Inter-rater agreement on PIVC-associated phlebitis signs, symptoms and scales

ARTICLE *in* JOURNAL OF EVALUATION IN CLINICAL PRACTICE · JULY 2015

Impact Factor: 1.58 · DOI: 10.1111/jep.12396 · Source: PubMed

6 AUTHORS, INCLUDING:

Gillian Ray-Barruel
Griffith University
19 PUBLICATIONS **55** CITATIONS
SEE PROFILE

Joan Webster
Royal Brisbane and Women's Hospital, Bris…
137 PUBLICATIONS **1,595** CITATIONS
SEE PROFILE

Claire M Rickard
Griffith University, Nathan, Australia
134 PUBLICATIONS **1,212** CITATIONS
SEE PROFILE

Journal of Evaluation in Clinical Practice
International Journal of Public Health Policy and Health Services Research

Journal of Evaluation in Clinical Practice ISSN 1365-2753

Inter-rater agreement on PIVC-associated phlebitis signs, symptoms and scales

Nicole Marsh RN BN MAdvPrac,[1,4] Gabor Mihala MEng (Mech) GCert (Biostat),[5,7] Gillian Ray-Barruel RN BSN BA (Hons) Grad Cert ICU Nursing,[5] Joan Webster RN BA,[2,6,8] Marianne C. Wallis RN PhD FACN[9] and Claire M. Rickard RN PhD[3,6]

[1]Nurse Researcher Intravascular Access, [2]Nursing Director, Research, [3]Professor of Nursing, Centre for Clinical Nursing, Royal Brisbane and Women's Hospital, Herston, Queensland, Australia
[4]Nurse Researcher Intravascular Access, [5]Senior Research Assistant, [6]Professor of Nursing, NHMRC Centre for Research Excellence in Nursing, Menzies Health Institute Queensland, Griffith University, Brisbane, Queensland, Australia
[7]Senior Research Assistant, School of Medicine, Griffith Health Institute, Griffith University, Meadowbrook, Australia
[8]Professor of Nursing, School of Nursing and Midwifery, University of Queensland, Brisbane, Australia
[9]Professor of Nursing, School of Nursing and Midwifery, University of the Sunshine Coast, Maroochydore, Queensland, Australia

Keywords
assessment, intravenous, measurement, phlebitis, psychometric assessment, scales, vascular access devices

Correspondence
Nicole Marsh
Centre for Clinical Nursing
Royal Brisbane and Women's Hospital
Level 2, Building 34
Butterfield Street
Brisbane, Queensland 4029
Australia
E-mail: Nicole.marsh@griffith.edu.au

Accepted for publication: 5 May 2015

doi:10.1111/jep.12396

Abstract

Rationale, aims and objectives Many peripheral intravenous catheter (PIVC) infusion phlebitis scales and definitions are used internationally, although no existing scale has demonstrated comprehensive reliability and validity. We examined inter-rater agreement between registered nurses on signs, symptoms and scales commonly used in phlebitis assessment.

Methods Seven PIVC-associated phlebitis signs/symptoms (pain, tenderness, swelling, erythema, palpable venous cord, purulent discharge and warmth) were observed daily by two raters (a research nurse and registered nurse). These data were modelled into phlebitis scores using 10 different tools. Proportions of agreement (e.g. positive, negative), observed and expected agreements, Cohen's kappa, the maximum achievable kappa, prevalence- and bias-adjusted kappa were calculated.

Results Two hundred ten patients were recruited across three hospitals, with 247 sets of paired observations undertaken. The second rater was blinded to the first's findings. The Catney and Rittenberg scales were the most sensitive (phlebitis in >20% of observations), whereas the Curran, Lanbeck and Rickard scales were the most restrictive (≤2% phlebitis). Only tenderness and the Catney (one of pain, tenderness, erythema or palpable cord) and Rittenberg scales (one of erythema, swelling, tenderness or pain) had acceptable (more than two-thirds, 66.7%) levels of inter-rater agreement.

Conclusions Inter-rater agreement for phlebitis assessment signs/symptoms and scales is low. This likely contributes to the high degree of variability in phlebitis rates in literature. We recommend further research into assessment of infrequent signs/symptoms and the Catney or Rittenberg scales. New approaches to evaluating vein irritation that are valid, reliable and based on their ability to predict complications need exploration.

Introduction

Peripheral intravenous catheter (PIVC) insertion for the administration of medications and intravenous (IV) fluids is the most common invasive procedure performed in acute care facilities worldwide. Up to 330 million PIVCs are sold in the United States each year [1]. One possible complication of catheterization is phlebitis, or inflammation of the vein, which may be chemical (caused by the infused substance), mechanical (caused by the device itself) or infective (due to microbial colonization of the catheter or IV site) [2,3]. Phlebitis may be associated with a range of patient-related (e.g. age, female sex, current infection and drugs infused) and catheter-related factors (e.g. PIVC size, insertion site, experience of inserter) [4].

Phlebitis can cause the patient severe discomfort and interrupt PIVC therapy resulting in a delay to treatment and the need for a PIVC resite [5]. Repeated instances of phlebitis can lead to difficulties with venous access and possibly result in the need for

central venous access [6]. PIVC failure as a result of phlebitis may lead to increased health care costs associated with equipment, staff time, prolonged hospital stay and blood stream infection [1]. Therefore, the timely detection of complications and removal of the cannula is essential. Post-infusion phlebitis may occur up to 48 h after PIVC removal [7], necessitating continued assessment of the site.

In addition to phlebitis, irritation of the vein may be accompanied by a variety of conditions including infiltration, extravasation, vein occlusion or PIVC blockage [8], and it can be challenging to differentiate between these conditions and phlebitis as they can produce similar signs and symptoms.

Phlebitis is diagnosed by observation of clinical *signs*, or when a patient reports various *symptoms*, and sometimes using severity assessment scoring tables or guidelines. However, a recent systematic review failed to locate a comprehensive inter-rater agreement study of phlebitis symptoms and scales [9]. Despite a plethora of existing phlebitis scales, the validity and reliability has not been established for most of these scales [9] or use in the clinical setting. Scales incorporate an array of symptoms and scoring measures, and consensus on definitions for phlebitis measures is lacking. This has likely contributed to the enormous disparity in phlebitis incidence rates reported in the literature, ranging from 0 [10,11] to 100% [12].

This paper reports the findings of a study examining levels of inter-rater agreement among registered nurses (including vascular access research nurses) for the most commonly used phlebitis symptoms and scales.

Methods

This study is based on a subset of the data collected for a large multicentre, randomized controlled trial comparing different regimens for PIVC replacement [13]. In that study (also known as the DRIP Trial), data were collected from May 2008 until September 2009 in three university-affiliated hospitals in Queensland, Australia. Hospitals 1 and 2 are large metropolitan tertiary hospitals managing a full range of general and specialist health care services, both of these hospitals had a dedicated PIVC insertion team during the recruitment period of this study. Hospital 3 is a large regional hospital without cardiac surgery or burns patients, or a PIVC insertion team.

Ethics committee approval was obtained from Griffith University and each hospital, and all participants provided written, informed consent prior to participation. Adult patients in medical and surgical units with a PIVC *in situ* and expected to require IV therapy for more than 4 days were eligible. Exclusion criteria included existing bloodstream infection, planned PIVC removal within 24 h or PIVC already *in situ* for more than 72 h. All study PIVCs (Insyte Autoguard; Becton Dickinson, Franklin Lakes, NJ, USA) were inserted into the arm or hand.

Data collection

In a subset of patients of the DRIP Trial, the daily assessments of phlebitis signs and symptoms (pain, tenderness, swelling, erythema, palpable venous cord, purulent discharge and warmth) were performed twice, a few minutes apart, by independent raters. The first raters were research nurses (registered nurses with a minimum of 10 years of clinical experience) with training in PIVC site assessment. These first observations were used for the DRIP Trial and also for this inter-rater study. Raters who made the second site assessment were also experienced registered nurses with expertise in PIVC site assessment. Their observations were used only in this inter-rater study. Data were collected using a convenience sample on the days that the two experienced registered nurses were available for PIVC site assessments. All patients on the trial that day were included.

Signs and symptoms

The term 'sign' refers to observations made by research nurses (e.g. palpable venous cord), whereas the term 'symptom' refers to patient reports (e.g. pain). The following five signs and two symptoms were used by Rickard *et al.* [13] based on commonly reported signs and symptoms [14–16]. The severity categories of the sign or symptoms were arbitrarily chosen to be tested during the DRIP Trial.

- *pain* (patient-reported symptom, severity levels: none, 1 out of 10, 2–4 out of 10, 5–8 out of 10 or 9–10 out of 10)
- *tenderness* (on palpation, patient-reported symptom, severity levels: none, 1 out of 10, 2–4 out of 10, 5–8 out of 10 or 9–10 out of 10)
- *swelling* (visual observation, severity levels: none, <1 cm, 1 to <2.5 cm, 2.5 to <5 cm or 5 cm or larger)
- *erythema* (or redness, visual observation, severity levels: none, <1 cm from exit site, 1 to <2.5 cm, 2.5 to <5 cm or 5 cm or larger)
- *palpable venous cord* (on palpation, severity levels: none, <7.5 cm or 7.5 cm or longer)
- *purulent discharge* (visual observation, categories: none, from site or with ulceration)
- *warmth* (on palpation, categories: no, yes)

These signs and symptoms were later found to be on the top six of most commonly used signs and symptoms (from a total of 15) for the diagnosis of phlebitis [9], with *warmth* the eighth.

It was expected that majority of observations would fall into the *no* or *none* categories, with the few positive observations spread across up to five severity levels. This would have led to zero frequencies in some of the severity levels, which would have made the analysis unnecessarily complicated (requiring the calculation of weighted kappa) and the interpretation of results potentially misleading. Because disagreement between positive (symptom present) and negative (symptom absent) observations was more important to measure and report than disagreement between two positive ratings with different severity levels, the non-zero severity levels were collapsed effectively turning all sign and symptom variables into dichotomous (i.e. no/yes) type. A sensitivity analysis was also performed when the second lowest levels (e.g. 1 out of 10) of *pain*, *tenderness*, *swelling* and *erythema* were grouped with the *none* observations.

Phlebitis scales

The phlebitis scales and definitions to be compared in this paper were selected based on a literature review of the Cochrane library, Ovid MEDLINE and EBSCO CINAHL until September 2013, for English-language randomized controlled trials, prospective cohort and cross-sectional studies, using the search terms infusion phle-

bitis, thrombophlebitis, peripheral IV catheter, phlebitis score, phlebitis grade and phlebitis assessment [9]. To be eligible for inclusion in this study, a scale had to have measured phlebitis using only the above listed (most common) signs and/or symptoms. The following 10 scales and definitions were selected (identified by common names or first author's name): Barker [17], Baxter [18], Catney [19], Curran [20], Lanbeck [21], Maki [16], Rickard [13], Rittenberg [22], Van Donk [23] and the Visual Infusion Phlebitis Score (VIP) scale [3,24]. We were unable to use scales and definitions (e.g. Infusion Nurses Society (INS)) if their criteria relied on signs/symptoms not collected for this study (e.g. induration, streak formation), or if the scale was not clearly defined in publications.

Some scales classify the severity of phlebitis on an ordinal scale (e.g. 0–5), often without recommendation on acceptable/unacceptable levels or when to initiate corrective action. In order to calculate the inter-rater agreement measures of such scales, the unacceptable levels of phlebitis had to be established for every scale, effectively turning the scales into dichotomous measures (similarly as for signs and symptoms, discussed earlier). According to the scale developer, phlebitis was present if the following conditions applied:

- *Barker*: at least two of pain, swelling, erythema, palpable cord and warmth
- *Baxter*: at least one of pain, swelling, erythema, palpable cord and purulence
- *Catney*: at least one of pain, tenderness, erythema and palpable cord
- *Curran*: at least one of erythema (≥2.5 cm) and purulence
- *Lanbeck*: erythema and swelling with either tenderness or pain
- *Maki*: at least two of pain, tenderness, swelling, erythema, palpable cord and purulence
- *Rickard*: at least two of erythema (≥1 cm), swelling (≥1 cm), palpable cord, purulence and [pain or tenderness (≥2 out of 10)
- *Rittenberg*: at least one of pain or tenderness, or swelling or erythema
- *Van Donk*: at least one of pain (≥2 out of 10), swelling (≥1 cm), erythema (≥1 cm), or at least two of pain, swelling, erythema and purulence
- *VIP*: at least two of pain, swelling and erythema

Calculation of clinical agreement measures

It has been shown that a proportion of agreement is more suitable to assess observer variation than the commonly used Cohen's kappa [25]. Proportions of positive and negative agreement are absolute measures appropriate for evaluating a measurement instrument, whereas reliability parameters (e.g. Cohen's kappa) are relative measures more suitable to assess measurement reliability. In this study, both absolute and relative measures were calculated and presented, as kappa is still frequently used in the literature.

Proportions of specific agreement (i.e. positive or negative) were calculated as described in de Vet *et al*. [25]. Observed and expected agreements, Cohen's kappa, the maximum achievable kappa and prevalence- and bias-adjusted kappa were also calculated [26]. Data management and analysis was completed with Stata 12.1 (StataCorp. 2011. Stata Statistical Software: Release 12. College Station, TX: StataCorp LP).

Results

A total of 210 patients were recruited across the three sites, and a total of 247 sets of paired observations (average 1.2 paired set of observations per patient) were recorded. There were no losses during the study, and no observations had to be excluded during analysis due to missing data. When compared with the other sites, participants in Hospital 1 appeared to have worse skin integrity as defined by the skin integrity tool used by the PIVC insertion team at Hospital 1 (see Table 1) and vein quality [27] more insertions by IV service, and more 20 gauge devices used, whereas Hospital 3 appeared to have more female participants, a lower nosocomial infection risk [28] and more insertions by clinical staff (see Table 2).

The most prevalent sign/symptom was *tenderness* [at least one rater reported tenderness in 47 (19%) of paired observations], whereas the least observed signs/symptoms were *purulence*, *warmth* and *palpable cord* (≤2% of paired observations). Raters in Hospital 3 found considerably higher proportions of *tenderness*, *erythema* and *swelling* when compared with Hospitals 1 and 2 (see Table 3).

Table 3 also shows the frequencies of positive phlebitis assessments using the various scales. Overall, the *Catney* and *Rittenberg* scales were the most sensitive (diagnosing phlebitis in an overall >20% of observations) while the *Curran*, *Lanbeck* and *Rickard* scales were the most restrictive (with ≤2% phlebitis rates). Many patients had very low levels of positivity for various signs and symptoms (e.g. pain at 1 out of 10). Results of the sensitivity analysis (not tabled here) where the lowest positive value was grouped with 'none' showed that the number of positive observations/assessments for the signs, symptoms and scales were reduced by at least 50%; for example, *tenderness* from a total of 19.0% down to 8.9%, and for the *Rittenberg* scale from 25.1% down to 11.3%.

The various absolute (agreement) and relative (reliability) measures are presented in Table 4. *Tenderness* had the highest level (81%) of positive agreement between raters, followed by *erythema* (64%), while there was no agreement (0%) on *palpable cord*, and results for *warmth* and *purulent discharge* were unable to be calculated due to very low or zero frequencies. The *Catney* and *Rittenberg* scales had the highest levels (>80%) of positive agreement between raters, while the *Barker* and *VIP* scales performed at the lowest level (20%), and results for the *Curran* scale were unable to be calculated due to very low or zero frequencies. Best performing signs/symptoms under the sensitivity analysis conditions (not tabled here) were similar (*erythema* and *tenderness* had 67% agreement on positive rating), and the best performing scales remained the *Rittenberg* (68%) and *Catney* (63%) scales.

Table 1 Definitions used to classify skin integrity

Skin condition	Definition
Good	Healthy, well hydrated, elastic
Fair	Intact, mildly hydrated, reduced elasticity
Poor	Papery, dehydrated, small amount or no elasticity

Table 2 Participant characteristics and insertion details

	Hosp. 1	Hosp. 2	Hosp. 3	Total
Per participant (*n* = 210)	91 (43.3)	82 (39.1)	37 (17.6)	210 (100.0)
Age (years)*	53 (17.7)	53 (19.0)	59 (17.6)	54 (18.3)
Sex (male)	59 (64.8)	73 (89.0)	15 (40.5)	132 (76.3)
Co-morbidities (two or more)	44 (48.4)	44 (53.7)	16 (43.2)	88 (50.9)
Admission type (surgical)	70 (76.9)	79 (96.3)	26 (70.3)	175 (83.3)
Skin integrity (good)	49 (53.9)	63 (76.8)	29 (78.4)	141 (67.1)
Vein quality (fair/good)	71 (78.0)	77 (93.9)	34 (91.9)	182 (86.7)
Infection risk (low)	41 (45.1)	42 (53.2)	29 (78.4)	112 (54.1)
Infection current (no)	76 (83.5)	67 (81.7)	35 (94.6)	178 (84.8)
Per insertion (*n* = 247)	101 (40.9)	102 (41.3)	44 (17.8)	247 (100.0)
Insertion per participant†	1 (3)	1 (4)	1 (2)	1 (4)
Trial arm (control group)	54 (53.5)	55 (53.9)	21 (47.7)	130 (52.6)
Inserting department (ward)	77 (76.2)	84 (82.4)	25 (56.8)	186 (75.3)
Insertion side (left)	54 (53.5)	57 (55.9)	28 (63.6)	139 (56.3)
Insertion location (forearm)	57 (56.4)	56 (54.9)	21 (47.7)	134 (54.3)
Inserted by				
Clinical staff	32 (31.7)	62 (60.8)	44 (100.0)	138 (55.9)
IV service	69 (68.3)	40 (39.2)	0 (0.0)	109 (44.1)
Device size				
≤18 gauge	15 (14.9)	8 (7.8)	17 (38.6)	40 (16.2)
20 gauge	84 (83.2)	29 (28.4)	19 (43.2)	132 (53.4)
≥22 gauge	2 (2.0)	65 (63.7)	8 (18.2)	75 (30.4)

n (%) shown unless otherwise indicated.
*Mean (standard deviation).
†Median (max).
IV, intravenous.

Table 3 Descriptive statistics of signs, symptoms and phlebitis scales (*n* = 247 paired OBS.)

	Hosp. 1		Hosp. 2		Hosp. 3		Total
	A	B	A	B	A	B	A or B*
Signs and symptoms							
Tenderness	15 (14.9)	22 (21.8)	9 (8.8)	11 (10.8)	11 (25.6)	11 (25.6)	47 (19.0)
Pain	3 (3.0)	4 (4.0)	1 (1.0)	4 (3.9)	2 (4.5)	1 (2.3)	12 (4.9)
Erythema	2 (2.0)	2 (2.0)	6 (5.9)	5 (4.9)	3 (6.8)	7 (15.9)	17 (6.9)
Swelling	1 (1.0)	0 (0.0)	2 (2.0)	4 (3.9)	4 (9.1)	8 (18.2)	15 (6.1)
Palp. cord	3 (3.0)	1 (1.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (2.3)	5 (2.0)
Warmth	1 (1.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (0.4)
Purulence	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Phlebitis scales							
Barker	0 (0.0)	0 (0.0)	1 (1.0)	4 (3.9)	0 (0.0)	5 (11.4)	9 (3.6)
Baxter	9 (8.9)	7 (6.9)	8 (7.8)	8 (7.8)	9 (20.5)	11 (25.0)	37 (15.0)
Catney	18 (17.8)	24 (23.8)	13 (12.7)	14 (13.7)	15 (34.1)	17 (38.6)	60 (24.3)
Curran	0 (0.0)	0 (0.0)	0 (0.0)	2 (2.0)	0 (0.0)	1 (2.3)	3 (1.2)
Lanbeck	0 (0.0)	0 (0.0)	1 (1.0)	2 (2.0)	0 (0.0)	1 (2.3)	3 (1.2)
Maki	6 (5.9)	5 (5.0)	3 (2.9)	5 (4.9)	2 (4.5)	8 (18.2)	23 (9.3)
Rickard	2 (2.0)	0 (0.0)	1 (1.0)	2 (2.0)	0 (0.0)	1 (2.3)	5 (2.0)
Rittenberg	17 (16.8)	23 (22.8)	14 (13.7)	15 (14.7)	18 (40.9)	18 (40.9)	62 (25.1)
Van Donk	1 (1.0)	1 (1.0)	3 (2.9)	6 (5.9)	2 (4.5)	8 (18.2)	16 (6.5)
VIP	0 (0.0)	0 (0.0)	1 (1.0)	4 (3.9)	0 (0.0)	5 (11.4)	9 (3.6)

n (%) of positive/non-zero/yes observations and phlebitis assessments shown; A = first observer; B = second observer.
**n* (%) in either A or B.
Palp., palpable.

Table 4 Agreement and reliability results

	Agreement (%)				Reliability		
	Pos*	Neg†	Obs‡	Exp§	κ (95% CI)	κ_{max}¶	PABAK**
Signs and symptoms:							
Tenderness	81.0	96.4	93.9	73.0	0.77 (0.65–0.90)	0.86	0.88
Pain	40.0	98.1	96.4	94.1	0.38 (0.26–0.50)	0.79	0.93
Erythema	64.0	98.1	96.4	90.4	0.62 (0.50–0.74)	0.87	0.93
Swelling	42.1	97.7	95.5	92.6	0.40 (0.28–0.52)	0.73	0.91
Palp. cord	0.0	99.0	98.0	98.0	–0.01 (–0.13–0.11)	0.80	0.96
Warmth			99.6	99.6	0.00 (–)		0.99
Purulence							
Phlebitis scales							
Barker	20.0	98.3	96.8	96.3	0.19 (0.12–0.27)	0.19	0.94
Baxter	57.7	95.0	91.1	81.2	0.53 (0.40–0.65)	1.00	0.82
Catney	81.2	95.2	92.3	67.4	0.76 (0.64–0.89)	0.89	0.85
Curran			98.8	98.8	0.00 (0.00–0.00)		0.98
Lanbeck	50.0	99.6	99.2	98.4	0.50 (0.39–0.60)	0.50	0.98
Maki	41.4	96.3	93.1	88.9	0.38 (0.26–0.50)	0.74	0.86
Rickard	33.3	99.2	98.4	97.6	0.33 (0.20–0.45)	1.00	0.97
Rittenberg	81.9	95.1	92.3	66.5	0.77 (0.65–0.89)	0.92	0.85
Van Donk	47.6	97.7	95.5	91.8	0.46 (0.35–0.57)	0.56	0.91
VIP[7]	20.0	98.3	96.8	96.0	0.19 (0.12–0.27)	0.19	0.94

*Specific agreement on a positive rating.
†Specific agreement on a negative rating.
‡Observed agreement.
§Expected agreement.
¶Max. attainable kappa with these raters.
**Prevalence and bias-adjusted kappa.
κ, kappa; hyphen, unable to be calculated; palp., palpable.
[7]Visual Infusion Phlebitis Score (VIP)

Discussion

A plethora of phlebitis definitions and scales, with varying measurement options, are currently available to clinicians and researchers. However, validity and reliability data have not been established for most of these scales [9]. This makes selecting a scale for research or clinical purposes difficult and contributes to the widely varying phlebitis rates in the literature. Our study begins to address this deficit; it is the first to compare inter-rater agreement of PIVC-associated signs and symptoms associated with phlebitis and a number of phlebitis scales.

Although the raters in our study were experienced nurse clinicians, it was surprising to find inconsistencies in their levels of agreement for some observations. For example, tenderness was the symptom with the highest level of positive agreement between raters at 81% but the level of agreement was zero for palpable cord. Generally speaking, however, the level of agreement matched the frequency of the problem, with higher agreement for the more commonly reported signs or symptoms. Similarly, the positive agreement level for the best performing phlebitis scales was 82% with the poorest at 20%. Given the variation between different scales in identifying phlebitis, it is little wonder that reported rates vary so much. Positive agreement values can be interpreted with an example: paraphrasing de Vet *et al*. [25]. 'Suppose the first rater observes *swelling*, what is the probability that the second rater also observes *swelling*?', and for this example the probability of positive agreement was 42% (Table 4). The positive agreement values appeared to be correlated with the kappa results. Although the interpretation of kappa is affected by sample size, positive agreement results are not subjected to hypothesis testing and therefore their interpretation is not affected by small sample sizes. Using an arbitrary level of 2 out of 3 (66.7%), the probability of positive agreement to separate the best performing signs/symptoms/scales from the rest, only *tenderness*, and the *Catney* and *Rittenberg* scales had acceptable inter-rater agreement levels (*erythema* performed slightly under this cut-off level. Under sensitivity analysis conditions *tenderness*, *erythema* and the *Rittenberg* scale performed above the cut-off point, with the *Catney* scale, was slightly below the cut-off level).

A strength of this study is that patients had a broad range of demographic and clinical characteristics and were recruited across three hospitals. A further strength was that assessments were performed immediately after each other, eliminating confounding factors, such as medication administration, time for signs and symptoms to become worse or resolve or removal of the PIVC. Taken together, these strengths support generalization of our results.

The number of paired observations was 247, but the prevalence rates were well below the required 10% for most signs, symptoms and scales (except *tenderness*, and the *Baxter*, *Catney* and *Rittenberg* scales), resulting in a potential type II error in the calculations of Cohen's kappa. The sample size for an outcome of only 5% prevalence and with a type II error of ≤10% would have needed to be at least 500 paired observations (estimated).

It cannot be ruled out that the levels of *pain* and *tenderness* experienced by the patient increased after the first observer had left due to palpation of the PIVC site and the specific questions asked by the first observer; therefore, the patient could have been more aware of these symptoms by the time the second observer commenced. This seems to have been confirmed by the results: all but one patient reported the same or higher level of *pain*, and all but three patients reported the same or higher level of *tenderness* for the second observer. Although it is a concern, the level of bias was low: 9% (1/11) of *pain* observations and 6% (3/47) of *tenderness* observations appeared to have been affected. The number of positive observations (prevalence) for *warmth* and *purulent discharge* in the study was less than 2, creating inconclusive inter-rater agreement results for these symptoms and for scales utilizing these symptoms. Any future inter-rater study investigating these signs would require a much larger sample size.

With regard to the overall assessment of phlebitis using scales, clinicians would have found phlebitis rates of less than 4% using the *Barker*, *Curran*, *Lanbeck*, *Rickard* and *VIP* scales, but at least 24% with the *Catney* or *Rittenberg* scales. This is a remarkable difference between the performances of the 10 scales used in this study. Rates were lower under the sensitivity analysis, but the difference was still considerable. Such variability can be attributed to the lack of an agreed scientific definition of phlebitis, and the fact that some phlebitis symptoms are difficult to quantify or subject to bias. Apart from the *VIP* score, none of the scales contain directions for action related to specific scores. Further research is required on the relationship between scale scores and outcomes.

Conclusions

Inter-rater agreement for phlebitis signs, symptoms and scales is generally low, and poor agreement likely contributes to the high degree of variability in phlebitis rates in the literature. Based on the results from this study, we would be unable to recommend the use of any particular phlebitis scale. The variation in performance (i.e. the number of positive phlebitis assessments) for various phlebitis scales is a concern for several reasons. Firstly, nurses are investing time in using tools that have not been adequately validated. Secondly, the lack of an accepted phlebitis definition and recommendations on how to follow up a positive assessment puts the use of many scales into question. At the very least, our findings demonstrate the need for further research into agreement for infrequent symptoms and more education for nurses regarding intravascular device complications. Until further research clarifies the usefulness of phlebitis scales, removal of the PIVC should be performed promptly at the completion of therapy or with any sign of PIVC complication. Finally, research is urgently required to explore new approaches to evaluating vein irritation that are practical, valid, reliable and based on evidence that they accurately predict clinical complications.

References

1. Hadaway, L. C. (2012) Short peripheral catheters and infections. *Journal of Infusion Nursing*, 35 (4), 230–240.
2. Campbell, L. (1998) IV-related phlebitis, complications and length of hospital stay: 1. *British Journal of Nursing (Mark Allen Publishing)*, 7 (21), 1304–1306.
3. Higginson, R. & Parry, A. (2011) Phlebitis: treatment, care and prevention. *Nursing Times*, 107 (36), 18–21.
4. Wallis, M. C., McGrail, M., Webster, J., *et al*. (2014) Risk factors for peripheral intravenous catheter failure: a multivariate analysis of data from a randomized controlled trial. *Infection Control and Hospital Epidemiology*, 35 (1), 63–68.
5. Dillon, M. F., Curran, J., Martos, R., *et al*. (2008) Factors that affect longevity of intravenous cannulas: a prospective study. *QJM: Monthly Journal of the Association of Physicians*, 101 (9), 731–735.
6. Hawes, M. L. (2007) A proactive approach to combating venous depletion in the hospital setting. *Journal of Infusion Nursing*, 30 (1), 33–44.
7. Hershey, C. O., Tomford, J. W., McLaren, C. E., Porter, D. K. & Cohen, D. I. (1984) The natural history of intravenous catheter-associated phlebitis. *Archives of Internal Medicine*, 144, 1373–1375.
8. Doellman, D., Hadaway, L., Bowe-Geddes, L. A., *et al*. (2009) Infiltration and extravasation: update on prevention and management. *Journal of Infusion Nursing*, 32 (4), 203–211.
9. Ray-Barruel, G., Polit, D. F., Murfield, J. E. & Rickard, C. M. (2014) Infusion phlebitis assessment measures: a systematic review. *Journal of Evaluation in Clinical Practice*, 20 (2), 191–202.
10. Birkenshaw, R., Ali, B. & Sammi, I. (1997) Local infections at cannula site. *Journal of Accident and Emergency Medicine*, 14 (3), 199.
11. May, J., Murchan, P., MacFie, J., *et al*. (1996) Prospective study of the aetiology of infusion phlebitis and line failure during peripheral parenteral nutrition. *British Journal of Surgery*, 83, 1091–1094.
12. Madan, M., Alexander, D. J. & McMahon, M. J. (1992) Influence of catheter type on occurrence of thrombophlebitis during peripheral intravenous nutrition. *Lancet*, 339 (8785), 101–103.
13. Rickard, C. M., Webster, J., Wallis, M. C., *et al*. (2012) Routine versus clinically indicated replacement of peripheral intravenous catheters: a randomised controlled equivalence trial. *Lancet*, 380 (9847), 1066–1074.
14. Monreal, M., Quilez, F., Rey-Joly, C., *et al*. (1999) Infusion phlebitis in patients with acute pneumonia: a prospective study. *Chest*, 115 (6), 1576–1580.
15. Lai, K. (1998) Safety of prolonging peripheral cannula and IV tubing use from 72 hours to 96 hours. *American Journal of Infection Control*, 26 (1), 66–70.
16. Maki, D. G. & Ringer, M. (1991) Risk factors for infusion-related phlebitis with small peripheral venous catheters. A randomized controlled trial. *Annals of Internal Medicine*, 114 (10), 845–854.
17. Barker, P., Anderson, A. D. & MacFie, J. (2004) Randomised clinical trial of elective re-siting of intravenous cannulae. *Annals of the Royal College of Surgeons of England*, 86 (4), 281–283.
18. Baxter Healthcare Ltd (1988). Principles and Practice of IV Therapy. Compton, Berks: Baxter Healthcare Ltd.
19. Catney, M. R., Hillis, S., Wakefield, B., *et al*. (2001) Relationship between peripheral intravenous catheter dwell time and the development of phlebitis and infiltration. *Journal of Infusion Nursing*, 24 (5), 332–341.
20. Curran, E. T., Coia, J. E., Gilmour, H., McNamee, S. & Hood, J. (2000) Multi-centre research surveillance project to reduce infections/phlebitis associated with peripheral vascular catheters. *The Journal of Hospital Infection*, 46 (3), 194–202.
21. Lanbeck, P., Odenholt, I. & Paulsen, O. (2002) Antibiotics differ in their tendency to cause infusion phlebitis: a prospective observational study. *Scandinavian Journal of Infectious Diseases*, 34 (7), 512–519.
22. Rittenberg, C. N., Gralla, R. J. & Rehmeyer, T. A. (1995) Assessing and managing venous irritation associated with vinorelbine tartrate (Navelbine). *Oncology Nursing Forum*, 22 (4), 707–710.

23. Van Donk, P., Rickard, C. M., McGrail, M. R. & Doolan, G. (2009) Routine replacement versus clinical monitoring of peripheral intravenous catheters in a regional hospital in the home program: a randomized controlled trial. *Infection Control and Hospital Epidemiology*, 30 (9), 915–917.
24. Trinh, T. T., Chan, P. A., Edwards, O., *et al*. (2011) Peripheral venous catheter-related Staphylococcus aureus bacteremia. *Infection Control and Hospital Epidemiology*, 32 (6), 579–583.
25. de Vet, H. C., Mokkink, L. B., Terwee, C. B., Hoekstra, O. S. & Knol, D. L. (2013) Clinicians are right not to like Cohen's kappa. *BMJ (Clinical Research Ed.)*, 346, f2125.
26. Sim, J. & Wright, C. C. (2005) The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical Therapy*, 85 (3), 257–268.
27. Webster, J., Morris, H. L., Robinson, K. & Sanderson, U. (2007) Development and validation of a Vein Assessment Tool (VAT). *The Australian Journal of Advanced Nursing: A Quarterly Publication of the Royal Australian Nursing Federation*, 24 (4), 5–7.
28. Tager, I. B., Ginsberg, M. B., Ellis, S. E., *et al*. (1983) An epidemiologic study of the risks associated with peripheral intravenous catheters. *American Journal of Epidemiology*, 118 (6), 839–851.

