Application of a bootstrap method to estimate the inter-lab agreement

F. Ingravalle[§], Crescio M.I.[¥], Abete M.C.[¥], Caramelli M.[§], Ru G.[§]

[§] CEA – National Reference Laboratory for Animal TSEs

[¥] CReAA – National Reference Laboratory for surveillance and monitoring in feeding stuff

Introduction

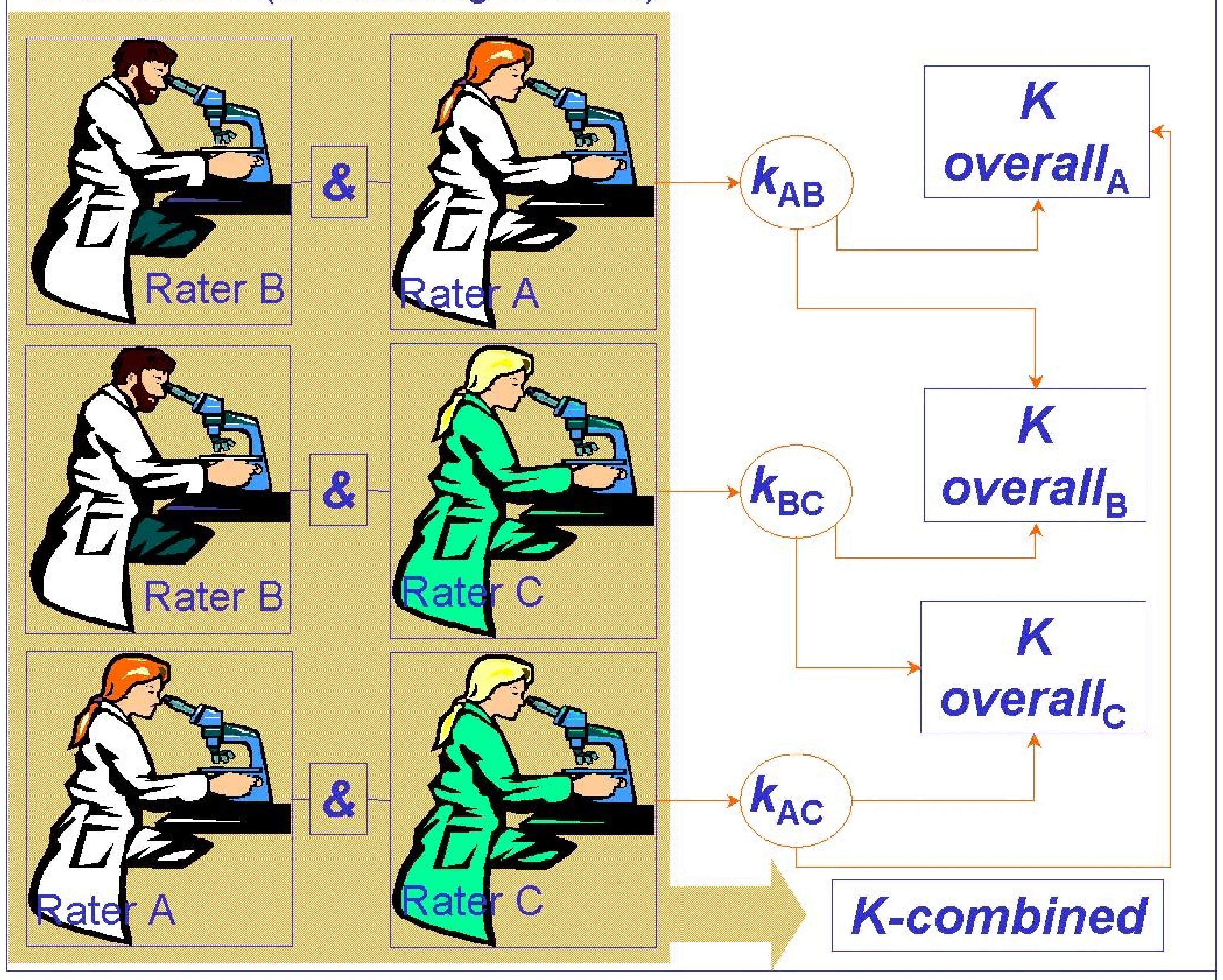
Reproducibility studies are usually designed to asses the agreement among results obtained by different raters (=analysts) applying a diagnostic test in the same conditions. In this work, a reproducibility study (ring test) concerning the official method for detection of animal derived particles in feeding stuff is described. Inter-rater agreement, measured by **Cohen's k** for each couples of raters, ***k-overall*** for each rater and ***k-combined*** for all the raters, are usually calculated to assess reproducibility of a method (see figure 1). On the contrary, it is not possible to calculate a summary measure of the agreement among different labs, based on the results obtained by each analyst (inter-laboratory agreement). **Aims of the study are**:

- **the assessment of inter-laboratory agreement (based on the results of individual raters), calculating *k-combined* and *k-overall* for each lab, applying a bootstrap method**
- **the evaluation of reproducibility of the official method for feeding stuff control (based on inter-rater agreement)**

Table 1: number of analysts employed in each lab

Code of the lab	Number of analysts
A	2
B	2
C	2
D	7
E	2
F	8
G	8
H	6
I	2
J	4
K	2
L	1
M	1
TOTAL	47

Figure 1: exemplification of difference between *k-overall* and *k-combined* (considering 3 raters).



Materials and methods

The study involved:

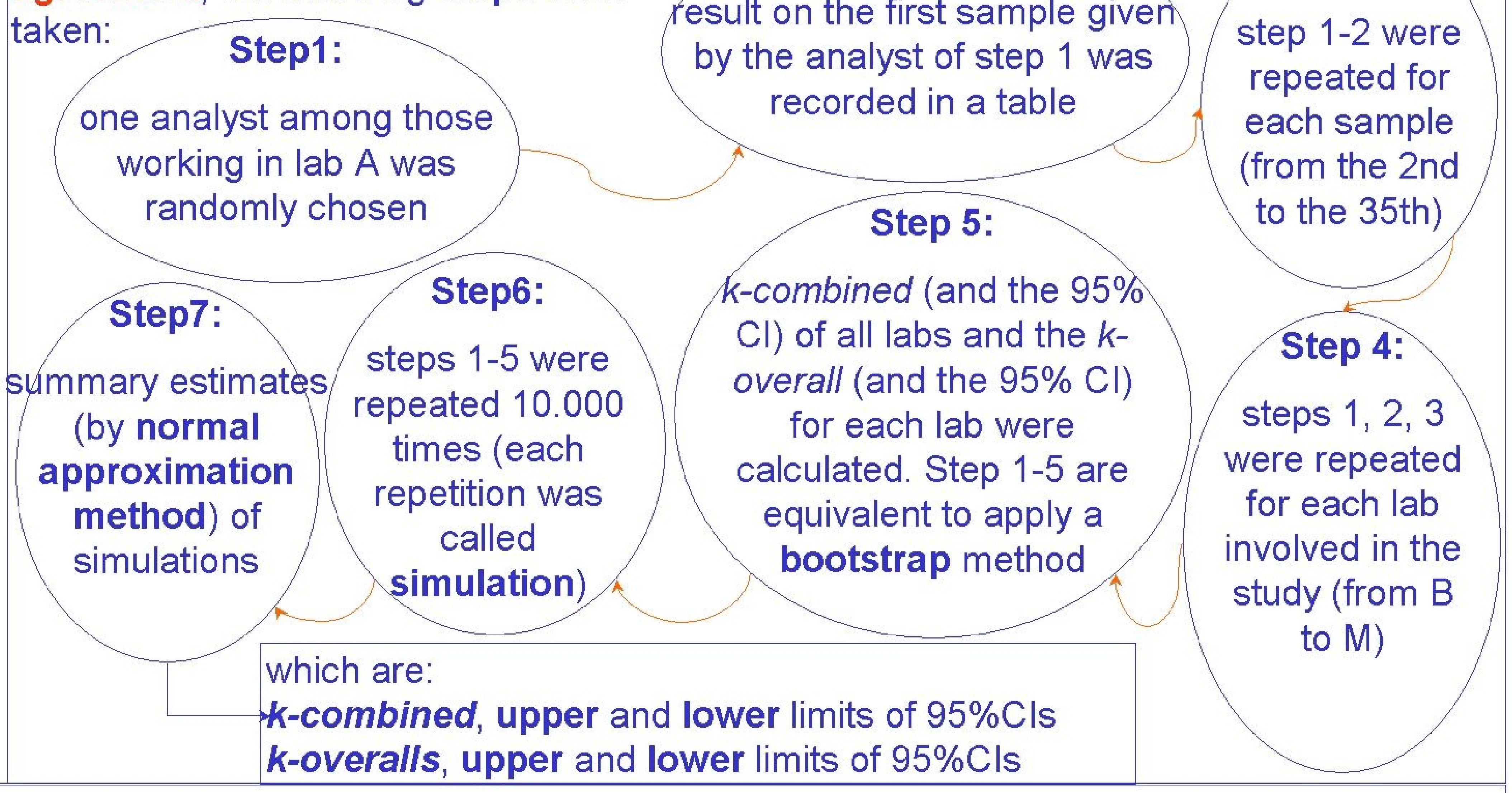
- **47** raters
- **13** different labs (the number of raters for lab varied from 1 to 8, see table 1)
- **35** samples of feeding (15 contaminated by mammal, poultry or fish derived particles and 20 not contaminated). Each rater, **independently**, tested samples and for each one gave a result as positive (presence of derived particles in the examined feeding stuff) or negative (absence of animal derived particles)

In order to assess the reproducibility of the method among raters, **inter-rater agreement** was evaluated by:

- ***Cohen's k*** for each couples of analysts
- ***k-overall*** for each analyst
- ***k-combined*** for the results given by the 47 analysts

(and 95% CIs)

In order to assess the **inter-laboratory agreement**, the following **steps** were taken:



Step1: one analyst among those working in lab A was randomly chosen

Step 2: result on the first sample given by the analyst of step 1 was recorded in a table

Step 3: step 1-2 were repeated for each sample (from the 2nd to the 35th)

Step 4: steps 1, 2, 3 were repeated for each lab involved in the study (from B to M)

Step 5: *k-combined* (and the 95% CI) of all labs and the *k-overall* (and the 95% CI) for each lab were calculated. Step 1-5 are equivalent to apply a **bootstrap** method

Step6: steps 1-5 were repeated 10.000 times (each repetition was called **simulation**)

Step7: summary estimates (by **normal approximation method**) of simulations

which are:
k-combined, **upper** and **lower** limits of 95%CIs
k-overalls, **upper** and **lower** limits of 95%CIs

Results and discussion

- *K-overall* and *k-combined* (and 95% CIs, i.e.: confidence intervals) showed **high inter-rater reproducibility** of the microscopic method throughout the Italian surveillance network (figure 2).
- Mean values for estimates (*k-overall, k-combined*, limits of 95%CIs obtained by normal approximation of simulations) were very high. Lower limits for all the labs were >0.80 (figure 3): it points out **very good reproducibility of the microscopic method among labs.**
- Finally, bootstrap method in reproducibility study allows the evaluation of agreement among complex structures (labs) when ratings (results) are given in term of **individual raters (analysts) nested in complex structures**. The described statistical procedure provides a **realistic scenario** of the performances of each lab.

Figure 2: k-overall (yellow bars) for each analyst and k-combined (blue bar) for all the analysts with 95% CIs (orange lines)

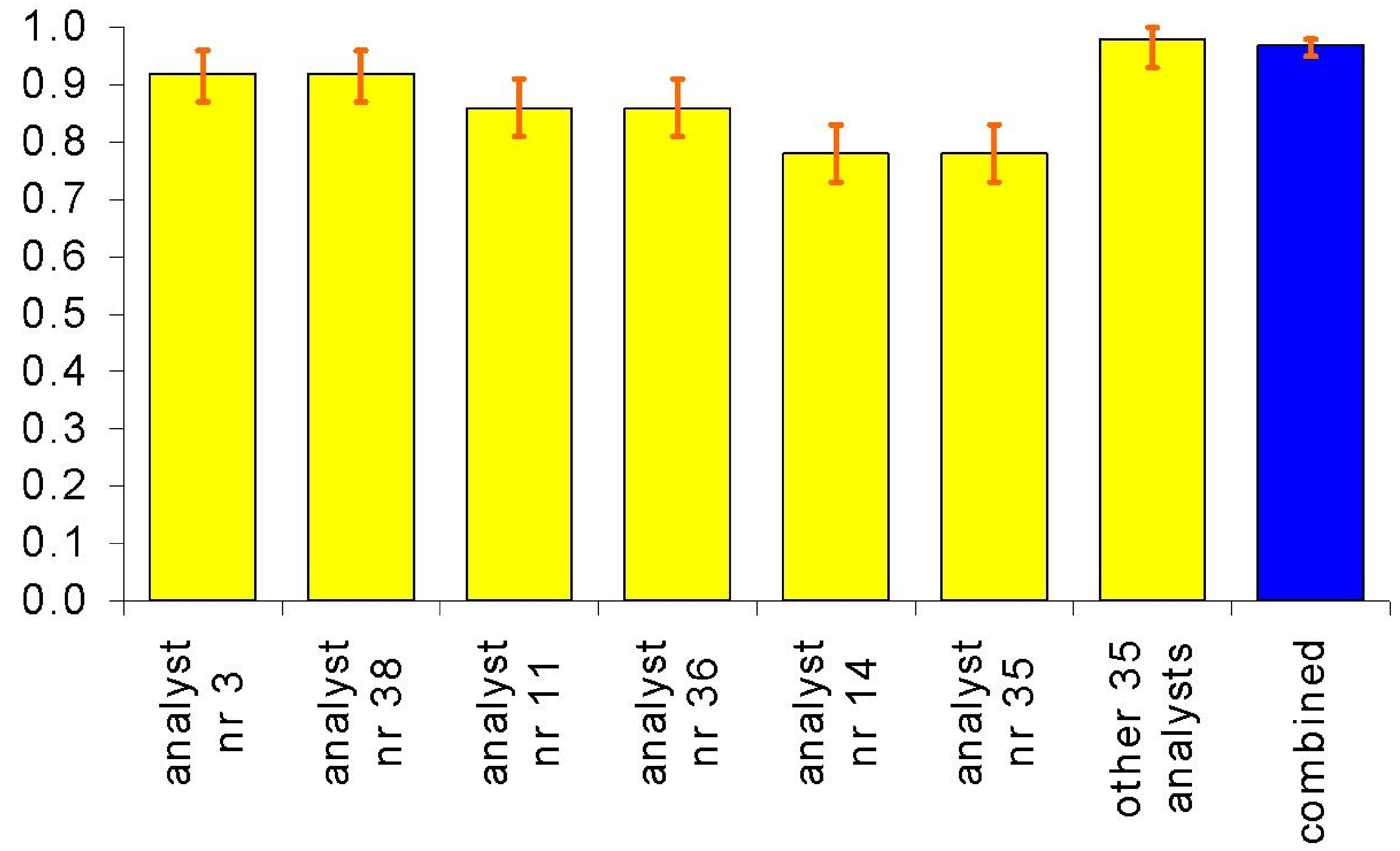
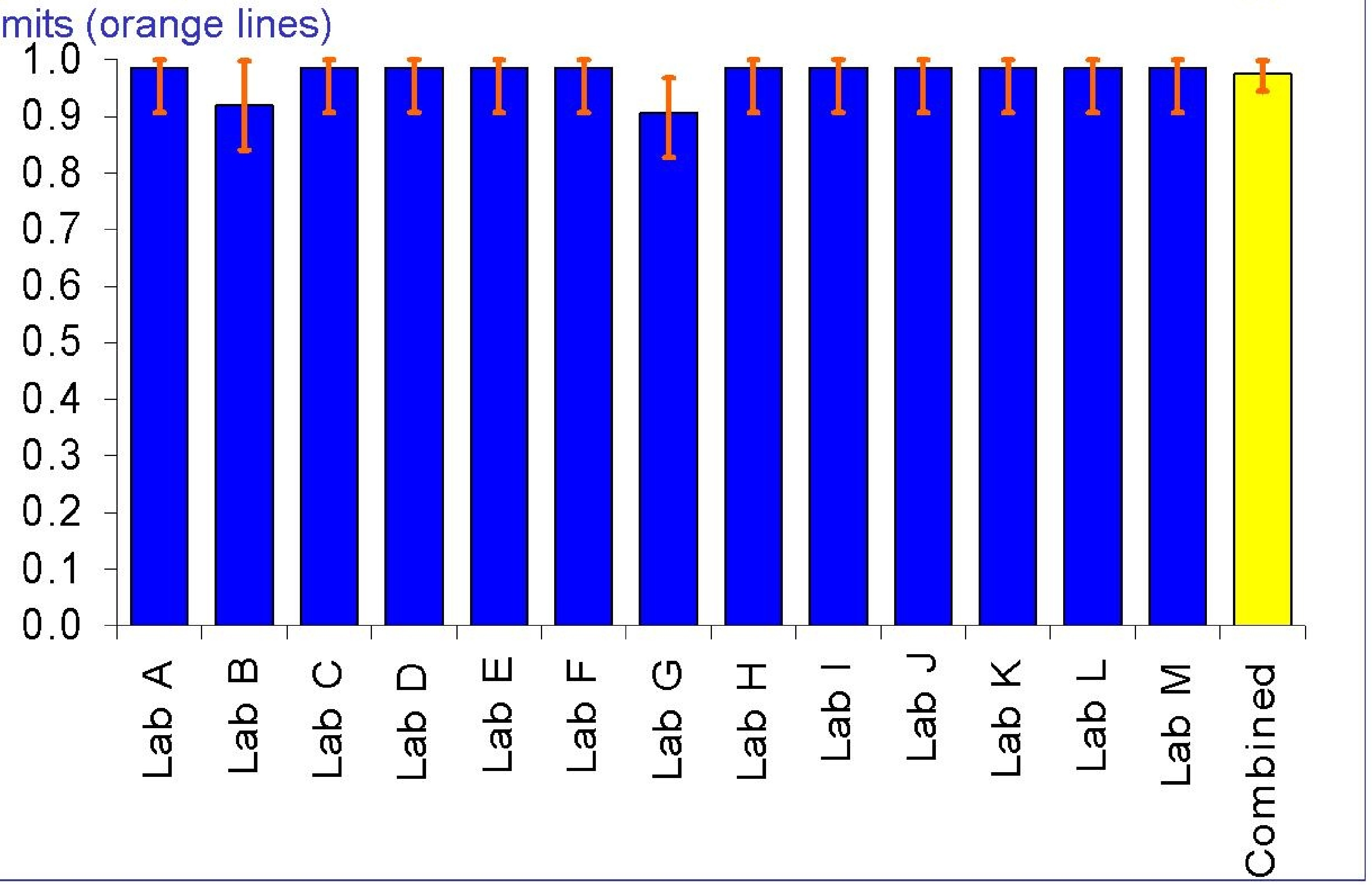


Figure 3: mean values of *k-overall* (blue bars) for each lab and *k-combined* (yellow bar) for all the labs with mean values of theirs 95% lower and upper limits (orange lines)


Lab A
Lab B
Lab C
Lab D
Lab E
Lab F
Lab G
Lab H
Lab I
Lab J
Lab K
Lab L
Lab M
Combined