Reliability (Inter-rater Agreement) of the Barthel Index for Assessment of Stroke Survivors
Systematic Review and Meta-analysis

Laura Duffy, MRCP; Shelley Gajree, MRCP; Peter Langhorne, PhD; David J. Stott, MD; Terence J. Quinn, MD

Background and Purpose—The Barthel Index (BI) is a 10-item measure of activities of daily living which is frequently used in clinical practice and as a trial outcome measure in stroke. We sought to describe the reliability (interobserver variability) of standard BI in stroke cohorts using systematic review and meta-analysis of published studies.

Methods—Two assessors independently searched various multidisciplinary electronic databases from inception to April 2012 inclusive. Inclusion criteria comprised: original research, human stroke participants, and inter-rater reliability data on equivalent methods of BI administration. Manuscripts were reviewed against prespecified inclusion criteria. Primary outcome for meta-analysis was reliability, measured by weighted $\kappa$ (kw).

Results—From 20210 titles, 306 abstracts were reviewed, 12 studies met inclusion criteria, and 10 were included in meta-analysis (n=543 participants; range of participants in studies, 7–21). There was substantial clinical heterogeneity with respect to method of BI application; population studied and assessors. Two papers were graded high quality. Overall interobserver reliability of standard administration of the BI was excellent ($\kappa$, 0.93; 95% confidence interval, 0.90–0.96 random effects modeling).

Conclusions—The BI has excellent inter-rater reliability for standard administration after stroke. However, included studies were modest in size, with clinical heterogeneity and variable methodological quality. Despite these limitations, standard BI seems an appropriate outcome measure for stroke trials and practice. (Stroke. 2013;44:462-468.)

Key Words: Barthel Index ■ clinimetrics ■ clinical trials ■ outcomes ■ scales ■ stroke
Participants
Study population was any age, human, stroke survivors. We included all clinical and pathological stroke subtypes (ischemic, hemorrhagic, transient ischemic attack). Where study population was mixed stroke and nonstroke, we included the study if population of stroke survivors was greater than an arbitrary cut point of 85%. We used no restrictions relating to the BI assessor and specifically did not exclude studies on the basis of background or training of observers.

Study Methodology
All studies purporting to measure BI reliability through patient interview (interobserver variability of BI scoring) were reviewed with no specific restriction on the basis of study design, intervention, or language. We excluded studies that compared differing BI measures, for example, reliance of structured telephone questionnaire versus unstructured face-to-face assessment. We did not prespecify exclusion criteria relating to numbers of participants or assessors.

Outcomes
We applied no restrictions on the basis of BI assessment methodology. We prespecified subgroup analysis based on method of BI assessment (face-to-face, telephone). We recognized that several we use the label BI, we included all papers that purported to assess BI but our primary analysis was limited to those scales based on the original 10-item scale—standard BI.1

Search Strategy
We devised a sensitive search strategy with assistance from an information scientist. We interrogated a comprehensive battery of cross-discipline electronic databases from inception to April 2012 inclusive: ARIF (University of Birmingham), Cochrane Database of Systematic Reviews (Cochrane Library), CINAHL (EBSCO), EMBASE (EMBASE, EMBASE Classic (OVID), LILACS (BIOMED), Google Scholar (Google), MEDION (Universities of Maastricht and Leuven), MEDLINE (OVID), OLDMEDLINE (OVID), and Web of Science (Thomson Reuters). We formulated keywords using National Library of Medicine Medical Subject Headings headings and study-specific terms; keywords were designed to be as inclusive as possible. We created search terms according to key concepts of: (1) functional assessment, (2) stroke, and (3) reliability assessment. If the electronic database allowed, we used explosion features for keyword searching (Figure 1). In addition to the electronic database search, we handsearched contemporary reviews, key reference works,10,11 and high impact journals in the field of stroke medicine (Age and Ageing, Cerebrovascular Diseases, International Journal of Stroke, Journal of the American Geriatrics Society, Stroke). To identify studies not yet in print, we handsearched proceedings of scientific meetings for the period January 2010 to May 2012 (American Stroke Association—International Stroke Conference; European Stroke Conference; World Stroke Organization—World Stroke Congress). We searched bibliographies of all retrieved articles for further references, and the process was repeated until no new articles were found.

From the list of included titles, we reviewed abstracts for appropriateness to the study question. We retrieved the full text of any article that either reviewer suspected may be relevant. Where potentially relevant data were not available in the published manuscript, we attempted electronic contact with the original authors. For those studies not published in English language, translation services were used.

We assessed internal and external validity of the search. A third researcher (T.Q.) checked a random selection of 1000 titles for inclusion and results were compared with the researchers’ consensus inclusion list. Before formulating the search strategy, we selected 5 papers relevant to BI properties.12,13,14 We cross-checked our list of titles from literature searching to ensure all papers were included. Our search strategy was developed independent of and blinded to the choice of 5 titles for validation.

Data Extraction
We extracted data to a prespecified and piloted proforma. Data extraction was performed independently and blinded by 2 researchers (L.D., T.Q.). Disagreement was resolved by discussion and consensus.

Quality Assessment and Risk of Bias
We assessed study quality and risk of bias using a prespecified bespoke tool based on the Guidelines for Reporting Reliability and Agreement studies (GRRAS).13 Important elements were: description of the study population (recruitment, demographics, stroke severity); description of the assessors (number, occupation, experience with functional assessment, or BI); description of time between first and second assessment; description of methodology of BI administration; description of training of the assessor in BI; description of sample size calculation; and description of blinding.

Each category was scored high quality if described in sufficient detail in the paper and no potential for bias. A priori we defined high quality of paper as a high quality score on ≥5 of the 7 criteria. We piloted the tool using 2 papers and revised. Quality assessment was performed independently and blinded by 2 researchers (L.D., T.Q.) with disagreement resolved by discussion and consensus.

Statistical Analysis
Reliability of BI is traditionally described using either κ statistics (this may be weighted to account for degree of difference between observers [kw]), intraclass correlation coefficient (ICC), Bland Altman method, or percentage agreement between observers. We included any of these descriptors and derived others if data allowed. ICC equates to kw when quadratic weighting is applied. κ Statistics range from κ=0.00 (no agreement other that that expected by chance) to κ=1.00 (complete agreement). We used standard definitions of poor (κ=0.00–0.20), fair (κ = 0.21–0.40), moderate (κ=0.41–0.60), good (κ=0.61–0.80), and very good (κ=0.81–1.00) agreement.16

For meta-analysis, we used a single-group descriptive analysis. We described standardized differences and corresponding 95% confidence intervals (95% CIs). We present both fixed and random effects model data. We assessed consistency (heterogeneity) by visual inspection of study data plots. We planned to explore potential publication bias by plotting κ against the inverse of its standard error to allow visual inspection of symmetry. All calculations were performed using Comprehensive Meta-analysis software (Version 2, Biostat, Englewood).

Results
From 20 210 original titles, 306 abstracts were eligible for review, 35 papers were initially considered for inclusion, and 12 studies involving 627 patients met our inclusion criteria.17–28 Of these, 10 (n=543 subjects) had data that allowed for meta-analysis.17–26 Study populations were exclusively stroke survivors. Four reports required translation (German, Dutch, and Portuguese).17–20 We contacted 3 authors for additional information and received additional data from 2 included studies were from various countries (n=9), using assessors of differing background and experience and assessing BI at various time points poststroke (range, 14 days to >1 year; Table 1)

Internal validity of our search strategy was confirmed as an independent third reviewer found no titles other than those already selected by the original 2 researchers. External validity of our search strategy was confirmed as all prespecified titles were included in the search results.

Variability in BI scoring was described using various measures; κ statistics (n=3), kw (n=6), ICC (n=4), and percentage agreement (n=2). Interobserver variability of BI varied from near perfect (kw≥0.99) to good (κ=0.62) in the original descriptions.(Table 2) To make use of the largest dataset, quadratic kw, and ICC scores were combined and used for the meta-analysis. For those studies that described results in terms of κ and agreement, there were insufficient
data presented in the original reports to allow back derivation of \( \kappa \)w or ICC.

Meta-analysis suggested overall very good reliability for standard face-to-face BI (\( \kappa \)w, 0.95; 95% CI, 0.94–0.96 fixed effects modeling; \( \kappa \)w, 0.93; 95% CI, 0.90–0.96 random effects modeling); there were insufficient suitable data to perform meta-analysis for other methods of BI assessment. (Figure 2).

Visual inspection suggested modest statistical heterogeneity. There was substantial clinical heterogeneity across the included studies, particularly with respect to BI assessors, population studied, and numbers included. Our quality assessment tool described 2 papers as high quality. The most common limitations in published descriptions of studies were lack of sample size calculation (n=11), poor description of blinding (n=10), and poor description of BI

---

**Bibliographic databases searched**

ARIF; CINAHL; Cochrane Central Register of Controlled Trials; Cochrane Database of Systematic Reviews; EMBASE; Google Scholar; LILACS; MEDION; MEDLINE; OLDMEDLINE; Web of Science

---

<table>
<thead>
<tr>
<th>Concept 1: Functional assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barthel Index*; Disability evaluation*; Assessment, Outcome*; Assessment, Patient Outcome*; Outcome Assessment, Patient; Outcome Measure*; Outcomes Assessment*; Outcomes Research*; Patient Outcome Assessment*; Rehabilitation Outcomes*; Activities*, Daily Living*; ADL; Chronic Limitation of Activity*; Self Care*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Concept 2: Stroke</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apoplexy*; Cerebral Stroke*; Cerebrovascular Accident*; Cerebrovascular Apoplexy*; Cerebrovascular Stroke; CVA; Cerebrovascular Accident*; Stroke*; Vascular Accident, Brain</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Concept 3: Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reproducibility of Results*; Reliability*; Validity; Reproducibility of Findings*; Validity of Result*; Observer Variation*; Bias, Observer*; Inter-Observer Variability*; Inter-Observer Variation*; Interobserver Variability*; Interobserver Variation*; Intra-Observer Variability*; Intra-Observer Variation*; Intraobserver Variability*; Intraobserver Variation</td>
</tr>
</tbody>
</table>

---

**Figure 1.** Profile of the systematic review.

---

**Figure 2.** A summary of the data presented.
training (n=7) (Figure 3). Given the small number of included studies in the meta-analysis with limited spread of standard error, we did not use funnel plot asymmetry to assess for publication bias.

Certain studies described reliability across individual items of the BI (n=7). The items bladder and bowel continence were consistently rated as having very good reliability (Table 3). We performed 1 post hoc subanalyses comparing those papers where raters are described as having training in BI application and administration (n=5 papers, 4 suitable for meta-analysis)\textsuperscript{19,20,22,26} with those papers that have no description of using BI training (n=7, 6 suitable for meta-analysis).\textsuperscript{17,18,21,23–25} Meta-analysis (random effects model) suggested a strong trend toward improved reliability in those papers that describe BI training:

\[ \kappa_w \text{ (trained) } 0.95 \text{ (95\% CI, 0.92–0.98)}; \kappa_w \text{ (no training) } 0.91 \text{ (95\% CI, 0.89–0.92)}. \]

Discussion
We studied published reports of reproducibility of BI assessment scores. Our results suggest excellent interobserver reliability of the standard BI as a stroke outcome measure.

The reliability of BI is often quoted as a particular strength of this outcome measure for use as a stroke trial end point.\textsuperscript{29} A previous systematic review of the reliability of BI is available,
but this study excluded stroke survivors.12 We present the first systematic review of the interobserver reliability of BI in stroke. Our review findings are broadly similar to the non-stroke analysis12—published accounts describe good reliability of BI but clinical heterogeneity and variable study quality preclude any definitive statements.

The preferred functional outcome measure in stroke research is the mRS.3 Interobserver variability in mRS is perceived as a potential limitation of this measure, although use of novel administration techniques may improve mRS reliability.30 Based on our data, BI would seem the more reliable outcome measure, although differences are modest: \( \kappa_w = 0.95 \) for BI, compared with \( \kappa_w = 0.90 \) for standard mRS in a recent meta-analysis.3 Reliability is only one of the important clinimetric properties to be considered when choosing a functional outcome measure. Other favorable clinimetric properties of BI include various measures of validity and acceptability.2 However, BI performs less well for other important properties, most notably responsiveness to clinical change and well-recognized floor and ceiling effects.2 In this respect, mRS remains preferable to BI14 and we would not suggest replacing mRS as the study end point of choice. As the 2 scales measure differing constructs, activities of daily living (BI) and global activity with a focus on walking (mRS), measurement of both may have some utility.

The utility of meta-analysis is dependent on the quality of the included papers. In this regard, our analysis is limited by the substantial clinical heterogeneity between included papers and problems with potential for bias. To limit heterogeneity, we focussed on studies that compared equivalent BI administration methods. We recognize that many good studies have described reliability of differing BI techniques, for example, structured telephone versus face-to-face interview.31 Study quality and potential for bias was a concern, only 2 of the included papers were graded as high quality and no papers fulfilled all of our prespecified quality criteria. Specific reporting guidance for reliability studies is relatively new, and we encourage future studies to make use of this resource.15 Our data were not suited to conventional assessment of publication bias, and it is possible that the favorable reliability described could be, in part, because of publication bias. However, unlike intervention trials, there is no reason to suspect that results of a psychometric study would impact on its chance of publication.

### Figure 2.
Forest plot describing Barthel Index reliability (\( \kappa_w \)) across studies. CI indicates confidence interval.

### Figure 3.
Quality assessment of included studies. n indicates no and y, yes.
It has been demonstrated that training improves consistency in application of stroke assessment scales. Training and certification resources are available for common stroke scales (mRS; National Institutes of Health Stroke Scale). These are recognized by regulatory authorities and have become industry standard for large-scale trials. Our subanalysis suggested that training may be associated with improved reliability of BI. This analysis comes with certain caveats. The meta-analysis was as comprehensive and generalizable as possible. Paper search was intended to capture only individual item reliability reports for BI. Several papers may have used training but not described this in the reliability of BI. This analysis comes with certain caveats. Nonetheless, the ideal study methodology would involve a large series from non-English and nonmedical sources, and several non-English language reports were included in the final analysis. Our literature searching strategy was as comprehensive and systematic as possible. The spread of reliability estimates obtained suggests no overt publication bias.

There were limitations to our study methodology. No universally accepted method for analysis of multiple k statistics from differing populations has been described. Analysis of k statistics is subject to numerous caveats. In general, reliability of a tool is a factor of the tool itself, the subjects, the assessor, and the setting, and so k will vary with population size/number of assessors. To give a summary of reliability across multiple studies, we performed a meta-analysis that made the fewest assumptions of the data. We used this to give a summary estimate, but recognize that some statisticians do not feel k statistics should be pooled. Similarly, our analyses of heterogeneity and publication bias were the analyses that made fewest assumptions of the data.

If we wish to improve our understanding of how BI performs as a clinical and trial outcome measure in contemporary practice, the ideal study methodology would involve a large series of trained observers of differing backgrounds and from differing international centers, assessing BI, using a standard methodology, on unselected patients at a predetermined time point after discharge. The importance of studying the properties of a scale in a setting that mirrors clinical practice can be seen from review studies describing mRS reliability. Many of the published studies reporting reasonable mRS reliability are single center, with small numbers of raters and stroke survivors. However, interobserver variation is considerable in those studies that are multicenter, with multiple raters from different professional backgrounds and large numbers of stroke survivors. This is the situation that more closely mirrors a contemporary stroke intervention study. We found no published papers with a design similar to our ideal. Thus, suitably powered, multicentre studies of BI properties will still provide useful data.

Our data suggest excellent reliability of BI as a stroke outcome measure across the published literature, although studies included in our meta-analysis may not be representative of contemporary, multicenter stroke trials. As there was a suggestion of improved reliability with BI training, we would encourage use of standardized methods of BI administration or training.

**Disclosures**

Dr Quinn has assisted with creation of training resources for stroke scales and has received payment from Training Campus to support this work. The other authors have no conflicts to report.

**References**


Reliability (Inter-rater Agreement) of the Barthel Index for Assessment of Stroke Survivors: Systematic Review and Meta-analysis
Laura Duffy, Shelley Gajree, Peter Langhorne, David J. Stott and Terence J. Quinn

Stroke. 2013;44:462-468; originally published online January 8, 2013;
doi: 10.1161/STROKEAHA.112.678615

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://stroke.ahajournals.org/content/44/2/462

Data Supplement (unedited) at:
http://stroke.ahajournals.org/content/suppl/2013/06/27/STROKEAHA.112.678615.DC1

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Stroke can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Stroke is online at:
http://stroke.ahajournals.org//subscriptions/
脳卒中生存者の評価に関するバーセル指数の信頼性（評価者間の一致）
系統的レビューおよびメタ解析

Reliability (Inter-rater Agreement) of the Barthel Index for Assessment of Stroke Survivors
Systematic Review and Meta-analysis

Laura Duffy, MRCP1; Shelley Gajree, MRCP1; Peter Langhorne, PhD2; David J. Stott, MD2; Terence J. Quinn, MD2
1 Department of Geriatric Medicine, Glasgow Royal Infirmary, Glasgow, United Kingdom; and 2 Institute of Cardiovascular and Medical Sciences, School of Medicine, University of Glasgow, Glasgow, United Kingdom.

Abstract

背景および目的: バーセル指数 (BI) は、日常生活動作に関する10項目の尺度であり、脳卒中診療、および脳卒中に関する臨床試験の転帰の尺度として用いられることが多い。これらの臨床試験に関する系統的レビューおよびメタ解析を用いて、脳卒中コホートにおける標準的なBIの信頼性（観察者間の誤差）を検討した。

方法: 2名の評価者がそれぞれ個別に2012年4月までを含めた様々な集学的電子データベースを検索した。選択基準は、新規の試験、脳卒中対象者、およびBI使用と同等の方法に関する評価者間信頼性のデータを含めた。事前に特定した選択基準に基づいて論文をレビューした。メタ解析の主要評価項目は、重み付けκ係数（κw）により評価した信頼性とした。

結果: 20,210件のタイトルから306件の抄録をレビューしたところ、12件の試験が選択基準と一致し、うち10件をメタ解析に組み入れた（症例n = 543、各試験の登録者数は7～21例の範囲）。BIの適用方法、試験対象集団、評価者に関して、かなりの臨床的不一致が認められた。質の高い研究は2件であった。BIの標準的使用に関する全体的な観察者間信頼性は、良好であった[κw = 0.93、95%信頼区間(CI): 0.90 ～ 0.96、変量効果モデル]。

結論: 脳卒中後のBI標準的使用は、評価者間信頼性において良好であった。しかし、選択した試験は規模が比較的小さく、臨床的不一致と方法論的な質のばらつきがあった。このような限界はあるものの、標準的なBIは、脳卒中試験および診療において適切な転帰尺度であると考えられる。

Stroke 2013; 44: 462-468

表3 各項目に関するバーセル指数の信頼性

<table>
<thead>
<tr>
<th>BI 項目</th>
<th>de Haan et al</th>
<th>Heuschmann et al</th>
<th>Küçükdeveci et al</th>
<th>Leung et al</th>
<th>Loewen et al</th>
<th>Oveisgharan et al</th>
<th>Schlote et al</th>
</tr>
</thead>
<tbody>
<tr>
<td>歩行</td>
<td>κw, 0.96</td>
<td>κw, 0.97(0.93 ～ 1.01)</td>
<td>κ = 0.67</td>
<td>κ = 0.90</td>
<td>κ = 0.66</td>
<td>ICC, 0.87</td>
<td>ICC, 1.00</td>
</tr>
<tr>
<td>入浴</td>
<td>κw, 0.96</td>
<td>κw, 0.92(0.82 ～ 1.01)</td>
<td>κ = 0.61</td>
<td>κ = 0.81</td>
<td>κ = 0.88</td>
<td>ICC, 0.63</td>
<td>ICC, 0.79</td>
</tr>
<tr>
<td>排尿</td>
<td>κw, 0.96</td>
<td>κw, 0.91(0.84 ～ 0.99)</td>
<td>κ = 0.97</td>
<td>κ = 0.91</td>
<td>κ = 1.00</td>
<td>ICC, 0.62</td>
<td>ICC, 1.00</td>
</tr>
<tr>
<td>排便</td>
<td>κw, 0.96</td>
<td>κw, 0.97(0.92 ～ 1.01)</td>
<td>κ = 0.95</td>
<td>κ = 0.91</td>
<td>κ = 1.00</td>
<td>ICC, 0.59</td>
<td>ICC, 1.00</td>
</tr>
<tr>
<td>更衣</td>
<td>κw, 0.96</td>
<td>κw, 0.92(0.86 ～ 0.97)</td>
<td>κ = 0.78</td>
<td>κ = 0.83</td>
<td>κ = 0.72</td>
<td>ICC, 0.78</td>
<td>ICC, 0.90</td>
</tr>
<tr>
<td>食事</td>
<td>κw, 0.96</td>
<td>κw, 0.89(0.80 ～ 0.98)</td>
<td>κ = 0.67</td>
<td>κ = 0.87</td>
<td>κ = 1.00</td>
<td>ICC, 0.75</td>
<td>ICC, 0.86</td>
</tr>
<tr>
<td>整容</td>
<td>κw, 0.97</td>
<td>κw, 0.94(0.87 ～ 1.02)</td>
<td>κ = 0.63</td>
<td>κ = 0.91</td>
<td>κ = 0.47</td>
<td>ICC, 0.77</td>
<td>ICC, 1.00</td>
</tr>
<tr>
<td>階段昇降</td>
<td>κw, 0.96</td>
<td>κw, 0.99(0.96 ～ 1.01)</td>
<td>κ = 0.77</td>
<td>κ = 0.81</td>
<td>κ = 1.00</td>
<td>ICC, 0.82</td>
<td>ICC, 1.00</td>
</tr>
<tr>
<td>移動</td>
<td>κw, 0.96</td>
<td>κw, 0.93(0.87 ～ 0.99)</td>
<td>κ = 0.82</td>
<td>κ = 1.00</td>
<td>κ = 0.59</td>
<td>ICC, 0.87</td>
<td>ICC, 0.98</td>
</tr>
<tr>
<td>トイレ動作</td>
<td>κw, 0.96</td>
<td>κw, 0.88(0.81 ～ 0.96)</td>
<td>κ = 0.84</td>
<td>κ = 0.90</td>
<td>κ = 0.74</td>
<td>ICC, 0.87</td>
<td>ICC, 1.00</td>
</tr>
</tbody>
</table>

BI:バーセル指数, ICC:級内相関係数, κw:重み付けκ係数。