SUPPLEMENTARY MATERIAL *- Effects on Clinical Outcomes of a 5-year Surgical Safety Checklist Implementation Experience: A Large-scale Population-based Difference-in-Differences Study***.**

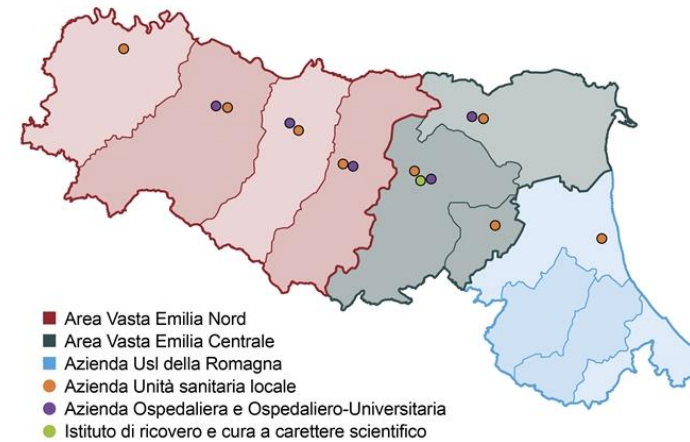**Figure 1 –** Emilia-Romagna, Italy (Source: Wikipedia)

Figure 2 – Emilia-Romagna, administrative areas: North (in red), Centre (green) and Romagna (blue). Source: ER Salute, http://salute.regione.emilia-romagna.it/ssr/servizio-sanitario-er (last access June 30, 2018)

Appendix 1 – Difference in differences (DID) approach: short description

Difference in differences (DID) is a method to estimate treatment effects with non-experimental data[28]. This approach requires data measured at two or more different time intervals, resulting in a time series - or sequence of values - related to a particular measure of interest and detected at regular intervals of time. Differently from an "interrupted time series" (ITS) - a conceptually related design - the DID approach includes a control group, not exposed to the intervention and followed over the same time period as the intervention group.

Figure A1 outlines this method by referring (for the sake of simplicity) to the ITS approach. In a time series, a sequence of measures might be split in at least two (or n) time periods by one (or more) point(s) of change, tracing the start (or several phases) of an event (i.e. a policy intervention or an experimentation): as a consequence, at least two (or n) "segments" are defined. For each "segment" the parameters of interest are identified. The *level* is represented by the value of the time series at the beginning of the observation period, (i.e. the value of the intercept y) for the first segment identified and, for each segment following the first, the value of the measure of interest immediately following each point of change. A statistically significant change in the level (or a "slip" of the segment, up or down) is a "sharp" effect (i.e. *Lev 2 – Lev 1*). The *trend* is represented by the rate of change of the measure in the period defined by each segment (i.e. the value of the angular coefficient). A variation of the trend is defined by a statistically significant increase or decrease of the value of the angular coefficient (i.e. post-implementation) with respect to the value observed in the previous segment, before the introduction of the intervention (pre-implementation, i.e. *Tr 2 – Tr 1*).

The DID approach is based on a double comparison of effects, measuring differences in time (pre- and post-implementation) and in subjects (the intervention group and the control group), thus allowing to isolate the effects of the intervention among others that may have occurred in the same time intervals (Figure A2). The assumption underlying this method is that no systematic differences exist between the groups as for the time trend of the variable of interest in the absence of the intervention. According to this assumption, observed changes in the comparison group reproduce those hypothetically occurred in the intervention group, had it not been exposed to the implemented intervention. Ultimately, it will be possible to analyze changes of the measure of interests between segments in different periods of time, traced by one or more points of change
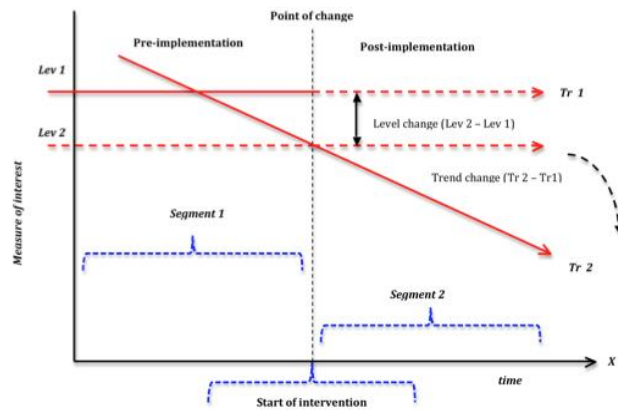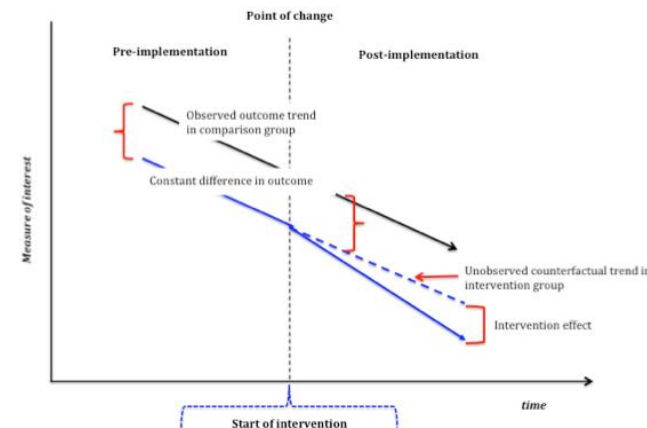

Figure A1

Figure A2

Table 1

Emilia-Romagna Region (Italy). SOS.net Project: average rates of full completion of the checklist, public hospitals (2012).

Checklist component	%
Sign In	97.5
Time Out	97.3
Sign Out	95.8

NOTE. Rates of full completion of the checklist showed variations among hospitals (min 68% max 100%); as reported in other studies (18, 30-32) the completeness was worse in *sign out*, while a better quality was observed in *sign in* and *time-out*.

In addition to the data on completeness, in 2012 we were able to measure a defined set of most common flaws detectable through the checklist (see text of the article, Table 1, form A and B mentioned): overall, at least a flaw was detected in 3.2% of all surgeries performed in public hospitals. Most of the flaws were reported in *sign in* (73%). Ten types of flaws accounted for 70% of the total number of detected flaws. "Unmarked site" was the most frequently reported flaw.