Revised Standards for Quality Improvement Reporting Excellence (SQUIRE 2.0)
September 15, 2015

Text Section and Item Name	Section or Item Description	
Notes to authors	- The SQUIRE guidelines provide a framework for reporting new knowledge about how to improve healthcare - The SQUIRE guidelines are intended for reports that describe system level work to improve the quality, safety, and value of healthcare, and used methods to establish that observed outcomes were due to the intervention(s). - A range of approaches exists for improving healthcare. SQUIRE may be adapted for reporting any of these. - Authors should consider every SQUIRE item, but it may be inappropriate or unnecessary to include every SQUIRE element in a particular manuscript. - The SQUIRE Glossary contains definitions of many of the key words in SQUIRE. - The Explanation and Elaboration document provides specific examples of well-written SQUIRE items, and an in-depth explanation of each item. - Please cite SQUIRE when it is used to write a manuscript.	
Title and Abstract		
1. Title	Magnitude of terminological bias in international health services research: A disambiguation analysis	Pg 1

2. Abstract	**Aims:** Health Services Research (HSR) is affected by a widespread problem related to service terminology including non-commensurability (using different units of analysis for comparisons) and terminological unclarity due to ambiguity and vagueness of terms. The aim of this study was to identify the magnitude of the terminological bias in health and social services research and health economics by applying an international classification system. **Methods:** This study, that was part of the PECUNIA project, followed an ontoterminology approach (disambiguation of technical and scientific terms using a taxonomy and a glossary of terms). A listing of 56 types of health and social services relevant for mental health was compiled from a systematic review of the literature and feedback provided by 29 experts in 6 European countries. The disambiguation of terms was performed using an ontology-based classification of services (Description and Evaluation of Services and DirectoriEs - DESDE), and its glossary of terms. The analysis focused on the commensurability and the clarity of definitions according to the reference classification system. Interrater reliability was analysed using Kappa **Results:** The disambiguation revealed that only 13 terms (23%) of the 56 services selected were accurate. Six terms (11%) were confusing as they did not correspond to services as defined in the reference classification system (non-commensurability bias), 27 (48%) did not include a clear definition of the target population for which the service was intended, and the definition of types of services was unclear in 59% of the terms: 15 were ambiguous and 11 vague. Kappa analyses were significant for agreements in unit of analysis and assignment of DESDE codes and very high in definition of target population. **Conclusions:** Service terminology is a source of systematic bias in health service research, and certainly in mental healthcare. The magnitude of the problem is substantial. This finding has major implications for the international comparability of resource use in health economics, quality and equality research. The approach presented in this paper contributes to minimize differentiation between services by taking into account key features like target population, care setting, main activities and type and number of professionals among others. This approach also contributes to support financial incentives for effective health promotion and disease prevention. A detailed analysis of services in terms of cost measurement for economic evaluations reveals the necessity and usefulness of defining services using a coding system and taxonomical criteria rather than by "text-based descriptions".	Pg 2

Introduction	Why did you start?	
3. Problem Description	Health services research (HSR), health economics and financing, and research of quality and equality in healthcare require comparable data on service provision (Husereau et al., 2013; Raine et al., 2016). However, the reporting methods can differ substantially, and HSR faces significant problems regarding the terminology of services, (Salvador-Carulla et al., 2013) its implication for the measurement of resource utilization (Thorn et al., 2013) and its monetary valuation (Moreno, Sanchez and Salvador-Carulla, 2008; Barnett, 2009). The two major terminological problems in HSR are the non-commensurability bias and terminological unclarity. Non-commensurability is due to research involving different units of analysis that are not comparable like-with-like. For example, it occurs when the costs of an outpatient psychotherapy unit (i.e. a "service") are compared to the costs of psychotherapy as an "intervention" in another setting. The problem of terminological unclarity is also widespread. For example the term "service" can refer to a range of elements such as the provider, the facility, an organizational unit within the facility, or a combination of functions, programs and resources provided in this clinical unit (Salvador-Carulla et al., 2013). Another problem refers to the lack of a formal definition of the target population the service has been designed for (i.e. diagnosis group and age group) (Salinas-Pérez et al., 2020), and the variability in the typology of services depending on location (different areas, regions, countries etc.) and time of evaluation (Salvador-Carulla et al., 2015).	Pg 1
4. Available knowledge	Twenty years ago, Maciejewski et al. identified significant terminology problems in the methods for HSR. To overcome these problems, these authors produced a list of terms commonly used in HSR methods following a scoping review of the literature and internal and external expert consultations (Maciejewski et al., 2002). These terminology problems have been also described in mental health service evaluation (Salvador-Carulla et al., 1999; Salvador-Carulla and Hernández-Peña, 2011) and are critical in the standardisation of international Resource Use Measurement (RUM) instruments (Thorn et al., 2013; Noben et al., 2016). Despite previous efforts, the terminology problems in HSR remains largely unnoticed and unaccounted for. For instance, in the U.S., the Institute of Medicine (IoM) prioritized different areas of comparative effectiveness research but did not mention this source of systematic bias (Iglehart, 2009). Likewise, there is a substantial degree of variation in the applied valuation methods in health economic studies and guidelines in Europe, but the terminological variability is rarely being mentioned (van Lier et al., 2018; Mayer et al., 2020).	Pg 2

5. Rationale	In the last decade, ontoterminology has been proposed in information technologies (Roche, 2012) and adapted to disambiguation in HSR (Castelpietra et al., 2021). Apart from providing an adequate framework for the analysis of terms in a given field, a classification using a hierarchical taxonomy with a coding system provides a reference framework to code definitions as acceptable, or as ambiguous, vague or confusing (i.e. wrong or mistaken) in a reproducible way (Castelpietra et al., 2021).	Pg 3	
6. Specific aims	The aim of the study was to identify the magnitude of the bias of non-commensurability and terminological unclarity in health and social services research by applying an international classification system for coding human services and adapting it as needed to the newly emerging requirements for health economics research from a societal perspective. A complementary objective of this study is to demonstrate the usability of the ontoterminology approach to disambiguation in complex topics in healthcare research.	Pg 3	
Methods	*What did you do?*		
7. Context	This study was part of the PECUNIA project (ProgrammE in Costing, resource use measurement and outcome valuation for Use in multi-sectoral National and International health economic evaluAtions) conducted from 2018 to 2021. PECUNIA was aimed at developing standardised multi-sectoral, multi-national and multi-person resource use measurement instruments, unit cost valuation templates, reference unit costs, and outcome assessment tools to improve the methodology and comparability of economic evaluations in the European Union with a special focus on mental health (Mayer et al., 2022). The PECUNIA consortium coordinated by the Medical University of Vienna consisted of ten partners for health economics and health systems research, located in six European countries (Austria, Germany, Hungary, Spain, The Netherlands and United Kingdom). (The PECUNIA group, 2018) Due to their high disease burden and economic relevance, three mental disorders (depression, schizophrenia and posttraumatic stress disorder) were chosen as reference disorders to analyse the applicability of the newly developed methods and tools. This study concentrated on the disambiguation of services in the health and social care cluster relevant for mental health and was carried out in parallel to other activities of the project.	Pg 4	
8. Intervention(s)	Not applicable		
9. Study of the Intervention(s)	Not Applicable		

10. Measures	We used an updated glossary of terms based on the Psicost and REFINEMENT glossaries (Montagni et al., 2018) and an international classification of human services, the Description and Evaluation of Services and Directories (DESDE system) (Salvador-Carulla et al., 2013) for the disambiguation of terms. The REFINEMENT glossary provides consensus-based operational definitions of the basic terms relevant for the disambiguation process in health services. DESDE has been used for the comparison of mental health service typologies across countries, (Alonso-Solís et al., 2020) analysis of disambiguation of complex terms in health care, such as psychotherapy, (Castelpietra et al., 2021) and for the content analysis of the national classifications system compared to an international standard (Rosen A, Rock, D, Salvador-Carulla, 2020). Previous research has shown that the DESDE instrument scores high in feasibility, consistency, inter-rater reliability as well as face, content and construct validity (Salvador-Carulla et al., 2013), as well as its applicability in health economics studies. (Romero-Lopez-Alberca et al., 2019)	Pg 6
11. Analysis	Steps I and II yielded a list of 56 key services relevant to mental health based on extensive literature review and selected by international expert panels. Terms and definitions were classified into the following categories: accurate (the term could be classified using one code); ambiguous (the term was labelled with more than one -typically two- code), vague (the term could be coded with a series of codes), confusing (the term was wrong or incomplete according to the reference classification system as it required additional significant interpretation from the experts). Definitions were analysed at different levels. In level 1, the two raters confirmed that the definitions corresponded to services and not to another unit of analysis in HSR such as procedures, interventions or professionals, to ensure the commensurability of the terms included in the listing (Salvador-Carulla et al., 2015). In level 2, the two raters analysed the information on the target population for which the service was intended. This included age, diagnosis group, functioning or other characteristics influencing health status and contact with health services (e.g. homelessness, domestic violence). Finally, level 3 of disambiguation included the definition of the service typology using DESDE taxonomy based on MTC codes.	Pg 7
12. Ethical Considerations	The listing of terms and definitions used in this study did not require ethical approval or consents in the participating countries as data were obtained from a review of the scientific literature and did not include information on individual patients.	

Results	*What did you find?*	
13. Results	LEVEL 1: Commensurability (the unit of analysis actually refers to services) Fifty terms (89%) were considered accurate at level 1 (table 1). Six items (11%) were considered confusing regarding the unit of analysis included in the definition. LEVEL 2: Target Population Clarity The definition of the target population was considered accurate for 29 terms (52%). Three terms (6%) were judged ambiguous as they referred to two not-linked population groups (defined by age and/or diagnosis) at the same time, without explaining the service specificities for each of these groups. LEVEL 3: Service Type Clarity The type of care provided by the service was judged accurate for the 23 terms (41%) that could be translated into a single MTC code. Fifteen terms (27%) were rated as ambiguous because they needed two codes or admitted several code ranges (…) Finally, six (11%) definitions were judged as confusing (…) The six terms classified as confusing in level 1 (different unit of analysis) required expert interpretation in level 3 (coding service type), this was expressed in italics. The comparison between service targeting generic health (ICD-10) and mental health (F0-F9) showed remarkable differences regarding accuracy (22% vs 42%).	Pg 7-8
Discussion	*What does it mean?*	
14. Summary	This study aimed to identify the magnitude of the bias of non-commensurability and terminological unclarity bias in HSR and health economics [2,42] by applying an international classification system to a set of services used by persons experiencing a mental health condition. The results are meant to be used for further processing of service items for the development of the multi-national, multi-sectoral costing tools in the PECUNIA project. The approach was not comparing variation country by country but identifying an international basic listing of services relevant for mental health care. Despite an extensive process of revision prior to disambiguation, 11% of the terms in the final listing were not services according to the definition provided by the DESDE system and the related glossary of terms. In addition, 43 terms were unclear, and could not be used for international comparability. Nearly half of the terms lacked a clear definition of the target population and around 60% had problems in the definition of service types that impeded matching them to an MTC code even though we opted for broad categories within the MTC taxonomy to facilitate matching.	Pg 8-9

15. Interpretation	Our findings indicate that the terminology problem in HSR is extensive. Surprisingly enough, health economic guidelines provide detailed information on the study designs, methods of analysis and interpretation of results but they do not mention this fundamental problem for regional or international comparability and for aggregation purposes (Simon, 2020). Similarly, the problem of service terminology is not even mentioned in international strategies that necessarily require comparison of service delivery such as the WHO Mental Health Gap Action Programme (mhGAP) (World Health Organization, 2008). A gap analysis cannot be conducted without a standardised description of local mental health services to allow aggregate comparisons of care systems across re¬gions and countries. The approach presented in this paper contributes to minimize differentiation between services and to support financial incentives for effective health promotion and disease prevention. Health economic studies on services and their utilization are key for resource use measurement and cost calculation for efficiency (cost-effectiveness), equality (access and utilization) and quality research.	Pg 9
16. Limitations	Firstly, the analysis of the terminology bias in healthcare is extremely challenging and may have problems with corroboration, even when we adapted a previously tested method (Maciejewski et al., 2002) and used a standardized procedure. Secondly, the findings cannot be fully generalized to all areas of healthcare. We selected mental health care as case study due to its highest complexity of care provision stretching across numerous sectors (Salvador-Carulla et al., 2006) including a mixture of health and social care services, the high variation and diversity in service provision, (Johnson and Salvador-Carulla, 1998) and its high ambiguity in key aspects such as diagnosis (Keil et al., 2016) and treatment interventions (Castelpietraa et al., 2017; Castelpietra et al., 2021). Finally, we opted for a broad approach to disambiguation selecting the lower level of granularity in the MTC taxonomy and avoiding a detailed definition of the different subtypes of ambiguity and vagueness (Castelpietra et al., 2021). The disambiguation data are related to one frame of reference (DESDE system) and cannot be generalized to other frames (e.g. Systems of Health Accounts 2.0, or SNOMED). However, the validity and the formal ontology conditions of the classification of services within these other frames have not been tested. Finally, we limited our analysis to English and did not account for the variation of terminology across other languages and contexts. In any case the reference tools ESMS and DESDE have been translated into Finnish, French, German, Italian, Polish, Portuguese, Norwegian and Spanish; and the reference coding system has been used across a wide variety of contexts in over 34 countries (Romero-Rodriguez-Alberca et al 2019).	Pg 9-10

17. Conclusions	Currently, the majority of comparative healthcare studies rely on official service names, without taking into account other key features of every service. Service health research, health economics, care gap analysis, quality and equality research should address terminological variability as a main source of systematic bias, particularly, but not only, in bottom-up international comparative studies. For example, cost-effectiveness and comparative effectiveness research should compare the same units of analysis of service provision, and use a common vocabulary, which is feasible with a coding system such as the one provided by DESDE. This bias is also relevant in equity studies as equal access is a critical component of equity (Raine et al., 2016). Finally, an international glossary of service terms and a classification of service should be incorporated into the WHO International Family of Classifications as related classifications. Likewise, national classifications of services should provide an analysis of their semantic interoperability with international standards.	Pg 10
Other information		
18. Funding	The PECUNIA project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 779292.	Pg 11