



Benchmarking factor selection and sensitivity: a case study with nursing courses.

Journal:	<i>Studies in Higher Education</i>
Manuscript ID	CSHE-2016-0540.R1
Manuscript Type:	Article
Keywords:	Metric benchmarks, machine learning, RandomForest, factor selection, sensitivity

SCHOLARONE™
Manuscripts

Benchmarking factor selection and sensitivity: a case study with nursing courses

There is an increasing requirement in higher education worldwide to deliver excellence. Benchmarking is widely used for this purpose, but methodological approaches to the creation of benchmark metrics vary greatly. Approaches require selection of factors for inclusion and subsequent calculation of benchmarks for comparison. We describe an approach using machine learning to select input factors based on their value to predict completion rates of nursing courses. Data from over 36 000 students, from nine institutions over three years were included and weighted averages provided a dynamic baseline for year on year and within year comparisons between institutions. Anonymised outcomes highlight the variation in benchmarked performances between institutions and we demonstrate the value of accompanying sensitivity analyses. Our methods are appropriate worldwide, for many forms of data and at multiple scales of enquiry. We discuss our results in the context of higher education management, highlighting the value of scrutinising benchmark calculations.

Keywords: Metric benchmarks, machine learning, RandomForest, factor selection, sensitivity.

Background

There is a sustained and growing impetus worldwide to measure performance in higher education (HE) through the use of comparative quality metrics (Hazelkorn 2015). However, there is a lack of consensus on the choice of factors and methods that should be used to produce such metrics (Tam 2001) and approaches used to ‘benchmark’ metrics can have significant effects on how institutions are publicly ranked. The term ‘benchmarking’ is widely used in private and public domains, there are a variety of interpretations and processes used to implement benchmarking in the educational literature (Draper and Gittoes 2004). Formal benchmarking (hereafter referred to simply as ‘benchmarking’) is a technique originating in the business sector used to analyse

1
2
3 performance between competitors by comparing operating costs, product features and
4
5 operating capabilities (Asif 2015). Here, benchmarking can be used to compare
6
7 performance outcomes, but can also be used to inform the management of improvement
8
9 relative to competitors. The concepts underlying benchmarking in business have
10
11 undergone a long history of development (Kumar and Chandra 2001), resulting in many
12
13 different definitions and approaches (Zairi 1998; Kyrö 2003), while the core process
14
15 typically involves a performance comparison and the identification of factors for
16
17 improvement in individual businesses (Stapenhurst 2010).
18
19

20
21 Many potential benefits of HE benchmarking have been proposed, including
22
23 strengthening institutional ability to self-assess performance, evaluating the reasons for
24
25 sector differences to inform policy, and to managing improvement in performance
26
27 (Hazelkorn 2015). The global nature of HE has led to reviews of international
28
29 benchmarks, highlighting how these measures influence competition for students, staff
30
31 and research funding (e.g. HESA 2011). Careful consideration of metrics like these are
32
33 required when they are used for strategic decision-making, particularly to account for
34
35 the high levels of heterogeneity that exists both within and between universities
36
37 (Agasisti and Bonomi 2014). Measurement of performance or efficiency in HE usually
38
39 relies on data relating to the output of degree courses, such as qualification rate, student
40
41 employment and levels of student satisfaction or engagement. The effective
42
43 identification and adoption of such performance indicators is necessary to fulfil the first
44
45 goal of most benchmarking projects, where performance is compared amongst
46
47 institutions. A second goal of many benchmarking projects is to identify key areas for
48
49 enhancement, most often in consultation with a range of stakeholders. This should
50
51 identify best practices and focus resources on particular areas for improvement. While
52
53 the first goal of merely performing benchmarking is widespread and on the rise, the
54
55
56
57
58
59
60

1
2
3 latter goal of acting on benchmarking results to improve quality is sometimes
4
5 challenging (e.g. Tillema 2010).
6

7
8 The education and health sectors have been subject to significant performance
9
10 comparisons, which has led to league tables of performances and pressures to score
11
12 highly in a suite of performance indicators (e.g. Draper and Gittoes 2004; Northcott and
13
14 Llewellyn 2005; Shober 2013; Hazelkorn 2015). Benchmarking in Higher Education
15
16 (HE) has been described as a process through which performance and practice are
17
18 analysed to provide a standard measurement (the 'benchmark') of effective performance
19
20 (e.g. 'league tables' or publication of 'standard performance indicators'; Draper and
21
22 Gittoes 2004). This is not a simple process and to provide equitability in HE
23
24 benchmarking systems, direct comparisons of institutions should account for inherent
25
26 differences between institutions, such as heterogeneity of students, staff and
27
28 programmes of study. This is a challenge for benchmarking in HE and relies on the
29
30 principle that there are established patterns of success for any given set of student or
31
32 institutional typologies. Weighting metrics relative to an expected average of
33
34 performance across these typologies, rather than simply using absolute measures, can
35
36 mitigate this issue and is an essential part of benchmarking in the business sector
37
38 (Stapenhurst 2009). Despite the advantage of adjustments of this type, the use of simple
39
40 ranked metrics is common in HE despite well-documented shortcomings (Saisana et al
41
42 2011). Williams and de Rassenfosse (2016) provide a valuable insights into the current
43
44 practices and pitfalls of the use of performance measures in higher education systems.
45
46
47
48

49
50 The choice of variables for such calculations and the method of analysis is
51
52 known to affect the outcome of rankings (Saisana et al 2011; Draper and Gittoes 2004).
53
54 Asif (2015) points out that the methodologies for selecting processes or factors for
55
56 benchmarking in HE are sometimes vague or subjective. Benchmarking methodology
57
58
59
60

1
2
3 should be objective and repeatable, but there is evidence that this is not always the case
4
5 (Partovi 1994; Garengo et al 2005; De Toni and Tonchia 2001). Thus, because the
6
7 success of metric benchmarking is governed by the identification of a small set of
8
9 appropriate factors that influence the measure of performance, factor selection can
10
11 involve a large number of possible permutations of factors requiring a rigorous process
12
13 to devise the most appropriate solution. To overcome challenges in factor selection for
14
15 benchmarking, it has been suggested that factor selection should incorporate data that
16
17 are: 1) routinely acquired by stakeholders, 2) highly predictive, 3) meaningful to the
18
19 stakeholders and are 4) relevant and necessary to add value to the benchmark (Hall and
20
21 Holmes 2003).
22
23

24
25 In this study, we describe a two phase approach to perform objective and
26
27 repeatable benchmarking in higher education, using data from nursing courses at nine
28
29 universities in the UK. In the first phase of our benchmarking, we used machine
30
31 learning to objectively choose variables of value to predict our measure of success,
32
33 student progression. Machine learning methods (such as classification and decision
34
35 trees) are increasingly used as a data reduction method and for variable selection
36
37 (Rokach and Mainon 2014). This results in a subset of informative and necessary
38
39 variables to be used in benchmarking modelling. The second phase of our study
40
41 involved using variable selected via machine learning to construct a benchmarking
42
43 model. The approach we used is similar conceptually to best-practice using
44
45 benchmarking, where we used direct standardization of variables to weight variables
46
47 within and between like student groupings (Draper and Gittoes 2004). Here, direct
48
49 standardisation is used to benchmark performance of specific student cohorts where the
50
51 benchmark *per se* is the difference between the actual performance of students at that
52
53 institution relative to the performance of the same student cohort at a generalized
54
55
56
57
58
59
60

1
2
3 regional 'average' university (see below for further explanation). We discuss the
4
5 method we implement here and how this approach can be used to inform quality
6
7 improvement in higher education.
8
9

10 11 **Metric benchmark development**

12
13 For the factor selection phase of this study, all analyses were carried out on the open
14
15 source statistical package 'R' (<https://www.r-project.org/>). The dependent variable for
16
17 benchmark calculations was the measure of student qualification, a binomial variable
18
19 indicating either qualification or withdrawal before course completion. In terms of the
20
21 accompanying factors available for analysis, routinely captured data were used from
22
23 'PETD' (Professional Education Training Database) data that had been provided on a
24
25 student-by-student basis for nursing courses by Health Education England (formerly
26
27 Health Education North West) for nine higher education institutions. Our data
28
29 represented cohorts of students that had finished their degree and who began their
30
31 degree during the period of 2008-2011 and would have qualified by 2014 at the earliest.
32
33 The database provided a range of factors associated with each individual student, such
34
35 as student age, gender, registered disability, ethnicity or whether the student had
36
37 suspended their studies. Home postcode was also provided and this was assumed to
38
39 represent the (parental) home of the individual. There were also two associated metrics
40
41 provided as surrogates of socioeconomic status: (i) youth participation rates in 'further'
42
43 education ('youthEd'), and (ii) adult participation rates in further and higher education
44
45 ('adultEd').
46
47
48
49
50

51
52 The approaches to benchmark the dataset drew on methods from Draper and
53
54 Gittoes who referred to the National Committee of Inquiry into Higher Education (or
55
56 the 'Dearing Committee') that led to the Performance Indicators Steering Group. Their
57
58 work focused on student progression data, with an underlying intention to provide more
59
60

1
2
3 reliable information on the performance of UK HEIs, in order to inform policy
4
5 developments and enhance public accountability. Draper and Gittoes (2004) dissected
6
7 the HEFCE approach, which was based on three main components: inputs (status of
8
9 individuals as they enter higher education), process (what happens to them when in
10
11 higher education) and outputs (outcome measures that can be evaluated, such as student
12
13 success). This framework underpins the input-output (IO) approach, is also known as
14
15 'provider profiling' in health areas and has been used in benchmarking for schools and
16
17 hospitals internationally. The IO approach is used here to provide a schematic overview
18
19 of our methods (Figure 1).
20
21

22 [Figure 1 near here]
23
24
25

26 ***Machine learning for factor selection*** 27

28
29 In light of the absence of machine learning methods in the literature about
30
31 benchmarking, we provide some detail to both justify and provide an overview of the
32
33 approach. Statistical analysis of data sets with high dimensionality (i.e. many potential
34
35 factors under consideration) is challenging as a consequence of processing and
36
37 analysing a large number of predictor variables. One approach to managing this
38
39 challenge is to reduce the dimensionality of the dataset, to reduce redundancy amongst
40
41 variables and leverage power to resolve the effects in individual variables (Rokach and
42
43 Maimon 2014). To achieve this, we used machine learning, a computational approach to
44
45 identify variables with high explanatory power within a modelling framework.
46
47
48

49 One implementation of machine learning is using classification and regression
50
51 trees (CART) in creating so-called 'random forests' (Breiman 2001). This approach is
52
53 robust to strict assumptions of data conformity and has become one of the most widely
54
55 used data analysis tools involving large and high-dimensional datasets (Liaw and
56
57 Wiener 2002). Practical applications of this technique can be found in a number of
58
59
60

1
2
3 research fields, including: psychology (Strobl et al 2009), ecology (Prasad et al 2006)
4
5 and HE student progression (Hardman et al 2013). RandomForest analysis in particular
6
7 is in wide use as a tool to make predictions based on variable associations, and also to
8
9 identify variables with predictive value by ranking the predictive importance of these
10
11 variables (Grömping 2009; Genuer et al 2010). A specific strength of the algorithm is
12
13 that it can be used to combine factorial and numerical data within a modelling
14
15 framework and is robust with data sets of high dimensionality, whilst also being free of
16
17 parametric assumptions of data conformity (Breiman 2001; Strobl et al 2009).
18
19

20
21 RandomForest was used to identify variables with the highest importance in
22
23 predicting whether (or not) students in nursing programmes become qualified in their
24
25 field of study. We calculated the ‘variable importance’ values of all predictor variables
26
27 provided in the dataset, generating 10,000 trees. Our first goal was to identify a subset
28
29 of the most important predictor variables to use in the construction of our benchmarking
30
31 tool, following Liaw and Wiener’s (2002) approach. In this way we ranked predictor
32
33 variables to inform the benchmarking tool, whilst objectively retaining the highest
34
35 amount of information possible. To avoid problems due to the uneven distribution of
36
37 subsets of data (e.g. across universities and programmes) and possible correlations
38
39 amongst variables, we used an area under the curve (AUC) correction approach with
40
41 variable and model validation (Strobl et al 2009).
42
43

44
45 Our second goal was to describe how the variables of high importance
46
47 (identified in the initial analysis) varied amongst universities and programmes in their
48
49 association with whether or not students ultimately became qualified. We calculated
50
51 variable importance using equally weighted measures of mean square error and the Gini
52
53 index (see Breiman 2001) for each of these data subsets, and partitioned data subgroups
54
55
56
57
58
59
60

1
2
3 using conditional inference trees. This approach is coherent with the methods in Strobl
4
5 et al (2009).
6

7 Interpretation of the modelling outcomes may be complex. For example,
8
9 distance of university from home postcode was identified as having high importance in
10
11 predicting whether a student became qualified. However, larger distances may be an
12
13 advantage to one university that is highly selecting and draws the best candidates
14
15 nationwide, whereas larger distances may be a disadvantage to another rural university
16
17 where this implies difficult commutes from home to placement/university. Thus, future
18
19 functional interpretations of the outcomes (such as the distance variable) would need to
20
21 be made by the users of the benchmarks who can account for characteristics of
22
23 individual universities and their constituent student populations.
24
25
26
27

28 ***Factor selection***

29
30
31 There were 52 routinely used factors available from the dataset made available for nine
32
33 (anonymised) universities. These included variables such as student 'age', 'gender',
34
35 'ethnicity', registered 'disability' and 'home address'. The distance of the home
36
37 postcode to the institution of study was also calculated ('distance'). We note that the
38
39 dataset available for analysis was limited due to data availability and data protection,
40
41 which is a practical constraint in any benchmarking project. For example, there is a
42
43 great deal of literature surrounding the role of student entry qualifications (e.g. Yorke
44
45 and Longden 2004). This type of data would be available at institutional level, but it
46
47 was not available from the PETD dataset and thus not available for our analyses. We
48
49 note that there is underlying debate concerning tariff as being a factor that is 'in the
50
51 control of the institution' which may erode its value as an input variable for
52
53 benchmarking (see Draper and Gittoes 2004).
54
55
56
57
58
59
60

1
2
3 Machine learning analysis using classification and regression trees was carried
4
5 out using all data for which we had complete cohorts (2008-2011) in order to identify
6
7 factors of high importance in predicting whether students become qualified in their field
8
9 of study after graduating. We first included the factors for the institution of study and
10
11 whether the student had suspended their studies (returning to study at a later date). Both
12
13 these variables were found to have greater influences than any other factor. The strong
14
15 'university effect' was anticipated as there were known differences in the qualification
16
17 rates between the institutions with associated known differences in entry qualifications.
18
19 Students that suspended during their studies were also more likely to withdraw. Indeed,
20
21 suspension was a stronger predictor of whether a student qualified than institution,
22
23 suggesting the need for future research to explain factors that underpin suspensions. For
24
25 further analyses, 'suspend' was excluded on the basis that it was an outcome of the
26
27 educational process (and not an input variable, such as gender) and because of its close
28
29 association with our dependent variable. Institutions were subsequently analysed
30
31 separately in order to explore consistency of the factor effects across institutions.
32
33
34
35

36 Individual factors to inform benchmarking ideally should represent unrelated
37
38 characteristics. Both youth and adult participation rates in education were included in
39
40 the preliminary analyses and both identified as having high importance. However, these
41
42 factors are strongly correlated with each other, and only youth participation (having a
43
44 higher importance value than adult participation) was included in the benchmark
45
46 modelling. As a result of our factor selection analysis, we selected the five factors of
47
48 highest average importance for inclusion in our benchmark model. These were 'age',
49
50 'gender', 'ethnicity', 'registered disability' and 'youth participation rate into higher
51
52 education'.
53
54
55
56
57
58
59
60

Calculation of metric benchmarks using weighted averages

Demographic segments

Here, we refer to a ‘segment’ as a unique grouping of demographic factors. Segments representing combined groupings of data generated by the analysis are a product of partitioning data between the many combinations of factors. One challenge with benchmarking modelling is to avoid the reduction into too many segments of data. This problem practically limits including a large number of variables in benchmarking. The reason for this is that when data are segmented some may have very few records, for example records which are for Male AND Young AND High Youth HE Participation. We avoided this by overwriting the segment qualification rate with the rate at the higher level in the hierarchy according a lower threshold we set (a standard solution to this problem). For example, instead of using the rate for the segment Male AND Young AND High Participation, the rate for Male AND Young only was used. In the test model, this the low threshold was set at a minimum of 30 students in the starting cohort as this threshold value is commonly used in statistical tests as the minimum number of cases required for the test to be reliable.

Another important consideration was missing data, as not all students provide all the information, particularly concerning protected characteristics such as ethnicity and disability. In the test model, missing data were treated as a characteristic in their own right. For example, gender had three characteristics: male, female and not known (i.e. not reported). This necessary addition resulted in larger number of small segment counts as there were a greater number of attributes.

Factor order

The order that factors are used in benchmark calculation affects benchmark outcomes,

1
2
3 both as a consequence of the segment threshold and because of missing data. In order to
4
5 measure the impact that factor order had, we used a sensitivity analysis approach. With
6
7 regards to data segmentation, if the segment count was below our threshold when
8
9 including all variables in a benchmark, then one factor would be dropped from the
10
11 analysis to alleviate this (where variable order impacted which variable was dropped).
12
13 To avoid factor order bias, all possible combinations of factor order were used to
14
15 compute benchmarks and the variation around resulting benchmark estimates was
16
17 evaluated in relation to actual performance (Figure 2).
18
19

20 [Figure 2 near here]
21
22

23
24 The average difference between the maximum and minimum calculated benchmark
25
26 score for each institution (based on the different orderings of the demographic factors)
27
28 was 6.52%. For most institutions (except institutions B and G), the actual performance
29
30 for the programme was either above or below benchmark boxplot whiskers. Thus,
31
32 regardless of the ordering of the five factors, the actual performance of most institutions
33
34 can be classified unambiguously as being above (or below) benchmark. For institutions
35
36 B and G, actual performance lay within the range of likely benchmark results and so is
37
38 not unambiguously different. This highlights the importance of exploration of the
39
40 effects of factor order in the development of benchmarks, and the need for informed
41
42 decision-making on the outcomes of this. A striking feature of our benchmark results is
43
44 that variation in benchmark estimates is very different for different institutions, where,
45
46 for example, institution B shows large variation in benchmarks due to factor order while
47
48 institutions D and H show very little variation in benchmarks. This effect may reflect
49
50 variation in segmentation between these intuitions, variation in missing data or a
51
52 genuine differential effect of individual factors on benchmark performance. Finally, we
53
54 point out that merely ranking actual performance in relation to average actual across
55
56
57
58
59
60

1
2
3 institutions may can be quite different in comparison to ranking performance based on
4
5 benchmarking. For example, relative to the average benchmark, institutions E, F and G
6
7 are on or above average in actual performance. However all of these institutions show
8
9 performance below that predicted by the benchmark. Likewise, while institution B
10
11 exhibits actual performance below the institutional average, it is performing to
12
13 expectation based on the benchmark estimates.
14
15

16 17 18 *Number of factors*

19
20 We also performed sensitivity analysis to measure the effect of factor number on the
21
22 outcome of benchmarking. The number of factors included affects benchmark variation
23
24 and magnitude (Figure 3). The pattern of median values for benchmark estimates is
25
26 similar to that seen when varying order alone for all five variables. The spread of
27
28 benchmark estimates in our sensitivity analysis is small compared to that for factor
29
30 alone (comparing Fig 3 to Fig 2). This suggests that, for these data, the model appears to
31
32 be (at least) relatively robust regardless of the number of factors included. The
33
34 interpretation that a particular institution is over- or under-performance appears robust
35
36 in that most institutions (except institution B) are clearly above or below the
37
38 benchmark.
39
40

41
42 [Figure 3 near here]
43
44

45 46 47 *Which years to include*

48
49 Another factor that can affect the model is the number of years of data that are used for
50
51 the calculation. In our models, three years of data were used to inform the benchmark
52
53 (2009, 2010 and 2011), where year is the starting year for whole cohorts of students
54
55 followed to degree completion. If fewer years of data were used, for example to
56
57 specifically calculate benchmarks for each year, then segments would have had smaller
58
59
60

1
2
3 counts that would affect the importance of factor order. When fewer years of data were
4
5 used, the benchmark range was, in some cases, much broader due to lower sample sizes
6
7 impacting segmentation issues. While the increased variability resulted in broader
8
9 benchmark estimates, using only the most recent year meant that the data were specific
10
11 to individual years which could be useful to inform, for example, year-specific effects
12
13 such as a change in policy or other measures expected to impact performance.
14
15

16 Overall, when taking into account both factor order and number of factors with
17
18 sensitivity analysis we describe here, the indirect benchmarking method appears to be
19
20 robust in that the interpretation of whether an institution is performing well or poorly
21
22 relative to the benchmark is generally little affected by specific choices for factor
23
24 ordering. The error around benchmark estimation that is produced using sensitivity
25
26 analysis in this way is a robust way to aid interpretation of benchmarking results that
27
28 can be used to overcome challenges involving factor choice, factor order and
29
30 segmentation.
31
32
33
34

35 **Discussion**

36
37
38 This study outlines the approaches used to calculate metric benchmarks for nursing
39
40 courses in the Northwest region of the UK, but the findings and approaches described
41
42 are relevant to benchmarking calculations worldwide and at many scales of enquiry. We
43
44 describe a conceptual input-output model (aligned to Draper and Gittoes 2004) that
45
46 identifies clear differences in institutional performances compared to benchmarks and
47
48 can be used to account for variation in learner compositions and assumptions associated
49
50 with the benchmarking process. We believe this approach combines a robust factor
51
52 selection approach with sensitivity analysis to mitigate many of the shortcomings
53
54 associated with HE rankings. The outcomes provide a valid and reliable method to help
55
56 both policy makers and universities for the identification of universities that are
57
58
59
60

1
2
3 performing above or below benchmark predictions, thus helping to target improvement
4 activities. It is noteworthy that this comparative information is potentially valuable but,
5 in isolation, metric benchmarks do not provide an understanding of the explanatory
6 factors that drive performance. Subsequent use of the outputs for performance
7 enhancement requires a layer of interpretation, preferably in consultation with other
8 stakeholders (Jackson and Lund 2000).
9

10
11 Differences in institutional benchmark outcomes were anticipated after the
12 preliminary analysis highlighted 'institution' as the most important explanatory variable
13 in the model (also see the exploration of heterogeneity of institutions in Agasisti and
14 Bonomi 2014). This factor was subsequently removed to allow institutions to be
15 compared and preventing the factor from masking comparatively subtle effects of other
16 factors in predicting student qualification. Similarly, records of student suspensions that
17 were available in the dataset were used in preliminary calculations were also strong
18 predictors of future failure to graduate (as may be anticipated; see Nonis and
19 Wright 2003) and were also not considered in the final benchmark calculations.
20
21

22
23 The factors selected for the benchmarking process were coherent with other HE
24 benchmarking documentation For example, the UK's National Student Survey (NSS)
25 benchmark has the intention of accounting for the mix of students and subjects at
26 different institutions when reporting on survey outcomes (see Leman 2011). The
27 respondent-related factors included in the NSS benchmarking process are similar to
28 those identified in this study and purported to have demonstrable, consistent effects on
29 survey outcomes and are also considered to be outside of institutional control, namely:
30 subject (not included in our study as limited to nursing courses), ethnicity (included);
31 age (included), mode of study (not included in our study as only full time students are
32 enrolled on these courses), sex (included as 'gender') and disability (included).
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 However, HESA (2011) also have a benchmarking process that incorporates a similar
4 suite of factors relating to entry qualifications of students, age and subjects of study.
5
6
7 The HESA benchmark accounts for the effects of subject profiles and the entry
8 qualifications of students for comparisons of the rates of student qualification. The
9
10 HESA report (*ibid.*) encapsulates the need for this process, suggesting that benchmark
11 outcomes provide information about the ‘sort of values that might be expected for a HE
12 provider’s indicator’. This assumes that other factors not used in the benchmark
13 calculation do not have significant effects on the output variable. If this assumption is
14 true, and the factors entered are all ‘input variables’, it suggests that differences in
15 performance relative to the benchmarks are either a direct result of HE provider
16 performance or are influenced by other factors not included in the benchmark.
17
18
19
20
21
22
23
24
25
26

27 The sensitivity analysis used provided a range of potential benchmark values
28 based on potential permutations of the numbers of factors used and the order they are
29 entered into the calculation. Our findings highlight the importance of this stage of
30 benchmark calculation, most markedly the order that factors are used in the benchmark
31 calculation, particularly for institutions that are close to their benchmark values. Our
32 findings concur with other authors that stress the importance of explicit sensitivity
33 analyses for calculating metric benchmarks, although this key component is not always
34 evident in benchmarking systems (see Reichmann & Sommersguter-Reichmann 2006).
35
36 Our approach contrasts others (e.g. HESA benchmarking of NSS results detailed
37 previously) which provides a single point of comparison. It is suggested that the visual
38 presentation of the actual result against the quartiles of benchmark options provides
39 ‘non-statistical’ users with a sense of confidence around the interpretation, as it is not an
40 over-simplification. This was an important basis for performance discussions with
41 universities in the current study.
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 The methodological approaches have been described here in detail, to illustrate
4 the underpinning rationale and highlight approaches such as sensitivity analysis and the
5 debate surrounding the number of factors for inclusion. We suggest that these
6 approaches can be applied more widely, for different subject areas, greater geographical
7 areas (e.g. national datasets) and beyond the educational arena. The novel use of this
8 machine learning technique is proposed as going some way to solve the difficult issue
9 of selecting the most appropriate factors for inclusion to allow for objective
10 benchmarking (e.g. Hall and Holmes 2003). The robust nature of the RandomForest
11 procedure allows objective, data-informed inclusion of many data types, is not limited
12 by data distributions or types, and is seen as a starting point to factor selection that
13 could be adapted year on year as data to include a wider dataset whenever new data
14 acquisition reaches a scale suitable for benchmarking purposes. Of course, the key
15 themes of benchmarking processes go far beyond the calculation and comparison of
16 metric benchmarks. Their importance lies in the subsequent identification and
17 implementation of new processes, for example based around the use of best practices
18 (Camp 1989).

19
20
21 In our case study of nursing programmes at nine UK universities, the choice of
22 factors was determined by the available dataset, in this case data that are routinely
23 collected and available to institutions. We acknowledge that this constraint precludes
24 many other potentially important factors such as aspects of learning and teaching
25 structures. This includes the clinical placement element of nursing courses that appears
26 to be of significant importance in this subject area (Hamshire et al 2011 and 2012).
27 However, we believe this example is a realistic representation of practical limitations in
28 benchmarking and we suggest our method can be used to improve current
29 benchmarking practice in HE, despite unavoidable limitations in data quality or
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 availability. Because formal benchmarking is recognized as an ongoing process
4
5 (Stapenhurst 2009), we suggest that our approach provide an opportunity for regular
6
7 recalculations to incorporate data as they become available and to evaluate their impact
8
9 on the outcomes through accompanying sensitivity analyses. Potentially, the machine
10
11 learning techniques described here could also be uses to identify the factors that predict
12
13 successful completion of other within-degree output variables such as placement
14
15 completions (e.g. as opposed or in addition to whether a student progresses or becomes
16
17 qualified in their field of study).
18
19

20
21 In conclusion, this study has outlined a new approach for factor selection and
22
23 sensitivity analysis of benchmarking programme performance in HE. This approach is
24
25 robust and objective, incorporating machine learning for factor selection. The software
26
27 used is open source and available to all stakeholders and the code used easily shared.
28
29 The benchmarks created are coherent with others in the educational sector due to the
30
31 factors included and our outcomes have highlighted the importance of sensitivity
32
33 analyses to clarify the effects of factor number and order on the outcomes. Our
34
35 outcomes could be built on and improved, for example by future acquisition of data
36
37 surrounding the clinical placement component of these courses. Thus, we have provided
38
39 baseline data and documented processes that could be developed further, and identified
40
41 issues with current data that are routinely available and valuable for benchmarking
42
43 processes. Ultimately, we have created and described a benchmarking approach that
44
45 could be further generalised to explore benchmark performance in other regional or
46
47 national datasets in order to improve quality in key HE outputs such as successful
48
49 completion of clinical placements. In a wider context, the approaches described have
50
51 relevance to the creation of benchmarks on a greater scale, such as those being devised
52
53 at the time of writing for the proposed 'Teaching Excellence Framework', a component
54
55
56
57
58
59
60

1
2
3 of the UK government's White Paper intended to reform higher education, that is
4
5 founded on a selection of benchmarked metrics (BIS 2016).
6
7
8
9

10 11 **References**

- 12
13 Agasisti, T. and F. Bonomi. 2014. "Benchmarking universities' efficiency indicators in
14 the presence of internal heterogeneity." *Studies in Higher Education* 39: 1237-
15 1255.
16
17
18 Ammons, D.N., C. Coe and M. Lombardo. 2001. "Performance comparison projects in
19 local government: participant's perspectives." *Public Administration Review*
20 61: 100-110.
21
22
23 Ammons, D.N. and W.C. Rivenbark. 2008. "Factors influencing the use of performance
24 data to improve municipal services: evidence from the North Carolina
25 benchmarking project." *Public Administration Review* 68: 304-318.
26
27
28 Anand, G. and R. Kodali. 2008. "Benchmarking the benchmarking models."
29 *Benchmarking: An International Journal* 15: 257-291.
30
31
32 Asif, M. 2015. "Determining improvement needs in higher education benchmarking."
33 *Benchmarking: An International Journal* 22: 56-74.
34
35
36 Asif, M. and A. Raouf. 2013. "Setting the course for quality assurance in higher
37 education." *Quality & Quantity* 47: 2009-2024.
38
39
40 Besterfield, D.H., Besterfiels-Michna, C., Besterfield, G.H. and Besterfield-Sacre, M.
41 1999. *Total Quality Management*. New Jersey: Prentice-Hall.
42
43
44 Bhutta, K.S. and F. Huq. 1999. "Benchmarking-best practices: an integrated approach."
45 *Benchmarking: An International Journal* 6: 254-268.
46
47
48 BIS. 2014. *Methodological issues in estimating the value added for further education,*
49 *higher education and skills: a review of relevant literature*. London: Department
50 for Business, Innovation and Skills, April.
51
52
53 BIS. 2016. *Success as a Knowledge Economy: Teaching Excellence, Social Mobility*
54 *and Student Choice*. (Department for Business, Innovation and Skills).
55 [https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/5](https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/523546/bis-16-265-success-as-a-knowledge-economy-web.pdf)
56 [23546/bis-16-265-success-as-a-knowledge-economy-web.pdf](https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/523546/bis-16-265-success-as-a-knowledge-economy-web.pdf)
57
58
59 Bowerman, M., G. Francis, A. Ball, and J. Fry. 2002. "The evolution of benchmarking
60 in UK local authorities." *Benchmarking: an International Journal* 9: 429-449.

- 1
2
3 Braadbaart, O. and B. Yusnandarshah. 2008. "Public sector benchmarking: a survey of
4 scientific articles, 1990-2005." *International Review of Administrative Sciences*
5 74: 421-433.
6
7
8 Breiman, L. 2001. Random forests. *Machine learning* 45: 5-32.
9
10 Camp, R.C. 1989. *Benchmarking: The Search for Industry Best Practices that Lead to*
11 *Superior Performance*. Madison: Quality Press.
12
13 Carpinetti, L.C.R. and A.M. De Melo, 2002. "What to benchmark: a systematic
14 approach and cases." *Benchmarking: An International Journal* 9: 244-255.
15
16 CHEMS. 1998. *Benchmarking in Higher Education: An International Review*. London:
17 Commonwealth Higher Education Management Service.
18
19 Dattakumar, R. and R. Jagadeesh. 2003. "A review of literature on benchmarking."
20 *Benchmarking: An International Journal* 10: 176-209.
21
22
23 De Toni, A. and S. Tonchia. 2001. "Performance measurement systems-models,
24 characteristics and measures." *International Journal of Operations and*
25 *Production Management* 21: 46-71.
26
27
28 Draper, D. and M. Gittoes. 2004. "Statistical analysis of performance indicators in UK
29 higher education." *Journal of the Royal Statistical Society Series A (Statistics in*
30 *Society)* 167: 449-474.
31
32
33 Drew, S.A.W. 1997. "From knowledge to action: the impact of benchmarking on
34 organizational performance." *Long Range Planning* Vol. 30, pp. 427-441.
35
36 ECPE (2011) *Education Criteria for Performance Excellence*. Gaithersburg: National
37 Institute of Standards and Technology (NIST).
38
39
40 Folz, D.H. 2004. "Service Quality and benchmarking the performance of municipal
41 services." *Public Administration Review* 64: 209-220.
42
43
44 Garengo, P., S. Biazzo, and U.S. Bititci. 2005. "Performance measurement systems in
45 SMEs: a review for a research agenda." *International Journal of Management*
46 *Reviews* 7: 25-47.
47
48
49 Genuer, R., J.M. Poggi, and C. Tuleau-Malot. 2010. "Variable selection using random
50 forests." *Pattern Recognition Letters* 31: 2225-2236.
51
52
53 Goldstein, H. 2001. "Using pupil performance data for judging schools and teachers."
54 *British Educational Research Journal* 27: 433-442.
55
56
57 Goncharuk, A.G. and J.P. Monat. 2009. "A synergistic performance management model
58 conjoining benchmarking and motivation." *Benchmarking: An International*
59 *Journal* 16: 767-784.
60

- 1
2
3 Grömping, U. 2009. "Variable importance assessment in regression: linear regression
4 versus random forest." *The American Statistician* 63: 308–319.
- 5
6 Hall, M.A. and G. Holmes. 2003. "Benchmarking attribute selection techniques for
7 discrete class data mining." *IEEE Transactions on Knowledge and Data
8 Engineering*, 15: 1437-1447.
- 9
10
11 Hamshire, C., T. Willgoss, and C. Wibberley. 2012. "Should I stay or should I go? A
12 study exploring why healthcare students consider leaving their programme."
13 *Nurse Education Today* 33: 889-895.
- 14
15
16 Hamshire, C., T. Willgoss, and C. Wibberley. 2011. "The placement was probably the
17 tipping point' – The narratives of recently discontinued students." *Nurse
18 Education in Practice* 12: 182-186.
- 19
20
21 Hardman, J., A. Paucar-Caceres, and A. Fielding. 2013. "Predicting Students'
22 Progression in Higher Education by Using the Random Forest Algorithm."
23 *Systems Research and Behavioral Science* 30:194-203.
- 24
25
26 Hazelkorn, E. 2015. *Rankings and the Reshaping of Higher Education: The Battle for
27 World Class Excellence*. Basingstoke: Palgrave Macmillan.
- 28
29
30 HESA. 2011. *International Benchmarking in UK Higher Education*. Higher Education
31 Statistics Agency report. [https://benchmarking.hesa.ac.uk/wp-](https://benchmarking.hesa.ac.uk/wp-content/uploads/2011/10/HESA_International_Benchmarking_report.pdf)
32 [content/uploads/2011/10/HESA_International_Benchmarking_report.pdf](https://benchmarking.hesa.ac.uk/wp-content/uploads/2011/10/HESA_International_Benchmarking_report.pdf).
- 33
34
35 Jackson, N. and H. Lund. 2000. *Benchmarking for Higher Education*. Open University
36 Press, Milton Keynes, UK.
- 37
38
39 Kouzmin, A., E. Löffler, H. Klages, and N. Korac-Kakabadse. 1999. "Benchmarking
40 and performance measurement in public sectors. Towards learning for agency
41 effectiveness." *The International Journal of Public Sector Management* 12: 121-
42 144.
- 43
44
45 Kumar, S. and C. Chandra. 2001. "Enhancing the effectiveness of benchmarking in
46 manufacturing organizations." *Industrial Management and Data Systems* 101:
47 80-89.
- 48
49
50 Kyrö, P. 2003. "Revising the concept and forms of benchmarking." *Benchmarking: An
51 International Journal* 10: 210-225.
- 52
53
54 Lema, N. and A. Price. 1995. "Benchmarking: performance improvement toward
55 competitive advantage." *Journal of Management in Engineering* 11: 28-37.
- 56
57
58
59
60

- 1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
- Leman, J. 2011. *Benchmarking and the NSS*. Presentation at the Higher Education Surveys for Enhancement Conference 2011.
<https://www.heacademy.ac.uk/resource/benchmarking-and-nss>.
- Liaw, A. and M. Wiener. 2002. "Classification and regression by RandomForest." *R news* 2: 18-22.
- Nedwek, B.P. and J.E. Neal. 1994. "Performance indicators and rational management tools: a comparative assessment of projects in North America and Europe." *Research in Higher Education* 35: 75-103.
- Nonis, S.A. and D. Wright. 2003. "Moderating effects of achievement striving and situational optimism on the relationship between ability and performance outcomes of college students." *Research in Higher Education* 44: 327-346.
- Noordhoek, M. 2013. *Municipal benchmarking: organisational learning and network performance in the public sector*. PhD Thesis, Aston University. Available at: <http://eprints.aston.ac.uk/19540/1/Studentthesis-2013.pdf>
- Northcott, D. and S. Llewellyn. 2005. "Benchmarking in UK health: a gap between policy and practice?" *Benchmarking: An International Journal* 12: 419-435.
- Partovi, F.Y. 1994. "Determining what to benchmark: an analytic hierarchy process approach." *International Journal of Operations and Production Management* 14: 25-39.
- Prasad, A.M., L.R. Iverson, and A. Liaw. 2006. "Newer classification and regression tree techniques: bagging and random forests for ecological prediction." *Ecosystems* 9: 181-199.
- Pryor, L.S. 1989. "Benchmarking: a self-improvement strategy." *Journal of Business Strategy* 10: 28-32.
- Reichmann, G. and M. Sommersguter-Reichmann. 2006. "University library benchmarking: An international comparison using DEA." *International Journal of Production Economics* 100: 131-147.
- Rokach, L., and O. Maimon, 2014. *Data mining with decision trees: theory and applications*. World Scientific. Series in Machine Perception and Artificial Intelligence 69. <http://www.worldscientific.com/worldscibooks/10.1142/6604>.
- Saisana, M., B. d'Hombres, and A. Saltelli. 2011. "Rickety numbers: Volatility of university rankings and policy implications." *Research Policy* 40: pp.165-177.
- Shober A. F. 2013. "Debate: Benchmarking inequality – driving education progress in the USA." *Public Money and Management* 33: 242-244.

- 1
2
3 Spendolini, M.J. 1992. *The Benchmarking Book*. New York: American Management
4 Association.
5
6 Stapenhurst, T. 2009. *The Benchmarking Book*. Oxon: Routledge.
7
8 Strobl, C., J. Malley, and G. Tutz, 2009. "An introduction to recursive partitioning:
9 rationale, application, and characteristics of classification and regression trees,
10 bagging, and random forests." *Psychological Methods* 14: 323-348.
11
12 Tam, M. 2001. "Measuring quality and performance in higher education." *Quality in*
13 *Higher Education* 7: 47-54.
14
15 Tillema, S. 2007. "Public sector organisations' use of benchmarking information for
16 performance improvement: theoretical analysis and explorative case studies in
17 Dutch water boards." *Public Performance and Management Review* 30: 496-
18 520.
19
20
21
22
23 Watson, G.H. 1994. "A Perspective on Benchmarking." *Benchmarking for Quality*
24 *Management and Technology* 1: 5-10.
25
26 Williams, R. and G. de Rassenfosse. 2016. "Pitfalls in aggregating performance
27 measures in higher education." *Studies in Higher Education* 41: 51-62.
28
29 Wynn-Williams, K.L.H. 2005. "Performance assessment and benchmarking in the
30 public sector: an example from New Zealand." *Benchmarking: An International*
31 *Journal* 12: 482-492.
32
33
34 Yorke, M. and B. Longden. 2004. *Retention and Student Success in Higher Education*.
35 Maidenhead: Society for Research in Higher Education/Open University Press.
36
37 Zairi, M. 1998. *Effective Management of Benchmarking Projects*. Oxford: Butterworth
38 Heinemann.
39
40
41
42
43
44
45

46 **Figure Captions**

47
48 **Figure 1.** Conceptual model of the input-output processes used to create benchmarks
49 from the available dataset (inputs). The simplified structure does not include the
50 iterations of the computations used for the metrics (adjusted outputs) such as sensitivity
51 analyses.
52
53
54

55
56 **Figure 2:** Performance of adult nursing courses compared with calculated benchmarks
57 when factor order varies. Benchmark outcomes are shown for all possible orderings
58
59
60

1
2
3 using all five factors. Benchmarks are shown as a boxplot, actual performance is shown
4 as a large dot and the mean actual performance across institutions is shown with a
5 dashed line.
6
7

8
9 **Figure 3:** Performance of adult nursing courses compared with calculated benchmark
10 when factor number varies. Benchmark outcomes are shown for all orderings for the
11 inclusion of 5, 4, 3, 2 or 1 factor(s). Benchmarks are shown as a boxplot, actual
12 performance is shown as a large dot and the mean actual performance across institutions
13 is shown with a dashed line.
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

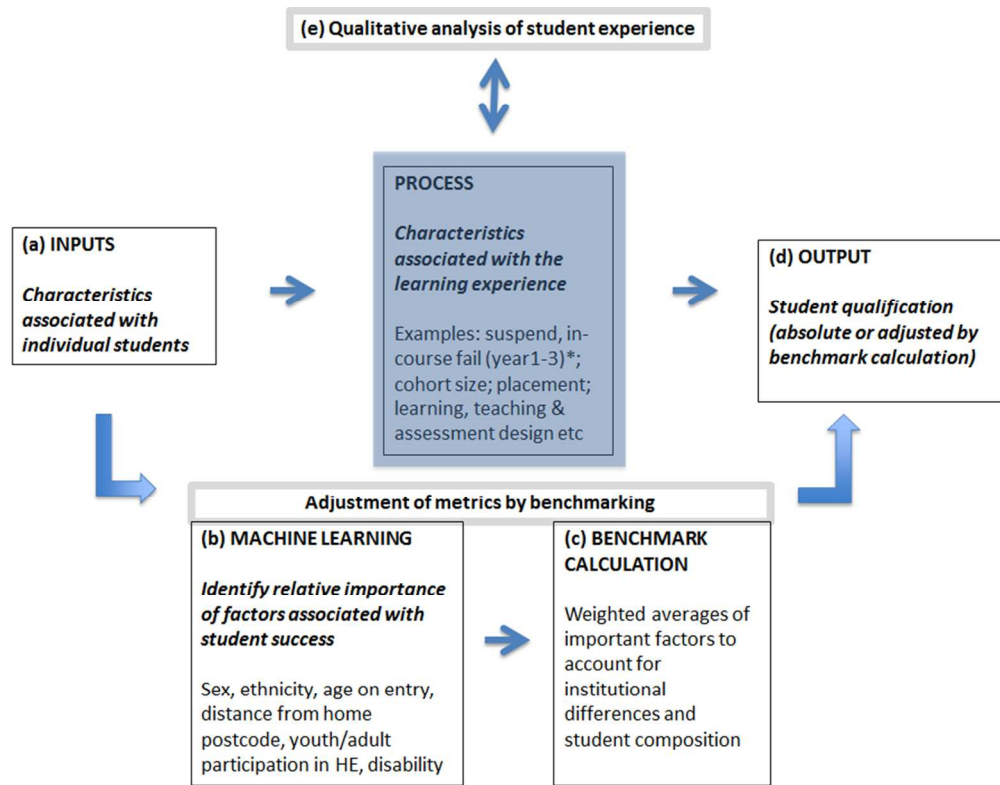


Figure 1. Conceptual model of the input-output processes used to create benchmarks from the available dataset (inputs). The simplified structure does not include the iterations of the computations used for the metrics (adjusted outputs) such as sensitivity analyses.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

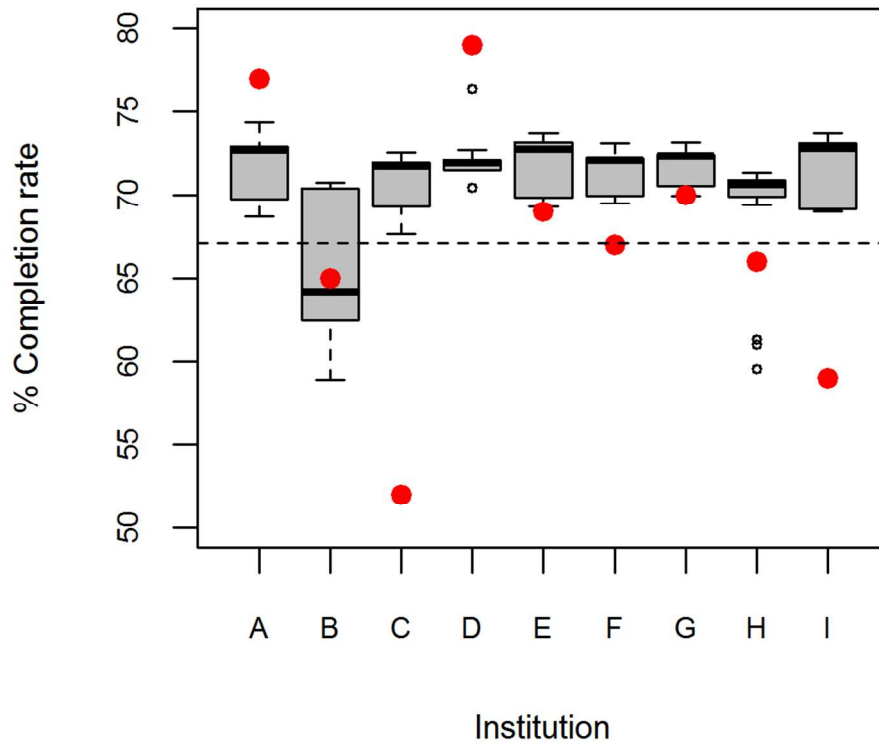


Figure 2: Performance of adult nursing courses compared with calculated benchmarks when factor order varies. Benchmark outcomes are shown for all possible orderings using all five factors. Benchmarks are shown as a boxplot, actual performance is shown as a large dot and the mean actual performance across institutions is shown with a dashed line.

101x101mm (300 x 300 DPI)



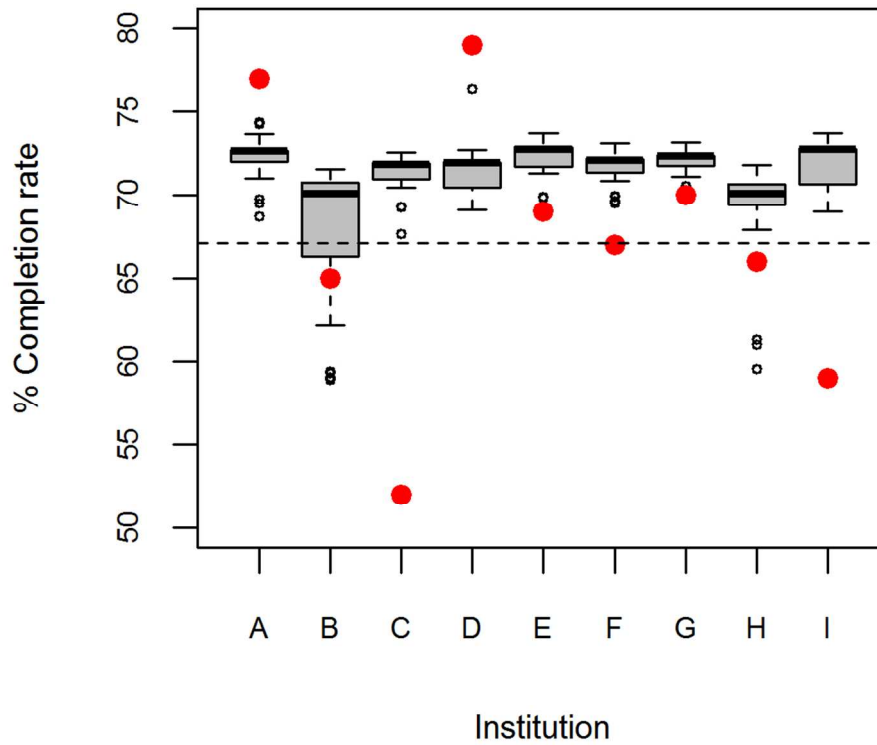


Figure 3: Performance of adult nursing courses compared with calculated benchmark when factor number varies. Benchmark outcomes are shown for all orderings for the inclusion of 5, 4, 3, 2 or 1 factor(s). Benchmarks are shown as a boxplot, actual performance is shown as a large dot and the mean actual performance across institutions is shown with a dashed line.

101x101mm (300 x 300 DPI)

