Barthel Index for Stroke Trials
Development, Properties, and Application

Terence J. Quinn, MD, MBChB (hons), BSc (hons), MRCP (UK); Peter Langhorne, PhD; David J. Stott, MD

Background and Purpose—Robust measures of functional outcome are required to determine treatment effects in stroke trials. Of the various measures available, the Barthel index (BI) is one of the more prevalent. We aimed to describe validity, reliability, and responsiveness (clinimetric properties) of the BI in stroke trials.

Methods—Narrative review of published articles describing clinimetric properties or use of the BI as a stroke trial end point.

Results—Definitive statements on properties of BI are limited by heterogeneity in methodology of assessment and in the content of “BI” scales. Accepting these caveats, evidence suggests that BI is a valid measure of activities of daily living; sensitivity to change is limited at extremes of disability (floor and ceiling effects), and reliability of standard BI assessment is acceptable. However, these data may not be applicable to contemporary multicenter stroke trials.

Conclusions—Substantial literature describing BI clinimetrics in stroke is available; however, questions remain regarding certain properties. The “BI” label is used for a number of instruments and we urge greater consistency in methods, content, and scoring. A 10-item scale, scoring 0 to 100 with 5-point increments, has been used in several multicenter stroke trials and it seems reasonable that this should become the uniform stroke trial BI. (Stroke. 2011;42:00-00.)

Key Words: activities of daily living ▪ Barthel index ▪ clinimetrics ▪ disability scales ▪ outcomes

Stroke is a disabling condition, with cerebrovascular diseases being the leading cause of disability in industrial countries. Thus, efficacy of stroke interventions is often described via measures of disability, ie, functional assessment.1 Stroke assessments that focus on basic activities of daily living, “chronic disease” hospitals of Baltimore, a “Maryland disability index” was developed.6 Dr Florence I. Mahoney and Dorothea W. Barthel modified this scale to produce “a simple index of independence, useful in scoring improvement in rehabilitation,” ie, the BI.7

BI has become a prevalent outcome measure for stroke, with substantial supporting research.1,3 This review discusses development and application of the BI in its many iterations, giving particular attention to “clinimetric” properties. Clinimetrics is the methodological discipline that focuses on quality of clinical measurements. Scales are traditionally assessed for validity, reliability, and responsiveness, and these are described in turn. Such analysis is of more than academic interest; even the best designed trials will produce meaningless data if assessment scales used are not “clinimetrically” fit for purpose.

Materials and Methods
The review is based on the authors’ clinical and research experience and is informed by a search of published literature. Electronic databases (Medline and Embase) were searched from inception to September 2010 inclusive, using the following truncated key words: “Barthel,” “activities of daily living,” “disability evaluation,” and “stroke or cerebrovascular.” In addition, key reference works2,4 and selected journals (Stroke, Age and Ageing, Archives of Physical Medicine and Rehabilitation) were manually searched for relevant articles. Particular attention was given to studies describing clinimetrics or use of BI in stroke trials. The intention was to provide a narrative overview and appraisal of the strengths and weaknesses of the BI as an outcome measure for stroke trials. It should be noted that although this critique is informed by published literature, it is not a fully comprehensive systematic review.

Results
Development of the BI
As rehabilitation became established as a medical discipline, many scales offering objective measures of recovery were described.9 These instruments were usually developed “in house” and often are not subject to further assessment. In the “chronic disease” hospitals of Baltimore, a “Maryland disability index” was developed.6

The scale described 10 tasks and was scored according to amount of time or assistance required by the patient. Total score was from 0 to 100, with lower scores representing greater nursing dependency. First used in approximately 1955, Barthel’s eponymous scale became popular in rehabilitation and was well-established by time of publication.6 Use of the BI spread quickly, such that it is now...
arguably the most popular ADL scale in clinical practice.3 Examples of BI assessment in studies of spinal injury, burns, cardiac disease, rheumatoid arthritis, amputations, and frail elderly are available.2

Although not designed for clinical trials and not specifically a stroke scale, BI has been used as a trial end point, either singly or as part of a “global” measure, in landmark studies of thrombolysis and acute stroke units, and now BI is second only to the modified Rankin scale (mRS) as stroke outcome measure of choice1 (Figure 2). BI use is international and native language versions are described for several countries.8 However, not all non-English language versions of BI have undergone the recommended forward-and-back translation process,9 and certain translated BI scales are thought to be inappropriate for the target population.10

Variations and Modification to the BI

Several authors have proposed modifications to Barthel’s original scale. These modifications have variously reordered scale items,11 changed or expanded on definitions,12 changed scoring,13 and added/removed items.14 Distinguishing between these BI scales is crucial, because even minor changes to scales can produce substantial differences in scoring.15 It is unfortunate that many of these BI variations maintain the descriptor “BI.” At least 4 stroke scales in common usage are described as Barthel (Figure 1). This confuses the literature and complicates attempts at comparative or meta-analysis. There is no consensus on the optimal version, although for consistency we urge that a single version of BI is adopted for stroke trials. The 10-item scale, scoring 0 to 100 with 5-point increments (Figure 1), has been used in several multicenter stroke trials, and in the absence of any clearly superior “Barthel” it seems reasonable that this should become the uniform stroke trial BI. This scale is equivalent in content to Collin and Wade’s BI (scored 0–20);11 the change in scoring values will not alter the other properties of the scale and so clinimetrics also should be equivalent. Regardless of scale chosen, it is good practice to describe the scale used and to reference the original descriptor. In the remainder of this article, we use the term BI to refer generically to scales based on Mahoney and Barthel’s tool; specific alternative versions of the scale are identified in the text.

Extended and truncated modifications of the BI are available and deserve comment because they have been used in stroke trials. An “extended BI” comprises BI with additional components from the functional independence measure.16

<table>
<thead>
<tr>
<th>Task</th>
<th>With help</th>
<th>Independent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feeding</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>Moving from wheelchair to bed</td>
<td>5-10</td>
<td>15</td>
</tr>
<tr>
<td>Personal toilet (wash, shave, comb)</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Getting on / off toilet</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>Bathing</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Walking on level surface</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>Ascend, descend stairs</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>Dressing</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>Controlling bowels</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>Controlling bladder</td>
<td>5</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Task</th>
<th>Dependental</th>
<th>Independent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drinks from cup/ feed from dish</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Dress upper body</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Dress lower body</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Don brace or prosthesis</td>
<td>-2</td>
<td>0</td>
</tr>
<tr>
<td>Grooming</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Wash or bathe</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Bladder continence</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Bowel continence</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Care of perineum/clothing at toilet</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Transfer chair</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Transfer toilet</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Transfer tub or shower</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Walk on level 50 yards</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>Up and down stairs for one flight</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Wheelchair 50 yards if not walking</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 1. Four “Barthel indices” in common usage.
Addition of measures pertaining to cognition, expression, social interaction, and vision makes intuitive sense because BI does not address these areas. However, by adding new items, the nature of the scale is fundamentally changed and validity cannot be assumed. Certain disciplines have taken BI and added items specific to that field. For example, in neurorehabilitation an early rehabilitation BI is used, comprising original BI and additional categories describing need for intensive care, tracheostomy, and ventilation. Again, validity of any “bespoke Barthel” is questionable unless corresponding clinimetric assessment is offered.

Shortened versions of BI, comprising 3-, 4-, and 5-item scales are also described. Using various techniques, trialists have removed all but the most discriminating items in the scale. Interestingly, the items included differ across the truncated scales. The value of this approach in clinical practice is open to question, because 10-item BI is already considerably shorter than many other scales.

**BI Methodology**

In addition to heterogeneity in “Barthel” content, there is further heterogeneity in the methodology used to administer the scale. In the original description, BI was assessed through interview and distant observation, and this remains the standard assessment. For stroke trials, a variety of other techniques have been used. Pertinent areas of heterogeneity for clinical trial purposes are use of training, choice of interview subject, and method of data collection.

An assessment using the BI should describe only what the patient can do at time of grading. However, there may be a temptation for assessors to adjust BI scores based on what the patient should be able to do or would like to do. To standardize the assessment, guidance notes for BI administration are available, and in certain trials bespoke training has been offered. However, unlike other stroke scales, there is no internationally recognized training resource for BI. It has been suggested that written guidance alone may not improve rigor; rather, training and explicit descriptions of scoring for each item are required.

When BI is scored using an interview, the interviewer may be a doctor, nurse, therapist, or professional researcher. There are some data to suggest that profession may influence grading, although differences between graders are modest. More important may be choice of interview subject, which again may include nurse, care giver, family, or therapist. We could find no data to recommend one interviewee over another and, in practice, interview with several parties may be required. However, patients probably should not be directly interviewed. Studies have described poor validity of self-reported BI, particularly in older and cognitively impaired patients and both are cohorts likely to account for substantial numbers in stroke trials.
Method of conducting BI interview can also vary; again, trials have used varying methodologies assuming clinimetric properties without robust testing. For trials, assessments performed remote of a testing center or even without the need for a researcher offer economic and time advantages. Literature describing telephone assessment is limited but studies suggest telephone disability scoring may be systematically lower than for direct interview. For BI based on postal questionnaire, results have varied, although there is agreement that good responses rates can be achieved. Because studies to date have been modest in size, conducted in small geographic areas, and have not used exclusively stroke survivor cohorts, we still lack definitive data on utility of these methodologies for clinical trial use.

Clinimetric Properties of the BI
The ideal outcome measure would be easy to administer, show consistency with repeated use and multiple users, would capture information relevant to patient and trialists, and detect small changes over time. No perfect outcome measure exists or is likely to ever exist. The relative importance of various clinimetric properties will depend on the proposed application. In a multicenter trial with outcome data collected at a single time point, validity and interobserver reliability are arguably the most important properties.

Reliability
Reliability describes measurement error associated with an instrument; it can be assessed across several domains. Internal consistency, traditionally measured with Cronbach’s α, is the extent to which all items in a scale measure a single factor. Higher values indicate greater consistency, although very high values may signal a degree of redundancy in the component items. For BI, internal consistency has been described as good (α = 0.80–0.89) to excellent (α = 0.93). The reliability of repeated BI measures (test–retest reliability) is important in clinical work because serial measures are used to chart progress. For a clinical trial, such considerations are less important because outcomes are likely to be measured during a limited number of predefined times. When data are available, test–retest reliability of BI is usually described as good. A review of several scales suggested BI had better test–retest reliability than scales measuring extended ADL.

Substantial literature of BI interobserver reliability (ie, do independent observers agree on scores for a given subject) is available and reference texts generally describe this property as a particular strength of the scale. However, most studies have used only modest numbers of raters/patients with heterogeneity in assessment methodology and quality. Applicability of these data to a contemporary multicenter trial is questionable.

In a systematic review of BI in the elderly, consistent findings included greater reliability at higher BI scores, varying reliability dependent on assessor and interviewee, and reliability varying across items of the scale. No equivalent systematic review of BI in stroke is available. Reports of BI reliability in stroke, described using Cohen’s κ, range from moderate (κ = 0.41–0.60) to good (κ = 0.61–0.80) to very good (κ = 0.81–1.00). To put these figures in context, recent meta-analysis of mRS reliability reported scores ranging from κ of 0.25 to 0.95, and overall reliability was κ of 0.46.

Validity
Validity describes the extent to which an instrument measures what it purports to measure. In the absence of a “gold standard” ADL measure, other methods to gauge validity are required. Face validity of BI seems original, originally formulated for neurological and musculoskeletal disease; it is intuitive that BI would be a valid measure in stroke. The content of BI includes those domains thought to be most important to ADL measurement, although lack of measures pertaining to communication, mood, and cognition can create the anomalous situation in which a dependent stroke survivor achieves a good BI score. Therefore, the BI is primarily of value as a measure of physical dependency and is not of use (other than as a basic patient descriptor) in studies that target speech disorder (including dysphasia), depression, or cognitive function.

Stroke care and rehabilitation has changed considerably in the decades since the Barthel scale was first described. Modifications to BI scoring guidance have recognized that with appropriate assistive devices a degree of independence is possible even if impairments remain. Thus, for example, a stroke survivor with urinary incontinence can still score independence if they are catheterized and can perform catheter care.

Concurrent validity is suggested by close association between BI and clinical data such as amount of nursing time required by patients, extent of motor loss, and radiological size of infarct. Likewise, favorable construct validity is suggested by close correlation with other measures of activity. In fact, such is the widespread popularity of BI that it is often used as the gold standard comparator in studies of novel ADL scales. Comparison of simultaneously collected BI and mRS scores further proves validity, although mRS is probably superior for describing extremes of disability.

ADL are a proxy of how the stroke survivor will function in the home environment. Thus, ability of BI to predict return home provides further evidence of validity. Lower BI is associated with greater future disability, longer time to recovery, and greater care needs to facilitate recovery. In fact, BI measured at time of admission to a rehabilitation setting may be a better predictor of return home than “clinical” measures. Change in BI over a set time may be an even more powerful predictive tool. The predictive utility of early BI is not clearly demonstrated and certain authors have argued that BI measured before day 5 after the event is suboptimal. Although BI has reasonable prognostic utility, across various analyses of the GAIN trials mRS was superior for prediction of discharge destination, health care costs, and time spent at home.

Responsiveness
To describe improvement or deterioration, the outcome measure must be responsive to change. Across a certain range of poststroke disability, responsiveness of BI is reasonable and,
with 10 graded items, BI is more sensitive to change than other common stroke scales.\textsuperscript{38}

The minimal degree of change that is thought to be clinically significant will vary according to the trial. However, even clinically modest improvements in functioning can have substantial meaning to patients and can be important at a population level. Literature on the minimal clinically important differences detected with BI suggest that a change of \( \pm 2 \) points (BI scored 0–20) is meaningful and beyond measurement error.\textsuperscript{25,39}

A scale should span the complete distribution of the concept to be measured; therefore, for a stroke trial, BI should measure and detect change across the range of possible functional outcomes. Here, a weakness of BI becomes evident; BI is not sensitive to change at extremes of ability\textsuperscript{40,41} (Figure 3). These “floor” and “ceiling” effects limit utility of BI and, in particular, make the scale less discriminating in patients with severe or minor stroke events. For longer-term assessment, BI on its own is unlikely to be sufficiently sensitive and should be replaced or used along with other scales. Floor and ceiling effects are not apparent with other prevalent functional outcome measures such as mRS.

BI scale modifications designed to improve responsiveness are described.\textsuperscript{13} Efforts to improve sensitivity are to be lauded; however, the success of certain modifications have not been consistently demonstrated. Changes may simply add complexity with no other advantages.\textsuperscript{41,42} Greater detail (and hence greater administration time) provides qualitative information for clinical use, but this level of detail is rarely required by trialists.

Acceptability
Although we could find no studies formally exploring BI acceptability, few would argue against acceptability of BI to patients and assessors. Standard BI requires no direct testing and should take only minutes, making BI among the quickest of the ADL instruments. Time required for testing is a major factor in determining acceptability of a scale to therapists.\textsuperscript{43} The simplicity of BI makes it particularly suited to clinical trials; however, even this relatively quick tool may present too great a burden in patients with severe or minor stroke events. These floor and ceiling effects are a particular issue for strokes. These floor and ceiling effects are a particular issue for the spread of data are increasingly applied.\textsuperscript{48}

Interpreting BI Data for Stroke Trials
Statistical manipulation of BI data are problematic. The ordinal nonhierarchal nature of the scale invalidates many “standard” comparative tests. The use of total BI score for analysis assumes that all items are measuring a common domain and can be summed without weighting or standardization. Whether this is true in stroke is debatable, with some studies suggesting BI is 1-dimensional\textsuperscript{45} and others suggesting that certain items do not show internal consistency.\textsuperscript{46}

A common approach has been to dichotomize total BI, defining cut-offs that represent favorable and nonfavorable outcomes. Again, there is no standard approach, with good BI arbitrarily defined as total scores ranging from BI of 50 to 95, with the most prevalent cut-off point at BI >95.\textsuperscript{47} It has been suggested that key scores are BI <40 (representing complete dependence on others), BI >60 (transition from complete dependence to assisted independence), and BI >85 (representing independence with minor assistance as could be reasonably provided in a community setting).\textsuperscript{38} Some may consider poorest outcome after stroke to be death; unlike mRS, the BI does not have a separate score to represent mortality.

Use of dichotomized BI has been criticized as inefficient, making use of only part of a complete trial dataset. For example, with a cut-off BI score of >85, patients starting with minor impairment can make clinically important recovery but not have impact on trial results, whereas patients with very low BI may recover substantially but not reach the cut-off point. Analytical methods that measure change across the range of possible outcomes, particularly in minor or more severe strokes. These floor and ceiling effects are a particular issue for stroke trials and limit the potential utility of the scale.

Conclusions
In a previous review of ADL tools, it was commented that BI possessed advantages of completeness, sensitivity, suitability for statistical manipulation, and familiarity.\textsuperscript{49} Based on the literature described, strengths of BI are widespread use and ease of application. However, sensitivity of BI is poor across the range of possible outcomes, particularly in minor or more severe strokes. These floor and ceiling effects are a particular issue for stroke trials and limit the potential utility of the scale.

Despite the limitations, BI continues to be used by trialists and, as such, attempts to improve the clinical application of the scale are welcome. Development of BI as a trial outcome is hindered by the substantial heterogeneity within scales described as Barthel and the methodologies used to administer the scale. For consistency, we urge that a single version of BI is adopted.
for stroke trials; the 10-item scale, scoring 0 to 100 with 5-point increments, is suggested as the uniform stroke trial BI.

Although there is extensive literature related to BI, questions regarding certain clinimetric properties remain. In particular, data relating to the reliability of BI for stroke trials and the optimal methodology for administration are lacking. Literature describing clinimetrics applicable to stroke trials and methods to improve properties of scales is emerging.20,30 There is an urgent need for clinimetric studies using BI after stroke.

Acknowledgments

The authors are grateful to the staff of Glasgow Royal Infirmary and document retrieval services.

Disclosure

Dr. Quinn has previously received grant funding to explore properties of the modified Rankin scale.

References
