Foundations for Studying Clinical Workflow: Development of a Composite Inter-Observer Reliability Assessment for Workflow Time Studies

Marcelo Lopetegui, MD, MS[1], Po-Yin Yen, RN, PhD[2], Peter Embi, MD, MS[3], Philip Payne, PhD[2]

[1] Instituto de Ciencias e Innovación en Medicina, Facultad de Medicina Clínica Alemana, Universidad del Desarrollo, Chile; [2] Institute for Informatics, Washington University School of Medicine, St. Louis, MO, [3] Regenstrief Institute, Indiana University School of Medicine, Indianapolis, IN, USA;

Abstract

The ability to understand and measure the complexity of clinical workflow provides hospital managers and researchers with the necessary knowledge to assess some of the most critical issues in healthcare. Given the protagonist role of workflow time studies on influencing decision makers, major efforts are being conducted to address existing methodological inconsistencies of the technique. Among major concerns, the lack of a standardized methodology to ensure the reliability of human observers stands as a priority. In this paper, we highlight the limitations of the current Inter-Observer Reliability Assessments, and propose a novel composite score to systematically conduct them. The composite score is composed of a) the overall agreement based on Kappa that evaluates the naming agreement on virtually created one-seconds tasks, providing a global assessment of the agreement over time, b) a naming agreement based on Kappa, requiring an observation pairing approach based on time-overlap, c) a duration agreement based on the concordance correlation coefficient, that provides means to evaluate the correlation concerning tasks duration, d) a timing agreement, based on descriptive statistics of the gaps between timestamps of same-task classes, and e) a sequence agreement based on the Needleman-Wunsch sequence alignment algorithm. We hereby provide a first step towards standardized reliability reporting in workflow time studies. This new composite IORA protocol is intended to empower workflow researchers with a standardized and comprehensive method for validating observers' reliability and, in turn, the validity of their data and results.

Background and Rationale.

Time-motion studies (TMS) have been widely adopted in biomedicine to study workflow[1]. Among the various approaches available, "workflow time studies" provide the most exhaustive approach for understanding clinical workflow[1]. In this technique, observers continuously follow a subject for a predefined period of time and record tasks as they occur, producing a sequence of time-stamped tasks[2,3]. This approach allows observers to track unexpected instances of tasks, accounting for task fragmentation, interruptions, and the real-world variability of clinical workflow. Although the introduction of electronic time capture tools has facilitated the recording process by allowing observers to direct their attention on the subjects being studied[4], there are still concerns of overburdened observers due to the complexity of the data capture process and its effect on data quality.

Like any other method requiring a human data collector, workflow studies are subject to variability and error in the data-capture process. Therefore, well-designed studies often conduct inter-observer reliability assessments (IORA) in order to acknowledge the amount of error introduced to the study due to the inconsistency among data collectors. Although IORA are essential for the overall validity and reliability of a study's results, IORA are not a common practice in most clinical workflow TMS[5]. IORA are systematically underreported and, when conducted, no specifications on the implementation are provided, raising concerns on the data quality produced by these TMS.

When attempting to conduct IORA in workflow time studies, researchers are faced with three major problems, including a) variability of the observable entity, b) multi-dimensionality of workflow analysis, and c) pairing observations.

a) Variability of the observable entity.

The most-cited method, the Kappa coefficient (κ), is a landmark in the development of agreement theory and one of the most-used methods to assess reliability in the biomedical sciences[6]. However, unlike most inter-rater reliability assessments (i.e. two radiologists diagnosing x-rays), in TMS the assessment is conducted on an entity that is variable and non-constant over time: clinical workflow can be different for every combination of patient, clinician

and location. Thus, IORA in workflow time studies must be conducted by having two[a] independent observers conduct a synchronous observation, following the same subject at the same time, and without interchanging data or thoughts until the observation is concluded. In view of the nature of clinical workflow, it is impractical to attempt IORA with more than two simultaneous observers due to the disruption of the workflow and an augmented Hawthorne effect caused by space constraints.

b) Multidimensionality of workflow analysis.

Besides the difficulties involved with conducting synchronous IORA sessions, the data schema resulting from these IORA sessions (Figure 1) raises concerns about the suitability of the described statistical methods, and several questions arise. Are workflow researchers analyzing the reliability of identifying the occurrence of relevant tasks? Or are they assessing the agreement in naming those tasks? Or the duration of the tasks? Most studies assess agreement in naming tasks, which results in only a partial assessment of reliability, by focusing only on one dimension of the workflow data and thereby losing the ability to identify other types of errors, like a late start of an observed event.


```
{
  IORA_session: {
    startTime: "04:15:22 pm",
    endTime: "04:30:54 pm",
    subjectName: "Bob",
    location: "Clinic A",
    observerGoldStandard: {
      observerName: "Eve",
      observationData: [
          { taskName : "task 2", startTime: "04:15:40 pm"},
          { taskName : "task 6", startTime: "04:17:55 pm"},
          { taskName : "task 2", startTime: "04:21:00 pm"},
          { taskName : "other", startTime: "04:22:08 pm"},
          { taskName : "task 5", startTime: "04:22:59 pm"},
          { taskName : "task 4", startTime: "04:25:14 pm"},
          { taskName : "task 4", startTime: "04:30:00 pm"}
        ]
      },
      observerTrainee: {
        observerName: "Alice",
        observationData: [
          { taskName : "task 2", startTime: "04:15:43 pm"},
          { taskName : "task 6", startTime: "04:18:00 pm"},
          { taskName : "task 3", startTime: "04:21:00 pm"},
          { taskName : "other", startTime: "04:22:10 pm"},
          { taskName : "task 4", startTime: "04:25:17 pm"}
        ]
      }
    }
}
```


Figure 1. JavaScript Object Notation describing an IORA session of "Eve" and "Alice" following nurse "Bob" simultaneously, and recording tasks independently to assess their reliability. It is common to record irrelevant tasks as "other" (i.e. when the subject is performing tasks not relevant to the study). This helps the researcher to establish a total time allocation for an activity including the time devoted to tasks not specified.

c) Pairing observations

In a typical implementation of Cohen's kappa for biomedical research, two observers need to "name" (categorize) a predefined number of "items" (petri dishes, tissue samples, etc.). In contrast, in workflow time studies observers must first recognize the existence of the "items" (detecting that a task is happening), name them, and accurately time-stamp them. This usually results in a different number of observations being recorded by independent observers (one observer identifies multiple tasks while the other observer only identifies one, like the case presented in Figure 1). Which tasks are considered for the paired analysis? And how are they paired?

Problem Statement.

As workflow researchers, we are interested in every aspect of clinical workflow: the sequence (order in which tasks occur), time of occurrence (time when tasks occur), count (number of task occurrences), and duration of tasks required to accomplish an activity or goal. The multi-dimensionality of the data produced by workflow time studies provides the required information for such comprehensive analyses. Thus, each aspect should be taken into account when conducting inter-observer reliability to ensure that we are maintaining the integrity of the data captured.

In a previous review, only 6% of TMS were aware of this issue and, in an initial effort towards conducting a more comprehensive IORA, they attempted to use a combination of two methods[7,8] (intra-class correlation for time and Kappa for categorization).

Given the current state of the art and the limitations encountered, we hypothesize that a composite score for assessing each dimension would provide a more meaningful and comprehensive methodology to train observers and report reliability in workflow time studies. We aim to develop a comprehensive IORA methodology for workflow time studies, contributing to advancing the standardization of workflow research.

[a] Usually, one of these observers is highly knowledgeable in the environment being studied and participates in the tasks definitions (specifying start and end milestones for each observable task), and is considered to be the gold standard to train less-experienced observers. There are no guidelines or consensus on how to define the gold standard.

Methods.

Research questions involving clinical workflow focus on different aspects of the workflow itself. Questions include: How much time do subjects devote to a task? What do subjects do at any given time? What is the duration of specific instances of tasks? What is the sequence of tasks required to complete an activity? When does a given task occur? Each of these questions requires data precision in different dimensions of the observations. Thus, an IORA should assure the reliability of identifying tasks, including *naming* and *timing* those tasks (the agreement on the start time and the duration of the task), and the *sequence* of tasks being captured.

Given the common use of Kappa in TMS, it could be considered as a sufficient, but limited, approach for assessing IORA in workflow time studies. Kappa provides a meaningful and interpretable score for the dimension being assessed, but can only assess one dimension, which opens a chance for bias and misleading IORA scores, jeopardizing the validity of study results.

To the best of our knowledge, the pairing process of the observations in workflow time studies has never been fully described in the literature. This is not trivial, because a) given that each observer can create a different number of observations during the activity, different pairing permutations are possible; and b) depending on the pairing approach, Kappa is measuring different aspects of the data.

Figure 2 describes two proposed computationally achievable pairing methods to prepare the data for Kappa. The first approach artificially breaks the observation into atomic events of one second of duration. This allows us to calculate an agreement based on the number of seconds that both observers are recording the same task name. We named this approach "proportion-kappa" (PK). In the second approach, researchers must arbitrarily match tasks from both observers. It might seem intuitive to perform the matching based on the order of the tasks; however, it fails with insertions/omission scenarios (when the trainee records more than one task for any given task from the gold standard,

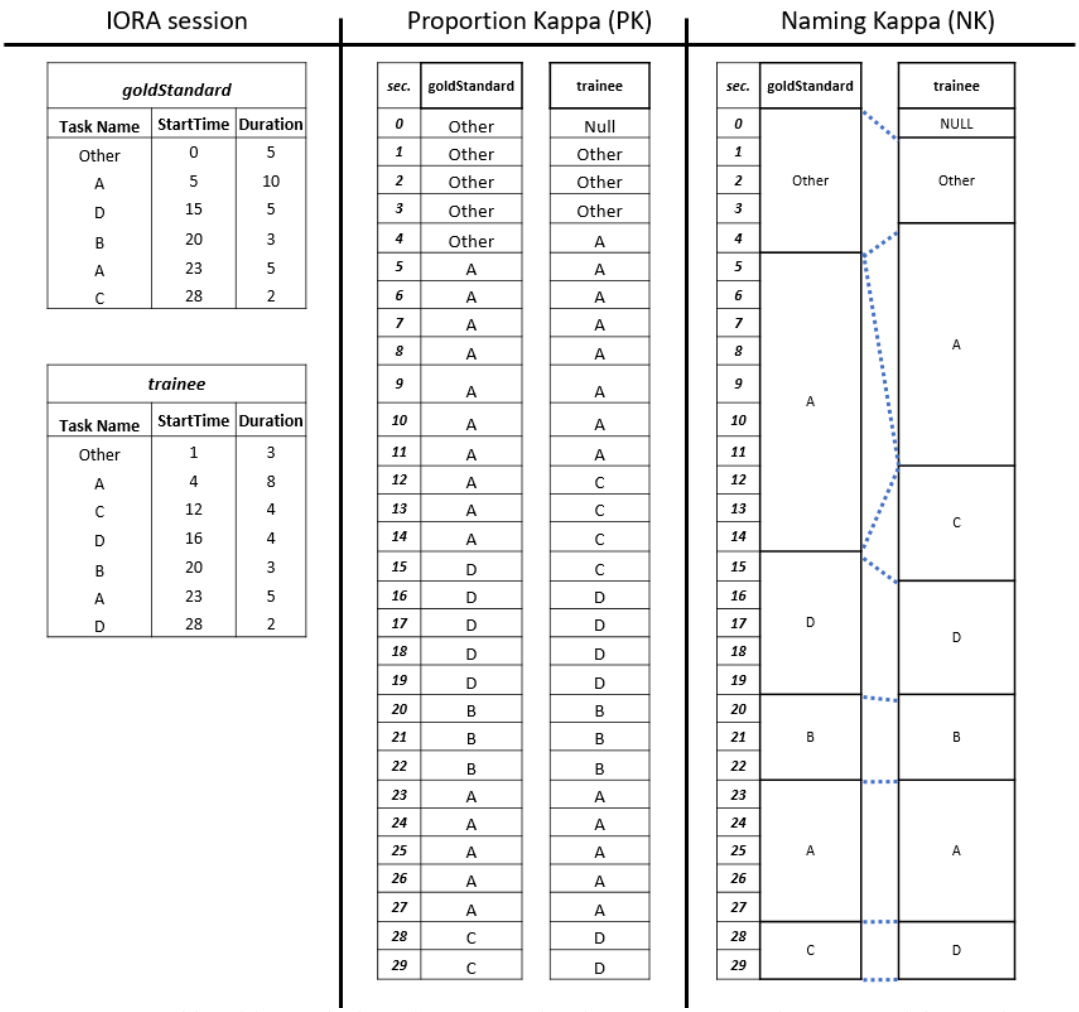

Figure 2: Computable pairing methods. Left: IORA session data. Center: Proportion Kappa. Right: Naming Kappa

or vice versa). Even if both observers end up with the same number of tasks recorded, it might be due to one of them creating a balanced number of insertions and omissions. To overcome these complex scenarios, the desired task matching is achieved by pairing any given task to the task it overlaps the most (time-wise). This approach seems more natural, since it compares each task as a unit, while being flexible with expected minor time-gaps in tasks' start times. Once pairs are created, researchers can evaluate the agreement on naming those tasks and/or evaluate the duration of the matched pairs. We called this approach "naming-kappa" (NK).

Although PK measures the agreement on *naming* artificially-created one-second tasks, it is technically evaluating the agreement on the number of seconds sharing the same task name: the overall length of time observers agree it took to complete a task. On the other hand, although in NK we are matching based on time (overlap), we are technically measuring the agreement on naming.

In both approaches, the Kappa coefficient accurately describes proportion agreement or naming agreement. But using only one dimension IORA, is insufficient to assure data quality. For example, PK could serve as a valid IORA technique if the research question focuses on overall time devoted to specific tasks. However, PK fails if the research question concerns the tasks' frequency or the average duration of tasks required to complete the activity. Furthermore, PK does not consider the task frequency, and is sensitive to tasks' duration, which allows bias if there are long idle times during the observation (captured as "others"). Thus, even if observers fail at identifying or timing tasks of interest, as long as they agree in "other" tasks, they might reach a good reliability (although Kappa contributes to reduce this bias compared to percentage agreement alone). On the other hand, NK completely disregards tasks' duration and is non-sensitive to idle time, but is insufficient if the research question focuses on "how much time subjects devote to a given task", since the pairing does not consider precise start times or precise durations.

Therefore, we designed a composite score comprising five metrics as described below, the first two of which use the Kappa statistic.

Metric 1: Proportion-Kappa (PK).

PK provides an appropriate reliability assessment to evaluate the agreement over time devoted to specific tasks. The result of this agreement provides a corrected estimate of the proportion of time two observers agree on the task name, which ensures data validity for studies focusing on time devoted to specific tasks. The abovementioned limitations of PK are tasks' frequency and overweighting idle time, which should be compensated by other types of agreement assessments.

Metric 2: Naming-Kappa (NK).

NK provides a means to analyze the agreement on naming tasks recorded by the observers, weighing each record equally, and not being influenced by the duration of the task. This balances the concern of overweighting long duration tasks on the agreement introduced by the PK method.

Metric 3: Duration Concordance Correlation Coefficient (D-CCC).

The only methods used to assess reliability of tasks' durations in published TMS are the Bland-Altman and the intra-class correlation coefficient[5]. Although the intra-class correlation coefficient stands as a more suitable approach to assess consistency of the tasks' durations compared to Pearson's correlation, like Bland-Altman, it assumes that the true value is unknown, thus only measuring the similarity among the measurements. We propose the use of a more suitable measure of agreement concerning continuous variables for non-exchangeable observers (i.e. having a gold-standard): the concordance correlation coefficient (CCC).[9] In basic terms, the CCC evaluates the degree to which pairs of observations fall on the 45° line through the origin. It is composed by the Pearson correlation coefficient (measuring how far each observation deviates from the best-fit line), and a bias correction factor (measuring how far the best-fit line deviates from the 45° line through the origin)[10]. It has been described as a robust test on as few as 10 pairs of data, improving the suitability for short observation periods. The incorporation of this method into our IORA protocol would provide the assessment of data validity for research questions concerning individual tasks' durations.

In order to implement the D-CCC, paired tasks are also required. Tasks need to maintain duration data, thus the PK second-by-second approach is inappropriate. The "most-overlapping task" pairing approach used in NK preserves tasks' duration. However, in this case, we are only interested in tasks where observers agree on the name, having in mind that the disagreement in naming is already being penalized in NK. Also, in the insertion or omission scenarios, the NK pairing approach creates multiple pairs of the longest task with the overlapping insertions. However, this metric attempts to provide an assessment of data for research questions concerning individual tasks' duration: hence each task should be paired only once. Once pairs are created with the NK "most-overlapping" approach, all pairs of tasks with non-matching names are dropped (penalized in NK). Where insertions/omissions occur, the gold

standard task is only paired to the task from the insertion block with the same name. If more than one task with the same name exists in the insertion block, the first instance of the task is matched (see Figure 3). Although we feel inclined to add the duration of all tasks with the same name from the insertion block and compare them against the long overlapping task, we consider that the PK pairing is already doing so. In this case, we are interested in the duration of each instance of a task. Thus, only one task from the insertion block should be paired and evaluated. Although this makes the metric very sensitive for insertion/omission scenarios, it provides an assessment of the impact of the insertion (how early it happened relative to the expected duration of the stopped task).

Metric 4: Sequence-Needleman-Wunsch (S-NW).

In order to assess reliability considering the sequence of tasks recorded, regardless of their start time and duration, we exploited the transdisciplinary nature of informatics. Sequence comparison is the most important primitive operation in computational biology, serving as a basis for many other, more complex, manipulations[11]. In the bioinformatics arena, assessing sequence alignment of similar strings is a common need, with existing working solutions that provide optimal alignments for similar DNA strings, usually complemented with a similarity score. This is used to study phylogenetic trees, gene mutations, or to reconstruct full DNA sequences based on a master string (among many other uses). Thus, the arrays of data produced by IORA sessions could be represented as two DNA strings. By removing the time and duration dimensions of the observation (taken care of in other scores), each task would represent a "nucleotide," thus an observation output could be thought of as a sequence of nucleotides, with insertions left unpaired and penalized, as a nucleotide insertion would be in a DNA alignment.

We propose the use of the Needleman-Wunsch algorithm: one of the most-used global comparison algorithms that relies on dynamic programming to compute the similarity between two sequences[12]. First, a similarity matrix is created, based on arbitrary scores for each match, mismatch, or gap. Then, the optimal alignment is reconstructed based on the similarity matrix, and a score is derived from a pre-computed matrix.

The first row and column of the matrix are initialized with multiples of the gap penalty: they represent the score received if only gaps were created (the alignment if the other string is empty). Then, the other cells are computed based on the three adjacent cells already populated [*gs(i*-1), *tr(j)*], [*gs(i-1),tr(j-1)*] and [*gs(i),tr(j-1)*]. This can be thought of as the three options to match the current "nucleotides-tasks":

a) Align *gs(i)* with *tr(j*-1) and pair a space with *tr(j)* [*tr(j)* being an insertion].

b) Align *gs(i*-1) and *tr(j*-1) and pair *gs(i)* with *tr(j)* [*gs(i),tr(j)* being either a match or mismatch].

c) Align *gs(i*-1) with *tr(j)*and pair *gs(i)* with a space [*gs(i)* being an insertion].

Thus, the score to populate on the cell [tr(i),gs(j)] is given by the equation[11]:

$$[\mathrm{tr(i)}, \mathrm{gs(j)}] = \max \begin{Bmatrix} [\mathrm{tr(i)}, \mathrm{gs(j-1)}] - \mathrm{G} \\ [\mathrm{tr(i-1)}, \mathrm{gs(j-1)}] + \mathrm{P}[\mathrm{tr(i)}, \mathrm{gs(j)}] \\ [\mathrm{tr(i-1)}, \mathrm{gs(j)}] - \mathrm{G} \end{Bmatrix}$$

Where G corresponds to the gap penalty, and P corresponds to either the "match" score or the "mismatch" score. Once the similarity matrix is computed, a recursive algorithm computes the optimal alignment. We empirically set the scores of -1 for the gap penalty, 0 for a mismatch, and +3 for a match.

In order to produce a meaningful score to be interpreted by the workflow researcher, we propose to transform linearly the theoretical maximum and minimum alignment scores to 0-1 boundaries. The theoretical maximum corresponds to "P-match" times the number of tasks of the longest string, while the theoretical minimum corresponds to negative difference of tasks between the two strings (only mismatches for the pairs, plus gap penalties for the remainder).

Metric 5: Timing Agreement (TA).

The timing agreement represents the degree of synchronization by both observers in recording the start time of a task. Given the lack of formal methods to assess this, we propose to report the median time-gap of paired tasks from the D-CCC assessment, grouped by task: the median time (in seconds) of the difference between start times of paired tasks (positive if the trainee is behind the gold standard, negative if the trainee is ahead). We believe this metric might not play a useful role in reporting agreement as a global score since the interpretation may vary depending on the variance of individual tasks (a median time-gap of 2 might be representing being 10 seconds ahead on task "A", and late 2 seconds on task "B" and 12 seconds on task "C"). However, this metric might be of great contribution to train observers and evaluate tasks based on the magnitude and the sign of the time-gap. We suggest to report the

median time-gap by task, accompanied by all the time-gaps for that task. This will provide a quick assessment of any potential difficulties on identifying start times for specific tasks.

In this paper, we synthesized a small data set of an IORA session, and used the proposed composite IORA scores to demonstrate their meaning and interpretation. Implementation of the new methods proposed were based on data set shown in Figure 3.

goldStandard		
taskName	**startTime**	**duration**
Other	0	5
A	5	10
D	15	5
B	20	3
A	23	5
C	28	18
A	46	10
B	56	7

trainee		
taskName	**startTime**	**duration**
Other	1	4
A	5	11
C	16	5
B	21	3
A	24	5
C	29	17
A	46	5
B	51	3
A	54	2
B	56	6

Figure 3: Synthesized scenario depicting a brief IORA session with a trained observer (goldStandard) and a trainee.

Results.

Application of the composite IORA protocol.
Proportion-Kappa (PK).

For the synthesized scenario (Figure 3), this score results in a Kappa coefficient of 0.744 and 82% agreement. This demonstrates an overall substantial agreement between the observers, which is particularly relevant if the research question focuses on how much time subjects devote to any given task.

Naming-Kappa (NK).

This score results in a Kappa coefficient of 0.718 and 80% agreement, demonstrating a substantial agreement on naming tasks between the observers, which is of interest if the research question relates to "what do subjects do at any given time?".

Duration Concordance Correlation Coefficient (D-CCC).

In the synthesized scenario, the trainee performed well, confusing task D with C, and creating an insertion of B while A was described by the gold standard. The "most overlapping" pairing approach produced 10 pairs of tasks. Following the proposed modifications, 3 pairs are dropped: D(5)-C(5) since observers failed at naming them, and A(10)-B(3) and A(10)-A(2) since they correspond to the left-over after pairing the insertion (see Figure 4).

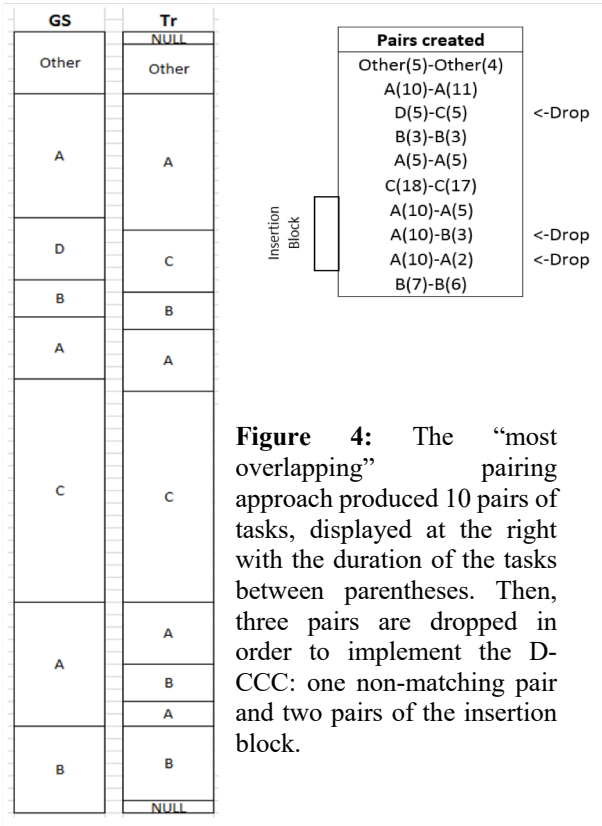

Figure 4: The "most overlapping" pairing approach produced 10 pairs of tasks, displayed at the right with the duration of the tasks between parentheses. Then, three pairs are dropped in order to implement the D-CCC: one non-matching pair and two pairs of the insertion block.

Thus, using Lin's[9] concordance correlation coefficient in STATA (*concor* command) results in 0.906. Unlike Kappa, there is no specification in the literature on a descriptive scale for the degree of agreement achieved with Lin's CCC, only existing a lower bound for an acceptable CCC of 0.75 (not universally accepted)[13].

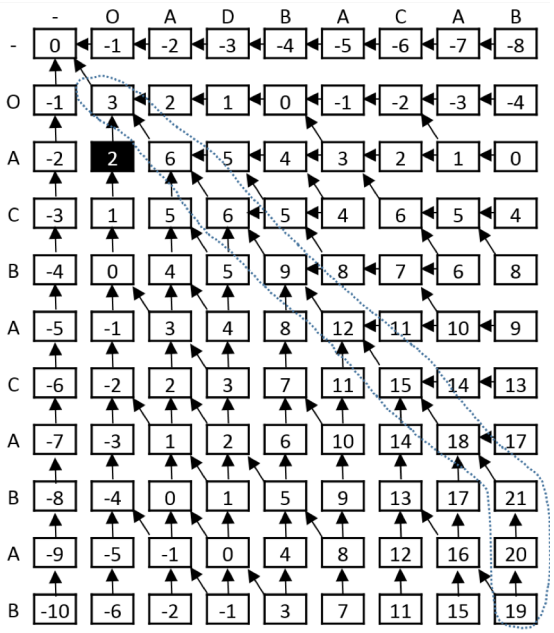

Figure 5: Similarity matrix for the synthesized scenario. The trainee string is on the left, and the gold standard is on the top. To compute the score for the highlighted cell, the three options are evaluated for the highest score:

1. Create a gap on the trainee string leaving A unpaired, producing a score of 2 [3 from OO plus a gap penalty].

2. Align PK producing a score of -1 [-1 from a previous O- gap plus a mismatch score].

3. Create a gap on the gold-standard string leaving O unpaired, producing a score of -3 [-2 from the previous 2 gaps, plus a gap penalty].

Once the highest score is computed (2, creating a gap on the trainee string), an arrow is drawn to show where the highest score comes from. The dotted line shows the final highest score path, which is used to create the following optimal alignment:

OADBACAB* *
OACBACABAB

Sequence-Needleman-Wunsch (S-NW).

Figure 5 shows the similarity matrix corresponding to the synthesized scenario, based on the strings gs= O-A-D-B-A-C-A-B and tr= O-A-C-B-A-C-A-B-A-B, for the gold standard and the trainee observer, respectively.

In the synthesized scenario, the computed maximum score is 19 (Figure 5). The theoretical maximum corresponds to 3x10, while the theoretical minimum corresponds to -2. Thus, the linear transformed final score for the current example would be 0.656 ([19+2]/ [30+2]). This is a quantified interpretation for a visual inspection: 7 out of the 10 tasks were correctly matched, one mismatch, and two insertions.

Timing Agreement (TA).

In the synthesized scenario, accuracy for timestamping each task was: 0 seconds for A (0;0;1), 0.5 seconds for B (0;1), and 1 second for C (1), based on the pairing achieved using the D-CCC pairing shown in Figure 4.

Expanding the impact of a meaningful IORA: contributing to an optimized training feedback.

The composite IORA metric intends to empower researchers with a means of ensuring reliability and validity of the data in a quantitative, comparable, and systematic manner. However, there is room for a further contribution: besides informing the researchers "how good" or "how bad" observers are doing based on the metrics calculated, we thought to include visualizations that would help better understand how the scores are calculated and provide an answer to "what" and "when" they disagree. For example, both Kappa metrics, the contingency table highlights discrepancies (any cell not in the diagonal). Also, a side-to-side visualization of both observations permits a global qualitative IORA, which, although not useful for reporting purposes, might be very helpful in training the observers, highlighting times and tasks that produce discrepancies. Similar to the clinical workflow analysis tool (CWAt)[14] and the sample report of Mache's tool[15], we developed a graphic representation for an IORA session (Figure 6).

Finally, a simple scatter plot of the durations assessed by the D-CCC helps identify where the timing discrepancies are occurring. The scatter plot depicts the concordance between the duration of each task captured by the gold standard and the trainee. Everything below the plot's purple diagonal represents a task captured by the trainee having a logged duration that is shorter than the gold standard; anything above represents a task captured by the trainee having a logged duration that is longer than the gold standard (Figure 6).

Discussion.

As previously described, studying workflow goes beyond simply calculating duration of events. In order to fully understand workflow, analyses focus on *what* happens, *when,* and for *how long*, and usually analyze the order in which those tasks occur (the sequence required to complete an activity). Having evidenced the limitations of using a single statistical test to ensure reliability of the observation, and given the lack of a single exhaustive test to accomplish comprehensive IORA, we postulate that using multiple instances of existing tests, one for each dimension, might provide IORA in a more meaningful manner.

Given our experience designing and conducting time-motion studies, achieving scores consistently high, like the one in our synthesized scenario, is not common, and reliability efforts shouldn't pursue optimal scores in every dimension. Although we recommend assessing every dimension, agreement goals should be directed to the specific research question under study.

Besides the contribution as a reporting tool, our proposed IORA stands as a major contribution to the standardization of the observers' training, which is usually only reported as total hours of training. With our IORA

Figure 6: Composite score example. Interpretation for this report: Overall, both observers have an almost perfect PK agreement (relevant if data intends to represent the overall time devoted to any given task). Strong agreement for NK. D-CCC: satisfactory (not directly interpretable, most paired tasks on 45-degree line). Sequence agreement: satisfactory (2 insertions out of 20 recordings). Training recommendation based on report: Review specifications on when "treatment" (blue) ends [avoid trainee early stop, that created the insertion]. Review when "education" (green) begins [improve timing agreement]

approach, researchers have powerful means to direct efforts on observers' training, by both visually examining details of the observation and by evaluating agreements trends over time in the training period. Thus, instead of reporting hours of observers' training alone, they will be able to report training until X agreement is achieved in Y dimension (the one of more relevance for the research question).

These metrics were implemented and calculated to produce meaningful and interpretable results, however the D-CCC and S-NW haven't been used before, and thus interpretations and extrapolation of the scores aren't straightforward. Although range and directionality of the scores are logical, further intensive testing and validation studies are required to define meaningful cut points. This will allow us to study IORA scores over long periods of observations (assessing observers' fatigue), study IORA at different points of the study, not just the training portion (assessing observers' drift), and many other interesting related issues such as the effect of the initial number of tasks on IORA.

Limitations.

The PK, NK, and D-CCC are statistical estimates, and confidence intervals for the estimates are dependent on the sample size. For the PK, a large number of pairs is usually created (number of seconds in the observation), and narrow confidence intervals can be achieved. However, for the NK, the number of pairs corresponds to the number of tasks recorded and, for the D-CCC, to the number of tasks recorded minus the discrepant tasks. Thus, when the number of tasks recorded is small, the NK and D-CCC are bounded by large confidence intervals, jeopardizing the significance and meaningfulness of the scores obtained.

The S-NW and the TA interpretation are not affected by sample size. The S-NW provides a normalized arithmetic score, interpretable independent of the sample size (number of tasks recorded in the observation): insertions are more heavily penalized in shorter observations and *vice-versa*. Likewise, the TA is a composite descriptive statistic (the median and quartiles), interpretable regardless of the sample size.

We recommend that, if only a small number of tasks are recorded during an observation, individual scores should be carefully considered. We believe the PK might be the most representative score to report in those cases.

Potential improvements to our proposed IORA include the use of weighted Kappa to calculate the agreements. Based on an ontology of commonly used tasks in healthcare, mismatches could be penalized differently depending on how similar or different the two concepts are.

Despite the limitations of a percentage agreement, we believe that, since observers are trained before the IORA session and little guessing is likely to exist, providing the Kappa statistics accompanied by the percentage agreements might be of aid in interpreting the agreement[16].

Conclusion.

Confirming that the establishment of IORA protocols and guidelines are a priority in validating continuous observation TMS, we evidenced the limitations of using a single metric IORA, and limitations related to the data manipulation required to implement it. We confirmed that assessing inter-observer reliability with Kappa, although partially useful, misrepresents and overestimates the real agreement, since it only focuses on one dimension of a multi-dimensional data set.

We proposed a composite IORA protocol, including a set of methods to assess each relevant dimension in workflow time studies: a proportion agreement, naming agreement, duration agreement, sequence agreement, and timing agreement. We demonstrated our proposed IORA methodology in a synthesized scenario.

We hereby provided a first step towards a standardized reliability reporting in workflow time studies. This new composite IORA protocol is intended to empower workflow researchers with a standardized and comprehensive method for validating observers' reliability and, in turn, the validity and representativeness of the data collected.

References.

1. Lopetegui M, Yen P-Y, Lai A, Jeffries J, Embi P, Payne P. Time Motion Studies in Healthcare: What are we talking about? J Biomed Inform. 2014 Mar;
2. Mache S, Busch D, Vitzthum K, Kusma B, Klapp BF, Groneberg DA. Cardiologists' workflow in small to medium-sized German hospitals: an observational work analysis. J Cardiovasc Med (Hagerstown). 2011 Jul;12(7):475–81.
3. Westbrook JI, Ampt A, Kearney L, Rob MI. All in a day's work: an observational study to quantify how and with whom doctors on hospital wards spend their time. Med J Aust. 2008 May 5;188(9):506–9.
4. Lopetegui M, Yen P, Lai AM, Embi PJ, Payne PRO. Time Capture Tool (TimeCaT): Development of a

Comprehensive Application to Support Data Capture for Time Motion Studies . AMIA Annu Symp Proc. 2012;
5. Lopetegui MA, Bai S, Yen P-Y, Lai A, Embi P, Payne PRO. Inter-observer reliability assessments in time motion studies: the foundation for meaningful clinical workflow analysis. AMIA Annu Symp Proc. 2013 Jan;2013:889–96.
6. Yang Z, Zhou M. Kappa statistic for clustered matched-pair data. Stat Med. 2014 Feb 16;
7. Cady R, Finkelstein S, Lindgren B, Robiner W, Lindquist R, VanWormer A, et al. Exploring the translational impact of a home telemonitoring intervention using time-motion study. Telemed J E Health. 2010 Jun;16(5):576–84.
8. Lindquist R, VanWormer A, Lindgren B, MacMahon K, Robiner W, Finkelstein S. Time-motion analysis of research nurse activities in a lung transplant home monitoring study. Prog Transplant. 2011 Sep;21(3):190–9.
9. Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics. 1989 Mar;45(1):255–68.
10. Concordance correlation coefficient [Internet]. Available from: http://www.medcalc.org/manual/concordance.php
11. Setubal, Meidanis. Introduction to Computational Molecular Biology. Boston: PWS Publishing Company; 1997.
12. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970 Mar;48(3):443–53.
13. Choudhary P, Nagaraja H. Measuring agreement in method comparison studies -- A review. In: Advances in Ranking and Selection, Multiple Comparisons, and Reliability. Boston: Birkhauser; 2005. p. 215–44.
14. Hanauer DA, Zheng K. Detecting workflow changes after a CPOE implementation: a sequential pattern analysis approach. AMIA Annu Symp Proc. 2008 Jan;963.
15. Mache S, Scutaru C, Vitzthum K, Gerber A, Quarcoo D, Welte T, et al. Development and evaluation of a computer-based medical work assessment programme. J Occup Med Toxicol. 2008 Jan;3:35.
16. McHugh ML. Interrater reliability: the kappa statistic. Biochem medica. 2012 Jan;22(3):276–82.