Online supplementary file 2. Key quality issues identified in included studies

- **Common data inconsistencies:** numbers not matching across the manuscript (age and gender, area under the curve (AUC), numbers of cases and controls), percentages in tables, numbers in text and tables not matching, overlapping age groups (e.g. 40-65 and ≥65).

- **Inconsistencies when describing phases/stages:** some papers reporting on more than one phase only had measures for different phases combined (not eligible). Others did not describe which stage was used to calculate the measures (if there was uncertainty, we did not extract the data). Others used the same cases and controls for different stages (not eligible). Finally, some studies had several measures of performance for the discovery phase (not eligible), but only a few for the validation phase.

- **Cut-off points:** several papers used a range of cut-off points (often not giving the rationale for them). Because different cut-off points performed better for different subgroups, it was not possible to choose the "optimal" cut-off point when reporting the findings.

- **Measures of performance:** papers reported on sensitivity but not on specificity, on negative predictive values but not on positive predictive values (or vice versa). Many only provided measures for a proportion of the biomarkers they proposed to investigate (this was not always justified). A confusion matrix was only provided by two studies.

- **Combined vs individual measures:** studies reporting on panels did not always provide individual measures of performance for each biomarker. When they did so, they often did not provide the same measures given for the panel. For example, data for the panel included sensitivity, specificity, and AUC, but only AUC was available for each individual biomarker.

- **Inconsistency in the use of diagnostic accuracy terms:** at times, expression was described as being the same as sensitivity, or diagnostic performance, or "predictive probability". It was not always clear whether the used terms were synonyms for our measures of interest; these terms included "positive detection rates", "positivity rates", "positive rates", and "detection values". One paper reported measures for sensitivity and specificity, and then swapped the values for each of these in a separate section in the paper. When it was not clear that we could extract the data we were looking for, we did not extract anything.

- **Characteristics of cases/controls:** At least 25% of studies had missing/incomplete information for age and sex of cases and controls. Data on age was not always clear (i.e. whether they were means or medians, ranges or interquartile ranges). At times, these data were only available for the full cohort, for different phases combined, according to different age groups, health conditions, or levels of biomarker expression.

- **Data availability:** At times, information described as being in supplementary data could not be found, or supplementary data could not be accessed. Some papers provided graphs for AUC but did not give any numbers. Furthermore, it was difficult to understand what to extract when several statistical analyses were carried out (and these were not clearly described).

- **Included biomarkers:** sometimes the title and abstract mentioned some biomarkers, methods mentioned others, and results added even more biomarkers (or did not report on previously mentioned biomarkers). Careful reading was required at all times.