Chapter 12
Understanding Clinical Workflow Through Direct Continuous Observation: Addressing the Unique Statistical Challenges

Scott R. Walter, William T. M. Dunsmuir, Magdalena Z. Raban, and Johanna I. Westbrook

12.1 Background

12.1.1 General Introduction

The nature of healthcare as a dynamic human process occurring within complex socio-technical systems means that there is no unique or standard way to examine its inner workings. Rather, a range of observational methods drawn from multiple disciplines have been used to study workflow *in situ* (McCurdie et al. 2017). Areview of methods used to study and model workflow across different industries, including healthcare, identified qualitative approaches such as ethnographic observation and interviews, along with quantitative methods including structured or timed observations, and surveys (Unertl et al. 2010).

Analogous to timed observations, the term *time and motion* is applied in many studies of workflow in healthcare. This umbrella term encompasses a range of methods and designs with the common feature of directly observing an individual's activities and recording aspects of that action, usually in a quantitative way. Zheng et al. (2011) reviewed time and motion studies used to assess the effect of interventions, especially technology-related interventions, on workflow in healthcare settings. From their synthesis, they developed the STAMP checklist (Suggested Time and Motion Procedures) to promote consistency in design, conduct and reporting of time and motion studies. Lopetegui et al. (2014) took this theme

S. R. Walter (✉) · M. Z. Raban · J. I. Westbrook
Centre for Health Systems and Safety Research, Australian Institute of Health Innovation, Faculty of Medicine and Health Sciences, Macquarie University, Sydney, NSW, Australia
e-mail: scott.walter@mq.edu.au; johanna.westbrook@mq.edu.au

W. T. M. Dunsmuir
Department of Statistics, School of Mathematics and Statistics, University of New South Wales, Sydney, NSW, Australia

K. Zheng et al. (eds.), *Cognitive Informatics*, Health Informatics,
https://doi.org/10.1007/978-3-030-16916-9_12

further by reviewing the distinct methods used in healthcare under the banner of 'time and motion studies'. The many variations they identified were categorized into three groups: those involving external observers shadowing participants, those using information self-reported by participants, and those that employed automated data recording such as GPS devices or accelerometers. Of the first type, they identified a method employing continuous observation and coined the term *workflow time study* to describe it as a distinct but increasingly common approach. This method constituted 26% of all time and motion studies reviewed, and over 60% of all studies that involved continuous observation by an external observer. Also, the proportion of studies employing continuous observation was noted to have increased over the review period.

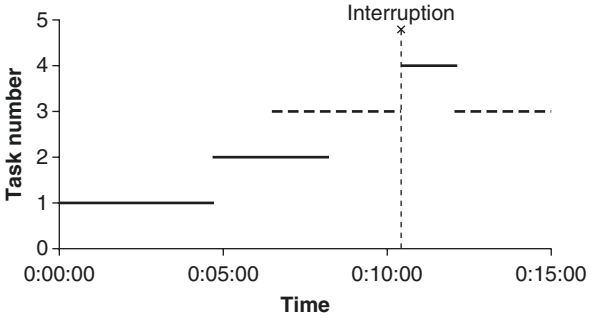
Although the workflow time study approach is one among many observational approaches, it offers many advantages over other quantitative methods, and its growing use in healthcare is a testament to this. This method itself involves observers shadowing individual clinicians and continuously recording time-stamped data about an individual's tasks and interactions (see Sect. 12.1.2 for more detail). Workflow time studies capture more of the fine-grained complexity of clinical work than methods such as work sampling, and the temporal continuity of the data forms the most complete record of an individual's workflow of any observational technique, barring audio-visual recording which is often not acceptable in a clinical environment. Workflow time studies have great potential to help us understand clinical work and workflow and can be applied to a diverse range of research questions and professional groups (Walter et al. 2015). This includes descriptive analyses that examine the way clinicians distribute their time between different tasks, between patients, between locations, and so on (Westbrook et al. 2008; Li et al. 2015; Richardson et al. 2016). It also supports assessment of the impact of interventions on workflow, such as the introduction of new technological systems, policies or practices (e.g. Georgiou et al. 2017). Furthermore, workflow time studies enable interrogation of more complex questions such as the way clinicians sequence, prioritize and interleave tasks. They can also examine associations between clinicians' work and safety-related outcomes, such as factors that contribute to errors of task omission and commission (e.g. Westbrook et al. (2018).

Capturing a more complete record of the complexity of workflow in healthcare settings is necessary to generate valid and relevant insights about everyday clinical work within a quantitative paradigm. However, this also introduces some unique methodological challenges in all aspects of the study process including design, data collection, analysis and interpretation of findings. Despite the importance of applying appropriate quantitative methods, methodology in the area is still evolving, and there is a tendency to apply conventional statistical methods to data that are inherently non-standard. This chapter examines the critical quantitative and statistical challenges with which workflow time studies are confronted, including reviewing methods applied in studies to date and suggestions for methodological improvements. Many of the aspects discussed in this chapter may also be relevant to the quantitative study of workflow more generally.

12.1.2 Defining Workflow Time Studies

The original definition of workflow time studies referred to those studies involving periods of continuous observation of a participant where "the observer records the occurrence and duration of unpredicted instances of tasks, producing a data schema of time-stamped tasks, which accounts for task fragmentation, interruptions and work variability" (Lopetegui et al. 2014). There are several features that distinguish this technique from other observational methods. First, the fact that observers continuously shadow participants sets it apart from approaches such as self-reporting of work activities (Ampt et al. 2007), work sampling or multimedia recording. Second, although carrying out detailed observations over extensive periods of time has parallels with ethnography, observers in workflow time studies apply predefined categories of task attributes at the time of observation, as distinct from ethnography where grouping of types of observed action into categories or themes occurs during the analysis phase (e.g. Malhotra et al. 2007). Third, the recording of time stamped intervals for each task generates data that represents a temporally complete record of the observed activity. In other words, at every time point during observation, action is assigned to one category or another, or, equivalently, no time in the workflow is unaccounted for. This contrasts with other methods where the observer may continuously shadow the participant but may only record data at certain times or on particular activities.

The data generated by workflow time studies is essentially a set of time intervals, each defined by a start and end time, and having any number of categorical attributes such as task type, location where the task was performed, with whom it was performed, and so on. Figure 12.1 provides a simple illustration of tasks plotted over time, in addition to one possible way to represent the raw data. The intervals can be contiguous where one task ends and another begins, as between tasks 1 and 2 in the figure; or they can overlap where two types of action occur in parallel (commonly called multitasking) as with tasks 2 and 3. When intervals represent fragmen-



Task ID	Start time	End time	Type	Interrupted by
1	0:00:00	0:04:40	A	
2	0:04:40	0:08:10	A	
3	0:06:30	0:10:25	B	4
4	0:10:25	0:12:05	A	
3	0:12:05	0:15:00	B	
...	...	...	...	

Fig. 12.1 Example of four tasks observed in a workflow time study, represented as intervals on a time line and as records in a dataset

Table 12.1 Examples of dimensions and categories used in workflow time studies

Dimension	Category
Task type	Direct care
	Indirect care
	Documentation
	Clinical communication
	Management communication
	Social communication
	Prescribing
	Other
With whom	Specialist (consultant)
	Fellow (registrar)
	Resident/intern
	Nurse
	Relative
	Patient
	Paramedic
	Other
	No one

tation of tasks that are suspended due to interruptions and later resumed, this can be indicated with categorical labels, as shown by the 'interrupted by' column in the figure. Some studies also augment with data from other sources such as patient load, self-reported measures or participant characteristics, in an effort to include factors at multiple system levels (see for example Westbrook et al. (2018).

The task attributes mentioned above are termed *dimensions*, each of which may have several *categories* (Westbrook and Ampt 2009). In workflow time studies, a dimension is an aspect of clinical work that is relevant to the research questions of a study. In the example in Fig. 12.1, 'type' is the main dimension which has categories 'A' and 'B'. In clinical settings, dimensions may be the type of task performed by the participant (usually the main dimension), the location where the task is performed, or with whom the participant interacts with while performing the task. In the language of quantitative analysis, dimensions can equivalently be thought of categorical variables, and the categories represent all the potential values that a variable can take on. Table 12.1 illustrates two dimensions and their categories from a study of emergency doctors in Australia (Walter et al. 2017).

12.2 Sampling Strategies

The first major methodological challenge in conducting a workflow time study is how to approach data sampling. The sampling strategy naturally depends on the study design. As it is impractical to cover the sampling strategies for all possible workflow time study designs within this chapter, we limit our discussion to the following three major study types: (1) descriptive studies that provide a snapshot of the

clinical work process, (2) intervention studies that assess change in workflow over time as a result of an intervention, and (3) association studies that aim to link aspects of clinical work to patient safety or quality of care outcomes.

One aspect of the sampling strategy that impacts all three study types is that there is a limit as to how much one observer can continuously observe without a break. However, much of health care, particularly critical care, occurs around the clock. Although in an ideal situation we may wish to observe all clinicians at all times throughout the study period, this is simply not practical. Thus, the data in workflow time studies are often collected across many separate observation sessions, wherein each session typically consists of a few hours of shadowing with a single participant. The data from these sessions are then combined together to form a collection of workflow samples on multiple participants.

The nature of clinical work varies with time-related factors: time of day, day of the week, time of year, etc. (Walter et al. 2014). It also differs between clinician roles or seniority (Westbrook et al. 2010), and between the idiosyncrasies of individuals (Walter et al. 2014). Oversampling at certain times or among certain roles can therefore influence the study results, underscoring the need for an appropriate sampling strategy to avoid biases. Descriptive studies generally aim to generate a set of samples that, when combined, are representative of clinical work in a certain setting, among a particular professional group, or during a given period of the working day. For example, Arabadzhiyska et al. (2013) studied the work of resident physicians on night shifts (10 pm to 8 am) on general hospital wards.

Generating a representative sample is usually accomplished by applying a time-based sampling scheme to collect approximately equal amounts of observation time balanced across known factors that may influence summary measures such as proportions and rates. To illustrate, the rate at which clinicians' work is interrupted is known to be higher for those who are more senior (Walter et al. 2017), during weekends (Richardson et al. 2016) and is related to workload (Weigl et al. 2012) which varies throughout the course of the day. If there is unintentional oversampling of senior clinicians, Saturdays and Sundays or busy periods, it could then inflate the interruption rate to be observed. In contrast, balancing observation time across such factors provides an interruption rate estimate that is more representative of the 'average' workflow within the study population.

Such a sampling scheme was used by Richardson et al. (2016) who conducted a descriptive study of junior physicians working on day shifts during the weekend. The study population was from a single professional group of the same seniority; and a sampling scheme was developed to ensure balance in observation hours over time of day (between 8 am and 5 pm), day of the week (Saturday and Sunday) and also over the 13-week observation period (Table 12.2).

Another major source of variation in workflow is between individuals. A study of how clinicians in three hospital settings respond to interruptions found that significant variation between individuals persisted after adjusting for many task-level and temporal factors (Walter et al. 2014). Attempting to average individual differences by balancing (as shown in Table 12.2) would mean an unrealistically large increase in required sample size and hence observation time. For example, the Richardson et al. study had 16 participants, so to observe each of them, during every time of the

Table 12.2 Sampling schedule used by Richardson et al. (2016) to study junior physicians working on day shifts over the weekend

Observation time	Saturday A Week 1, 3, 5, 7, 9, 11, 13	Sunday A Week 1, 3, 5, 7, 9, 11, 13	Saturday B Week 2,4, 6, 8, 10, 12	Sunday B Week 2,4, 6, 8, 10, 12
0800–0950	Observing			Observing
0950–1140	Resting	Observing	Observing	Resting
1140–1330	Observing	Resting	Resting	Observing
1330–1520	Resting	Observing	Observing	Resting
1520–1710	Observing	Resting	Resting	Observing
1710–1900		Observing	Observing	

day, day of the weekend and week of the study period, it would require an increase of the total observation time from 132 h to more than two thousand hours. Randomisation offers a way to average out the effects of temporal factors and individual differences with a more realistic sample size. For each observation session the participant is randomly selected, as is the time of day, day of the week, and so on. Sessions can be assigned in this way until a sufficiently large sample is attained.

In practice, it is not always possible to implement either a balanced or randomised sampling scheme exactly as planned. Finding a certain participant at a particular time can be difficult, especially in a hospital setting where staff rosters change and clinicians swap shifts at the last minute. While it is important to have a sampling plan, it may be necessary to modify it over the course of the study period to compensate for imbalances introduced by unanticipated deviations from the schedule. If logistical constraints cause the final sample to be unbalanced, it is possible to adjust for this in the analysis phase using multivariate regression. For example, to calculate the interruption rate across task type categories (as in Table 12.1) when there has been oversampling of senior clinicians, Poisson regression could be applied with the main covariate as task type, but also including, say, time of day and participant seniority as additional variables. This does not preclude the need for a sampling plan, but rather provides a way to mitigate the effects of compromised implementation of the plan.

For studies assessing the impact of an intervention using a pre-post design, an additional consideration is to use a consistent sampling strategy for each time period. While studies of this type should ideally use a control group to capture any pre-post changes not attributable to the intervention, the controls may not necessarily capture pre-post differences due to sampling. For example, if senior clinicians are oversampled post-intervention for the intervention group, but not for the control group, then the intervention effect will be muddied with sampling effects, with no completely satisfactory way to separate them during the analysis.

For association studies, the sampling priority is somewhat different as the aim is not to generate representative summary measures of workflow, but to assess statistical associations between aspects of clinical work. Where descriptive studies use a sampling strategy based on observation time, association studies build sampling around the units of analysis (tasks, events, etc.). To examine associations in an

observational study it is necessary to adjust for confounding factors (in the epidemiological parlance) to derive the least biased estimate of the association of interest, usually done through multivariate modelling. The variables generated by workflow time studies are typically categorical, so an important consideration is whether there will be sufficient outcome data in each category. Small numbers in certain categories may cause issues with model fitting, so it may be desirable to oversample certain times of day, certain professional groups, and so on, to avoid this issue. In a study by the authors (Walter et al. 2017) on physicians' response strategies for dealing with external prompts (i.e. interruptions), the original analysis plan involving both categorical outcome and covariates was not possible due to some outcome categories never occurring at the same time as certain covariate categories. This caused implausible or nonsensical model outputs for some variables even after collapsing of some categories, and an alternative analysis approach was necessary. Therefore, for association studies, the sampling strategy must necessarily be developed in parallel with dimensions and categories.

12.3 Inter-observer Reliability

A fundamental aspect of generating high quality data from observations of clinical work is to ensure consistent application of dimensions and their categories between different observers. This is often called inter-rater reliability, a term taken from psychology, although in this context we use the term inter-*observer* reliability (IOR) since we are interested in observations as a more varied set of judgements, as opposed to ratings which tend to involve assigning a single value or category at a time. The fact that workflow data recorded at task-level have time stamps, involve temporal order and feature multiple categorical attributes makes it rather complex to compare between two or more observers who are following the same participant. To date, there has been persistent use of simple methods borrowed from other contexts that are not well suited for their purpose, and this is somewhat of an 'elephant in the room' in quantitative observational studies of clinical workflow.

A range of methods have been applied in workflow time studies to assess IOR and a review of these identified seven different approaches among the 27% of studies that provided some details of their IOR assessment (Lopetegui et al. 2013). The most common was Cohen's kappa, a well-known method used in psychology to quantify the level of agreement between two or more raters assigning units to a set of categories, such as assigning exam papers to either pass or fail (Cohen 1960). In workflow time studies this approach seems to be treated as somewhat of a gold standard, while at the same time most studies gloss over the details of its application to IOR assessment (Lopetegui et al. 2013). There are several issues with kappa, and other similar measures, that mean assessments of IOR are limited at best, and may even be misleading in that high kappa scores can be achieved even though significant observer differences are present.

Table 12.3 Example data from two hypothetical observers shadowing the same participant

Observer	Task ID	Start time	End time	Task type	Performed with nurse
1	1	0:00:00	0:04:30	A	0
1	2	0:04:30	0:07:00	B	1
1	3	0:06:30	0:10:25	B	0
1	4	0:10:25	0:12:05	A	1
1	5	0:12:05	0:15:00	B	0
1	6	0:15:00	0:20:00	A	1
2	1	0:00:00	0:04:40	A	1
2	2	0:04:40	0:08:30	A	1
2	3	0:06:30	0:10:25	B	1
2	4	0:10:25	0:12:05	A	0
2	3	0:12:05	0:15:00	B	0
2	5	0:15:00	0:20:00	A	0

The first main limitation is that for time-stamped and time-ordered tasks with multivariate attributes, identifying pairs of tasks from two observers that refer to the same observed action cannot be done with any certainty. Table 12.3 shows some example data from two observers shadowing the same physician. Task 2 recorded by the first observer lasted two and a half minutes, was of task type B, was performed with a nurse, and overlapped with the next task for 30 s. In contrast, task 2 recorded by observer 2 lasted almost 4 min, was of type A, was performed with a nurse and overlapped with the next task for 2 min. Given the disagreement on several attributes, it is not possible to conclusively decide if task 2 for each observer refers to the same observed action, and to decide they *do* agree based on only some agreeing attributes introduces unreasonable assumptions, or even outright guessing.

The second main limitation is that most methods used for assessing IOR only apply to one variable at a time. This may be acceptable for descriptive studies reporting summary measures of individual variables but is likely inadequate for association studies involving multivariate analyses. In one of our prior studies (Walter et al. 2014), a reanalysis of the data collected from three hospital settings found significant observer effects in multivariate models despite high univariate IOR scores.

12.3.1 Nonparametric Hypothesis Testing for IOR Assessment

In this chapter we look at two broad approaches to addressing these limitations. The first approach compares summary measures at an aggregated level using hypothesis tests. For example, the proportion of time spent performing tasks direct care tasks could be compared between observers shadowing the same participant. This method ignores temporal order and thus does not require matching at either task or time

window level, making it applicable only for descriptive studies where reliability at such an aggregate level is sufficient. This approach assumes that the data from different observers should be the same and that any observed difference in summary measures is due to observer effects. Rather than generating an IOR score, this method provides a p-value where we hope to find a non-significant (large) value indicating no evidence of a difference in time proportions for data collected by different observers (as in Westbrook et al. (2018).

Proportions of time are the most common measure in descriptive workflow time studies, however, since these are proportions of a continuous variable they require unique methods (see Sect. 12.4.2.1 for more details). For this purpose, nonparametric resampling tests, specifically permutation tests, offer several advantages over conventional parametric options. Of the parametric tests, it is possible to aggregate the data into subgroups or clusters (e.g. by observation sessions) and to use a logistic transformation on the proportion for each group. This is appropriate where the subgroups or clusters are fixed (Warton and Hui 2011), however, in workflow time studies the choice of subgroups, such as observation sessions or individual participants, is not necessarily clear.

Permutation tests avoid the issues with distributional assumptions and sampling units. This approach involves reordering observer labels in the task-level data, cycling through all possible combinations and calculating the statistic of interest each time (such as the difference between proportions for two observers). These resampled values form the null distribution against which the actual difference can be compared. The proportion of null values more extreme than the 'true' difference provides the p-value. For large samples, the Monte Carlo permutation test uses many random shuffles of the labels to generate a p-value without having to calculate every possible label combination, thus reducing computation time. Good (2010) provides a comprehensive discussion of these methods. Applying a permutation test to the data in Table 12.3 to compare proportions of time spent on task types A and B and time spent working with a nurse yielded p-values of 0.61, 0.73 and 0.45, respectively. In other words, there was no evidence of a difference between observers in terms of time proportions.

12.3.2 Conventional IOR Measures Applied to Time Windows

The second approach addresses the time alignment issue by reformatting the task-level data into small time windows. This idea originated with Bakeman et al. (2009) who discussed applying Cohen's kappa in this way for timed-event sequential data, which is similar to workflow time study data. When comparing data from two observers shadowing the same participant, we can assume that during a given small time window they were observing the same activity, and this circumvents the issue with temporal alignment at the level of tasks described earlier in this section. Existing IOR methods, such as Cohen's kappa, can then be applied to the aligned time windows.

The time window approach then allows us to encompass the multivariate nature of data from workflow time studies. Janson and Olsson (2001) developed an IOR assessment method analogous to Cohen's kappa that is applicable to multivariate categorical data (pp. 282–283). When applied to two observers and one variable it is equivalent to Cohen's kappa, but can be generalised to any number of observers and variables. When applied to time windows, this is the best currently available approach for IOR assessment in workflow time studies. It is represented by the Greek letter iota, ι, (the letter before kappa).

Applying univariate kappa to the example data shown in Table 12.3 with time windows of 1 s (i.e. 1200 windows) we get scores of 0.57 for 'task type' and −0.45 for 'performed with nurse', indicating 'good' agreement for the former and moderate disagreement for the latter. If we apply Janson and Olsson's method to both variables we get a score of $\iota = 0.04$. This can easily be extended to include a third binary variable that represents multitasking (yes or no) in each time window. This has a univariate kappa score of 0.38, while the iota score for all three variables is 0.08.

The results for the 'task type' variable were consistent between the two methods, but were contradictory for the 'performed with nurse' variable. Also, the low agreement shown by the multivariate iota score did not concur with the high univariate kappa score for 'task type' alone. These results from the two general approaches highlight some key points about IOR assessment. First, the utility of any IOR measure must be considered relative to the analysis. The motivation behind assessing IOR is to identify and minimise observer biases in the data, however, IOR measures do not necessarily quantify the extent to which results are biased due to observer differences. For example, if there is good agreement on the overall proportions of individual categories between observers, but poor agreement at task level when multiple task attributes are considered together, then an analysis that aims to simply summarise proportions would not be biased, while a multivariate regression model would be. A corollary of this issue is that IOR measures have limited comparability between studies, such that it only makes sense to compare IOR results when the IOR method *and* the analysis are the same.

Second, a high univariate IOR score, as is typically reported in workflow time studies, does not tell us much about agreement levels in the whole dataset. Unless the analysis only uses one variable, it is imperative to take a multivariate approach to IOR assessment and to pursue development of customised methods for workflow time studies. More generally, it is therefore important to move away from the idea that any existing approach is the gold standard for IOR assessment, to have more transparent reporting of IOR in workflow time studies, and to have more open discussions of the limitations of existing methods and how they can be improved.

A final consideration is that IOR is not the same as accuracy, as a high IOR score could simply mean two observers are both wrong in the same way. The lack of a true record of the observed activity necessitates assessment of IOR, but also makes it impossible to assess accuracy. While we would expect some correlation between IOR and accuracy, there will always be uncertainty about data accuracy that cannot be overcome by any IOR method.

12.4 Analysis

12.4.1 Summary Statistics

The descriptive studies discussed in this chapter use a range of measures to characterise observed workflow. Of these, we focus on the most commonly used measures: proportions of time, and rates of events per unit time.

12.4.1.1 Proportions of Time

Proportions of time are a key metric in workflow time studies, providing an indication of how participants distribute their time across various activities, locations, or between the different people with whom they interact. They are a mainstay of descriptive studies but are also useful in intervention studies as an indicator of changes in work patterns. The summation of time intervals tends to be non-trivial, due to the presence of multitasking which creates overlap and hence multiple counting of time. While sums of time are not usually reported directly, they are part of the calculation of other frequently used measures such as proportions and rates.

Quantifying the uncertainty around estimated proportions in the form of confidence intervals (CIs) is important for interpreting results. For proportions of countable units, such as people or events, constructing a CI is a well-trodden path described in most statistics textbooks: the CI for a binomial proportion. However, for proportions of time—a continuous measure—the binomial methods do not apply. Surprisingly, there is little methodology for calculating CIs for proportions of continuous variables. In the early 1980s Gilchrist (1982) noted the lack of discussion in the literature despite such proportions occurring frequently, and this is still the case more than 30 years later. Only a few papers to date have discussed analysis of continuous proportions using parametric assumptions (Warton and Hui 2011; Stephens 1982), but they do not directly tackle CIs. A simple modification of the CI for the mean of a normally distributed variable has often been used (Li et al. 2015; Arabadzhiyska et al. 2013), which is expressed in the following form:

$$\frac{T_c}{T} \pm z_{1-\alpha/2} \frac{s_c \sqrt{n_c}}{T}$$

where T_c is the time spent doing tasks from category c, T is the total observation time, s_c is the sample standard deviation of task times for category c, and n_c is the number of tasks in that category. In addition, $z_{1-\alpha/2}$ is the standard score from a normal distribution, for example for a 95% CI, this would have the value $z_{0.975} \approx 1.96$.

A drawback of this method is that what constitutes a task depends on the definitions of dimensions and categories and to some extent on interpretation of those definitions during observation. For example, if a task is completed in two fragments due to an interruption, should this be counted as one task or two? That is, choices

regarding task definition affect the term n_c, and hence the CI width is at the whim of these choices. Also, the normal assumption is only likely to be satisfied when samples of tasks (T_c) are at least 30, and in some cases it may generate values for the CI that are outside the plausible range, e.g. below zero or above one.

A natural alternative is to take a nonparametric approach, namely to use bootstrap CIs (as in Bellandi et al. 2018). This does not require parametric assumptions, which addresses the limitations just mentioned, making it an optimal choice for continuous proportions. DiCiccio and Efron (1996) offered a thorough discussion of the various approaches that can be used to construct bootstrap CIs. Below, we provide a brief description of the basic method.

For a dataset with n tasks, a random selection of n of these is drawn with replacement. Even though the new sample has the same number of tasks as the original data, it will not necessarily be the same dataset since the random selection *with replacement* means that in the new sample some tasks will appear multiple times while others may not appear at all. The proportion of interest for the resampled data is then calculated. This procedure is repeated many times to generate a large number of resampled proportions. The simplest way to generate an interval is to then take the 2.5th and 97.5th percentile of the resampled proportions (for a 95% CI) as the lower and upper limits of the confidence interval.

We use a simulation study to illustrate the utility of the bootstrap approach by comparing the normal approximation method to the simple bootstrap. We also apply the bias-corrected and accelerated (BC_a) bootstrap which accounts for asymmetry in the CI. A sample of tasks was drawn with time durations from either an exponential, gamma or normal distribution. A random subset of 5, 10 or 20% of tasks was selected to represent some category of interest. For that 'category' the proportion of time was calculated along with its CI according to the three methods. This was repeated 1000 times and the proportion of CIs containing the true value, the coverage probabilities, are shown in Table 12.4. By definition, a 95% CI should cover the true proportion 95% of the time for a large number of repeated studies (or simulations in this case), so the expected coverage probability is then 0.95.

Table 12.4 Coverage probabilities for confidence intervals of proportions of time generated via three methods

Total tasks	'True' proportion	Normal approximation			Simple bootstrap			BC_a bootstrap		
		Exp	Gamma	Normal	Exp	Gamma	Normal	Exp	Gamma	Normal
10	0.05	0.070	0.049	0.013	0.384	0.398	0.406	0.391	0.404	0.404
10	0.5	0.786	0.782	0.501	0.892	0.905	0.925	0.938	0.931	0.946
10	0.95	0.987	0.982	0.946	0.375	0.394	0.396	0.386	0.398	0.394
100	0.05	0.671	0.654	0.381	0.830	0.881	0.905	0.851	0.895	0.925
100	0.5	0.934	0.882	0.587	0.940	0.951	0.950	0.948	0.952	0.954
100	0.95	0.999	1.000	0.980	0.831	0.867	0.903	0.853	0.882	0.924
1000	0.05	0.830	0.732	0.452	0.933	0.947	0.942	0.946	0.950	0.945
1000	0.5	0.946	0.877	0.584	0.948	0.941	0.952	0.951	0.942	0.956
1000	0.95	1.000	1.000	0.993	0.926	0.927	0.948	0.931	0.939	0.949

Both bootstrap approaches appear to perform better than the normal approximation method when the true proportion is near the lower boundary of the possible range of values (true proportion $\pi = 0.05$) or in the middle of the range ($\pi = 0.5$), especially for small and medium samples. The normal approximation performs particularly poorly for small proportions and small samples, with coverage probabilities less than 0.1. Towards the upper end of the range ($\pi = 0.95$), however, the normal approximation seems to perform better for small to medium samples, although proportions of this magnitude are rarely reported in the literature. Study samples are typically in the several thousands, and the results generated by the bootstrap method are consistently closer to the expected coverage probability of 0.95 for samples of that size. This suggests that the bootstrap CI is generally preferable to the normal approximation, which can be quite inaccurate. Further, the BC_a method consistently has slightly better coverage for all scenarios compared to the simple bootstrap and hence represents a better choice for calculating CIs of time proportions among the methods considered here.

12.4.1.2 Rates of Events Per Unit Time

Discrete events occurring at different points in time are common in clinical work and can be easily captured in workflow time studies. The most common example is interruptions. Since the number of such events is proportional to the length of time observed, they are generally analysed as rates per unit time, such as interruptions per hour. This quantifies the intensity of events while being independent of the amount of observation time. Descriptive studies tend to report rates in this form along with their CIs (Li et al. 2015; Walter et al. 2014; Westbrook et al. 2010). A common and simple approach for generating CIs is to assume that event counts, λ, are drawn from a Poisson distribution and to then generate a normal approximation CI in the form of:

$$\left(\lambda \pm z_{1-\alpha/2}\sqrt{\lambda}\right)/T$$

where T is the observation time. However, the Poisson assumption that the mean and variance are equal is not always met in workflow time study data and once again bootstrap CIs provide a more robust alternative.

We illustrate this through another set of simulations comparing the normal approximation method to both simple and BC_a bootstrap. This was done for task lengths drawn from two different distributions (exponential and normal), for small and large samples (n = 10 and n = 1000), for two different rates representing low and high rates relative to the typical range that appears in the literature on interruptions. We also simulated events to arrive according to either a Poisson or negative binomial distribution, where the former assumes that mean and variance are equal while the latter does not.

Table 12.5 Coverage probabilities for confidence intervals of rates per unit time generated via three methods

Total tasks	'True' rate[a]	'True' event distribution	Normal approximation		Simple bootstrap		BC_a bootstrap	
			Exp	Normal	Exp	Normal	Exp	Normal
10	3	Poisson	0.546	0.550	0.541	0.538	0.535	0.529
10	30	Poisson	0.919	0.921	0.868	0.904	0.865	0.903
1000	3	Poisson	0.939	0.960	0.940	0.961	0.938	0.961
1000	30	Poisson	0.932	0.948	0.930	0.944	0.933	0.944
10	3	NB[b]	0.533	0.567	0.529	0.561	0.521	0.553
10	30	NB	0.818	0.865	0.841	0.874	0.843	0.876
1000	3	NB	0.935	0.931	0.943	0.938	0.944	0.939
1000	30	NB	0.862	0.920	0.944	0.959	0.945	0.959

[a]Events per hour
[b]*NB* negative binomial

In the first part of Table 12.5, the simulated data satisfy the assumptions of all three methods and thus there is minimal difference between the three methods. The coverage probabilities are markedly lower for the small sample size scenarios, particularly when the underlying rate is also low. In the lower section of the table, the simulated events follow a negative binomial distribution. The differences in coverage between the three methods due to sample size and rate are similar, but a key difference can be seen for the scenario with large sample and high rate, in which the coverage for the normal approximation is lower than 0.95 while for the bootstrap method it is very close to the expected value of 0.95. This difference is amplified with increasing rate, such that for a rate of 300 events per hour the coverage for the normal approximation drops to 0.63 at best, compared to 0.96 for both bootstrap methods (data not shown in table). While the performance is comparable across most of the scenarios considered, the fact that the bootstrap approach is at least as good as, and in some cases clearly better than, the normal approximation method suggests that it may be considered a better choice to calculate CIs of rates.

12.4.2 Assessing Associations

12.4.2.1 Two Group Comparisons

Comparing outcomes between two groups is another common research goal in workflow time studies. For example, Richardson et al. (2016) (Table 3) compared both proportions of time and interruption rates between three studies of physicians, where each study used similar observational methodology and task definitions. Such comparisons in workflow time studies come with some important caveats, and some unique considerations are required for calculating significance.

Hypothesis testing was developed within the experimental paradigm in which factors extraneous to the effect of interest are controlled, such as randomly assigning subjects to one group or another. Any remaining difference in the outcome measure can then be attributed to the main effect. In other words, confounding is controlled through design. In observational studies of clinical work, this level of control is not possible, which means that the data represent a mixture of effects from many different factors, both known and unknown. When applying two group comparison tests to such data, it becomes difficult to definitively attribute the effect to any one factor. A study of physicians and nurses in surgical units (Bellandi et al. 2018) made such comparisons (adjusted for multiple testing), however, the authors appropriately refrained from attributing apparently significant differences to particular factors. Two-group comparisons in workflow time studies thus must be applied with caution.

As seen with calculating CIs, there is little methodology for analysing proportions of continuous measures. The calculations for parametric hypothesis tests involve the sample size, which, as seen several times in this chapter, can be open to interpretation. In the case of hypothesis testing, choices about what constitutes a task can then influence the sample size in the calculations and consequently the level of significance, which could result in incorrect conclusions, whether unconsciously or not.

Following on from the hypothesis testing approach used to assess IOR in Sect. 12.3.1, a way around these issues is, once again, through nonparametric methods. Permutation tests, or their Monte Carlo variation (Good 2010), can not only be applied to comparisons of typical measures in workflow time studies such as proportions of time and rates per unit time, but also to comparing means and counts. Rather than resampling the data as in the bootstrap method, the permutation tests randomly shuffle the group labels and calculate the difference between groups for each shuffle, e.g. the difference between proportions. This generates a null distribution for the observed difference and a p-value can then be determined as the proportion of permuted differences larger than the observed difference.

Again, we use a simulation to illustrate the efficacy of this approach. Tasks with durations following an exponential distribution were generated for two separate groups. For each group, a certain proportion of tasks (the 'true' proportion) were assigned to the category of interest and the difference between the group-level proportions of time for that category was calculated. The Monte Carlo permutation test was then applied to derive a p-value for the observed difference. This process was repeated 1000 times, from which the proportion of significant results was obtained using $\alpha = 0.05$. When there is a true difference, this proportion represents the power of the test. For a fixed proportion ($p1$) in the first group, the proportion in the second group ($p2$) was varied through a range of values and the power calculated each time as described above. This was done for $p1 = 0.05$ and $p1 = 0.2$, and also for sample sizes of 100 tasks (50 per group) and 1000 tasks (500 per group).

Figure 12.2 shows the estimated power for these four scenarios. Both plots show that power increases with greater true difference between groups and that this increase is more rapid for higher proportions (dotted lines for $p1 = 0.2$ versus solid

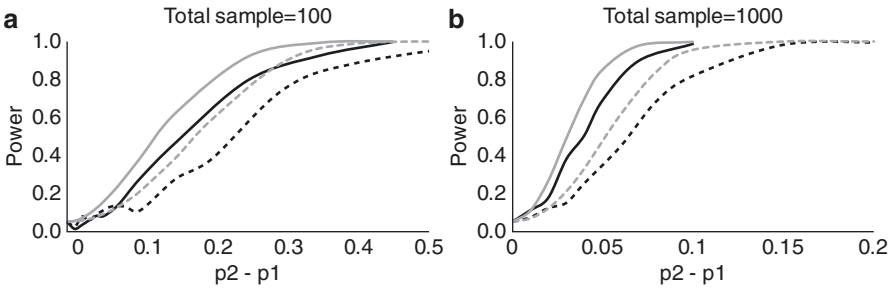


Fig. 12.2 Simulated power of the Monte Carlo permutation test to detect difference between two proportions of a continuous variable, for (**a**) a total sample of 100 tasks and (**b**) a total sample of 1000 tasks. The solid black line represents $p1 = 0.05$; the dashed black line represents $p1 = 0.2$. The computed power for equivalent differences in binomial proportions is shown as grey lines for reference

lines for $p1 = 0.05$), and for larger samples (plot **b** versus plot **a**). The two groups were simulated to have equal sample size. In additional simulations, it was found that keeping the same total sample size but allowing imbalance in group size reduced the power. The grey lines indicate power curves for the difference between two independent binomial proportions generated using the G∗Power program (Faul et al. 2007). While there is clear similarity, the power for the simulated permutation tests (black lines) are systematically lower. Nevertheless, the fact that they are in the same region and that the permutation test is applicable to proportions of continuous variables while binomial proportion methods are not, supports the permutation test as a reasonable choice for comparing proportions of time in workflow time studies.

An alternative testing approach, as outlined in Sect. 12.3, is to aggregate the data into subgroups. A proportion can be calculated for each subgroup, then the set of subgroup-level proportions can be analysed as continuous data, using methods such as *t*-tests or linear regression. We assessed this approach through simulation and compared it to permutation testing. To replicate a two-group comparison, we simulated 500 tasks per group (with exponentially distributed task duration) and divided the task in each group into either 10 subgroups of 50 tasks each, 50 subgroups of 10 tasks each, or six subgroups of eight or nine tasks each. In one group the underlying proportion of interest was set at 20% and for the other group this varied between 20 and 40%, that is, the difference between groups ranged from 0 to 20%. A *t*-test was applied to the subgroup-level proportions and the whole process was repeated 1000 times to obtain power estimates for the range of group differences.

The results of these simulations are shown in Fig. 12.3 where the power curves for *t*-tests applied at different levels of subgroup aggregation are relatively similar (all black lines). Although having fewer subgroups reduces the effective sample size of the tests, this seems to be counteracted by a proportional decrease in variance. The somewhat surprising result of which is that the power is not greatly affected by the level of aggregation. The grey line in the plot shows the power for the permuta-

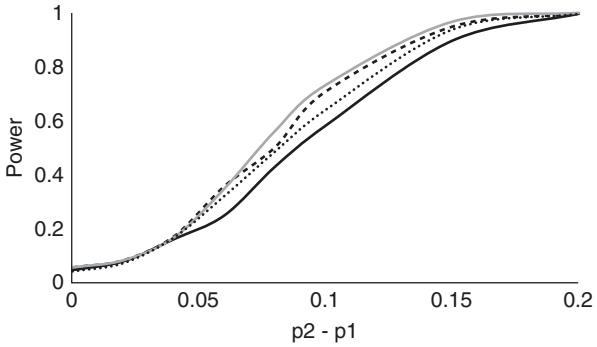


Fig. 12.3 Simulated power for *t*-tests applied to subgroups-level proportions for 50 subgroups of 10 tasks each (solid black line), 10 subgroups of 50 tasks each (dashed black line), and 6 subgroups of 8 or 9 tasks (dotted black line). The total sample of tasks was 1000 (500 per group), the underlying proportion of the group 1 was $p_1 = 0.2$ and proportions for group 2 ranged from 0.2 to 0.4. The power for a permutation test is shown for comparison (solid grey line)

tion testing approach. This is consistently as good or better than *t*-tests applied to aggregated data. The choice of units over which to aggregate data (e.g. observation sessions, clinicians, etc.) is not necessarily obvious in workflow time studies. Combined with the fact that permutation tests are at least as powerful, then once again a nonparametric approach is the better option.

12.4.2.2 Multivariate Analyses

There are many ways to apply multivariate methods in workflow time studies. Indeed, there is a strong case to make that most association studies should take a multivariate approach to better understand the factors operating at multiple system levels and minimise the bias in particular effects by adjusting for other influential factors. We have discussed general considerations of multivariate analysis in workflow time studies in our previous work (Walter et al. 2015). In this section we extend the theme of nonparametric analysis into the multivariate arena.

There are several ways to apply nonparametric methods to multivariate analyses. First, when fitting garden variety parametric models, such as linear regression, it is possible to use bootstrapping to determine the significance of the model estimates or to generate CIs for the estimates. This is essentially an extension of what we have discussed earlier regarding CIs and hypothesis tests, and similarly this may be an appropriate alternative when the data do not satisfy parametric model assumptions, as is often the case.

Second, there is a wide range of nonparametric multivariate modelling techniques that do not rely on assumptions about the distributional form (normal, Poisson, etc.) of the data. Some can be used as explanatory models, such as generalised additive models or spline regression, that can describe non-linear associations. In the study of

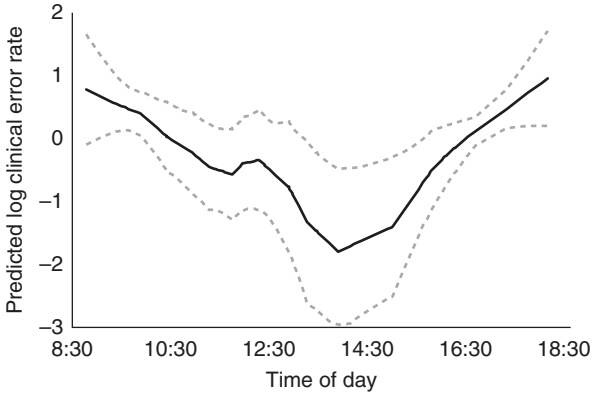


Fig. 12.4 Nonparametric estimate (LOESS smoother) of the relationship between time of day and clinical prescribing errors. The black line represents the predicted clinical prescribing error rate on the log scale and the dotted grey lines are the 95% confidence limits. This smoothing component for time of day had a p-value of 0.014

prescribing errors among ED physicians, Westbrook et al. (2018) found no evidence of an effect of time of day (categorised into 2-h blocks) on error rates using a Poisson regression model. However, Fig. 12.4 shows that fitting a nonparametric model (LOESS smoother) reveals a significant and distinctly non-linear relationship. Another explanatory approach is the classification tree, a version of which was used by Walter et al. (2017). In that study, discussed at the end of Sect. 12.2, the lack of data in certain categories necessitated a change from the original analysis plan. The alternative analysis used was a nonparametric model called a conditional inference tree, which iteratively splits the data into groups such that each group has a distinct outcome profile. Finally, in the area of predictive nonparametric models there is now a vast and growing collection of methods, such as Bayesian networks and random forests, that would be applicable to answering appropriately framed research questions in workflow time studies.

12.5 Discussion

Workflow time studies are an important type of research for generating knowledge about both the functioning of clinical work and workflow at a fine-grained level, and about the workflow-related factors that influence patient safety and quality of care. The data generated by such studies, and likely other types of time and motion studies, are not always amenable to conventional statistical methods. In this chapter we have highlighted some of the non-standard aspects of the data and offered alternative approaches that draw heavily from the family of nonparametric analysis techniques.

This chapter is somewhat technical, and it may be tempting for readers to form the impression that workflow time studies are overly complicated. The basic concept of these studies is, in fact, straightforward, but the complexity largely comes from the contexts in which they are applied. Clinical work is undeniably complex, and to understand its inner workings and interrelationships we must embrace that complexity into study design and data analyses, challenging as it may be. To design studies and analyses that fit within conventional approaches is to essentially shy away from or ignore those challenges. The methodological discourse in this chapter takes some steps towards tackling the intricacies of conducting quantitative studies of clinical work but is intended as a starting point for ongoing discussions rather than a definitive account of best practices.

Some recent studies have begun to employ more sophisticated methods such as multilevel models (Walter et al. 2014; Grundgeiger et al. 2010), transition state models (Carayon et al. 2015; Myers and Parikh 2019), and nonparametric models (Walter et al. 2017). However, explicit discussion of quantitative methodology appropriate for workflow time studies remains relatively rare. As we have highlighted in this chapter, there is an imperative to develop innovative approaches even for fundamental analyses such as IOR assessment, confidence intervals and hypothesis tests. Improving both our understanding of clinical workflow and the integrity of the workflow time study literature will require ongoing methodological innovation.

References

Ampt A, Westbrook JI, Creswick N, Mallock N. A comparison of self-reported and observational work sampling techniques for measuring time in nursing tasks. J Health Serv Res Pol. 2007;12(1):18–24.

Arabadzhiyska PN, Baysari MT, Walter SR, Day RO, Westbrook JI. Shedding light on junior doctors' work practices after hours. Internal Med J. 2013;43(12):1321–6.

Bakeman R, Quera V, Gnisci A. Observer agreement for timed-event sequential data: a comparison of time-based and event-based algorithms. Behav Res Methods. 2009;41(1):137–47.

Bellandi T, Cerri A, Carreras G, Walter SR, Mengozzi C, Albolino S, et al. Interruptions and multitasking in surgery: a multicentre observational study of the daily work patterns of doctors and nurses. Ergonomics. 2018;61:40–7.

Carayon P, Wetterneck TB, Alyousefa B. Impact of electronic health record technology on the work and workflow of physicians in the intensive care unit. Int J Med Inform. 2015;84:578–94.

Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37–46.

DiCiccio TJ, Efron B. Bootstrap confidence intervals. Stat Sci. 1996;11(3):189–228.

Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods. 2007;39:175–91.

Georgiou A, McCaughey EJ, Tariq A, Walter SR, Li J, Callen J, et al. What is the impact of an electronic test result acknowledgement system on Emergency Department physicians' work processes? A mixed-method pre-post observational study. Int J Med Inform. 2017;99:29–36.

Gilchrist R. An analysis of continuous proportions. In: Caussinus H, Ettinger P, Tomassone R, editors. COMPSTAT 1982 5th Symposium held at Toulouse. Heidelberg: Physica; 1982.

Good PI. Permutation, parametric and bootstrap tests of hypotheses. 3rd ed. New York: Springer; 2010.

Grundgeiger T, Sanderson P, Venkatesh B, MacDougall HG. Interruption management in the intensive care unit: predicting resumption times and assessing distributed support. J Exp Psychol Appl. 2010;16(4):317–34.
Janson H, Olsson U. A measure of agreement for interval or nominal multivariate observations. Educ Psychol Meas. 2001;61(2):277–89.
Li L, Hains I, Hordern T, Milliss D, Raper R, Westbrook JI. What do ICU doctors do? A multisite time and motion study of the clinical work patterns of registrars. Crit Care Resusc. 2015;17:159–66.
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;2013:889–96.
Lopetegui M, Yen PY, Lai A, Jeffries J, Embi P, Payne P. Time and motion studies in healthcare: what are we talking about? J Biomed Inform. 2014;49:292–9.
Malhotra S, Jordan D, Shortliffe E, Patel VL. Workflow modeling in critical care: piecing together your own puzzle. J Biomed Inform. 2007;40(2):81–92.
McCurdie T, Sanderson P, Aitken LM. Traditions of research into interruptions in healthcare: a conceptual review. Int J Nurs Stud. 2017;66:23–36.
Myers RA, Parikh PJ. Nurses' work with interruptions: an objective model for testing interventions. Health Care Manag Sci. 2019;22(1):1–15. https://doi.org/10.1007/s10729-017-9417-3.
Richardson LC, Lehnbom EC, Baysari MT, Walter SR, Day RO, Westbrook JI. A time and motion study of junior doctor work patterns on the weekend: a potential contributor to the weekend effect? Int Med J. 2016;46(7):819–25.
Stephens MA. Use of the von Mises distribution to analyse continuous proportions. Biometrika. 1982;69(1):197–203.
Unertl KM, Novak LM, Johnson KB, Lorenzi NM. Traversing the many paths of workflow research: developing a conceptual framework of workflow terminology through a systematic literature review. J Am Med Inform Assoc. 2010;17:265–73.
Walter SR, Li L, Dunsmuir WTM, Westbrook JI. Managing competing demands through task-switching and multitasking: a multi-setting observational study of 200 clinicians over 1000 hours. BMJ Qual Saf. 2014;23:231–41.
Walter SR, Dunsmuir WTM, Westbrook JI. Studying interruptions and multitasking in situ: the untapped potential of quantitative observational studies. Int J Hum Comput Stud. 2015;79:118–25.
Walter SR, Raban MZ, Dunsmuir WTM, Douglas HE, Westbrook JI. Emergency doctors' strategies to manage competing workload demands in an interruptive environment: an observational workflow time study. Appl Ergon. 2017;58:454–60.
Warton DI, Hui FKC. The arcsine is asinine: the analysis of proportions in ecology. Ecology. 2011;92(1):3–10.
Weigl M, Muller A, Vincent C, Angerer P, Sevdalis N. The association of workflow interruptions and hospital doctors' workload: a prospective observational study. BMJ Qual Saf. 2012;21:399–407.
Westbrook JI, Ampt A. Design, application and testing of the Work Observation Method by Activity Timing (WOMBAT) to measure clinicians' patterns of work and communication. Int J Med Inform. 2009;78S:S25–33.
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;188:506–9.
Westbrook JI, Coiera E, Dunsmuir WTM, Brown BM, Kelk N, Paoloni R, Tran C. The impact of interruptions on clinical task completion. Qual Saf Health Care. 2010;19:284–9.
Westbrook JI, Raban MZ, Walter SR, Douglas HE. Task errors by emergency physicians are associated with interruptions, multitasking, fatigue and working memory capacity: a prospective, direct observation study. BMJ Qual Saf. 2018;27:655–63.
Zheng K, Guo MH, Hanauer DA. Using the time and motion method to study clinical work processes and workflow: methodological inconsistencies and a call for standardized research. J Am Med Inform Assoc. 2011;18:704–10.