Supplemental Material 2: Full statistical analysis plan

General considerations:
A flow-chart will describe the flow of patients during the study, from inclusion until Day-30, by randomization arm and by type of scheduled postoperative ward, either geriatric perioperative unit or another type of ward (stratification variable).

All variables will be described overall and by randomization arm. Categorical variables will be described by frequencies for all categories, and percentages; continuous variables will be described by their min, max, mean, standard deviation, and interquartile range. Number of missing values will also be described for all variables.

In this randomized control trial, the principal and secondary analyses will follow the intention-to-treat (ITT) principle.

Stratification variables are center and postoperative wards. The last one is a dichotomic variable: either geriatric perioperative unit or another type of ward.

Following the ITT principle, not operated patients will be analyzed according to their initial randomization group. Reasons for not operating include switching to palliative care or transferring to other facilities. For those patients, treatment will be stopped and data will be censored at the decision date of do not operate.

The maximum duration between inclusion and intervention will be of 5 days. Beyond that duration, treatment is stopped. For principal analyses, occurrence of delirium of patients with stopped treatment will be censored at Day-5. If eventually they are operated, sensibility analyses will include their data.

For coma operated patients, for whom CAM or CAM-ICU cannot be assessed, the delirium occurrence will be imputed by a most pejorative outcome of CAM scores on the day of coma onset.

All effect size will be presented together with their 95% confident intervals.

The significant level of all analyses is fixed to a bilateral (alpha) Type I error of 5%.

All statistical analyses will be performed using SAS software (SAS Institute Inc., Cary, NC) v. 9.4 or later, or R software (R Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org/) v. 4.0 or later.

Primary endpoint analysis:
Delirium occurrence at time T is defined as the first CAM or CAM-ICU assessment that retained the diagnosis of delirium.
In this population of elderly patients, the risk of death in the context of postoperative acute surgery of fractured lower limb is of 7% at Day-30. Therefore, it will be estimated the concurrent distribution of mortality in both groups while estimating the incidence of the treatment by melatonin on delirium.

Postoperative delirium cumulative incidence function (CIF) will be compared at Day-10 between the two randomized arms (melatonin vs. placebo) by means of a competing risks framework, using the Fine and Grey model [1] that allows to estimate CIF on the presence of concurrent causes of failure, deceased in our study, altering the probability of experiencing the event of interest, delirium. Death is the competing risk under consideration.

Patients alive on Day-10 without delirium occurrence will be censored at Day-10. Those lost to follow up will be censored on the date of latest news. Non operated patients will be censored on the decision date of do not operate.

Secondary endpoints analyses:

1- Following recommendations, cause-specific hazard will be calculated and compared between the study arms for both delirium and mortality at Day-30.
2- Proportion of days where the CAM or CAM-ICU are positive will be compared between randomized arms at Dat-10 using a linear regression adjusted by the stratification variables.
3- Administration at any postoperative time of sedative or antipsychotic drugs, evaluated at Day-10, or at the day of leave if shorter, will be compared between groups using a logistic regression adjusted by the stratification variables.
4- Occurrence of postoperative physical restrain at Day-10, or at the day of leave if shorter, will be compared between groups using a logistic regression adjusted by the stratification variables.
5- Occurrence of postoperative falls at Day-10, or at the day of leave if shorter, will be compared between groups using a logistic regression adjusted by the stratification variables.
6- The Mini Mental State Examination score at Day-10, or at the day of leave if shorter will be compared between randomized arms using a linear regression adjusted by center and type of postoperative. The MMSE score runs from 0 point (worst result) to 30 points (best result). Score of deceased patients will be interpreted as 0 point. Transferred patients will be evaluated before leaving the hospital.
7- Duration of hospital stay at Dat-30 will be compared using a linear regression adjusted by center and type of postoperative ward. Number of days of deceased or transferred patients will be counted until the day of leave.
8- Katz Index of activities of daily living is a score running from 0 point (best result) to 6 points (worst result). Scores of deceased patients will be interpreted as 6 points (worst scores). The score at Day-30 will be compared between randomized arms using a linear regression adjusted by center and type of postoperative ward.
9- Quality of life (utility values derived from the score at EQ-5D-5L questionnaire) will be compared between the randomized arms using a linear regression adjusted by center and type of postoperative ward (see questionnaire in supplementary material 1).
10- Occurrence of side effects at Day-10, or at the day of leave if shorter, will be compared between groups using a logistic regression adjusted by the stratification variables.
11- In order to evaluate morphine consumption and administration of anticholinergic drugs, since those drugs might work as confounding factors, the morphine consumption and anticholinergic drugs until D-10, or the day of leave if shorter, will be used. With respect

to morphine, cumulated morphine consumption will be standardized in equivalent of oral morphine according to the following table [2].

Medication name	Equivalent of oral morphine (mg)
Oral morphine (mg)	1 : 1
Subcutaneous morphine (mg)	1 : 2 (SC morphine x 2 = oral morphine equivalent)
Intravenous morphine IV (mg)	1 : 3 (IV morphine x 3 = oral morphine equivalent)
Oral Oxycodone (mg)	1 : 1,5 (oxycodone x 1,5 = oral morphine equivalent)
Oral hydromorphone (mg)	1 : 5 (hydromorphone x 5 = oral morphine equivalent)
Transdermal buprénorphine (ug/h)	1 : 1,7 (patch dosing x 1,7 = oral morphine equivalent for 24 hours)
Transdermal Fentanyl (ug/h)	1 : 2,4 (patch dosing x 2,4 = oral morphine equivalent for 24 hours)

Morphine and anticholinergic drugs consumption will be compared between groups using a logistic regression adjusted by the stratification variables. If this endpoint is statistically significant between groups, it will be used as adjusting factor for the secondary other endpoints. In-depth analyses about their consumption will be done between groups.

12- **Compliance to treatment** will be assessed using records of unused packaging and medical files. We will confront the dose actually absorbed by the patient to the prescribed and predicted dose (which will consider possible discontinuation of treatment). Comparisons will be performed between randomized arms using a linear regression adjusted by center and type of postoperative ward.

Univariate and then multivariate models will be performed to determine the relative contributions of factors to delirium (potential confounders, in the perspective of further adjustment, if necessary). The selection of variables for these models will be done considering the number of events, significant factors in univariate analyses and those clinically relevant. Of particular interest are: the automatic restraint of patients (or its part contributing to the center effect), age, chronic alcoholism, antipsychotic or sedative treatment.
A sensitivity analysis may be performed based on the time frame of the preoperative period.

Economic evaluation

The outcome measure for the economic evaluation will be the estimation of the cost-effectiveness of melatonin. Because of the short duration of the follow up (30 days) the difference in quality-adjusted life year (QALY) is likely to be small; we will therefore add a measure of clinical effectiveness based on a composite of the primary and secondary clinical outcomes: incidence of delirium, need for sedatives, need for physical restraints, fall, and death. Quality of life will be assessed using the EuroQol EQ-5D-5L questionnaire as now recommended by the French national health authority.

The analysis will follow the French Health Authority and the Consolidated Health Economic Evaluation Reporting Standards (CHEERS) guidelines on economic evaluation in health care. [3,4]

Trial follow-up: Quality of life (assessed at hospital discharge and at 1 month, postoperative complications and use of hospital resources (emergency room visits, hospital admissions) will be obtained by collection of data during the initial hospital stay and then by telephone call from a clinician or a study nurse, at 1month. In addition, a double checking of the use of hospital-related resources after the initial discharge will be made via hospital databases.

Calculation of costs: Cost calculation will include all the hospital-related costs over a one-month period. Owing to the impact of a delirium on the use of hospital resources (ICU and in-hospital length of stay), the estimation of costs will focus on hospital costs. Primary care-related costs will not be analysed. The rates for the hospital stay will be calculated with respect to the Diagnosis Related Groups, adjusted for the patients' length of stay. Unit costs are presented in the following table.

Type of service/ product	**Unit cost (€)**	**source**
Melatonin	1€ per tablet	French red book (Vidal)
Surgery ward day	700-800	scansante
REA (resuscitation supplement)	804	scansante
STF (intensive care supplement)	402	scansante

Calculation of QALYs: EQ5D5L scores will be converted into utilities using the French value set [5]. QALYs will be calculated using the area under the curve approach. Missing data on EQ-5D score and health-care costs will be imputed with multiple imputations by chained equations, assuming data missing at random.

The value of the melatonin treatment will be determined with respect to 1) the extra cost and 2) its potentially beneficial impact on both quality of life and adverse events. QALY, the commonly used generic measure of disease burden, will be used. The overall cost difference between the standard- and melatonin treatments will be assessed via the calculation of the incremental cost-effectiveness ratio (ICER), expressed as € per QALY:

$$ICER = \frac{C\ melatonin - C\ standard\ management}{QALY\ melatonin - QALY\ standard\ managment}$$

To refine the 95% confidence interval of these parameters, the bootstrapping technique will be used. We will compare the result to the usually applied thresholds of €50,000-100,000/QALY and calculate the probability of cost-effectiveness from the bootstrapped probabilistic sensitivity analysis.

A complete case analysis will be performed on the population for whom all cost and effectiveness (EQ5D5L and 1-month clinical outcomes) and data are available. Secondarily, after imputation of missing data, an intention-to-treat analysis will be performed.

Of note, the sample size calculation has been based upon clinical outcomes, mostly for ethical reasons: the clinical outcome takes precedence over the efficiency of the allocation of healthcare resources. Hence, we input the 718 patients sample size into Glick's formula [6]. This sample size will allow testing for the existence of a difference of €800 and 0.04 QALYs at the €100,000/QALY threshold between standard management and melatonin, respectively [7].

Hospital length of stay, ICU length of stay, will be analysed using Cox proportional-hazards models.

State whether subjects who exit the study prematurely will be replaced and in what proportion: Patients exiting the study prematurely will not be replaced.

Anticipated level of statistical significance: The statistically significant level is fixed for all analyses as a bilateral (alpha) Type I error of 5%.

Statistical criteria for termination of the study: Not applicable.

Method for considering missing, unused or invalid data: Missing data will be described for all variables globally and by treatment group. Missing data for the principal endpoint will be censored at the last date of follow-up or at the day of latest news.

If eventually one of the 10 assessments on delirium is missing, but the following is available and consistent with the previous assessment, the missing data will be imputed by the closest value. If the available data before and after and missing data are discordant, imputation will follow two scenarios (absence or presence of delirium).

Management of modifications made to the analysis plan for the initial strategy: All major modifications to the planned analysis will be submitted to approval of the scientific committee and of the ethics committee (CPP).

Selection of populations: All analyses will follow the intention-to-treat principle.

The primary analysis will be repeated on population per protocol: only patients in the intervention group who absorbed at least 90% of the prescribed treatment before delirium occurrence or competitive event, and patients of the control group who have not received melatonin (or less than 10% of an equivalent dose of melatonin) will be included in this analysis.
Patients wrongly included, secondary excluded or lost to follow-up will not be considered in this analysis. No patient will be reclassified.

References

1 Fine JP, Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association* 1999;94:496–509. doi:10.1080/01621459.1999.10474144
2 Caraceni A, Hanks G, Kaasa S, *et al.* Use of opioid analgesics in the treatment of cancer pain: evidence-based recommendations from the EAPC. *Lancet Oncol* 2012;13:e58-68. doi:10.1016/S1470-2045(12)70040-2
3 Haute Autorité de Santé. Choices in Methods for Economic Evaluation. 2012.https://www.has-sante.fr/portail/upload/docs/application/pdf/2012-10/choices_in_methods_for_economic_evaluation.pdf
4 Husereau D, Drummond M, Petrou S, *et al.* Consolidated Health Economic Evaluation Reporting Standards (CHEERS) statement. *BMJ* 2013;346:f1049. doi:10.1136/bmj.f1049
5 Andrade LF, Ludwig K, Goni JMR, *et al.* A French Value Set for the EQ-5D-5L. *Pharmacoeconomics* 2020;38:413–25. doi:10.1007/s40273-019-00876-4
6 Glick HA. Sample size and power for cost-effectiveness analysis (part 1). *Pharmacoeconomics* 2011;29:189–98. doi:10.2165/11585070-000000000-00000
7 Téhard B, Detournay B, Borget I, *et al.* Value of a QALY for France: A New Approach to Propose Acceptable Reference Values. *Value Health* 2020;23:985–93. doi:10.1016/j.jval.2020.04.001