Inter-rater reliability of STOPP (Screening Tool of Older Persons' Prescriptions) and START (Screening Tool to Alert doctors to Right Treatment) criteria amongst physicians in six European countries

GALLAGHER, Paul, et al.
Inter-rater reliability of STOPP (Screening Tool of Older Persons’ Prescriptions) and START (Screening Tool to Alert doctors to Right Treatment) criteria amongst physicians in six European countries

SIR—Inappropriate prescribing (IP) encompasses the use of medicines where the risk of an adverse drug event (ADE) outweighs the clinical benefit, particularly when safer or more effective alternatives are available [1, 2]. IP also includes the use of medicines that increase the likelihood of drug–drug and drug–disease interactions, the mis-prescribing of medicines (incorrect dose, frequency and duration) and the under-use of clinically indicated medicines [3–5]. IP is highly prevalent in older people and has been associated with preventable ADEs, hospitalisation, institutionalisation, death and resource wastage [6–12]. With increasing proportions of older people worldwide, quality and safety of prescribing are becoming a global healthcare concern [5, 13].

One way of identifying IP is to use prescribing indicators such as the recently validated STOPP (Screening Tool of Older Persons’ Prescriptions) and START (Screening Tool to Alert doctors to Right Treatment) criteria [14]. STOPP comprises 65 indicators for potentially inappropriate prescribing including drug–drug and drug–disease interactions, therapeutic duplication and drugs that increase the risks of cognitive decline and falls (Appendix 1 in the supplementary data at Age and Ageing online) [14]. START incorporates 22 evidence-based indicators for prescribing omissions in older people (Appendix 2 in the supplementary data at Age and Ageing online) [14]. STOPP/START criteria are organised according to physiological systems for ease of use. Their content validity was established by a Delphi consensus process in which 18 experts in geriatric pharmacotherapy from Ireland and the United Kingdom participated [14]. A recent study showed that 35% of 715 acutely ill older patients requiring hospitalisation were regularly prescribed at least one potentially inappropriate medication according to STOPP criteria and 12% of admissions were directly attributable to avoidable ADEs, hospitalisation, institutionalisation, death and resource wastage [6–12]. With increasing proportions of older people worldwide, quality and safety of prescribing are becoming a global healthcare concern [5, 13].

Prospective randomised controlled trials are needed to test whether routine clinical application of STOPP/START criteria can significantly improve prescribing appropriateness and reduce drug-related morbidity. However, before demonstrating effects on patient outcome, a screening tool must be generalisable and reliable. Inter-rater reliability of STOPP/START criteria was substantial when tested between two researchers (kappa coefficient 0.75 STOPP criteria and 0.68 START criteria) [14]. Further evaluation of reliability between health professionals practicing in different countries is warranted to determine if STOPP/START criteria are generalisable. Accordingly, the aim of this study was to
Research letters

determine the inter-rater reliability of STOPP and START criteria between multiple physicians practicing in different European centres.

Methods

Twenty datasets were selected from a cohort of 200 consecutive patients aged ≥65 years admitted acutely to the general medical services of a university teaching hospital in Ireland who were participating in a concurrent randomised controlled trial designed to evaluate the effect of an intervention on prescribing appropriateness. The 20 datasets were specifically selected to represent patients with complex comorbidities and an appreciable incidence of potentially inappropriate prescriptions according to STOPP/START criteria. Ethical approval was obtained for anonymous use of these datasets in this inter-rater reliability study.

Each dataset was compiled from chart review and patient and/or carer interview at the time of hospitalisation, with supplementary information on diagnoses and prescriptions being sought from the patient’s general practitioner and/or community pharmacist when required. Datasets comprised age, gender, current and past diagnoses, detailed medication history, drug allergies, blood pressure profile, electrocardiograph results, serum biochemistry, glucose, lipid profile, urinalysis and estimated creatinine clearance using the Cockcroft–Gault equation [17], chosen instead of the Modified Diet in Renal Disease formula [18–20] as the latter is not well validated in renal disease.

Table 1. Inter-rater reliability of STOPP and START criteria between 9 hospital physicians on 20 datasets with 181 medications

<table>
<thead>
<tr>
<th>Rater combination</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>Ppos</th>
<th>Pneg</th>
<th>Kappa (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>STOPP criteria</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater 1 + rater 2</td>
<td>1,255</td>
<td>4</td>
<td>0</td>
<td>41</td>
<td>0.99</td>
<td>0.95</td>
<td>0.95 (0.91–0.99)</td>
</tr>
<tr>
<td>Rater 1 + rater 3</td>
<td>1,254</td>
<td>5</td>
<td>3</td>
<td>38</td>
<td>0.99</td>
<td>0.90</td>
<td>0.90 (0.83–0.97)</td>
</tr>
<tr>
<td>Rater 1 + rater 4</td>
<td>1,254</td>
<td>5</td>
<td>3</td>
<td>38</td>
<td>0.99</td>
<td>0.90</td>
<td>0.90 (0.83–0.99)</td>
</tr>
<tr>
<td>Rater 1 + rater 5</td>
<td>1,255</td>
<td>4</td>
<td>0</td>
<td>41</td>
<td>0.99</td>
<td>0.95</td>
<td>0.95 (0.91–0.99)</td>
</tr>
<tr>
<td>Rater 1 + rater 6</td>
<td>1,258</td>
<td>2</td>
<td>1</td>
<td>39</td>
<td>0.99</td>
<td>0.96</td>
<td>0.96 (0.92–1)</td>
</tr>
<tr>
<td>Rater 1 + rater 7</td>
<td>1,257</td>
<td>2</td>
<td>1</td>
<td>40</td>
<td>0.99</td>
<td>0.96</td>
<td>0.96 (0.92–1)</td>
</tr>
<tr>
<td>Rater 1 + rater 8</td>
<td>1,253</td>
<td>3</td>
<td>3</td>
<td>38</td>
<td>0.99</td>
<td>0.89</td>
<td>0.89 (0.82–0.96)</td>
</tr>
<tr>
<td>Rater 1 + rater 9</td>
<td>1,250</td>
<td>0</td>
<td>0</td>
<td>41</td>
<td>0.99</td>
<td>0.90</td>
<td>0.90 (0.83–0.96)</td>
</tr>
<tr>
<td>Median (IQR)</td>
<td>0.99</td>
<td>0.93</td>
<td>0.93 (0.90–0.96)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>START criteria</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater 1 + rater 2</td>
<td>417</td>
<td>3</td>
<td>2</td>
<td>18</td>
<td>0.99</td>
<td>0.88</td>
<td>0.87 (0.76–0.98)</td>
</tr>
<tr>
<td>Rater 1 + rater 3</td>
<td>417</td>
<td>3</td>
<td>1</td>
<td>17</td>
<td>0.99</td>
<td>0.85</td>
<td>0.84 (0.72–0.97)</td>
</tr>
<tr>
<td>Rater 1 + rater 4</td>
<td>418</td>
<td>2</td>
<td>1</td>
<td>19</td>
<td>0.99</td>
<td>0.92</td>
<td>0.92 (0.84–1)</td>
</tr>
<tr>
<td>Rater 1 + rater 5</td>
<td>417</td>
<td>3</td>
<td>0</td>
<td>20</td>
<td>0.99</td>
<td>0.93</td>
<td>0.93 (0.84–1)</td>
</tr>
<tr>
<td>Rater 1 + rater 6</td>
<td>416</td>
<td>4</td>
<td>1</td>
<td>17</td>
<td>0.99</td>
<td>0.83</td>
<td>0.82 (0.69–0.95)</td>
</tr>
<tr>
<td>Rater 1 + rater 7</td>
<td>415</td>
<td>5</td>
<td>5</td>
<td>15</td>
<td>0.98</td>
<td>0.75</td>
<td>0.74 (0.58–0.89)</td>
</tr>
<tr>
<td>Rater 1 + rater 8</td>
<td>413</td>
<td>7</td>
<td>1</td>
<td>19</td>
<td>0.99</td>
<td>0.83</td>
<td>0.82 (0.69–0.94)</td>
</tr>
<tr>
<td>Rater 1 + rater 9</td>
<td>414</td>
<td>6</td>
<td>0</td>
<td>20</td>
<td>0.99</td>
<td>0.87</td>
<td>0.86 (0.75–0.97)</td>
</tr>
<tr>
<td>Median (IQR)</td>
<td>0.99</td>
<td>0.86</td>
<td>0.85 (0.82–0.91)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A, both raters agreed criterion not fulfilled; B, rater 1 scored criterion not fulfilled and rater 2 scored criterion as being fulfilled; C, rater 1 scored criterion as fulfilled and rater 2 scored criterion as not fulfilled; D, both raters scored criterion as being fulfilled; ppos, proportion of positive agreement; pneg, proportion of negative agreement; CI, confidence interval; IQR, interquartile range.

The 20 patients’ mean (±SD) age was 76.8 (±5.4) years, and 50% were female. The total number of prescribed medications was 181, median 9, interquartile range 6–11. Two of the STOPP/START originators (PG, DO’M) from the coordinating centre in Ireland discussed the 20 datasets in detail and reached complete agreement in terms of prescribing appropriateness according to STOPP/START criteria. This combined level of agreement (labelled ‘rater 1’) was set as the standard against which other physicians’ ratings would be compared. Nineteen datasets had at least one potentially inappropriate prescription according to STOPP criteria (median 2; range 0–5). Eleven datasets had at least one prescribing omission according to START criteria (median 1; range 0–4).

Eight hospital physicians (labelled ‘raters 2–9’) with no prior experience of using STOPP/START criteria participated in the study. These physicians were based in teaching hospital geriatric medicine units in Belgium (n = 1), Czech Republic (n = 2), Italy (n = 3), Spain (n = 1) and Switzerland (n = 1). STOPP/START criteria were translated from English into Czech, French, Italian and Spanish (available on request from the corresponding author) to facilitate local application of the criteria. The 20 datasets were also translated. A teleconference facilitated by the coordinating centre in Ireland afforded all raters the opportunity to resolve any difficulties with translation or interpretation of the criteria before application. Two criteria were clarified: (i) therapeutic and maintenance doses of proton pump inhibitors (STOPP C4) and (ii) inclusion criteria for drug-class duplication (STOPP J1). All physicians then independently assessed the incidence of 65 STOPP and 22 START criteria in each of the 20 datasets and were invited to give written comments if necessary.

Responses of raters 2–9 were cross-tabulated with those of rater 1. Inter-group responses between physicians from Italy and the Czech Republic were also evaluated, to determine reliability independent of the STOPP/START originators. Data were analysed using SPSS 15.0 for Windows (SPSS Inc., Chicago, IL, USA). Inter-rater reliability analysis using the kappa statistic (chance corrected measure of agreement) was performed to determine consistency between raters [23]. The kappa statistic was interpreted as poor if ≤0.2, fair if 0.21–0.40, moderate if 0.51–0.6, substantial if 0.61–0.8 and good if 0.81–1.0 [23]. Proportions of positive and negative agreements were calculated [24]. Written comments were analysed to determine whether disagreements were due to misjudgement of appropriateness, difficulty with criteria or case interpretation.

Results

Columns A, B, C and D in Tables 1 and 2 indicate the status of agreement between raters. For example, raters 1 and
2 agreed that STOPP criteria were not identified in 1,255 instances (column A). In four instances, rater 1 did not identify a STOPP criterion, but rater 2 did (column B). There were no instances where rater 2 identified a STOPP criterion that rater 1 did not (column C). In 41 instances, both raters 1 and 2 identified a STOPP criterion (column D). The median (IQR) kappa coefficient between raters was 0.93 (0.90–0.96) for STOPP criteria and 0.85 (0.82–0.91) for START criteria.

Disagreement occurred with six STOPP criteria. Two raters disagreed that aspirin was potentially inappropriate without coronary, cerebral or peripheral arterial occlusive symptoms or risk factors (STOPP A13). Three raters misclassified second-generation antihistamines as first-generation antihistamines (STOPP B13 and H3). Loop diuretics were prescribed without indication in four cases; however, six raters inferred an indication of ankle oedema or hypertension, thereby resulting in disagreement over STOPP criteria A2 and A3. A typical neuroleptic was deemed inappropriate by one rater (STOPP B8), although it was prescribed for behavioural and psychological symptoms of dementia.

One rater judged it appropriate to omit a statin in one patient with coronary artery disease (START A5) and severe heart failure, as it was ‘unlikely to alter outcome’. Four raters did not apply this criterion, as they ‘did not have sufficient clinical information to measure life expectancy’. Two raters wanted more clinical details before applying the START criterion C2 (anti-depressant with depressive symptoms lasting >3 months) and START criterion E1 (disease modifying anti-rheumatic drug with moderate–severe rheumatoid arthritis). Five raters identified the START criterion D2 (fibre supplement for chronic symptomatic diverticular disease with constipation) in two patients with chronic constipation, though diverticular disease was not specifically documented. Medications without indication were sometimes interpreted as indicating an underlying disease for which another medication was omitted e.g. one rater interpreted nitrates as indicating underlying angina and recommended a beta-blocker (START A8).

### Discussion

Inter-rater reliability of STOPP/START criteria is good when tested between multiple physicians across six European centres. The more comprehensive clinical and medication details used in this study are likely to account for the higher level of inter-rater reliability than reported previously [14]. Disagreements in a minority of instances reflected the fact that details on functional status and life expectancy were not provided with the cases, though clearly, these are important considerations when applying STOPP/START criteria. Differences in prescribing guidelines and formularies between countries could cause disagreement with STOPP/START criteria, but no physician in this study reported this.

A high level of familiarity is required to efficiently apply ‘pencil and paper’ versions of STOPP/START criteria in clinical practice. This reality emphasises the need for computerised automation of STOPP/START whereby linkage of specific diseases or symptoms with specific medicines would lead to rapid identification of potentially inappropriate prescriptions according to STOPP criteria and omission of indicated drugs according to START criteria. The good inter-rater reliability demonstrated by this study indicates that results of studies on the prevalence of potentially inappropriate prescribing identified by STOPP/START criteria are comparable between countries.

### Key points
- Inter-rater reliability of STOPP and START criteria is good when tested between multiple physicians practicing independently in different European centres.
- STOPP and START criteria are generalisable across different European countries and languages.

### Conflicts of interest

No conflicts of interest.

### Funding

Health Research Board of Ireland (Clinical Research Training Fellowship CRT/2006/029) and the Czech Ministry of Health Internal Grant Agency (Grant Number 10029-4).
Supplementary data

Supplementary data are available online at Age and Ageing online.

PAUL GALLAGHER1,4, JEAN-PIERRE BAYEYENS2, EVA TOPINKOVA3, PAVALA MADLOVA3, ANTONIO CHERUBINI4, BEATRICE GASPERINI5, ALFONSO CRUZ-JENTOFF5, BEATRIZ MONTERO5, PIERRE OLIVIER LANG6, JEAN-PIERRE MICHEL6, DENIS O’MAHONY1

1 Department of Geriatric Medicine, Cork University Hospital, Wilton, Cork, Ireland
2 Department of Geriatric Medicine, AZ Damiaan Oostende, Oostende, Belgium
3 Department of Geriatric Medicine, First Faculty of Medicine, Charles University, Prague, Czech Republic
4 Department of Clinical and Experimental Medicine, Institute of Gerontology and Geriatrics, University of Perugia Medical School, Perugia, Italy
5 Servicio de Geriatria, Hospital Universitario Ramón y Cajal, Madrid, Spain
6 Rehabilitation and Geriatric Department, Geneva Medical School and University Hospitals, Geneva, Switzerland

*To whom correspondence should be addressed

References


Vitamin D supplementation and type 2 diabetes: a study of a randomised placebo-controlled trial in older people (RECORD trial, ISRCTN 51647438)

SIR—Studies in animals show that vitamin D deficiency is associated with impaired insulin sensitivity, and that insulin