Computer-generated STOPP/START recommendations for hospitalised older adults: evaluation of the relationship between clinical relevance and rate of implementation in the SENATOR trial

Abstract

Background

Findings from a recent qualitative study indicate that the perceived clinical relevance of computer-generated STOPP/START recommendations was a key factor affecting their implementation by physician prescribers caring for hospitalised older adults in the SENATOR trial.

Aim

To systematically evaluate the clinical relevance of these recommendations and to establish if clinical relevance significantly affected the implementation rate.

Methods

A pharmacist-physician pair retrospectively reviewed the case records for all SENATOR trial intervention patients at Cork University Hospital, and assigned a degree of clinical relevance for each STOPP/START recommendation based on a previously validated six-point scale. The chi-square test was used to quantify the differences in prescriber implementation rates between recommendations of varying clinical relevance, with statistical significance set at $p < 0.05$.

Results

In 204 intervention patients, the SENATOR software produced 925 STOPP/START recommendations. Nearly three quarters of recommendations were judged to be clinically relevant (73.6%); however, nearly half of these were deemed of '*possibly low relevance*' (320/681; 47%). Recommendations deemed of higher clinical relevance were significantly more likely to be implemented than those of lower clinical relevance ($p < 0.05$).

Conclusions

A large proportion (61%) of the computer-generated STOPP/START recommendations provided were either of potential '*adverse significance*', of '*no clinical relevance*', or of '*possibly low relevance*'. The adjudicated clinical relevance of computer-generated medication recommendations significantly affects their implementation. Meticulous software refinement is required for future interventions of this type to increase the proportion of recommendations that are of high clinical relevance. This should facilitate their implementation, resulting in prescribing optimisation and improved clinical outcomes for multimorbid older adults.

INTRODUCTION

Potentially inappropriate prescribing (PIP) remains to be highly prevalent in hospitalised older adults [1, 2]. Computerised interventions have been shown to reduce PIP in this patient cohort, but their benefit in routinely improving patient outcomes has not yet been established [3, 4]. The intervention in a recent multi-centre randomised controlled trial (RCT), as part of the SENATOR project (https://www.senator-project.eu/), involved the provision of computer-generated pharmacological and non-pharmacological recommendations to physician prescribers caring for hospitalised older adults, with the primary aim of reducing in-hospital adverse drug reactions (ADRs). The pharmacological recommendations in the SENATOR intervention were based on the STOPP/START criteria (version 2) [5], drug-drug interactions, and drug-disease interactions. Interim data analysis from the trial showed that prescriber implementation rates of the STOPP/START recommendations were lower than expected across the six trial sites. O'Connor *et al*. had previously shown in a single-centre RCT that high prescriber implementation rates of STOPP/START recommendations can significantly reduce in-hospital ADRs [6]. Thus, a qualitative interview study was run alongside the RCT to explore the reasons behind the low implementation rates observed [7]. Interviewees perceived that the clinical relevance of the recommendations was one of the key factors affecting their implementation, suggesting the SENATOR software was producing a high proportion of recommendations that were of low or doubtful clinical relevance for individual patients. However, rather than simply accepting these qualitative findings at face value, it would be of great significance to quantitatively corroborate a clear association between the relevance of recommendations and their rate of implementation. Therefore, the aim of this study was to systematically evaluate the clinical relevance of the computer-generated STOPP/START recommendations in the SENATOR trial and examine if the relevance of recommendations was associated with their rate of implementation.

METHODS

Context and Study setting

The SENATOR RCT was conducted in six large acute teaching hospitals in six European countries. All patients recruited were multimorbid older adults (≥65 years) who consented to their enrolment in the trial within 60 hours of hospital admission, who were prescribed medication for ≥3 active chronic medical disorders, and who had an expected length of hospital stay >48 hours. More details on patient eligibility criteria and other pertinent trial information are published elsewhere [8, 9].

This study evaluating the clinical relevance of SENATOR's STOPP/START recommendations was conducted in the RCT's lead recruitment site only, Cork University Hospital (CUH) - an 810-bed tertiary referral centre in southern Ireland. All patients who were randomised to the intervention arm at this site were included in the present study. In CUH, the SENATOR software generated a paper-based report detailing the STOPP/START recommendations, which was provided in each intervention patient's paper-based clinical record, and was also sent via email to the consultant with responsibility for clinical care of the patient. Of the 114 STOPP/START criteria (version 2), recommendations based on 3 criteria were excluded from our analysis: STOPP A1, START I1, and START I2, with reasons for exclusion provided in the Supplementary Material.

Data collection

A pharmacist and physician independently and retrospectively reviewed all CUH intervention arm patients' medical records, drug chart, laboratory test results, and STOPP/START recommendations. Through consensus agreement, the pharmacist-physician pair then assigned a degree of clinical relevance for each STOPP/START recommendation based on a previously validated six-point scale with the following categories: 0: '*adverse significance*', 1: '*no clinical relevance*', 2: '*possibly low relevance*', 3: '*possibly important relevance*', 4: '*possibly very important relevance*', and 5: '*possibly life-saving*' [10]. The pharmacist (KD) and physician (DC) were very familiar with the STOPP/START criteria and SENATOR's computerised algorithms, and, at the time of the reviews,

had three years and ten years post-qualification experience respectively in optimising the pharmacotherapy of hospitalised older adults.

Inter-rater reliability (IRR) was determined among a sample of three pharmacists and three physicians (one consultant geriatrician, and two specialist registrars - i.e. senior residents - in geriatric medicine) in applying the scale to independently assign a degree of clinical relevance to STOPP/START recommendations from twenty randomly-selected intervention cases. The study design for this IRR assessment is provided in the Supplementary Material.

Data analysis

Statistical analysis was performed using SPSS® Version 22 and Microsoft® Excel. Data on prescriber implementation were extracted from the RCT's electronic case report form, whereby implementation was defined as the prescriber discontinuing or initiating a medication in accordance with the recommendation at any point prior to hospital discharge. The percentage prescriber implementation rates were calculated for recommendations at each degree of clinical relevance. The chi-square test was used to determine if there were any significant differences between i) the proportion of recommendations and ii) the prescriber implementation rates of recommendations at varying degrees of clinical relevance, with differences considered statistically significant at $p < 0.05$.

In the assessment of IRR, the Fleiss kappa statistic was used to determine the agreement between all raters and across the subgroups of raters (i.e. pharmacists and physicians). Cohen's kappa statistic was used to determine the level of agreement between the individual raters. The kappa statistic was interpreted according to the following ranges: slight if 0.01-0.2, fair if 0.21-0.4, moderate if 0.41-0.6, substantial if 0.61-0.8, and almost perfect if 0.81-0.99 [11].

RESULTS

In CUH, there were 204 SENATOR intervention patients (51% male), with a mean age of 77.4 years (standard deviation [SD] 6.91; Range 65-92). In total, the SENATOR software generated 925 STOPP/START recommendations (mean 4.5/patient; SD 2.9; range 0-17), i.e. 563 STOPP recommendations (mean 2.8/patient; SD 2.3; range 0-13), and 362 START recommendations (mean 1.8/patient; SD 1.5; range 0-7).

Clinical Relevance Evaluation

Almost three quarters (73.6%) of recommendations were deemed to be clinically relevant i.e. assigned to categories 2, 3, or 4 – '*possibly low relevance*', '*possibly important relevance*', or '*possibly very important*' relevance (**Table 1**). The remaining 26.4% of recommendations were either category 1, i.e. of '*no clinical relevance*' (21.5%), or category 0, i.e. of possible '*adverse significance*' to the patient if implemented (4.9%). No recommendations were judged to be '*possibly life-saving*'.

When comparing the clinical relevance of STOPP and START recommendations in **Table 2**, there was a statistically significantly greater proportion of START recommendations i) of possible '*adverse significance*' (7.2% versus 3.4%; $p < 0.05$), and ii) of '*possibly very important relevance*' (12.2% versus 5.7%; $p < 0.05$). Conversely, there was a statistically significantly greater proportion of STOPP recommendations of '*possibly low relevance*' (37.7% versus 29.8%; $p < 0.05$).

Prescriber Implementation Rates

Data on prescriber implementation were available for 884/925 (95.6%) recommendations; reasons for unavailable data are provided in the Supplementary Material. **Table 1** illustrates the prescriber implementation rates for the recommendations according to each assigned category of clinical relevance. As the clinical relevance of recommendations increases, so too does the implementation rate, with statistically significant differences in implementation rates between recommendations of all categories identified ($p < 0.05$), the only exception being between recommendations of potential '*adverse significance*' and recommendations of '*no clinical relevance*' (6.7% versus 11.7%; $p = 0.33$).

Inter-rater Reliability Results

When assessing IRR in choosing the same degree of clinical relevance for recommendations, the Fleiss kappa coefficient was found to be fair (0.24). Kappa was higher among pharmacists (0.27) than among physicians (0.17). The mean Cohen's kappa coefficient between individual raters was also found to be fair (kappa = 0.24).

DISCUSSION

This is the first study to evaluate the clinical relevance of computer-generated STOPP/START recommendations. The key finding is that increasing clinical relevance of recommendations associated with significantly higher implementation rates by prescribers. Our results from this sample of acutely ill hospitalised multimorbid older patients show that nearly three quarters of STOPP/START recommendations were deemed to be clinically relevant (73.6%), whilst approximately one quarter of the recommendations were of '*no clinical relevance*' or of potential '*adverse significance*' (26.4%). Although most STOPP/START recommendations were deemed 'clinically relevant', we acknowledge that nearly half of the clinically relevant recommendations were deemed to have '*possibly low relevance*', i.e. category 2 on the six-point clinical relevance scale (320/681; 47%). Whilst these recommendations were correctly triggered by the SENATOR software, they may have been addressing issues that were of minor significance at the time of hospital admission, when the focus may have been on the patient's acute illness. For example, nearly half of all benzodiazepine-related recommendations were judged to be of possibly low relevance. Although it is well-known that this drug class is a common contributing factor to ADRs (principally falls) in older adults [12], deprescribing benzodiazepines may not have been a priority at the time the recommendations were provided. Recommendations like these may have been more clinically relevant later in the admission (such as pre-discharge), or in another setting (such as primary care or ambulatory care), where the patient may have been more stable, and it may have been easier to implement medication changes. Thus, the care setting and timing of the intervention must be key considerations for future studies.

The proportion of clinically relevant computer-generated recommendations can vary widely depending on the healthcare setting and the medications targeted [13-16]. However, there are few studies in the literature that have evaluated the clinical relevance of computer-generated recommendations concerning medication appropriateness in hospitalised older adults. One research group has previously reported findings similar to ours in a pilot study, with 74.5% of computerised alerts deemed clinically relevant [14]. However, when medication alerts in one of their subsequent

studies were based on a broader set of Beers criteria [17], it was found that only 30% of the alerts were clinically relevant in the intervention group [15]. In contrast, we have shown in the present study that the STOPP/START version 2 recommendations, i.e. another broad set of criteria, had a substantially higher proportion of clinically relevant recommendations.

Many of these previous studies have simply judged the computer-generated recommendations in a dichotomous manner - clinically relevant or not clinically relevant [14-16]. However, in the present study, we considered it important to transcend this and qualify clinical relevance in a more nuanced fashion, i.e. to assess the *degree* of clinical relevance, by applying a defined scale. Beaudoin *et al*. used a five-point Likert scale, ranging from 1 (not relevant) to 5 (very relevant), to evaluate the clinical relevance of computerised rule-based alerts concerning antimicrobials [18]; however, a limitation to this Likert scale is that it does not explicitly consider the possibility that the recommendations may be potentially inappropriate, and thereby have the potential to cause harm to the patient. The scale chosen for use in our study had been previously employed to assess the clinical relevance of pharmacist recommendations in a Belgian hospital, which found low agreement between evaluators (range of kappa values: 0.15-0.25) [10]. Similar agreement was found between the raters in our study (kappa = 0.24). Furthermore, Bech *et al*. found only slight agreement between raters when assessing the clinical relevance of drug-related problems among older patients using a five-point scale [19]. This lack of agreement among healthcare professionals in evaluating clinical relevance highlights the complexities associated with selection of appropriate pharmacotherapy in older adults [20].

Our study is important in that we did not merely indicate whether the recommendations were relevant or not, but rather we also qualified their degree of clinical relevance. If we know that recommendations pertaining to certain criteria or particular drug classes are more likely to be clinically relevant, then we can prioritise these recommendations in future interventions. For example, in a hospitalised patient presenting with falls, the software should prioritise recommendations relating to deprescribing of benzodiazepines over those relating to proton pump inhibitors (PPIs). Provision of the most clinically relevant recommendations only, or ensuring that these recommendations appear as

priorities, should help reduce the phenomenon of '*alert fatigue*' [21]. However, designing the software to take account of competing influences on clinical decision making, and ranking the recommendations in order of priority is a significant technical challenge [22]. This study has provided evidence on which recommendations may be more clinically relevant than others, and thus may inform ranking systems within future computerised algorithms.

The present study corroborates the findings from the contemporaneous SENATOR qualitative study, which indicated that the clinical relevance of the computer-generated STOPP/START recommendations was a key influence on their implementation by prescribers [7]. Our results show a clear association between these two factors – recommendations of higher clinical relevance had a greater probability of being implemented by prescribers. Previous research has shown that computer-generated recommendations that were inappropriate or erroneously triggered were unlikely to be adopted by physicians or were overridden within the software programme [23]. However, the potential risk remains that some users may blindly follow inappropriate recommendations; this increases the risk of error and possible patient harm [24, 25]. Therefore, such computer-generated recommendations should complement good clinical judgment, not replace it.

Our results indicate that a significantly greater proportion of START recommendations were either of possibly very high clinical relevance or of possible adverse significance in comparison to STOPP recommendations. Thus, certain START recommendations had the potential to be of great benefit in some patients, but could have caused serious harm if implemented in other patients. This indicates a lack of specificity in the computerised algorithms, resulting in the identification of more supposed instances of PIP than actual instances [26]. However, this lack of specificity is not purely an algorithm issue – it could also have originated from the criteria themselves. For example, previous research has highlighted that some of the STOPP/START criteria contain broad definitions, e.g. START A3 criterion (version 2) refers to "*...a documented history of coronary, cerebral or peripheral vascular disease*". Whilst this phrasing allows the criteria to be applicable to a large proportion of older adults, broad definitions such as this are more susceptible to clinician interpretation [27]. Thus, some criteria may not be as explicit as they should be for the purposes of designing computerised algorithms, and

previous research groups have outlined some of the complexities encountered in this process [27, 28]. Further iterations of STOPP/START criteria will likely need to be much more specific, especially if the intention is to incorporate them into computerised algorithms, which should facilitate the production of more clinically relevant recommendations that are tailored to individual patients.

However, simply producing clinically relevant STOPP/START recommendations does not guarantee their uptake; the medium through which the recommendations are delivered to prescribers also significantly affects their implementation [29]. In the present study, we have found that even the recommendations deemed to have '*possibly very important relevance*' were not implemented 30% of the time. One reason for this may have been due to the production of a high proportion of recommendations that were of potential adverse significance, not clinically relevant, or of low clinical relevance (61% of recommendations). These may, unwittingly, have undermined the trustworthiness of the SENATOR advice reports, and resulted in decreased engagement by clinicians with the most important recommendations [7]. Increasing the proportion of recommendations of higher clinical relevance will be essential in minimising user fatigue with future computerised interventions, and enhancing the likelihood of clinically important recommendations being implemented.

As with many studies of this type, they are limited by their retrospective design, and the subjectivity of raters must be considered as a potential source of bias. This is the first study that the authors are aware of which determines the IRR among healthcare professionals in evaluating the clinical relevance of computer-generated STOPP/START recommendations. However, failure to achieve high IRR may have been due to the scale used; it has been previously shown that rating scales with poor IRR are likely to result in low estimates of IRR in subsequent studies [30]. Furthermore, a scale with fewer categories or more specific categories would allow less room for discrepancy between raters, and should produce a higher IRR kappa value. Agreement may have been affected by raters simply interpreting the scale differently [31].

CONCLUSIONS

This quantifiably substantiates the findings from a recent qualitative study, which suggested that the clinical relevance of the STOPP/START recommendations in the SENATOR intervention was one of the key influences affecting their implementation. The present study shows that a large proportion (61%) of the STOPP/START recommendations provided were either of potentially adverse significance, irrelevant, or of low clinical relevance for the individual patients at the point of hospital admission. Recommendations of higher clinical relevance had significantly enhanced prescriber implementation rates. This study has also indicated the types of recommendations, based on the different physiological systems and drug classes, which are more likely to be of high clinical relevance; these findings may aid in the ranking of medication recommendations in future research. Future computerised interventions aimed at medication optimisation in multimorbid older adults must be meticulously designed to provide tailored advice specific to individual patients' pharmacotherapy, thereby minimising the number of recommendations that are irrelevant or of low clinical relevance. Achieving greater proportions of recommendations that are of high clinical relevance should facilitate implementation by prescribers, resulting in the resolution of PIP issues and improved clinical outcomes for older adults.

REFERENCES

1. Gallagher P, Lang PO, Cherubini A, et al. Prevalence of potentially inappropriate prescribing in an acutely ill population of older patients admitted to six European hospitals. Eur J Clin Pharmacol 2011; 67: 1175–88.
2. O'Connor MN, Gallagher P, O'Mahony D. Inappropriate prescribing: criteria, detection and prevention. Drugs Aging 2012; 29: 437–52.
3. Dalton K, O'Brien G, O'Mahony D, Byrne S. Computerised interventions designed to reduce potentially inappropriate prescribing in hospitalised older adults: a systematic review and meta-analysis. Age Ageing 2018; 47: 670–8.
4. Clyne B, Bradley MC, Hughes C, Fahey T, Lapane KL. Electronic prescribing and other forms of technology to reduce inappropriate medication use and polypharmacy in older people: a review of current evidence. Clin Geriatr Med 2012; 28: 301–22.
5. O'Mahony D, O'Sullivan D, Byrne S, O'Connor MN, Ryan C, Gallagher P. STOPP/START criteria for potentially inappropriate prescribing in older people: version 2. Age Ageing 2015; 44: 213–8.
6. O'Connor MN, O'Sullivan D, Gallagher PF, Eustace J, Byrne S, O'Mahony D. Prevention of hospital-acquired adverse drug reactions in older people using screening tool of older persons' prescriptions and screening tool to alert to right treatment criteria: a cluster randomized controlled trial. J Am Geriatr Soc 2016; 64: 1558–66.
7. Dalton K, O'Mahony D, Cullinan S, Byrne S. Factors affecting prescriber implementation of computer-generated medication recommendations in the SENATOR trial—a qualitative study. Int J Pharm Pract 2019; 27: 25–6.
8. Lavan AH, O'Mahony D, Gallagher P, et al. The effect of SENATOR (software ENgine for the assessment and optimisation of drug and non-drug therapy in older peRsons) on incident adverse drug reactions (ADRs) in an older hospital cohort—trial protocol. BMC Geriatr 2019; 19: 40.

9. O'Mahony D, Gudmunsson A, Soiza R, Petrovic M, et al. Prevention of adverse drug reactions in hospitalized older patients with multi-morbidity and polypharmacy: the SENATOR∗ randomized controlled clinical trial. Age Ageing 2020. doi: 10.1093/ageing/afaa072. In press.
10. Somers A, Robays H, De Paepe P, Van Maele G, Perehudoff K, Petrovic M. Evaluation of clinical pharmacist recommendations in the geriatric ward of a Belgian university hospital. Clin Interv Aging 2013; 8: 703–9.
11. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med 2005; 37: 360–3.
12. Markota M, Rummans TA, Bostwick JM, Lapid MI. Benzodiazepine use in older adults: dangers, management, and alternative therapies. Mayo Clin Proc 2016; 91: 1632–9.
13. Weingart SN, Toth M, Sands DZ, Aronson MD, Davis RB, Phillips RS. Physicians' decisions to override computerized drug alerts in primary care. Arch Intern Med 2003; 163: 2625–31.
14. Arvisais K, Bergeron-Wolff S, Bouffard C, et al. A pharmacist-physician intervention model using a computerized alert system to reduce high-risk medication use in elderly inpatients. Drugs Aging 2015; 32: 663–70.
15. Cossette B, Ethier JF, Joly-Mischlich T, et al. Reduction in targeted potentially inappropriate medication use in elderly inpatients: a pragmatic randomized controlled trial. Eur J Clin Pharmacol 2017; 73: 1237–45.
16. Garcia-Caballero TM, Lojo J, Menendez C, Fernandez-Alvarez R, Mateos R, Garcia-Caballero A. Polimedication: applicability of a computer tool to reduce polypharmacy in nursing homes. Int Psychogeriatr 2018; 30: 1001–8.
17. American Geriatrics Society 2015. Updated beers criteria for potentially inappropriate medication use in older adults. J Am Geriatr Soc 2015; 63: 2227–46.
18. Beaudoin M, Kabanza F, Nault V, Valiquette L. Evaluation of a machine learning capability for a clinical decision support system to enhance antimicrobial stewardship programs. Artif Intell Med 2016; 68: 29–36.
19. Bech CF, Frederiksen T, Villesen CT, et al. Healthcare professionals' agreement on clinical relevance of drug-related problems among elderly patients. Int J Clin Pharm 2018; 40: 119–25.

20. Spinewine A, Schmader KE, Barber N, et al. Appropriate prescribing in elderly people: how well can it be measured and optimised? Lancet 2007; 370: 173–84.
21. Cash JJ. Alert fatigue. Am J Health Syst Pharm 2009; 66: 2098–101.
22. Sittig DF, Wright A, Osheroff JA, et al. Grand challenges in clinical decision support. J Biomed Inform 2008; 41: 387–92.
23. Tsai CY, Wang SH, Hsu MH, Li YC. Do false positive alerts in naive clinical decision support system lead to false adoption by physicians? A randomized controlled trial. Comput Methods Prog Biomed 2016; 132: 83–91.
24. Campbell EM, Sittig DF, Guappone KP, Dykstra RH, Ash JS. Overdependence on technology: an unintended adverse consequence of computerized provider order entry. AMIA Annu Symp Proc/AMIA Symp 2007; 94–8.
25. Coiera E, Ash J, Berg M. The unintended consequences of health information technology revisited. Yearb Med Inform 2016; 163–9.
26. Coleman JJ, van der Sijs H, Haefeli WE, et al. On the alert: future priorities for alerts in clinical decision support for computerized physician order entry identified from a European workshop. BMC Med Inform Decis Mak 2013; 13: 111.
27. Huibers CJA, Sallevelt B, de Groot DA, et al. Conversion of STOPP/START version 2 into coded algorithms for software implementation: a multidisciplinary consensus procedure. Int J Med Inform 2019; 125: 110–7.
28. Anrys P, Boland B, Degryse JM, et al. STOPP/START version 2-development of software applications: easier said than done? Age Ageing 2016; 45: 589–92.
29. Dalton K, O'Mahony D, O'Sullivan D, O'Connor MN, Byrne S. Prescriber implementation of STOPP/START recommendations for hospitalised older adults: a comparison of a pharmacist approach and a physician approach. Drugs Aging 2019; 36: 279–88.
30. Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial. Tutor Quant Methods Psychol 2012; 8: 23–34.
31. Bosma L, Jansman FG, Franken AM, Harting JW, Van den Bemt PM. Evaluation of pharmacist clinical interventions in a Dutch hospital setting. Pharm World Sci 2008; 30: 31–8.

Table 1: Prescriber implementation rates of recommendations categorised according to their degree of clinical relevance.

Degree of clinical relevance	0 - Adverse significance	1 - No clinical relevance	2 - Possibly low relevance	3 - Possibly important relevance	4 - Possibly very important relevance	5 - Possibly life-saving
Number of recommendations (% total)	45 (4.9%)	199 (21.5%)	320 (34.6%)	285 (30.8%)	76 (8.2%)	0 (0%)
Prescriber Implementation* (% implemented)	3/45 (6.7%)	20/171 (11.7%)	60/319 (18.8%)	119/273 (43.6%)	53/76 (69.7%)	-
Difference in implementation†	0 versus 2,3,4: $p < 0.05$‡	1 versus 2,3,4: $p < 0.05$‡	2 versus 0,1,3,4: $p < 0.05$	3 versus 0,1,2,4: $p < 0.05$	4 versus 0,1,2,3: $p < 0.05$	-
Most common type of STOPP/START recommendation within the category	START A3: *Antiplatelet therapy (aspirin or clopidogrel or prasugrel or ticagrelor) with a documented history of coronary, cerebral or peripheral vascular disease.* (n = 22; 48.9%)	STOPP J3: *Beta-blockers in diabetes mellitus with frequent hypoglycaemic episodes (risk of suppressing hypoglycaemic symptoms).* (n = 40; 20.1%)	STOPP A2: *Any drug prescribed beyond the recommended duration, where treatment duration is well defined.* (n = 104; 32.5%)	START A6: *Angiotensin Converting Enzyme (ACE) inhibitor with systolic heart failure and/or documented coronary artery disease.* (n = 24; 8.4%)	START A1: *Vitamin K antagonists or direct thrombin inhibitors or factor Xa inhibitors in the presence of chronic atrial fibrillation.* (n = 13; 17.1%)	-
Possible reason for assigning this degree of relevance:	Recommendation to start an antiplatelet but patient already prescribed an anticoagulant – increased risk of bleeding.	Recommendation triggered for all diabetic patients prescribed beta-blockers. Patient not presenting with frequent hypoglycaemic episodes – therefore, not relevant.	Recommendation to stop long-term high-dose PPI. Not of high clinical relevance in a patient who may have a more serious acute issue to be dealt with.	Recommendation may be possibly important in reducing the risk of cardiovascular events in those with coronary artery disease.	Recommendation to start an anticoagulant in a patient with atrial fibrillation may be possibly very important in the prevention of future stroke.	

* Includes all recommendations with data available regarding prescriber implementation

† Difference in prescriber implementation rates between categories of clinical relevance; statistically significant difference observed where $p < 0.05$

‡ No statistically significant difference observed between the implementation rates of recommendations of potential '*adverse significance*' (category 0) and recommendations of '*no clinical relevance*' (category 1).

Table 2: Comparison between clinical relevance of STOPP and START recommendations

Category	0 - Adverse significance	1 - No clinical relevance	2 - Possibly low relevance	3 - Possibly important relevance	4 - Possibly very important relevance	Total
STOPP Recommendations (% Total STOPP)	19 (3.4%)	129 (22.9%)	212 (37.7%)	171 (30.4%)	32 (5.7%)	563
START Recommendations (% Total START)	26 (7.2%)	70 (19.3%)	108 (29.8%)	114 (31.5%)	44 (12.2%)	362
STOPP/START Recommendations (% Total STOPP/START)	45 (4.9 %)	199 (21.5%)	320 (34.6%)	285 (30.8%)	76 (8.2%)	925
Difference between proportion of START and STOPP at different categories of relevance*	$p = 0.0086$	$p = 0.1964$	$p = 0.0147$	$p = 0.7191$	$p = 0.0005$	-

* Statistically significant difference where $p < 0.05$

Supplementary Appendix 1 – Reasons for criterion exclusion in evaluation of clinical relevance

As part of the intervention, recommendations START I1 and START I2 (suggesting to ensure patients received influenza and pneumococcal vaccinations) appeared on all reports. These recommendations were excluded from the assessment of clinical relevance and implementation as it was not documented if all these patients had been vaccinated or not.

STOPP A1 (a recommendation suggesting to stop "*Any drug prescribed without an evidence-based clinical indication*") was also written on the report but this too could not be assessed for clinical relevance or implementation, as the indication was not clear for all medications.

Therefore, of the 114 STOPP/START criteria (version 2), recommendations based on 3 criteria were excluded from our analysis.

However, it should also be noted that two slight software modifications were made a few weeks into patient recruitment:

- START A2 ("*Aspirin [75 mg – 160 mg once daily] in the presence of chronic atrial fibrillation, where Vitamin K antagonists or direct thrombin inhibitors or factor Xa inhibitors are contraindicated*") was initially triggering as a stand-alone recommendation, but later appeared as a joint recommendation along with START A1 ("*Vitamin K antagonists or direct thrombin inhibitors or factor Xa inhibitors in the presence of chronic atrial fibrillation*"). Therefore, relevance and implementation data are only available for START A1 from that point on.

- STOPP A3 (a recommendation to stop "*Any duplicate drug class prescription*") was initially appearing on reports. However, due to high numbers of STOPP A3 recommendations being produced that were not clinically relevant, this recommendation trigger was ceased by trial researchers.

Supplementary Appendix 2 – Inter-rater Reliability Assessment

A convenience sample of three pharmacists and three physicians (one consultant geriatrician, and two specialist registrars – i.e. senior residents – in geriatric medicine) were invited to participate. Raters were purposively selected on the basis of their involvement with the SENATOR project and/or the OPERAM project, another multi-centre RCT in which the intervention similarly included the provision of computer-generated medication recommendations based on STOPP/START criteria version 2 (https://operam-2020.eu/index.php?id=1488).

Twenty intervention cases were selected at random, representing approximately 10% of intervention patients recruited at CUH. Details of this random selection can be found below. The study's objectives were explained to each rater, and all raters were supplied with instructions on how to assess the clinical relevance of the recommendations, whereby the rater had to independently assign a code of 0-5 for each SENATOR-generated STOPP/START recommendation based on the clinical relevance categories described previously. Three sample cases (all based on real intervention patients) were provided with the clinical relevance codes already assigned to the recommendations, and with a rationale given as to why each code was chosen by the pharmacist-physician pair for each patient. The twenty clinical cases were presented in a standardised format (see Supplementary Appendix 3) to include age, sex, comorbidities, medicines prescribed at the time of randomisation, laboratory test results, and any other important information required to facilitate the raters in evaluating the clinical relevance of the STOPP/START recommendations.

Random Selection of 20 Intervention Cases for Inter-rater Reliability Assessment

A list of all intervention patients in the study site was divided into four according to the date of recruitment to ensure patient cases were obtained from different times during the RCT.

An independent researcher (external to this study) rearranged the four lists of patient numbers into a random order. The first five patients with at least three STOPP/START recommendations in each list were chosen as the cases in the inter-rater reliability study. Thus, twenty intervention patients' cases were selected at random.

Supplementary Appendix 3 – Standardised Case Format for Inter-rater Reliability Assessment

Age: 80

Sex: Female

Date of recruitment: 11/2017

Presenting Condition: Patient presenting with urosepsis. Patient fell during the night before admission to hospital when on her way to the toilet with bruising to right arm and leg, but no fracture.

Medical History:

1. Hypertension
2. Hypercholesterolemia
3. Chronic ischaemic heart disease
4. Osteoarthritis
5. Osteoporosis
6. Neck of femur fracture 2014

Medications:

1.	Tinzaparin 3500 units OD SC	On since admission
2.	Piperacillin/Tazobactam 4.5g TDS IV	On since admission
3.	Aspirin 75mg OD	
4.	Ramipril 2.5mg OD	
5.	Atorvastatin 20mg OD	
6.	Zolpidem 10mg NOCTE	On for 6 – 12 months
7.	Paracetamol 1g QDS IV/PO PRN	

Laboratory Parameters:

Sodium (mmol/L)	140
Potassium (mmol/L)	4.1
Corrected calcium (mmol/L)	2.31
Creatinine (micromoles/L)	63
eGFR (MDRD) ml/min/1.73m^2	113
Haemoglobin (g/dl)	-
Platelets x 10^9	370
INR	-

Other relevant information:

- Blood Pressure: 124/79 mmHg
- Heart Rate: 76 beats/min
- Patient does not have a history of recurrent falls.
- On at home but not charted:
  Calcium/Vitamin D_3 500mg/400 units 1 tablet BD
  Risedronate sodium 35mg once weekly

STOPP Recommendations:

Drug	Recommendation	Clinical Relevance
Zolpidem	Hypnotic Z-drugs e.g. zopiclone, zolpidem, zaleplon (may cause protracted daytime sedation, ataxia).	
	Any drug prescribed beyond the recommended duration, where treatment duration is well defined.	

START Recommendations:

Recommendation	Clinical Relevance
Beta-blocker with ischaemic heart disease.	
Vitamin D and calcium supplement in patients with known osteoporosis and/or previous fragility fracture(s) and/or (Bone Mineral Density T-scores more than -2.5 in multiple sites).	
Bone anti-resorptive or anabolic therapy (e.g. bisphosphonate, strontium ranelate, teriparatide, denosumab) in patients with documented osteoporosis, where no pharmacological or clinical status contraindication exists (Bone Mineral Density T-scores -> 2.5 in multiple sites) and/or previous history of fragility fracture(s).	

Supplementary Appendix 4

Supplementary Table 1: Degree of clinical relevance of individual STOPP and START recommendations

Rule	Adverse significance	No clinical relevance	Possibly low relevance	Possibly important relevance	Possibly very important relevance	Total
STOPP A2	1	8	104	17	4	134
STOPP A3	-	20	-	-	-	20
STOPP B1	-	2	-	-	-	2
STOPP B3	-	2	-	1	-	3
STOPP B5	-	5	1	-	-	6
STOPP B6	2	2	-	1	-	5
STOPP B7	7	10	6	3	-	26
STOPP B8	-	-	3	2	1	6
STOPP B9	-	1	-	-	-	1
STOPP B11	-	-	2	1	1	4
STOPP B12	-	-	2	-	-	2
STOPP C3	3	8	1	5	-	17
STOPP C5	2	-	-	2	-	4
STOPP C6	1	4	-	1	-	6
STOPP C10	-	-	-	1	5	6
STOPP C11	-	-	-	-	1	1
STOPP D2	-	1	1	1	-	3
STOPP D4	-	1	1	-	-	2
STOPP D5	-	-	13	9	2	24
STOPP D8	-	-	5	13	-	18
STOPP D9	-	1	-	-	-	1
STOPP D10	-	2	1	1	1	5
STOPP D11	-	4	-	-	-	4
STOPP D12	-	-	-	9	-	9
STOPP D14	-	-	1	20	-	21
STOPP E3	-	1	1	-	-	2
STOPP E6	-	-	-	-	1	1
STOPP F2	1	-	15	-	-	16
STOPP F3	-	-	3	2	3	8
STOPP G1	-	1	-	-	-	1
STOPP G2	-	-	1	-	-	1
STOPP G4	-	-	1	-	-	1
STOPP H4	-	2	1	-	-	3
STOPP H5	-	2	-	1	-	3
STOPP H6	-	-	1	-	-	1
STOPP H7	-	-	-	1	-	1
STOPP H8	-	-	-	7	-	7
STOPP H9	-	-	-	2	-	2
STOPP I1	-	-	-	10	-	10

STOPP I2	-	1	-	4	-	5
STOPP J3	-	40	1	-	-	41
STOPP K1	-	-	13	10	8	31
STOPP K2	-	1	2	3	1	7
STOPP K3	1	1	3	11	-	16
STOPP K4	-	-	10	5	1	16
STOPP L1	-	4	5	2	-	11
STOPP L2	1	2	4	10	2	19
STOPP L3	-	2	6	4	1	13
STOPP N	-	1	4	12	-	17
START A1	1	-	1	12	13	27
START A2	-	1	-	1	1	3
START A3	22	15	3	4	3	47
START A4	-	5	4	4	3	16
START A5	-	6	12	11	4	33
START A6	-	2	21	24	3	50
START A7	2	1	5	5	-	13
START A8	-	-	8	5	1	14
START B1	-	1	1	-	2	4
START B2	-	1	1	2	2	6
START B3	-	5	-	-	-	5
START C1	-	-	-	-	1	1
START D2	-	-	1	-	-	1
START E1	-	3	3	1	-	7
START E2	-	-	-	-	1	1
START E3	-	-	-	5	2	7
START E4	-	2	6	5	1	14
START E5	-	-	-	8	3	11
START E6	-	-	1	2	-	3
START F1	-	1	1	10	1	13
START G1	-	8	-	1	-	9
START G2	-	12	8	1	-	21
START H1	-	4	3	-	1	8
START H2	1	3	29	13	2	48
All rules	**45**	**199**	**320**	**285**	**76**	**925**

Supplementary Appendix 5

Supplementary Table 2: Clinical relevance of recommendations based on drug class

Drug Class	Adverse significance	No clinical relevance	Possibly low relevance	Possibly important relevance	Possibly very important relevance	Total
Antithrombotics	29	33	6	32	17	117
PPIs	2	8	96	-	-	106
Benzodiazepines	-	-	37	27	11	75
ACE-Inhibitors and/or ARBs	1	3	25	38	5	72
Beta blockers	2	41	14	10	1	68
Opioids	1	13	19	16	7	56
Laxatives	1	5	29	13	2	50
Diuretics	9	13	11	6	1	40
Antihistamines	-	-	3	35	-	38
Statins	-	6	12	11	4	33
Antipsychotics	-	4	3	21	2	30
Z-drugs	-	-	18	8	2	28
Bone anti-resorptive, anabolic agents	-	2	7	11	2	22
5-alpha reductase inhibitors	-	12	8	1	-	21
Alpha-1 Receptor blockers	-	10	-	9	-	19
Vitamin D +/- Calcium	-	-	-	13	5	18
Drugs for obstructive airways disease	-	9	2	2	4	17
Antihypertensives (START A4)	-	5	4	4	3	16
Anticholinergics	-	1	3	11	-	15
Miscellaneous cardiac drugs	-	8	2	2	-	12
Other miscellaneous agents	-	1	4	3	3	11
Antidepressants	-	3	4	1	-	8
Corticosteroids	-	4	3	1	-	8
NSAIDs	-	-	-	2	6	8
Calcium channel blockers	-	2	2	3	-	7
DMARDs	-	3	3	1	-	7
Anti-diabetic drugs	-	4	-	-	1	5
Drugs to treat gout or hyperuricaemia	-	-	3	2	-	5
Oxygen	-	5	-	-	-	5
Anti-dementia drugs	-	4	-	-	-	4
Drugs for urinary frequency/incontinence	-	-	2	2	-	4
TOTAL	**45**	**199**	**320**	**285**	**76**	**925**

PPI: Proton pump inhibitors; ACE: Angiotensin Converting Enzyme; ARB: Angiotensin Receptor Blocker; NSAIDs: Non-steroidal anti-inflammatory drug; DMARD: Disease-modifying anti-rheumatic drug.

Supplementary Appendix 6 – Reasons for lack of data on prescriber implementation

Data on prescriber implementation was unavailable for 41/925 (4.4%) recommendations, which includes the following:

- all STOPP A3 recommendations, which were later removed from the intervention (n = 20),
- two patients without implementation data in the electronic case report form (n = 15),
- all START B3 recommendations, as there was no data on oxygen prescribing (n = 5), and
- one STOPP N recommendation that triggered inappropriately (n = 1).