Research Evaluation, 25(1), 2016, 94–106
doi: 10.1093/reseval/rvv023
Advance Access Publication Date: 27 October 2015
Article

OXFORD

Past performance does not guarantee future results: lessons from the evaluation of research units in Portugal

Ana Ramos[1,*] and Cláudia S. Sarrico[2,3]

[1]FCT Fundação para a Ciência e a Tecnologia, Av. D. Carlos I 126, 1249-074 Lisboa, Portugal, [2]ISEG Lisboa School of Economics and Management, Universidade de Lisboa, Rua do Quelhas 6, 1200-781 Lisboa, Portugal and [3]CIPES Centre for Research in Higher Education Policies, Rua 1° Dezembro 399, 4450-227 Matosinhos, Portugal

*Corresponding author. Email: ana.ramos@fct.pt

Abstract

Research units in Portugal undergo a formal evaluation process based on peer review which is the basis for distributing funding from the national research council. This article analyzes the evaluation results and asks how good they are at predicting future research performance. Better research evaluations mean the institution receives more funding, so the key question is to what extent research evaluations are able to predict future performance as measured by bibliometric indicators. We use data from the peer evaluation of units in 2007–08, and analyze how well it is able to predict the results of a bibliometric study of the units' Web of Science publications in the period 2007–10. We found that, in general, units that had better peer ratings, and thus more funding, as well as an increased capacity to attract extra funding, were not necessarily those that ended up producing more excellent research. The results provide an empirical contribution to the discussion regarding whether science can be measured and how, and reinforce the importance of evaluations where the use of quantitative data is defined and the differences between areas are accounted for. This analysis provides a snapshot of Portugal's recent scientific performance. Chemistry and physics are among the subfields with higher output and impact, which agrees with a traditional preferential funding of these areas. Institutions also excel in areas that may be assuming an increased relevance (Plant Sciences, Food Science and Technology, Neurosciences and other health-related subfields), which should be taken into account when implementing future science policies.

Key words: research performance; evaluation; funding policies; bibliometrics.

1. Introduction

1.1 The Portuguese R&D system

The last two decades have seen a remarkable growth in Portuguese science indicators, namely in expenditure on R&D, number of researchers and publications. In 1999, the overall expenditure on R&D accounted for 0.68% of gross domestic product (GDP), still far from the European Union (EU) average of 1.74%; in 2006 it reached 1% and by 2009 it had reached 1.58%, with half representing business expenditure (DGEEC 2014).

Portugal's accession to the EU in 1986 opened the way for participation in the Framework programs, which have subsequently had a profound effect on the internationalization of Portuguese science. During this period, public policies that fostered the formation of human resources were implemented: in the early 90 s, the European Structural Funds supported several new programs, such as the *Ciência Program* that included funding for building physical infrastructure as well as for individual research fellowships. This led to a notable increase in the number of researchers.

The expansion of the system gave rise to a set of new research entities envisaged in the Decree-Law 125/99 (*centros de investigação*) and funded directly by the national research council, *Fundação para a Ciência e a Tecnologia* (hereafter FCT). These university-based institutions have a multiannual funding program based on periodic international evaluations and benefit from more flexible rules to apply for

projects and manage funds when compared to universities. In 1999, a special statute of *Laboratório Associado* (Associated Laboratory) was granted to some of those institutions that demonstrated scientific excellence; the first of these units were established in 2000 and, by 2011, there were 26 Associated Laboratories (ALabs). Since one of the main objectives was to increase the number of science-based positions, the average number of researchers in ALabs was much higher than in the other R&D units (RUnits). Moreover, in ALabs, there was a clear goal to create critical mass in each and every scientific area by bringing together large research consortia built around thematic networks across a number of institutions (Heitor and Horta 2012). As a result of this explicit policy, many ALabs resulted from the cooperation between two or more research institutions that could be hosted by the same or by different universities.

By 2010, Portugal had reached 8.2 researchers per thousand work force, above the EU and OECD averages. Between 2003 and 2008, the percentage growth rate of researchers per thousand work force was the highest in Europe, well above the European average of 14%. The number of doctorate degrees awarded or recognized by Portuguese universities increased substantially (74%) between 2000 and 2010 and the annual scientific production (indexed in the Citation Indexes of Web of Science—WoS) rose from 3,792 publications to 10,081 in the same period (Heitor et al. 2014).

1.2 Evaluation and funding of research units

According to Heitor and Horta (2012), the Portuguese scientific system was effectively established in the mid-90s when independent and international evaluations were implemented. This included the evaluation of the RUnits by panels organized by scientific field and based on the institutions' reports and activity plans as well as on direct contact of the evaluators with researchers during on-site visits. After each evaluation, the institutions are awarded a qualitative rating on a five-point scale from excellent, very good, good, fair, to poor, which determines the level of funding. Until 2014, base funding was indexed to the number of research staff with a doctoral degree and to the rating obtained in the evaluations. In addition, some units received strategic funding to meet specific needs.

RUnits were evaluated in 2007, and ALabs were evaluated in 2008; both were evaluated again in 2014. The guidelines for the 2007 evaluation of RUnits stated that 'the rating must reflect the unit's performance in the past and the future research proposal' (FCT 2007). The report submitted to this evaluation had to be organized in individual research groups under the leadership of a responsible investigator, each group representing a different research area of the unit. The parameters used to evaluate the research groups were 'Productivity' (the total output of the group in its many different forms), 'Relevance' (the scientific, technical, and/or socioeconomic impact of the work carried out), 'Feasibility' (the capacity of the group to transform interesting plans into practical projects relevant at the international level), and 'Training' (PhD and Master students and participation in graduate programs). These four items were rated on a scale of 1 to 5, and the final numerical rating of each group was calculated using predefined weightings.

Table 1 shows the criteria used in 2007 for translating the results of the individual research groups to the overall grading of the unit. The overall unit report should reflect the consensus of the panel, integrating the reports and recommendations about the different research groups into a single document. Despite the fact that research groups were evaluated on four discrete criteria, the unit received a single rating, which in effect grades perceived research excellence on a scale from 'poor' to 'excellent', i.e. 'internationally recognized outstanding research'. Implicitly, it is being assumed that all four criteria considered when assessing research groups work as indicators of research performance. By the end of this evaluation, 293 RUnits were funded (units which were rated by peers as 'fair' or 'poor' did not receive any funding).

The evaluation of ALabs in 2008 was also conducted by panels of experts. The report submitted to the evaluators had to cover the period 2003–07 and be organized in research lines, each line being sub-divided in research groups under the leadership of a principal investigator. In addition, an activity plan for 2008–10 had to be submitted. The evaluators were required to take into account 'the success of the scientific and technological activities undertaken, their internationalization, the relevance of the institution's research and technological development and its pertinence to the objectives of national policy for S&T' (FCT 2008b). ALabs considered 'excellent' were able to maintain the status for an additional period of 5 years.

The evaluation that took place in 2014 resulted in a substantial reorganization of the system described above: units could keep the previous organization or reorganize to better achieve their strategic goals. This reorganization included the creation of new units, as well as the merger or closure of existing units. All units were evaluated on an equal level, irrespective of their legal status (i.e. whether a RUnit or an ALab), and with the same basis for funding. A six-point scale was adopted—ranging from 'poor' to 'exceptional'—and, again, only those institutions rated as 'good' or above obtained funding.

1.3 Peer review and bibliometrics

The reliability of tools employed by policy makers and funding agencies for evaluating a research unit's performance is increasingly scrutinized, since the results of these evaluations are used for allocating funds. The identification of 'excellence' has become more and more important as the crisis led to budget cuts and enhanced the need for transparency and accountability. Finding a reliable and common platform for the evaluation of research institutions with different profiles is a great challenge, and two main approaches have been used: peer review and bibliometrics. The former, considered the gold standard, is based on perceptions of well-informed experts about different dimensions of research and has the possibility of including a number of qualitative aspects into the final result. However, peer review is inherently subjective, as the peers' judgment can be influenced by their individual context (Nederhof and van Raan 1993; Langfeldt 2004; Fedderke 2013; Ahlgren and Waltman 2014), and can be extremely costly and provide little in terms of cost benefit, at least for areas well covered by bibliometrics (Abramo et al. 2013). Bibliometrics can provide quantitative data, but also has shortcomings. Since bibliometrics is restricted to the written output, the results are affected by the measurement method and depend on the publication and citation habits of the different scientific fields. Several studies (Aksnes and Taxt 2004; Allen et al. 2009; Abramo et al. 2011; Taylor 2011; Cabezas-Clavijo et al. 2013) have concluded that citation analysis and bibliometric indicators may be used as a complementary tool to peer review, in a process designated *informed peer review* (van Raan 1996). Derrick and Pavone (2013) have even argued that informed peer review is a way of democratizing research evaluation.

In 2007, the evaluators of the Portuguese RUnits were instructed to evaluate the 'Productivity' of research groups as follows: 'Total

Table 1. Criteria for establishing the overall grade of a research institution

Grades	Description
Excellent	Unit in which one or more groups carried out internationally recognized outstanding research which contributed to the advancement of the field while most others did high-quality international research which lead to some contributions to their specific fields.
Very good	Unit in which most groups did high-quality international research which lead to some contributions to the field and the remaining did good, solid research at international level, leading to incremental contributions to their fields.
Good	Unit in which one or few groups did high-quality international research which leads to some contributions to the field while most groups did good, solid research at international level, leading to incremental contributions to the field.
Fair	Unit in which few groups did good solid research at the international level, leading to incremental contributions to the field, while most groups did satisfactory research which will not necessarily lead to any significant contributions to the field.
Poor	Unit in which few groups did satisfactory research at the international level which will not necessarily lead to recognized contributions to the field and most groups carried out research that is unsatisfactory and unlikely to contribute to advancement of the field at any level.

Source: FCT 2007.

output of the group in its many different forms, including publications, patents, prototypes or products. Consider the output in terms of human resources. For those research areas in which bibliometric parameters are available they will be used' (FCT 2007). The publicly available documents contained neither guidance about the bibliometric indicators to be used nor the source of data.

In contrast, the use of bibliometrics was discouraged in the evaluation of ALabs in 2008: the quality and relevance of the outputs, as judged by independent experts will be favored to the detriment of simplistic metrics such as counting publications, citations, or impact factors (FCT 2008b, translation of the authors). It should be noted that despite those guidelines, a significant number of references to bibliometrics can be found in the evaluation reports of ALabs, as shown in the following transcripts (FCT 2008a):

'A brief overview of the scientific publication list suggests that [ALab] must increase the quality and impact factor of its scientific publications';

'The publication and citation record is also good to very good, with about 130 papers per year in refereed international journals, some of them published in highly cited journals (impact factor >5)';

'The principal investigators achieved good or even excellent citation records and h-indexes';

'The overall scientific quality steadily improved . . . with an increase in the impact factor by 50 per cent and the number of citations by approx. 32 per cent/yr';

'Very few publications are in journals with high impact': in this ALab, the impact factor of journals was one of the reasons supporting the recommendation for non-renewal of the contract.

The examples above illustrate how the panels resorted to a non-homogeneous set of 'indicators', including the questionable practice of using journal rankings or impact factor as proxies of research impact. To avoid these problems, clear rules for the use of citation data by the evaluators must be set *a priori*. An example of good practice is the Research Excellence Framework (REF) 2014 in the UK: each of the four main panels (organized around broad scientific domains) established a common set of criteria and working methods for its sub-panels. In REF 2014, expert review was the primary means of evaluating outputs, but the sub-panels could use citation data—provided by the REF team—to inform their judgments. However, it was clear that 'no panel will make use of journal rankings or journal impact factors in the assessment' (HEFCE 2012).

In 2012, FCT commissioned the Centre for Science and Technology Studies (CWTS) of Leiden University to produce a bibliometric study of the publication output (2007–10) of all the institutions that received funding following the 2007/8 evaluations. The data provide the first detailed bibliometric characterization of research institutions in Portugal. Although restricted to the outputs indexed in WoS, it is an invaluable tool to obtain a snapshot of the country's scientific production and impact in recent years. We are aware that another limitation of our study derives from the fact that bibliometric data are not available to all evaluated entities, but only for those rated as 'good' or above. Nonetheless, it provides an opportunity to reflect on the methodology of the past evaluations, and to understand which were their advantages and limitations. This set of data allows us to analyze how well the peer ratings attributed in the 2007 evaluation of FCT-funded units (i.e. those deemed to be at least 'good'), which referred to past performance (based on the institutional reports from 2003 to 2006/7) but also future prospects of the units evaluated (from the activity plans between 2007/8 and 2010), relate to research performance measured by bibliometrics for the subsequent period of 2007–10.

In a recent paper, Bornmann (2011) defends that unlike the results for journal peer review, where a good agreement exists between peers' perception and the future citation impact of the publications, contradictory results emerge in research on fellowship or grant peer review. While some studies confirm the predictive validity of peer review, others leave room for doubt (p. 223). Our exercise analyzes the predictive validity of peer evaluations, namely how well peers are able to discern the best 'investment opportunities' for FCT.

The present study aims to address the following questions:

- Could research evaluations based on peer review predict a units' future performance as measured by bibliometric indicators?
- What are the strengths and weaknesses of the output of Portuguese research institutions?

These questions take on increased relevance in a context of crisis with generalized cuts in funding. The public and private R&D expenditure reached 1.58% of GDP in 2009 but declined to 1.37% of GDP in 2012 (DGEEC 2014). Consequently, proper evaluations are crucial for decision makers to make informed decisions about the allocation of funding, to provide accountability for public investment in research and to produce evidence of the benefits of this investment.

2. Data sources and methodology

This study analyzed the results of the bibliometric exercise commissioned from CWTS of the publication output indexed in WoS for institutions that received FCT funding in the wake of the evaluations that took place in 2007–8. The lists of publications were supplied to FCT following a request sent to 319 institutions (293 RUnits and

26 ALabs) and included the references indexed in the Science Citation Index Expanded (SCI-EXPANDED), Social Sciences Citation Index (SSCI), Arts & Humanities Citation Index (A&HCI), Conference Proceedings Citation Index—Science and Conference Proceedings Citation Index—Social Science & Humanities. For each reference, the WoS Accession Number and the publication year had to be indicated. In addition, the publications had to meet both of the following criteria: at least one of the authors was a full member of the institution in the period 2007–10 and the publication must have at least one Portuguese address. The period covered was 2007–10 with citations gathered with a 5-year window (with 2012 as the last year for citations when the 5 years cannot be reached).

The results were collated in Excel files that contained bibliometric indicators per institution and also per WoS subject category. The average number of Full Time Equivalents (FTE) in 2008–10 in each RUnit was added by FCT as ancillary information. The files included data for 278 research institutions (corresponding to 87% of the total number of institutions queried by FCT). Only publications (articles and reviews) indexed in the SCI-EXPANDED, SSCI, and A&HCI were considered; conference proceedings, although supplied by the research units, were not included. CWTS also calculated bibliometric indicators per WoS subject category for the following sets of publications: (1) all publications containing at least one Portuguese address; (2) all FCT funded publications; (3) all non-FCT funded publications.

The present study uses only some of the research units' bibliometric indicators calculated by CWTS (for detailed descriptions of the indicators please see Waltman et al. 2012). The first indicator is the 'Number of publications' (P) for the unit being analyzed during the entire period (only articles and reviews); a whole counting method was used by CWTS, i.e. each organization listed on a paper received a whole count for that paper. Double occurrences were excluded within each unit of analysis: a paper co-authored by two or more researchers belonging to the same institution was counted only once, and one paper authored by two or more institutions was counted once at a higher aggregation level (Portugal). The second indicator is the 'Mean field normalized citation score' (MNCS), the average number of citations for publications of a unit normalized for differences between fields (WoS subject categories), publication years, and document types. The expected number of citations is based on the worldwide average citation score without self-citations (a citation is classified as a self-citation if the citing publication and the cited publication have at least one author name in common) of all papers belonging to the same field(s), year, and document type. A field normalized score is calculated for each paper, and the MNCS indicator is computed for each unit of analysis by taking the average of the field-normalized citation scores for individual papers. A value above 1 indicates that the mean impact of the unit is above world average, whereas a value below 1 indicates the opposite.

One should be aware that the field delineation used by CWTS is based on the classification of scientific journals into subject categories used by Thomson Reuters. In the majority of the cases, the output of a unit is distributed over two or more fields and a weighted average value is calculated, the weights being determined by the total number of papers the unit has produced in each field. Although this methodology for assigning publications to subject area(s) has flaws, at present there is not an alternative classification that better fits the purpose of these studies (van Leeuwen and Moed 2012).

The next indicator is the 'Percentage of highly cited publications' (PPTOP10), which indicates the proportion of publications that belong to the 10% most frequently cited in the same field(s), type of document, and publication year. The percentage of publications that result from institutional collaboration is measured by the PP(collab) indicator, and PP(int collab) corresponds to the subgroup that originates from collaboration with international entities (publications that have at least one non-Portuguese address). In our work, the units' scores for two impact indicators MNCS and PPTOP10 will be shown, but the latter is considered by CWTS to be the most reliable impact indicator, as MNCS can be very influenced by a low number of extremely highly cited publications (Waltman et al. 2012). CWTS also used an indicator called 'Coverage': since the bibliographic references can be considered the knowledge base on which the researchers build, this is a surrogate of the importance of WoS publications to the researchers in the units studied. Coverage is the number of WoS references cited over the total number of references cited in the unit's publications.

For our analysis, institutions with less than 25 publications were excluded, as the reliability of bibliometric indicators is low for small samples. From the universe of 278 research institutions included in the bibliometric study, 175 had 25 or more articles or reviews indexed in WoS in 2007–10. This group included the 26 ALabs. As stated above, to estimate the importance of WoS publications for each institution, CWTS analyzed their reference lists. Of the 175 units with 25 or more publications, 136 (22 ALabs and 114 RUnits) had a WoS coverage of 60% or above (60% coverage means that 60% of the references cited in the unit's publications are indexed in WoS); a percentage between 60 and 80% is generally regarded as 'good', whereas a coverage above 80% is considered 'excellent' (Moed 2005). Only the units with good or excellent coverage were included in our study. For institutions that do not meet this criterion, the WoS data would have to be complemented with information about other publications, either periodical literature or non-periodical output, such as conference proceedings, books, book chapters, and monographs.

When analyzing the relationship between peer-review results and bibliometrics, the Kruskal–Wallis (KW) test was used to test for differences in the values of the bibliometric indicators between different peer rating groups of research units. When there was statistical evidence of difference between groups, the KW test was followed by the post hoc Tukey test to test for differences between specific pairs of groups.

3. Results

3.1 Portugal versus FCT-funded institutions

In the period considered (2007–10), the publications of FCT-funded units accounted for 80% of all publications with at least one Portuguese address (32,540). The MNCS of all Portuguese publications was 1.03 and the PPTOP10 was 10.0%. The FCT-funded output (corresponding to 26,116 publications) and the remaining 6,424 publications (non-FCT funded) have an MNCS of 1.04 and 1.02, respectively, and PPTOP10 of 9.9% and 10.3%; hence, for the three sets of documents the field-normalized impact equals world average. The 30 WoS subject categories of higher production are shown in Fig. 1 for Portugal (A) and for FCT-funded research (B). The subject areas and the respective number of publications and impact are very similar in both graphs. *Neurosciences* are the exception; there is a significant production outside the group of FCT-funded units. Among the 6,424 publications that are not affiliated with any

FCT-funded unit, there are some subject areas with high impact: *Biochemistry & Molecular Biology*, *Clinical Neurology*, *Food Science & Technology*, *Immunology*, and *Neurosciences*, all with a publication output above 150 during 2007–10 (not shown).

Looking at the publication and impact distribution of the 30 subject categories for FCT-funded units (Fig. 2), three main groups can be distinguished: one with higher production that includes *Biochemistry & Molecular Biology*, *Environmental Sciences*, *Multidisciplinary Material Sciences*, and *Physical Chemistry*. A second group corresponds to subject areas of intermediate production and high impact: *Applied Chemistry*, *Astronomy & Astrophysics*, *Chemical Engineering*, *Food Science & Technology*, and *Multidisciplinary Physics*. The third cluster comprises subject areas with impact between 0.8 and 1.2 and low to intermediate output.

3.2 FCT-funded institutions

Table 2 compares the bibliometric indicators for ALabs and RUnits. There are some striking differences in size between the two types of units: the number of FTE and the publication output is respectively four- and fivefold higher in ALabs than in RUnits. Productivity (measured by P/FTE) is also higher in ALabs. Regarding impact, median values for MNCS and PPTOP10 are lower than means, i.e. the distributions are skewed to the right and a relatively low number of institutions have a very high citation impact. In this case, the best parameter for analyzing differences is the median. The median MNCS of ALabs and RUnits is 1.05 and 0.94, respectively, and median PPTOP10 values are 10.6% and 8.6%.

3.2.1 Associated laboratories.

All ALabs by definition had achieved a peer rating of 'excellent' in the 2008 evaluation. The names and abbreviations of the 22 ALabs included in our study are shown in Table 3. The four ALabs excluded for having a WoS coverage below 60% were *Instituto de Ciências Sociais* (ICS), *Centro de Estudos Sociais* (CES), *INESC—Lisboa*, and *Instituto de Telecomunicações* (IT). ICS and CES are institutions focusing on Social Sciences where books and national literature are important but not indexed in WoS, and the activity of INESC—Lisboa and IT is centered on *Electrical & Electronic Engineering*, *Telecommunications*, and *Computer Science* for which conference papers are important but not indexed in WoS or indexed in WoS but not included by CWTS in the bibliometric study.

Table 4 reveals a considerable heterogeneity among the ALabs: the number of FTE ranges from 44.33 to 232.98; P varies between 200 and 1,300, P/FTE between 2.8 and 9.9, MNCS from 0.79 to 1.64, and PPTOP10 from 7.0 to 20.3%. The lowest percentage of papers in collaboration is 57.2%, whilst the percentage of international collaboration can be as low as 31.0% or as high as 90.8%. This heterogeneity is also evident in Fig. 3 that depicts the number of publications and PPTOP10 for the 22 ALabs. Note, however, that publications in Table 4 and Fig. 3 do not include Conference Proceedings, which are relevant for some institutions.

A scatter plot of impact and production (Fig. 4) shows that most ALabs have an MNCS ranging from 0.8 to 1.2; six (27.3%) attained MNCS scores at least 20% above world average (>1.2): CICECO, ICVS/3B'S, IMM, ITQB, LIP, and LSRE. Two of these units have a number of publications in the lower range (less than 400) and two have a high publication record (above 1,200).

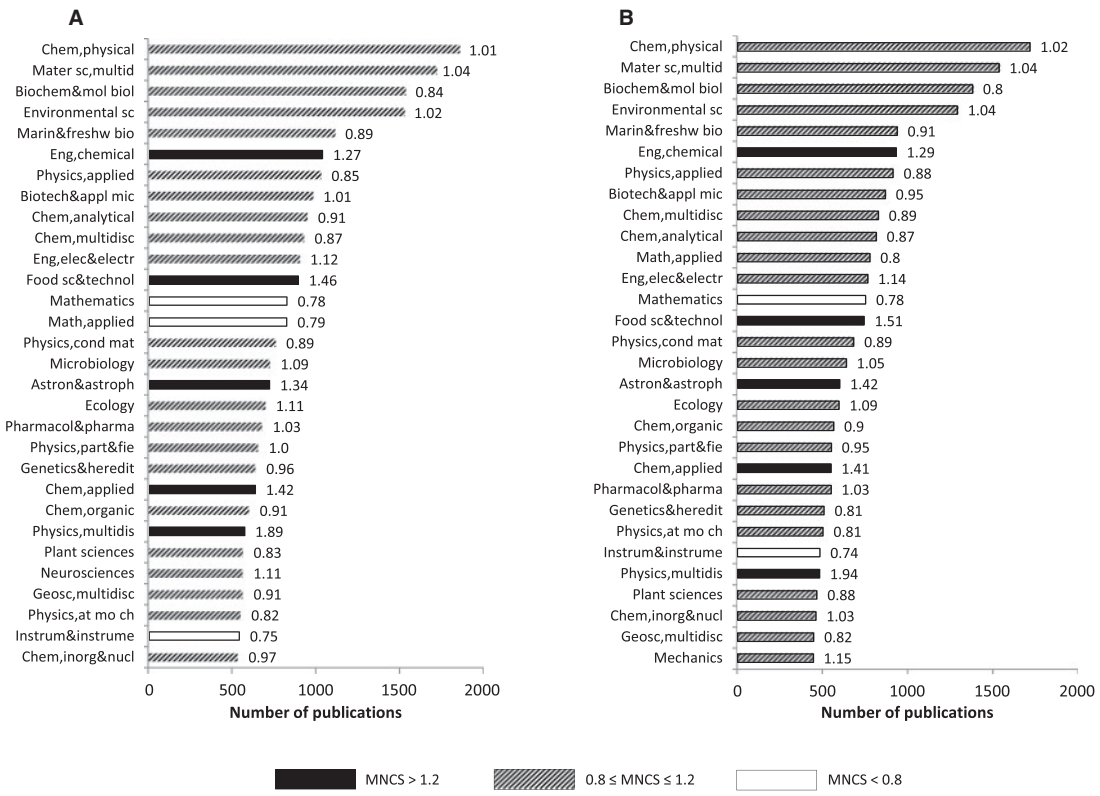

Figure 1. Subject areas of higher production in Portugal (A) and FCT-funded units (B).

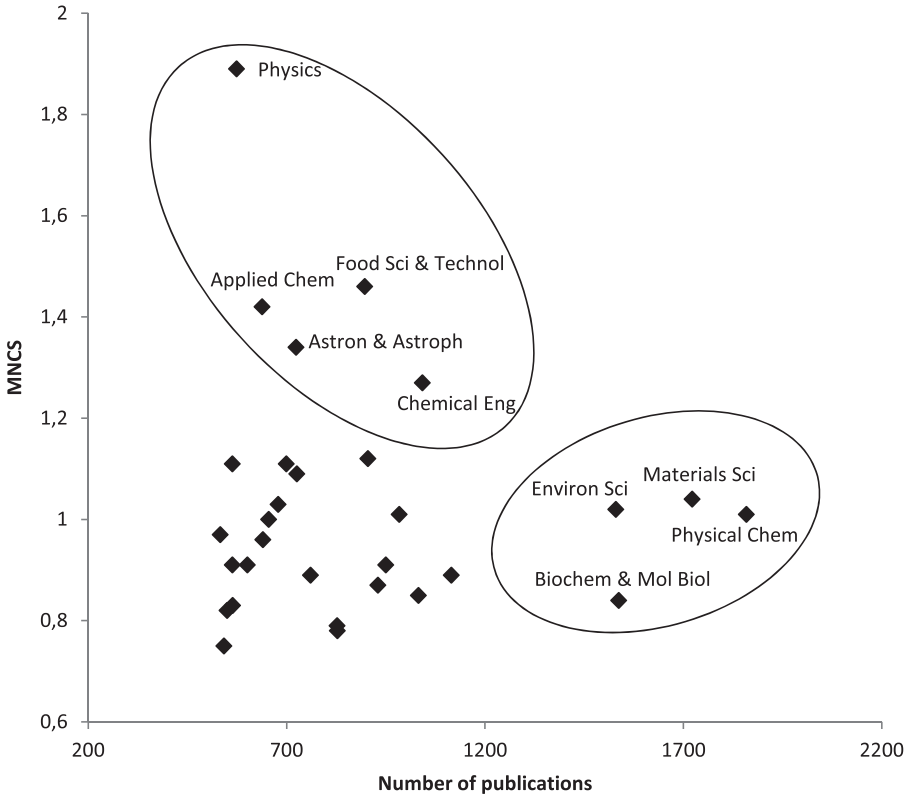

Figure 2. Impact distribution of the subject categories of higher production (Portugal).

Table 2. Statistics for Associated Laboratories ($N = 22$) and Research Units ($N = 114$)

	Associated Laboratories		Research Units	
	Mean	Median	Mean	Median
P	622.6	497.0	129.1	99.5
FTE	128.58	122.26	36.18	31.0
P/FTE	5.0	4.8	3.9	3.2
MNCS	1.11	1.05	1.03	0.94
PPTOP10	11.4%	10.6%	9.2%	8.6%

The number of publications does not include Conference Proceedings.

When the number of publications in the TOP10 is related to the expected number taking into account the output (Fig. 5), it can be concluded that 13 ALabs have a ratio above the expected (>1.0). In this group, nine institutions have a ratio higher than 1.2 (i.e. the number of top papers is at least 20% higher than expected): CICECO, ICVS/3B'S, IMM, IPATIMUP, IPFN, ITQB, LIP, LSRE, and InBIO. Conversely, in nine ALabs (41%) the number of publications in the TOP10 is lower than anticipated. Furthermore, despite the fact that the ALabs had to demonstrate 'excellence in research' to the evaluation panels, 68.2% have an MNCS between 0.8 and 1.2 and for one ALab the impact is more than 20% below international average impact (Table 5). For this institution, coverage (only 63.2%) may be a reason for the discrepancy between peer judgment and the bibliometric result. A correlation analysis of the data shown in Table 4 revealed a strong positive association between MNCS and PPTOP10, as could be foreseen ($r = 0.88$) and a weaker correlation between P and the number of FTE ($r = 0.74$). There is no correlation between measures of output (P) or productivity (P/FTE) and measures of impact (MNCS or PPTO10), or between impact and the percentage of publications in international collaboration (PP int collab).

Among the 25 pairs ALab/WoS subject category with higher output (between 221 and 80 publications), only two pairs with an MNCS significantly below world average were found (Fig. 6). Pairs with an impact significantly above world average (MNCS > 1.2 or PPTOP10 > 12%) were identified in several categories of chemistry, materials science, and physics, as well as in the subject categories *Chemical Engineering*, *Fisheries*, *Food Science & Technology*, and *Microbiology*.

Our further analysis examined the pairs ALab/subject category with a publication output between 80 and 25 (see Supplementary Data—Fig. S1). In this lower output range, several additional high impact areas (MNCS > 1.2 or PPTOP10 > 12%) were identified: *Biomedical Engineering*, *Cell Biology*, *Clinical Neurology, Ecology, Mechanics, Meteorology & Atmospheric Sciences*, *Nanosciences & Nanotechnologies*, *Neurosciences*, *Oceanography*, *Oncology*, *Pathology*, *Pharmacology & Pharmacy*, *Plant Sciences*, *Polymer Sciences*, *Toxicology*, *Urology*, and *Nephrology*.

3.2.2 Research units.

A plot of MNCS and P (Fig. 7) shows that the majority (53.5%) of the RUnits have an impact ranging from 0.8 to 1.2; nevertheless, a considerable number (26) have an impact well above the international average, mainly in the lower limit of the output range (less than 200 publications). Twenty-seven RUnits have an MNCS below 0.8 (the RUnits are not identified in Fig. 7, but the full list and the

Table 3. List of Associated Laboratories included in this work

Name	Abbreviation
Ctr de Biotecnologia e Química Fina	CBQF
Ctr de Estudos do Ambiente e do Mar	CESAM
Ctr de Investigação em Materiais Cerâmicos e Compósitos	CICECO
Ctr de Investigação Marinha e Ambiental	CIMAR
Ctr de Neurociências e Biologia Celular	CNC
INESC Tecnologia e Ciência	INESC TEC
Inst de Biologia Molecular e Celular. Inst Nacional de Engenharia Biomédica	IBMC INEB
Inst de Biotecnologia e Bioengenharia	IBB
Inst de Ciências da Vida e da Saúde/Biomaterials, Biodegradables and Biomimetics Group	ICVS/3B's
Inst de Medicina Molecular	IMM
Inst de Nanoestruturas, Nanomodelação e Nanofabricação	I3N
Inst de Nanotecnologias	IN
Inst de Patologia e Imunologia Molecular da Universidade do Porto	IPATIMUP
Inst de Plasmas e Fusão Nuclear	IPFN
Inst de Tecnologia Química e Biológica	ITQB
Inst Dom Luís	IDL
Lab Associado de Energia, Transportes e Aeronáutica	LAETA
Lab Associado para a Química Verde	REQUIMTE
Lab de Instrumentação e Física Experimental de Partículas	LIP
Lab de Processos de Separação e Reação	LSRE
Lab de Robótica e Sistemas em Engenharia e Ciência	LARSyS
Rede de Investigação em Biodiversidade e Biologia Evolutiva	InBIO

main bibliometric indicators can be found in the Supplementary Data—Table S1).

Table 6 shows the distribution of the ratings of FCT-funded RUnits obtained by peer evaluation in 2007 and relates them with their subsequent performance in terms of productivity and impact indicators. Median P/FTE, MNCS, and PPTOP10 values in the group rated as 'very good' are higher than in the group rated as 'good', but no discernible differences were found between the 'very good' and 'excellent' institutions. For the whole sample (N = 114), no relation was found between the rates attributed by peers and subsequent research performance measured by productivity or impact indicators. The Spearman rank correlation coefficients between peer rating and P/FTE, MNCS, PPTOP10 were 0.215, 0.067, and 0.065, respectively. The peer judgment was also not related to the degree of internationalization, measured by the percentage of publications in international collaboration (0.145). This suggests that the panels were not able to 'predict productive and high-impact research institutions'. Similar findings have been reported in previous studies that compared the outcome of peer evaluation with bibliometric data (Moed 2005: 244). Table 7 shows the distribution of RUnits by five citation impact classes and gives further insight into this issue: only one of the three RUnits that has developed research with very high impact (>2.0) had been rated as 'excellent' in the evaluation. On the opposite side of the impact distribution, one unit with research of very low impact (<0.5) had been considered 'very good' and seven RUnits with low impact (between 0.5 and 0.79) had been rated by peers as 'excellent'. These situations were analyzed in more detail. The RUnit with very low impact, despite the fact that it had been attributed a rating of 'very good', produced 62 WoS publications with an MNCS of 0.41. The fact that its main production falls in the WoS subject categories *Mathematics* and *Applied Mathematics*, both with coverage of only around 60%, may help to explain the discrepancy between peer judgment and bibliometric data. It might be the case that peers took into consideration very good work that is

Table 4. Main bibliometric indicators for Associated Laboratories

	P	FTE 2008–10	P/FTE	MNCS	PPTOP10 (%)	Coverage (%)	PP(collab) (%)	PP(int collab) (%)
CQFB	200	44.33	4.5	1.03	8.3	82.5	64.0	31.0
CICECO	1,300	130.96	9.9	1.38	16.1	87.6	79.9	55.9
CNC	483	125.18	3.9	1.15	11.8	94.2	70.8	44.1
CESAM	845	132.20	6.4	1.02	10.6	79.0	74.8	42.1
CIMAR	969	200.67	4.8	0.92	8.4	81.8	80.7	52.4
IBMC.INEB	714	206.67	3.5	1.00	9.8	91.6	80.7	47.3
ICVS/3B's	511	99.50	5.1	1.26	13.5	92.4	90.6	54.4
INESC TEC	338	98.11	3.5	0.95	9.9	61.3	80.2	39.1
IBB	1,036	232.98	4.5	1.05	10.1	86.5	71.5	42.4
IMM	466	147.01	3.2	1.22	12.2	92.9	81.8	45.7
I3N	625	119.33	5.2	1.01	9.5	84.7	82.2	60.5
IN	445	58.17	7.7	0.90	9.1	90.7	88.8	45.8
IPATIMUP	386	60.67	6.4	1.12	14.4	92.7	92.0	69.4
IPFN	423	72.33	5.9	1.00	12.4	84.7	83.9	76.8
ITQB	1,219	166.45	7.3	1.25	13.5	90.8	88.3	56.5
IDL	244	66.67	3.7	0.94	7.2	70.7	91.0	67.6
LAETA	633	228.87	2.8	1.10	10.5	64.9	70.1	32.4
LIP	273	67.22	4.1	1.56	13.6	75.2	96.7	90.8
LSRE	327	60.00	5.5	1.64	20.3	86.0	57.2	31.8
LARSyS	590	108.51	5.4	0.79	7.0	63.2	66.4	52.5
InBIO	371	126.75	2.9	1.19	12.5	71.3	92.2	72.8
REQUIMTE	1,299	276.08	4.7	1.04	9.7	88.9	70.7	38.6

The number of publications does not include Conference Proceedings.

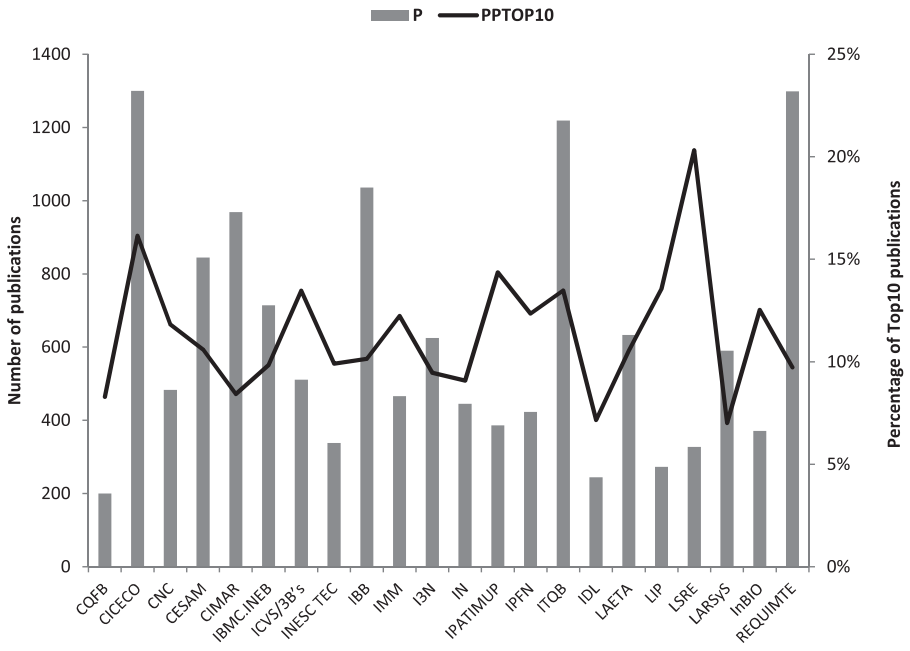

Figure 3. Alabs number of publications and percentage of publications in the TOP10.

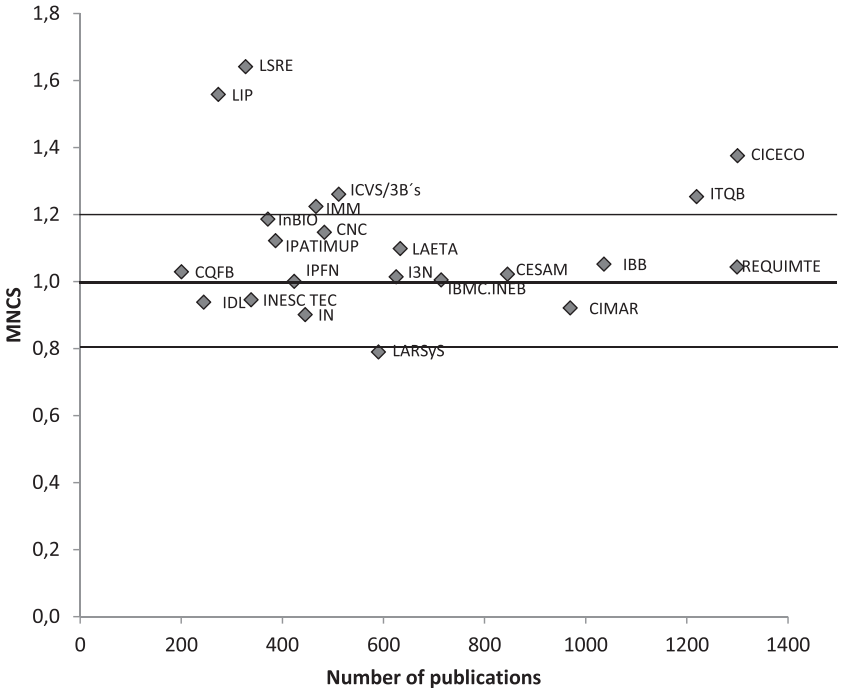

Figure 4. Impact of ALabs (MNCS) compared to the world average.

not indexed in WoS. The institution with impact >2.0 that had been rated by peers as 'excellent' has both high impact and output in *Astronomy & Astrophysics*. The other RUnit of high impact that had been rated as 'very good' also publishes mainly in the WoS subject category *Astronomy & Astrophysics*, but its output is low. It is worth mentioning that these two units were evaluated by different panels, Physics and Environment, respectively. Among the seven units that had obtained the highest ratings despite having developed research of low impact, their subject categories fall into different areas, such as chemistry, physics, mathematics, virology, geology, or economics. Their publication output ranges from 28 to 209 and the coverage is between 61.3% and 91.5%.

For the subject category analysis, the same sample of RUnits was used, but no coverage threshold was applied. Among the 25 pairs RUnit/subject category with higher publication output, (number of publications ranging from 55 to 210) only five have a high impact

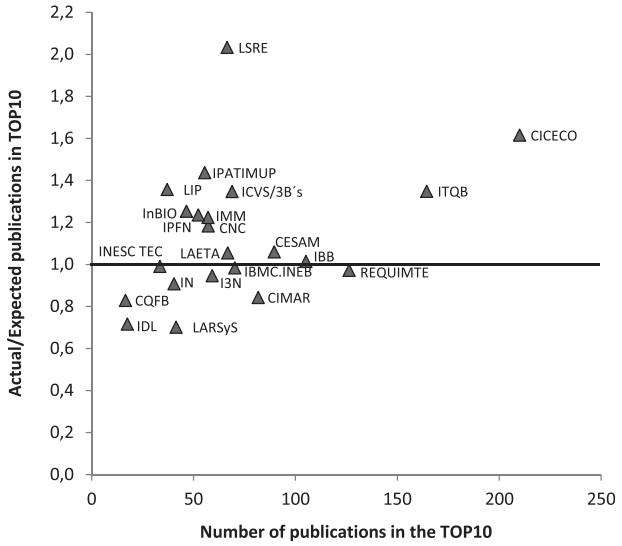

Figure 5. ALabs actual/expected number of publications in the TOP10.

Table 5. Distribution of Associated Laboratories by citation impact (MNCS) classes

	Citation impact class		
	Low (0.5–0.8)	Average (0.8–1.2)	High (1.2–2.0)
Nr (%) of units	1 (4.5)	15 (68.2)	6 (27.3)
Median P	590.0	483	488.5
Median FTE	108.5	125.2	115.2
Median P/FTE	5.4	4.5	5.3
Median MNCS	0.79	1.02	1.32
Median PPTOP10 (%)	7.0	9.9	13.5
Median coverage (%)	63.2	84.7	89.2

The number of publications does not include Conference Proceedings.

(MNCS > 1.2); 12 have an average impact (MNCS between 0.8 and 1.2), and the impact of those remaining is below world average (Fig. 8). One has to take into account that there are coverage differences among the subject categories, and for some pairs RUnit/area the indicators must be interpreted with caution. 'Top' areas defined by having MNCS > 1.2 or PPTOP10 > 12% were pinpointed: *Astronomy & Astrophysics* (contribution of two RUnits), *Forestry*, *Physical Chemistry*, and *Physics of Condensed Matter* (Supplementary Data—Fig. S2).

3.2.3 Differences in bibliometric indicators according to peer rating and field of science.

Since the evaluation of ALabs and RUnits was organized by discipline—but more importantly because performance measures are subject dependent, in particular those that are not normalized, such as productivity (Abramo and D'Angelo 2014)—we have tried to further detect relationships between the rating attributed by peers and bibliometric indicators relating to research subsequently developed (P/FTE, MNCS, PPTOP10, and PP int collab) for different fields of science (Table 8). The fields of science considered are those used by FCT to aggregate units. Each field of science has its own scientific council, composed of renowned researchers, who advise FCT on matters related to their respective field.

We have put together the 114 RUnits and 22 ALabs with at least 25 publications and 60% coverage, and included the ALabs in the group of units rated as 'excellent'.

It is found that for most cases there is no statistical difference between the bibliometric indicators for different peer rating groups. No differences are detected at all for Natural and Environmental Sciences. For Exact Sciences and Engineering, differences are only detected for the PP int collab indicator, and only between 'good' and 'very good' units, and 'good' and 'excellent' units; no discernible difference is detected between 'very good' and 'excellent' units. Differences are indeed detected for Life and Health Sciences for the impact measures MNCS and PPTOP10, but only between the 'good' and 'excellent' groups. Discrimination between 'good' and 'very good', and 'very good' and 'excellent' rated groups is not discerned. No differences are detected for productivity as measured by P/FTE.

4. Discussion and conclusions

This article addresses two questions: firstly, whether the research evaluation based on peer review is able to predict the units' subsequent performance measured by bibliometric indicators, and secondly, what the strengths and weaknesses of the recent output of Portuguese research institutions as measured by bibliometric indicators are.

Several authors have described reasonable correspondences between results from peer review and bibliometrics in research evaluation exercises (Rinia et al. 1998; Abramo et al. 2013). In editorial peer review, a good agreement exists between peers' perception and the future impact of the publications. However, contradictory results were found for fellowship or grant peer review: some studies confirm the predictive validity of peer review, while others do not (Bornmann 2011: 223). The bibliometric data available for this study were restricted to entities funded by FCT, i.e. those rated as 'good' or above, and not the whole spectrum of peer evaluation. Here we are not interested in comparing peer review and bibliometrics in the evaluation of research units as such, but how well peer review identifies good 'investment opportunities' for the funding council, i.e. whether the results of evaluation by peer review predict subsequent research performance. The data suggest that peer evaluation could not distinguish between subsequent very good and 'top' research. This may be due to several factors: on the one hand, peers did not have 'formal' access to bibliometric indicators and have appraised the units' past performance and future prospects from reports and activity plans. On the other hand, when the field of study was taken into consideration, differences in subsequent research were detected for Life and Health Sciences, although only between 'good' and 'excellent' groups, for two of the bibliometric indicators: MNCS and PPTOP10; and for PP int collab in Exact Sciences and Engineering.

Our findings show that the peers' predictive ability was somewhat better in fields for which bibliometrics has been traditionally used. The fact that bibliometric indicators were not formally supplied to the panels in 2007, does not mean that citation data did not play a role, since peers could directly consult bibliometric databases or make use of data included in the institutions' reports, and there is evidence that some did precisely that.

Nearly 20% of the RUnits with 'high' or 'very high' impact had been rated as 'good' or 'very good' and only 7% had simultaneously 'high' or 'very high' impact and had 'excellent' peer evaluation. If one assumes that the citation indicators are a measure of excellence,

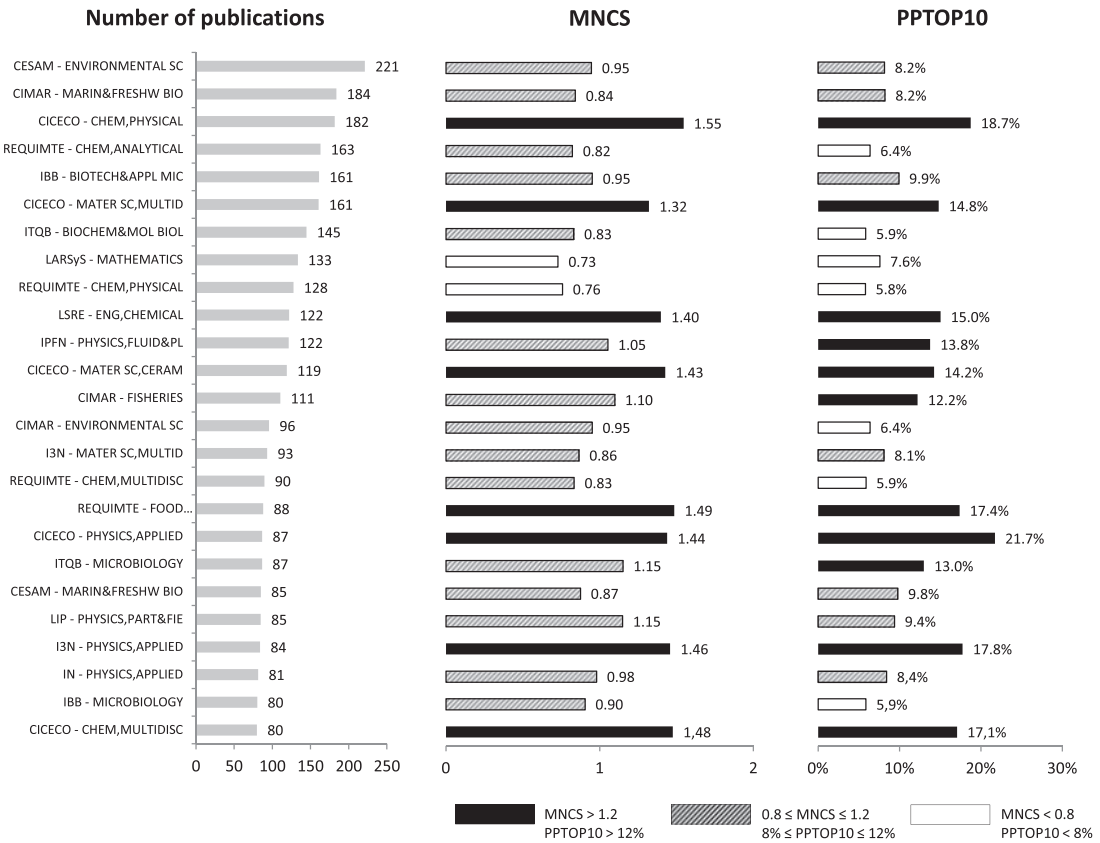

Figure 6. Bibliometric indicators of the 25 pairs ALab/Subject category with higher output.

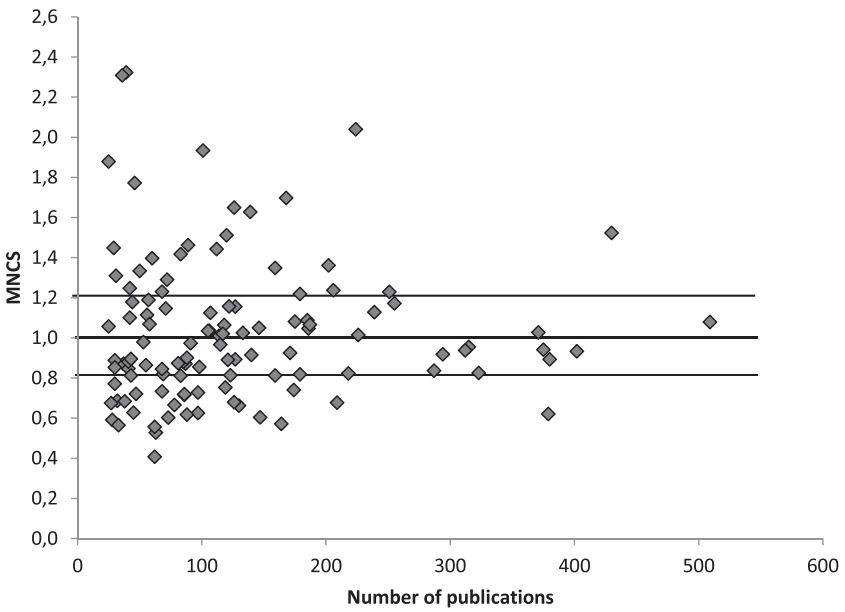

Figure 7. Impact of the RUnits (MNCS) compared to the world average.

it follows that the panels had difficulty in predicting subsequent top research, which is in agreement with the results of previous studies reported by Moed (2005: 242).

Although an equivalent analysis for ALabs could not be performed (because all these institutions were considered 'excellent'), it became clear that the evaluators did not succeed in predicting institutions performing less well in terms of citation impact and, more importantly, that different bibliometric indicators were used in an arbitrary fashion, with journal rankings or impact factor used as proxies for research quality. In fact, peer review

Table 6. RUnits—Productivity and citation impact in relation to peer rating

Peer rating	*N*	%	Mean P/FTE	Median P/FTE	Mean citation impact	Median citation impact	Mean PPTOP10 (%)	Median PPTOP10 (%)
Good	33	28.9	3.0	2.4	0.99	0.89	8.9	7.1
Very good	50	43.9	4.2	3.8	1.03	0.94	9.2	8.6
Excellent	31	27.2	4.3	3.6	1.04	0.94	9.0	8.6
Total	114	100.0	3.9	3.2	1.03	0.94	9.2	8.6

Table 7. RUnits—Peer rating and distribution by citation impact classes (MNCS)

Peer rating	Citation impact class				
	Very low (<0.5)	Low (0.5–0.8)	Average (0.8–1.2)	High (1.2–2.0)	Very high (>2.0)
Good	0	7	20	5	1
Very Good	1	12	25	11	1
Excellent	0	7	16	7	1
Total	1	26	61	23	3

and bibliometrics are never completely independent processes (Rinia et al. 1998).

One could argue that a good correlation between peer evaluation and bibliometrics was not to be expected, since the reviewers were not strictly asked to evaluate research excellence. However, although they were asked to evaluate the different research groups within research institutions on four criteria—productivity, relevance, feasibility, and training—they had to give a unique overall rating to the units on a scale that made explicit reference to research excellence. This means that policy makers were implicitly assuming that the four criteria used in evaluation of groups, would conflate in a measure of research excellence at the level of the research unit, i.e. those four criteria were determinants of research performance.

For instance, in the case of training, one would expect those units with stronger training programs to perform better both in terms of productivity and impact, as one would expect their doctoral and post-doctoral researchers to contribute to the subsequent productivity and impact of the funded units. As to the feasibility criteria, i.e. the capacity of the group to transform interesting plans into practical projects relevant at the international level, one would also expect this to materialize in both productivity and impact, since we are looking at research performed in the period after the units were evaluated. In our opinion, the fact that we could not discern a clear association between the peer review undertaken in 2007 and bibliometric indicators encompassing productivity and impact for the post-evaluation period raises questions as to the reliability of the peer-review-only method.

The funding of units is divided between basic funding (directly proportional to the peer rating) and strategic funding (related to what the unit promises to do and the evaluators think is feasible). Following the 2014 evaluation, the strategic funding for a number of units surpassed basic funding. This is a significant investment by the funding council based on peer review. To make it as robust as possible it seems wise to increasingly have bibliometric-informed peer review.

Evaluation of research institutions plays a crucial role in resource allocation, is determinant for the competitiveness of institutions and researchers and, ultimately, for the progress of a country. The establishment, in the mid-1990s, of international peer review evaluation was a decisive step for the Portuguese scientific system; however, as evident from the present analysis, actions are needed to correct biases and prevent distortions in future evaluations. One such measure was implemented by FCT for the 2014 evaluation of research institutions; the panels had access to bibliometric indicators produced by an independent provider from the lists of publications validated by each institution. This strategy is aligned with the practice in a growing number of countries, where peer review, site visits, and institutional reports are complemented with reliable bibliometric data. Regrettably, the relative importance of bibliometrics and other indicators was not clearly settled in the evaluation guide published by FCT (2013).

Mapping the output of Portuguese research institutions using bibliometric analysis helped to shed light on the strengths and weaknesses of the national R&D system. However, some limitations must be acknowledged: first, this landscape derives only from the WoS indexed publications and different pictures could emerge if other sources were used; secondly, small units were not considered due to the threshold applied to the number of publications; thirdly, almost all institutions whose main activity is in the social sciences and humanities or in some areas of engineering were excluded. This 'bibliometric landscape' is, therefore, confined to the exact, natural, life and health sciences, as well as some engineering areas. In this respect, the developments in altmetrics may be a fruitful avenue that is worth exploring as the field progresses and matures (Costas et al. 2015; Zahedi et al. 2014).

The impact of Portugal in 2007–10 equals the world average; in spite of the lack of previous directly comparable studies, data published by the European Commission showed that the impact of Portuguese institutions was well below average some years ago (EC 2003). This means that the remarkable growth in the number of publications (Heitor et al. 2014) was accompanied by an increase of the country's citation impact.

Among the subfields of higher output (Fig. 1), there is a clear dominance of chemistry and physics. The history of science policy in Portugal helps to understand the dominance of exact sciences (Heitor and Horta 2012): they were preferentially funded, both before Portugal joined the EU in 1986, and then subsequently through the scholarships awarded using European structural funds (exact sciences, natural sciences, and engineering accounted for two thirds of all fellowships approved by FCT between 1997 and 2008), and also in the programs organized by FCT to hire researchers (*Programas Ciência* in 2007 and 2008). *Biochemistry & Molecular Biology*, the third subject area in terms of output, has quite a low impact (Fig. 2) and this should be the subject of careful analysis by

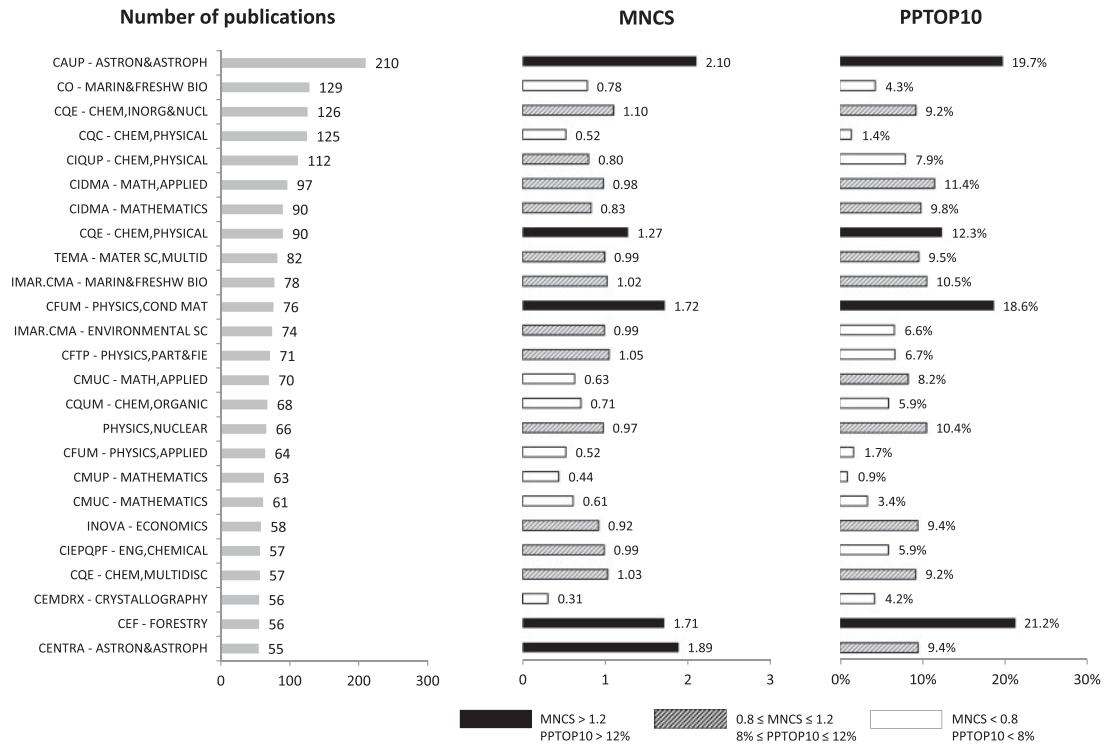

Figure 8. Bibilometric indicators of the 25 pairs RUnit/Subject category with higher ourput.

Table 8. Differences in bibliometric indicators between peer rating groups by field of science

Peer rating: Good (G); Very good (VG); Excellent (E)	P/FTE	MNCS	PPtop10	PPintCollab
Life and Health Sciences	KW: P = 0.709	KW: P = 0.050* TK: G-VG: 0.814 G-E: 0.048* VG-E: 0.114	KW: P = 0.032* TK: G-VG: 0.655 G-E: 0.025* VG-E: 0.106	KW: P = 0.256
Exact Sciences and Engineering	KW: P = 0.053	KW: P = 0.677	KW: P = 0.943	KW: P = 0.001* TK: G-VG: 0.000* G-E: 0.014* VG-E: 0.091
Natural and Environmental Sciences	KW: P = 0.065	KW: P = 0.256	KW: P = 0.449	KW: P = 0.397

Note: Kruskal–Wallis test (KW) for the comparison of three groups; Tukey test (TK) for the comparison of multiple orders.
*Significant at 5%.

researchers, institutions, and funding agencies. Some health-related areas with good performance, such as *Neurosciences*, *Clinical Neurology*, or *Immunology*, are important outside the FCT-funded units; this is probably due to the role of privately funded research carried out in hospitals or other, non-public, institutions.

A high percentage of the publications (38% if all FCT units are considered) included in the bibliometric study result from international collaboration. It would be useful to determine, per subject area, the number and impact of the international publications in which the corresponding author belongs to a Portuguese institution; this indicator of scientific leadership (Moya-Anegón et al. 2013) would give a more detailed picture of the country's strengths and weaknesses.

The results show that globally ALabs perform better than RUnits, in the sense that they have higher productivity (P/FTE), citation impact, and more publications in the top10%. Still, some RUnits are in areas that contribute to the overall impact of Portugal, such as *Astronomy & Astrophysics*, *Condensed Matter Physics*, *Forestry*, and *Physical Chemistry*. Not surprisingly, top-performing subject areas (high output and impact) both in ALabs and RUnits fall within chemistry, physics, and some engineering-related fields, such as material science or food technology. Furthermore, both types of institutions also excel in areas of lower output outside of the 'traditional' physics, chemistry, or materials science; these areas may be emerging or assuming an increased relevance and should be regarded in terms of strategic planning and implementation of future policies. The better performance of ALabs was to be expected, given the level of funding and the possibility of hiring researchers. In 2009, the 26 ALabs received more than 50% of the FCT budget for the research institutions.

Science is increasingly heterogeneous and interdisciplinary, imposing new challenges for research evaluation, and constantly raising the question 'Can science be measured?' (van Raan 2005). The results shown here are an empirical contribution to answering this question, and reinforce the need for stabilized and transparent evaluation processes where the use of quantitative data is clearly defined and the differences between scientific areas are correctly accounted for.

Acknowledgments

We wish to thank Maria Arménia Carrondo for critically reading this manuscript and Isabel Machado for help with the data.

Supplementary data

Supplementary data is available at *REEVAL* Journal online.

References

Abramo, G. and D'Angelo, C. A. (2014) 'How Do You Define and Measure Research Productivity?', *Scientometrics*, 101/2: 1129–44.

——, D'Angelo, C. and Di Costa, F. (2011) 'National Research Assessment Exercises: A Comparison of Peer Review and Bibliometrics Rankings', *Scientometrics*, 89/3: 929–41.

——, Cicero, T. and D'Angelo, C. A. (2013) 'National peer-Review Research Assessment Exercises for the Hard Sciences Can be a Complete Waste of Money: the Italian Case', *Scientometrics*, 95/1: 311–24.

Ahlgren, P. and Waltman, L. (2014) 'The Correlation between Citation-based and Expert-based Assessments of Publication Channels: SNIP and SJR vs. Norwegian Quality Assessments', *Journal of Informetrics*, 8/4: 985–96.

Aksnes, D. W. and Taxt, R. E. (2004) 'Peer reviews and bibliometric indicators: a comparative study at a Norwegian university', *Research Evaluation*, 13/1: 33–41.

Allen, L., Jones C., Dolby, K., Lynn D. and Walport, M. (2009) 'Looking for Landmarks: The Role of Expert Review and Bibliometric Analysis in Evaluating Scientific Publication Outputs', *PLoS One*, 4/6: e5910.

Bornmann, L. (2011)´Scientific Peer Reviеẃ, *Annual Review of Information Science and Technology* 45/1, 197–245.

Cabezas-Clavijo, Á., Robinson-García N., Escabias M., and Jiménez-Contreras E. (2013) 'Reviewers' Ratings and Bibliometric Indicators: Hand in Hand When Assessing Over Research Proposals?', *PLoS ONE*, 8/6: e68258.

Costas, R., Zahedi, Z. and Wouters, P. (2015) 'Do Altmetrics Correlate with Citations? Extensive Comparison of Altmetric Indicators with Citations from a Multidisciplinary Perspective', *Journal of the Association for Information Science and Technology*, 66/10: 2003–19.

Derrick, G. E. and Pavone, V. (2013) 'Democratising Research Evaluation: Achieving Greater Public Engagement with Bibliometrics-Informed Peer Review', *Science and Public Policy*, 40/5: 563–75.

DGEEC (2014) *Inquérito ao Potencial Científico e Tecnológico Nacional - 1982 a 2012: 3 décadas de indicadores de I&D em Portugal*. Direcção-Geral de Estatísticas de Educação e Ciência: Lisboa.

EC (2003) *Third European Report on Science & Technology Indicators*. Brussels: European Commission.

FCT (2007) *Evaluation of Research Units 2007: Guidelines for Evaluators and Evaluation Forms*. Lisbon: Fundação para a Ciência e a Tecnologia.

—— (2008a) *Avaliação de Laboratórios Associados 2008: Resultados da Avaliação*. Lisboa: Fundação para a Ciência e a Tecnologia.

—— (2008b) *Avaliação dos Laboratórios Associados 2008*. Lisboa: Fundação para a Ciência e a Tecnologia.

—— (2013) *Evaluation Guide – FCT Evaluation of R&D Units 2013*. Lisbon: Fundação para a Ciência e a Tecnologia.

Fedderke, J. W. (2013) 'The Objectivity of National Research Foundation Peer Review in South Africa Assessed Against Bibliometric Indexes', *Scientometrics*, 97/2: 177–206.

HEFCE (2012) *REF2014: Panel Criteria and Working Methods*. Bristol: Higher Education Funding Council for England.

Heitor, M. and Horta H. (2012) 'Science and Technology in Portugal: From Late Awakening to the Challenge of Knowledge-Integrated Communities'. In: Neave, G. and Amaral, A. (eds.) *Higher Education in Portugal 1974-2009*, pp. 179–226. Dordrecht: Springer.

Heitor, M., Horta, H. and Mendonça, J. (2014) 'Developing Human Capital and Research Capacity: Science Policies Promoting Brain Gain', *Technological Forecasting & Social Change*, 82: 6–22.

Langfeldt, L. (2004) 'Expert Panels Evaluating Research: Decision-making and Sources of Bias', *Research Evaluation*, 13/1: 51–62.

Moed, H. F. (2005) *Citation Analysis in Research Evaluation*. Dordrecht: Springer.

Moya-Anegón, F., Vicente, G. B., Bornmann, L. and Moed, H. F. (2013) 'The Research Guarantors of Scientific Papers and the Output Counting: A Promising New Approach', *Scientometrics*, 97/2: 421–34.

Nederhof, A. J. and van Raan, A. F. J. (1993) 'A Bibliometric Analysis of Six Economics Research Groups: A Comparison with Peer Review', *Research Policy*, 22/4: 353–68.

Rinia, E. J., van Leeuwen, N., Th, van Vuren, H. G. and van Raan, A. F. J. (1998) ´Comparative Analysis of a Set of Bibliometric Indicators and Central Peer Review Criteria. Evaluation of Condensed Matter Physics in the Netherlandś, *Research Policy*, 27/1: 95–107.

Taylor, J. (2011) 'The Assessment of Research Quality in UK Universities: Peer Review or Metrics?', *British Journal of Management*, 22/2: 202–17.

van Leeuwen, T. N. and Moed, H. F. (2012) 'Funding Decisions, Peer Review, and Scientific Excellence in Physical Sciences, Chemistry, and Geosciences', *Research Evaluation*, 21/3: 189–98.

van Raan, A. F. J. (1996) 'Advanced Bibliometric Methods as Quantitative Core of Peer Review Based Evaluation and Foresight Exercises', *Scientometrics*, 36/3: 397–420.

—— (2005) 'Measuring Science'. In: H. F., Moed, W., Glänzel and U., Schmoch (eds) *Handbook of Quantitative Science and Technology Research*. Dordrecht: Kluwer Academic Publishers.

Waltman, L., Modena, C. C., Kosten J., Noyons, E. C. M., Tijssen, R. and Jan van Eck, N. (2012) 'The Leiden Ranking 2011/2012: Data Collection, Indicators, and Interpretation', *Journal of the American Society for Information Science and Technology*, 63/12: 2419–32.

Zahedi, Z., Costas, R. and Wouters, P. (2014) 'How Well Developed are Altmetrics? A Cross-disciplinary Analysis of the Presence of 'Alternative Metrics' in Scientific Publications', *Scientometrics*, 101/2: 1491–513.