For interrater reliability assessments, we calculated percent agreement, kappa, and weighted kappa for all interview variables that were assessed (N=119) at item and summary score levels. The kappa statistic is a robust measure of agreement since it assesses agreement occurring beyond chance alone. The weighted kappa statistic is particularly useful when scores are ordered (e.g., more than a dichotomous response) and allows for disagreements to be weighted by degree of disagreement.[34]

For CAM ratings of delirium, reliability was assessed for the overall rating, and for the 10 individual CAM features. For the overall rating, agreement required exact agreement for the presence or absence of acute change and overall delirium (yes/no). For individual CAM features, agreement was required on the exact level (not present, mild, marked) for each of the 10 features of delirium (e.g., disorganized thinking, altered level of consciousness, etc.). For the neuropsychological test battery, agreement was examined for each individual test. For the Hopkins Verbal Learning Test, agreement for immediate recall required the two raters record the same total number of recalled words over the three trials. For the Visual Search and Attention Task, the two scores were added and agreement on total scores was compared across raters. For Trails A and B, time in seconds was compared and exact agreement was required. For verbal fluency, the total number of words generated for the letters F, A and S were compared and exact agreement was required. For digit span, scores for forward and backwards were added and compared. For the digit symbol substitution test, the category fluency test and the Boston Naming Test, exact agreement on the total correct score was required. For ADLs and IADLs, exact agreement was required for level of dependency (no help needed, help needed, completely unable to perform task) for each task.

## RESULTS

Participants were an average of 77 years old (standard deviation = 5.2 years), 58% were female and 93% were white (**Table 2**). The sample was highly educated, with only 30% having a high school education or less. The majority (81%) of participants underwent orthopedic surgery. At baseline (prior to surgery), 8% of participants reported dependencies in performing ADL, 30% reported dependencies in IADL, and 33% were hearing impaired. The cohort had high cognitive function with an average score on the General Cognitive Composite (GCP) of 58, indicating that the global cognitive function of the cohort, on average, was nearly one standard deviation above the age-matched general U.S. population.

Postoperative delirium occurred in 21% of the cohort (n=116) based on the CAM ratings, in 10% (n=57) based on the chart-based ratings and in 24% (n=135) based on the combined CAM and chart-based information.[14] The distribution of missing data at baseline are provided in Table 2.

### Follow-up interviews

Since the study is ongoing, the remainder of this report will focus on the 18-month follow-up, which are completed. In-person or phone interviews occurred 8 times over 18 months following hospital discharge. The dropout rate was less than 1% at each follow-up time point. Participants who dropped out cited time commitment, declining health and memory, or family member concerns with participation as reasons for discontinuation. Between 95% -98% of eligible patients were interviewed at each follow-up point and over 99% of the interviews have complete data on primary outcomes (**Table 3**). The blood collection rate was between 95% - 99% across the four time points and neuroimaging was completed on 86% of the participants at one-year follow-up (**Appendix Table 3**). The MRI sequence completion rate was between 90% - 100% at baseline and between 85% - 100% at the one-year follow-up (**Appendix Table 4**). The variables with the highest percentage of missing data were education level of participants' father and mother (12% and 9%, respectively), family income (9%) and occupation (6%) (**Appendix Table 5**). All other study variables were missing in <2% of participants, and the majority were missing in <1%.

### Interrater Reliability

Approximately 20 paired interviews were completed per year. Interrater reliability was calculated on individual items and summary scores for the main study variables (**Table 4**). Percent agreement, kappa and weighted kappa were high for all variables collected. Weighted kappa was 0.92 for overall rating of delirium and ranged from 0.66 (perceptual disturbance) to 0.98 (acute change) for individual features of delirium. For the 4 core features of delirium that are part of the CAM diagnostic algorithm, weighted kappas were all above 0.90, with the exception of disorganized thinking at 0.81 (**Table 4**).

Agreement was calculated for each test in the neuropsychological test battery and the overall GCP (**Table 4**). Weighted kappas were high, ranging from 0.94 - 1.00. Agreement was 100% for all ADLs and IADLs with the exception of 'Doing Housework' with a weighted kappa of 0.90.

## DISCUSSION

The study of delirium and long-term cognitive and functional outcomes is inherently challenging. The SAGES study[13] represents an important advance with a comprehensive preoperative evaluation and longitudinal follow-up of older persons undergoing major surgery. This paper provides the first descriptive summary of the full cohort. The sample includes community-dwelling older persons who had relatively high cognitive functioning at baseline, as measured by neuropsychological testing (e.g., mean of 0.8 standard deviations above the U.S. population mean). The study's data quality procedures, resulted in few losses to follow-up, minimal missing data, and high interrater reliability (kappa =0.92-1.0) on key

summary variables. Variables that typically have a high degree of missing data in self-report studies (e.g., income) had <10% missing values.

While there are many longitudinal studies of older adults in the literature, few studies have provided details about their quality assurance procedures. Therefore, this manuscript represents a valuable addition to the literature and serves as a primary reference source for future SAGES work. This paper may provide a useful guide for researchers in the field of aging, and also for clinicians who are trying to identify features of study data quality that might influence their clinical practice.

Strengths of the SAGES study include the measurement of novel risk factors for delirium, including genetic and plasma biomarkers, neuroimaging markers, life-course factors (e.g., early childhood factors such as family income) and reserve markers (e.g., occupational complexity). State-of-the-art approaches were used for measurement of these risk factors, as well as for delirium and long-term cognitive and functional outcomes. In addition to functional outcomes, other important patient-centered outcomes, such as depression and quality of life, were collected longitudinally with little attrition over 18 months. Since SAGES examines a cohort that was free of dementia at enrollment, the study provides a unique opportunity to disentangle the independent contribution of delirium to cognitive and functional outcomes, independent of the effects of pre-existing dementia. The study will provide a complete and valuable data set and biorepository for future work. Finally, the cohort description and data quality procedures will lay the groundwork for future SAGES work.

Several limitations of this study are worthy of mention. Frailty, an important risk factor for postoperative delirium, was not included as part of the original SAGES study; however, we plan to evaluate frailty in our future studies. The sample enrolled patients who were healthy enough to undergo elective surgery, from a single geographic area. In addition, the SAGES cohort was highly educated with limited ethnic and racial diversity. This lack of overall diversity limits the generalizability of findings and requires that the study be replicated with a more diverse cohort. However, these limitations pose no threat to the internal validity of the results. For some outcomes, the study findings might be more conservative in this selected cohort in comparison with a more generally representative sample.

This paper will serve as an important reference source for future studies using the SAGES cohort. We describe the SAGES cohort and data quality procedures to ensure a high retention rate with little missing data for key study variables, and high interrater reliability of all key study variables. The SAGES study will help to advance our understanding of delirium, a complex multifactorial problem and help to clarify risk factors, pathophysiology, and outcomes that ultimately may facilitate the development of targeted prevention and intervention strategies for surgical patients.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

This work is dedicated to the memory of Joshua Bryan Inouye Helfand.

**Sponsor's Role:** This study was funded by grants P01AG031720 (SKI) and K07AG041835 (SKI) from the National Institute on Aging. Dr. Saczynski was supported in part by grant K01AG33643 from the National Institute on Aging. Dr. Marcantonio was supported in part by grant K24AG035075 from the National Institute on Aging. Dr. Inouye holds the Milton and Shirley F. Levy Family Chair. The funding sources had no role in the design, conduct or reporting of this study.

## Abbreviations

| | |
|---|---|
| **BIDMC** | Beth Israel Deaconess Medical Center |
| **BWH** | Brigham and Women's Hospital |
| **HMS** | Harvard Medical School |
| **HSL** | Hebrew SeniorLife |
| **MGH** | Massachusetts General Hospital |
| **PI** | principal investigator |
| **UCONN** | University of Connecticut Health Center |

## REFERENCES

1. Saczynski JS, Marcantonio ER, Quach L, et al. Cognitive trajectories after postoperative delirium. N Engl J Med. 2012; 367:30–39. [PubMed: 22762316]
2. Inouye SK, Westendorp RG, Saczynski JS. Delirium in elderly people. Lancet. 2014; 383:911–922. [PubMed: 23992774]
3. Marcantonio ER, Goldman L, Mangione CM, et al. A clinical prediction rule for delirium after elective noncardiac surgery. JAMA. 1994; 271:134–139. [PubMed: 8264068]
4. Rudolph JL, Schreiber KA, Culley DJ, et al. Measurement of post-operative cognitive dysfunction after cardiac surgery: a systematic review. Acta Anaesth Scand. 2010; 54:663–677. [PubMed: 20397979]
5. Franco K, Litaker D, Locala J, et al. The cost of delirium in the surgical patient. Psychosomatics. 2001; 42:68–73. [PubMed: 11161124]
6. Sockalingam S, Parekh N, Bogoch, et al. Delirium in the postoperative cardiac patient: a review. J Card Surg. 2005; 20:560–567. [PubMed: 16309412]
7. Gottesman RF, Grega MA, Bailey MM, et al. Delirium after coronary artery bypass graft surgery and late mortality. Ann Neurol. 2010; 67:338–344. [PubMed: 20373345]
8. Rudolph JL, Inouye SK, Jones RN, et al. Delirium: An Independent Predictor of Functional Decline After Cardiac Surgery. J Am Geriatr Soc. 2010; 58:643–649. [PubMed: 20345866]
9. Agency for Healthcare Research and Quality A. [November 10, 2014] National quality clearinghouse measure: delirium: proportion of patients meeting diagnostic criteria on the confusion assessment method (CAM). Available at: http://www.qualitymeasures.ahrq.gov/content
10. Shekelle PG, MacLean CH, Morton SC, et al. Acove quality indicators. Ann Intern Med. 2001; 135:653–667. [PubMed: 11601948]
11. Krumholz HM. “Post-Hospital Syndrome” An Acquired, Transient Condition of Generalized Risk. N Engl J Med. 2013; 368:100–102. [PubMed: 23301730]
12. Inouye SK, Charpentier PA. Precipitating factors for delirium in hospitalized elderly persons. Predictive model and interrelationship with baseline vulnerability. JAMA. 1996; 275:852–857. [PubMed: 8596223]

13. Schmitt EM, Marcantonio ER, Alsop DC, et al. Novel risk markers and long-term outcomes of delirium: the successful aging after elective surgery (SAGES) study design and methods. J Am Med Dir Assoc. 2012; 13:818 e811–810. [PubMed: 22999782]
14. Saczynski JS, Kosar CM, Xu G, et al. A tale of two methods: chart and interview methods for identifying delirium. J Am Geriatr Soc. 2014; 62:518–524. [PubMed: 24512042]
15. Inouye SK, Kosar CM, Tommet D, et al. The CAM-S: development and validation of a new scoring system for delirium severity in 2 cohorts. Ann Intern Med. 2014; 160:526–533. [PubMed: 24733193]
16. Jones RN, Rudolph JL, Inouye SK, et al. Development of a unidimensional composite measure of neuropsychological functioning in older cardiac surgery patients with good measurement precision. J Clin Exp Neuropsychol. 2010; 32:1041–1049. [PubMed: 20446144]
17. Gross A, Jones RN, Inouye SK. Development of an expanded measure of phsyical functioning for older persons in epidemiologic research. Res Aging. 2014
18. Gross AL, Jones RN, Fong TG, et al. Calibration and validation of an innovative approach for estimating general cognitive performance. Neuroepidemiology. 2014; 42:144–153. [PubMed: 24481241]
19. Galea S, Tracy M. Participation rates in epidemiologic studies. Ann Epidemiol. 2007; 17:643–653. [PubMed: 17553702]
20. Inouye SK, van Dyck CH, Alessi CA, et al. Clarifying confusion: the confusion assessment method. A new method for detection of delirium. Ann Intern Med. 1990; 113:941–948. [PubMed: 2240918]
21. Albert MS, Levkoff SE, Reilly C, et al. The delirium symptom interview: an interview for the detection of delirium symptoms in hospitalized patients. J Geriatr Psychiatry Neurol. 1992; 5:14–21. [PubMed: 1571069]
22. Wei LA, Fearing MA, Sternberg EJ, et al. The Confusion Assessment Method: a systematic review of current usage. J Am Geriatr Soc. 2008; 56:823–830. [PubMed: 18384586]
23. Katz S, Downs TD, Cash HR, et al. Progress in development of the index of ADL. Gerontologist. 1970; 10:20–30. [PubMed: 5420677]
24. Lawton MP, Brody EM. Assessment of older people: self-maintaining and instrumental activities of daily living. Gerontologist. 1969; 9:179–186. [PubMed: 5349366]
25. Ware JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care. 1992; 30:473–483. [PubMed: 1593914]
26. Yesavage JA, Sheikh JI. Geriatric Depression Scale (GDS) Recent Evidence and Development of a Shorter Version. Clin Gerontol. 1986; 5:165–173.
27. Runge PE. Eduard Jaeger's Test-Types (Schrift-Scalen) and the historical development of vision tests. Trans Am Ophthalmol Soc. 2000; 98:375–438. [PubMed: 11190034]
28. Macphee GJ, Crowther JA, McAlpine CH. A simple screening test for hearing impairment in elderly patients. Age Ageing. 1988; 17:347–351. [PubMed: 3068972]
29. Charlson ME, Pompei P, Ales KL, et al. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis. 1987; 40:373–383. [PubMed: 3558716]
30. Knaus WA, Draper EA, Wagner DP, et al. APACHE II: a severity of disease classification system. Crit Care Med. 1985; 13:818–829. [PubMed: 3928249]
31. Ciulla TA, Sklar RM, Hauser SL. A simple method for DNA purification from peripheral blood. Analytical biochemistry. 1988; 174:485–488. [PubMed: 3239751]
32. Blake R, Mangiameli P. The effects and interaction of data quality and problem complexity on classification. Journal of Data and Information Quality. 2011; 2:1–28.
33. Parssian A. Managerial decision support with knowledge of accuracy and completeness of the relational aggregate functions. Decision Support Systems. 2006; 42:1494–1502.
34. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull. 1968; 70:213–220. [PubMed: 19673146]

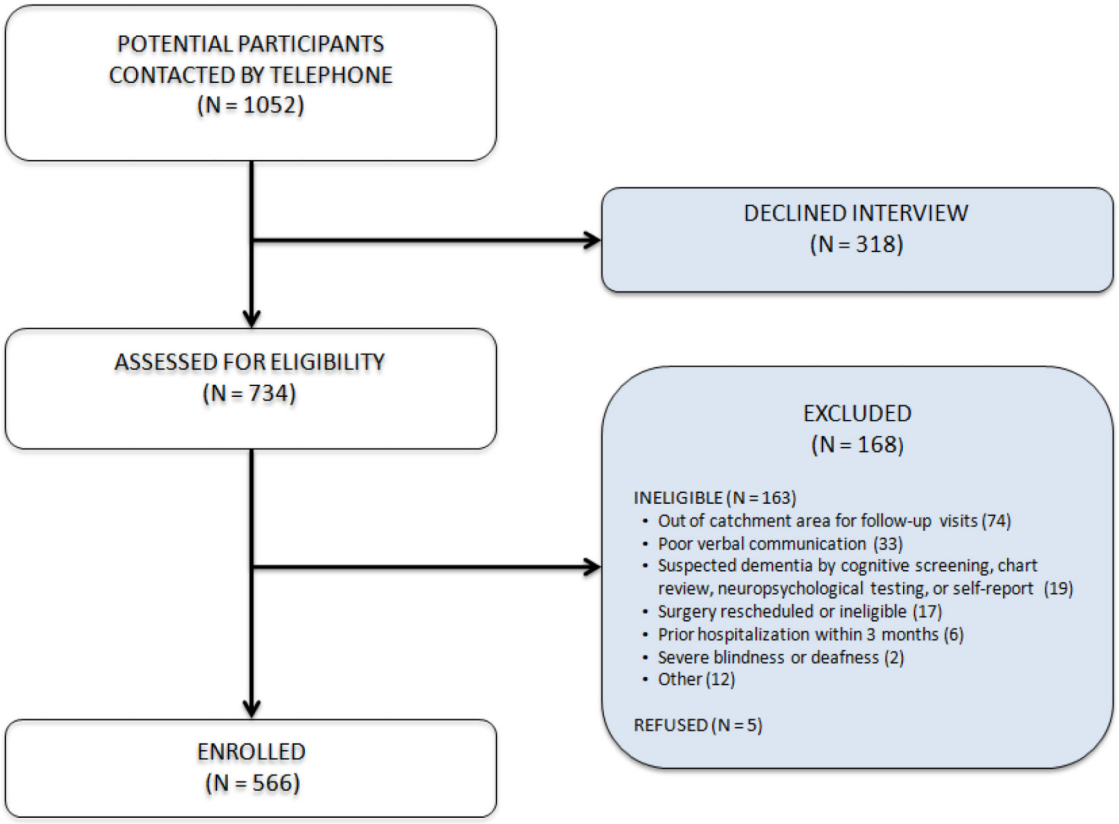


**Figure 1. Summary of SAGES Enrollment**

Eligible participants were age 70 years and older, English speaking, scheduled to undergo elective surgery with an anticipated length of stay of at least 3 days. Exclusion criteria included evidence of dementia, delirium, hospitalization within 3 months, terminal condition, legal blindness, severe deafness, history of schizophrenia or psychosis, and history of alcohol abuse. Patients who would be out of the catchment area at time of follow-up were also excluded from enrollment.

**Table 1**

Data Quality Procedures for the Dimensions of Completeness, Accuracy, Consistency and Timeliness

| | Completeness | Accuracy | Consistency | Timeliness |
|---|---|---|---|---|
| **Screening and Recruitment** | | | | |
| Medical record eligibility screening (EDC) | Flag in database if case not complete (built-in logic) | Computer based eligibility calculations and prevention of most nonsensical data, adjudication | Double check of all eligible cases | Daily tracks of frequency and volumes per site |
| Recruitment letters and enrollment phone calls | Daily sent letters/phone calls report | Addresses rechecked against United States Postal Service Systems | Standardized letters/phone call scripts | Enrollment date report |
| **In-person and phone interviews, medical record abstraction** | | | | |
| Interviewer/rater (data collection level) | Scoring re-check by second person | Team coding/scoring sessions | Interrater reliability interviews | Interview date/ time documentation |
| **Assays** | | | | |
| Blood collection | Report of missing samples | Study IDs and dates on all tubes | Time between blood draw and processing time | Standard processing protocol and metrics to ensure timeliness |
| **Neuroimaging** | | | | |
| Scan collection | Check for correct number of images per sequence | Visual assessment of coverage and artifacts* | Date alignments with other study dates | Scan date/time documentation |
| **Data Management** | | | | |
| Hard copy data entry | Data entry reports | Databases with built-in prevention of nonsensical data entry | Double data entry | Data collection and entry completion reports |
| Data tracking | Completion and data back-up reports | Checks for out of range or illogical values | Consistency checks of data within participants over time | Out-of-date data collection report |
| Data cleaning/freezing | Missing data definition and reports | Variable definition sheets# | Independent double-coding of derived variables | Standardized dates to create a ‘frozen’ compendium of validated data (semiannual) used for analyses |
| Preparation of analysis files | Table shells with mock data for outcomes data | Standardized manual with coding of key variables (‘Code book’) | Data quality report | Report deadlines |

*Artifacts include motion, wrapping/ringing/stripping, blurring, shadowed arc, ghosting, radio frequency noise/spiking, susceptibility artifacts and unexpected inhomogeneities.

#File with formal definition of the derived variable, description of the handling of missing data, citations of publication first using the variable, and a log of coding decisions.

## Table 2

Baseline Characteristics of Study Cohort (N=566)

| Characteristics | n (%)* | Missing Data |
|---|---|---|
| Age, mean years (SD) | 76.7 (5.2) | 0 |
| Female sex | 330 (58) | 0 |
| Race | | 0 |
| Asian | 5 (1) | |
| African American | 29 (5) | |
| White | 528 (93) | |
| Other | 4 (1) | |
| Hispanic ethnicity | 7 (1) | 0 |
| Education, mean years (SD) | 15.0 (2.9) | 0 |
| 0-12 Years | 165 (29) | |
| 13-16 Years | 234 (41) | |
| 17+ Years | 167 (30) | |
| Surgery Type | | 0 |
| Orthopedic | 460 (119) | |
| Vascular | 35 (6) | |
| Gastrointestinal | 71 (13) | |
| Married | 335 (60) | |
| Lives alone | 169 (30) | 0 |
| Visual impairment, <20/70 corrected binocular vision | 3 (.5) | 2 |
| Hearing impairment, <6/12 on Whisper Test | 185 (33) | 1 |
| Geriatric Depression Scale-15, mean score (SD) | 2.5 (2.5) | 2 |
| Charlson Comorbidity Index, mean score (SD) | 1 (13) | 0 |
| 3MS, mean score (SD) | 93.4 (5.4) | 1 |
| General cognitive function composite, mean score (SD) | 57.5 (7.4) | 0 |
| MOS SF-12 physical function composite, mean score (SD) | 35.3 (10.0) | 3 |
| Any IADL Impairment | 157 (28) | 0 |
| Any ADL Impairment | 42 (7) | 0 |
| Postoperative Apache II, mean score (SD) | 11.9 (3.0) | 0 |

MOS = Medical Outcomes Study; IADL = Instrumental Activities of Daily Living; ADL = Activities of Daily Living; Whisper: range (0-12), lower is worse; GDS: range (0-15), higher is worse; 3MS:range (0-100), higher is better

*All values are n (%) unless otherwise noted. Proportion of missing values was less than 0.5% on all items.

## Table 3

Participant Disposition by Study Visit, and Completeness of Data Collection for Selected Major Outcomes

| | Participant Disposition | | | | | Major Outcomes: Proportion non-missing, % | | |
|---|---|---|---|---|---|---|---|---|
| **Visit/Method** | **Potential Interviews[*] (N)** | **Completed Interviews N (%)[†]** | **Refused or Unobtainable[**] (N)** | **Deaths[#] (N)** | **Drop-Outs[#]** | **GCP** | **ADL** | **IADL** |
| 1 Month; In-Person | 566 | 556 (98) | 2 | 1 | 7 | 99.6 | 99.8 | 99.8 |
| 2 Month; In-Person | 558 | 543 (97) | 10 | 1 | 4 | 99.4 | 99.8 | 99.8 |
| 4 Month; Phone | 553 | 537 (97) | 10 | 3 | 3 | -- | -- | -- |
| 6 Month; In-Person | 547 | 537 (98) | 10 | 0 | 0 | 99.2 | 99.6 | 99.6 |
| 9 Month; Phone | 547 | 521 (95) | 19 | 3 | 4 | -- | -- | -- |
| 12 Month; In-Person | 540 | 517 (96) | 23 | 0 | 0 | 99.5 | 100 | 100 |
| 15 Month; Phone | 540 | 502 (93) | 32 | 3 | 3 | -- | -- | -- |
| 18 Month; In-Person | 534 | 508 (95) | 26 | 0 | 0 | 99.4 | 99.7 | 99.8 |

Note. Each row describes participants available (eligible) for interview. Reasons for dropout include: time commitment, gatekeeper preference, health status changes, declining memory.

GCP = General Cognitive Composite; ADL = Activities of Daily Living; IADL = Instrumental Activities of Daily Living

[*] Eligible defined as all enrolled participants (N=566) minus those who have died or dropped out in the prior time periods

[†] Proportion (%) complete defined as the ratio of interviews completed to those potentially available for interview at the relevant visit

[**] Those refusing interview may be available at later timepoints

[#] Lost to death or follow-up (will not contribute to later timepoints)

**Table 4**

Interrater Reliability for Key Study Variables

| Variable | Agreement % | Kappa | Weighted Kappa |
|---|---|---|---|
| Confusion Assessment Method (N=71) | | | |
| Acute Change | 99% | 0.98 | 0.98 |
| Inattention | 93% | 0.89 | 0.92 |
| Disorganized Thinking | 92% | 0.82 | 0.81 |
| Altered Level of Consciousness | 99% | 0.89 | 0.95 |
| Disorientation | 97% | 0.92 | 0.95 |
| Memory Impairment | 97% | 0.96 | 0.95 |
| Perceptual Disturbance | 97% | 0.66 | 0.66 |
| Psychomotor Agitation | 97% | 0.74 | 0.79 |
| Psychomotor Retardation | 87% | 0.68 | 0.72 |
| Sleep-Wake Cycle Disturbance | 85% | 0.77 | 0.80 |
| Overall Delirium | 97% | 0.92 | 0.92 |
| **Cognitive and Physical Functioning (N=48)** | | | |
| Neuropsychological Exam | | | |
| HVLT-R Total Recall | 91% | 0.91 | 0.99 |
| HVLT-R Delayed Recall | 93% | 0.92 | 0.98 |
| Visual Search and Attention Test | 97% | 0.97 | 1.00 |
| Trail Making Test A | 95% | 0.95 | 0.97 |
| Trail Making Test B | 93% | 0.92 | 0.98 |
| Digit Symbol Substitution | 100% | 1.00 | 1.00 |
| Digit Span Test | 87% | 0.86 | 0.97 |
| Verbal Fluency | 65% | 0.64 | 0.96 |
| Category Fluency | 83% | 0.81 | 0.97 |
| Boston Naming Test | 100% | 1.00 | 1.00 |
| Overall GCP | 98% | 0.70 | 0.94 |
| Basic Activities of Daily Living | | | |
| Bathing | 100% | 1.00 | 1.00 |
| Grooming | 100% | 1.00 | 1.00 |
| Dressing | 100% | 1.00 | 1.00 |
| Feeding | 100% | 1.00 | 1.00 |
| Getting from bed to chair | 100% | 1.00 | 1.00 |
| Using the bathroom | 100% | 1.00 | 1.00 |
| Walking across small room | 100% | 1.00 | 1.00 |
| Dependency Score | 100% | 1.00 | 1.00 |
| Instrumental Activities of Daily Living | | | |
| Telephone | 100% | 1.00 | 1.00 |
| Getting places out of walking distance | 100% | 1.00 | 1.00 |
| Shopping for groceries | 100% | 1.00 | 1.00 |
| Preparing meals | 100% | 1.00 | 1.00 |

| Variable | Agreement % | Kappa | Weighted Kappa |
|---|---|---|---|
| Doing housework | 98% | 0.90 | 0.90 |
| Managing money | 100% | 1.00 | 1.00 |
| Managing Medications | 100% | 1.00 | 1.00 |
| Dependency Score | 98% | 0.93 | 0.97 |

Boston Naming range (0-15), higher is better; Dependency Score range (0-14), higher is worse; Digit Span Test range (0-30), higher is better; Digit Symbol Substitution range (0-89), higher is better; GCP = General Cognitive Performance; HVLT-R = Hopkins Verbal Learning Test - Revised; HVLT-R Total recall, range (0-36), higher is better; HVLT-R Delayed recall, range (0-12), higher is better, Trails A range (0-180 seconds), higher is worse; Trails B range (0-300 seconds), higher is worse.