The “surprise question” for predicting death in seriously ill patients: a systematic review and meta-analysis

James Downar MDCM MHSc, Russell Goldman MD MPH, Ruxandra Pinto PhD, Marina Englesakis MLIS, Neill K.J. Adhikari MDCM MSc

■ Cite as: *CMAJ* 2017 April 3;189:E484-93. doi: 10.1503/cmaj.160775

ABSTRACT

BACKGROUND: The surprise question — “Would I be surprised if this patient died in the next 12 months?” — has been used to identify patients at high risk of death who might benefit from palliative care services. Our objective was to systematically review the performance characteristics of the surprise question in predicting death.

METHODS: We searched multiple electronic databases from inception to 2016 to identify studies that prospectively screened patients with the surprise question and reported on death at 6 to 18 months. We constructed models of hierarchical summary receiver operating characteristics (sROCs) to determine prognostic performance.

RESULTS: Sixteen studies (17 cohorts, 11 621 patients) met the selection criteria. For the outcome of death at 6 to 18 months, the pooled prognostic characteristics were sensitivity 67.0% (95% confidence interval [CI] 55.7%–76.7%), specificity 80.2% (73.3%–85.6%), positive likelihood ratio 3.4 (95% CI 2.8–4.1), negative likelihood ratio 0.41 (95% CI 0.32–0.54), positive predictive value 37.1% (95% CI 30.2%–44.6%) and negative predictive value 93.1% (95% CI 91.0%–94.8%). The surprise question had worse discrimination in patients with noncancer illness (area under sROC curve 0.77 [95% CI 0.73–0.81]) than in patients with cancer (area under sROC curve 0.83 [95% CI 0.79–0.87; *p* = 0.02 for difference]). Most studies had a moderate to high risk of bias, often because they had a low or unknown participation rate or had missing data.

INTERPRETATION: The surprise question performs poorly to modestly as a predictive tool for death, with worse performance in noncancer illness. Further studies are needed to develop accurate tools to identify patients with palliative care needs and to assess the surprise question for this purpose.

The surprise question (SQ) was developed more than a decade ago and has been suggested as a simple test to identify patients who might benefit from hospice and palliative care (HPC).[1] It involves a clinician reflecting on the question, “Would I be surprised if this patient died in the next 12 months?”. It was thought that the SQ would correct for a physician’s tendency to overestimate prognosis[2] by asking the physician to consider whether death in the coming year is possible rather than probable. The surprise question has been widely promoted[3,4] and adopted into frameworks for assessing hospice and palliative care needs.[5,6]

In the past few years, several studies have reported on the accuracy of the SQ for a different purpose: as a prognostic test of intermediate-term death in different patient populations. These studies sought to determine whether an answer of “no” (hereafter SQ+) predicts intermediate-term death. We conducted a systematic review of the literature to determine the performance characteristics of the SQ in predicting death and the methodologic characteristics of these studies.

Methods

Search strategy

We searched MEDLINE (from 1946 to week 2 of October 2016), MEDLINE in process (to Oct. 19, 2016), Embase (1947 to Oct. 19, 2016), Cochrane Central Register of Controlled Trials (to September 2016), Cochrane Database of Systematic Reviews (from 2005 to Oct. 19, 2016), PsycINFO (from 1806 to week 2 of October 2016), Cumulative Index to Nursing and Allied Health Literature (CINAHL; from 1961 to Oct. 20, 2016), Web of Science (Oct. 19, 2016), SCOPUS, PubMed and Google Scholar. Details of the search strategy are available in Appendix 1 (available at www.cmaj.ca/lookup/suppl/doi:10.1503/cmaj.160775/-/DC1).

Study selection

Two authors (J.D. and R.G.) independently screened citations to identify potentially relevant articles or abstracts. Potentially relevant citations were retrieved and reviewed independently in full text by the same 2 authors to determine whether they met inclusion criteria: a prospective cohort study with SQ asked of study participants, with prospective follow-up for the primary outcome (death at least 6 months after SQ asked). We did not consider designs where the SQ was asked retrospectively because of the potential for biased ratings based on knowledge of patient outcome. Reviewers were not masked to the study author, institution or journal. Because we did not wish to limit calculations to sensitivity, we excluded studies that provided data for outcomes only for patients who were SQ+, but we attempted to contact authors of these studies to determine whether outcomes were available for patients identified as SQ– (yes answer to SQ).

Data abstraction and methodologic quality

Two authors (J.D. and N.A.) abstracted in duplicate the following data from each study: setting; population; proportion of eligible patients enrolled; proportion of eligible patients with outcomes data; number of SQ evaluators for each patient and measure of agreement, if any; and outcomes (true positives, false positives, true negatives and false negatives). A true positive was defined as an answer of no to SQ for a patient who died. Two reviewers (J.D. and N.A.) also assessed the risk of bias of each study using the Quality in Prognosis Studies tool.[7,8] Disagreements between reviewers at any stage of the review were resolved by consensus.

Statistical analysis

For each study, we constructed a 2 × 2 table (predictor, SQ; outcome, death) and calculated the incidence of death in the study, sensitivity, specificity, positive likelihood ratio (LR+), negative likelihood ratio (LR–), positive (PPV) and negative (NPV) predictive values, and diagnostic odds ratio (DOR). The DOR is the ratio of odds of a prediction of death (SQ+) in a patient who dies divided by the odds in a patient who lives, calculated as LR+ divided by LR–. In studies reporting more than 1 clinician rating the SQ we used the study's definition of SQ+, which could require consensus or at least 1 rater.

We constructed a summary receiver operating characteristic (sROC) curve using a Rutter–Gatsonis hierarchical sROC model.[9] The ROC curve plots the true positive rate (sensitivity) against the false-positive rate (1 – specificity), and the area under the ROC curve (AUC) is a measure of discrimination, with 1 denoting perfect discrimination and 0.5 denoting no better than chance. Discrimination is the probability, given a randomly selected pair of patients, of whom one lives and the other dies, that the survivor will be coded as SQ– and the decedent coded as SQ+. The hierarchical model was appropriate given that the scale parameter β, which provides for asymmetry in the sROC model, was not significantly different from zero, and the distribution of the random errors was normal on visual inspection.

Based on the hierarchical sROC model, we determined the meta-analytic summaries of DOR, LR+ and LR–, sensitivity and specificity. Using the DOR, we derived the AUC and its SE based on the formula described by Walter.[10] We also presented summary PPV and NPV for illustration, and acknowledged the following limitations: heterogeneity may be greater than for test characteristics of sensitivity and specificity, and average predictive values relate to test utilization at some average, but unknown, incidence of death.[11]

For the predefined subgroup analyses of patients with cancer and those with noncancer illness, there were fewer studies; therefore, we used univariable random effects models, with each study weighted by the inverse of its variance for the given parameter. Parameters for these subgroups were compared using the z test. We report heterogeneity of summary estimates of diagnostic performance using the I^2 measure, which is the percentage of total variability across studies that is attributable to heterogeneity rather than chance, and used published guidelines for low (I^2 25%–49%), moderate (I^2 50%–74%), and high ($I^2 \geq$ 75%) heterogeneity.[12] To calculate I^2, we used univariable models for all studies together and for the subgroups with and without cancer. Heterogeneity measures were not available for AUC estimates, because they were derived from the DOR.

We assessed for publication bias using the regression test of asymmetry described by Deeks and colleagues,[13] which is based on a plot of the inverse of the square root of the sample size versus the log of the DOR. The analysis was performed in SAS version 9.3 (SAS Institute) using the MetaDAS macro for the hierarchical sROC model and R version 3.2.0 for the univariable meta-analysis. The sROC plot was generated using RevMan version 5.3 (Cochrane Community). Ethics approval was not required for this study.

Results

We identified 792 citations, of which 26[1,14–38] were potentially relevant (Figure 1). Of these studies, 7 did not include outcomes data for patients who were SQ–[14–19,36] that were confirmed in the publication[15–18,36] or after contact with study authors.[14,19] In 3 studies, the SQ was used only to predict 7- and/or 30-day mortality.[29,37,38] Sixteen remaining studies of SQ to predict death met inclusion criteria[1,20–28,30–35] (Table 1) and were included in the meta-analysis. One study provided data for a derivation cohort in the manuscript.[20] We also obtained data from a validation cohort from the study authors, which we analyzed separately in the numerical description and meta-analyses but considered with the derivation cohort as a single study otherwise.

The 16 studies (17 cohorts) enrolled a total of 11 621 patients (mean, 683 patients per cohort, range 49–4779; Table 1). All were prospective studies enrolling patients in 1 centre or group[1,20,22–24,26–28,30–35] or several clinics.[21,25] Five included only patients with cancer,[25,26,28,33,35] 7 included patients with renal failure,[1,20,22,27,30–32] 2 included patients with end-stage heart[21] or lung[23] disease, 1 included a heterogeneous population of patients with critical illness[24] and 1 involved a primary care practice.[34] The SQ was answered by physicians in 13 studies (including oncologists,[26,33,35] nephrologists,[20,22,30–32] respirologists,[23] intensivists,[24] and general practitioners[21,25,34]), specialist nurses in 1 study,[1] either a physician or a nurse in 1 study[28] and a multidisciplinary team of 2 physicians and a nurse in another study.[27] In each case, the assessors were familiar with the patient. Thirteen studies asked a 12-month SQ,[1,21–23,25–28,30–32,34,35] 2 asked a 6-month SQ[20,24] and 1 asked an 18-month SQ.[33]

In 13 studies reporting the number of patients enrolled and the eligible population,[1,20–22,24,26,27,30–35] the median participation rate was 100% (interquartile range [IQR] 98.5–100). Three studies reported incomplete availability of the SQ result for enrolled patients (85%,[20] 92%[21] and 97%[1]). Of note, 1 study reported that the 12% of eligible patients receiving hemodialysis without a recorded SQ response (because the physician did not answer the question) had higher mortality than those with a response.[20] Among 2 studies reporting provider participation, 1 reported that 8 of 50 (16%) general practitioners declined to participate,[25] and another found that 5 of 16 (31%) general practitioner clinics declined.[21] No study reported incomplete availability of status of death for patients with an SQ response.

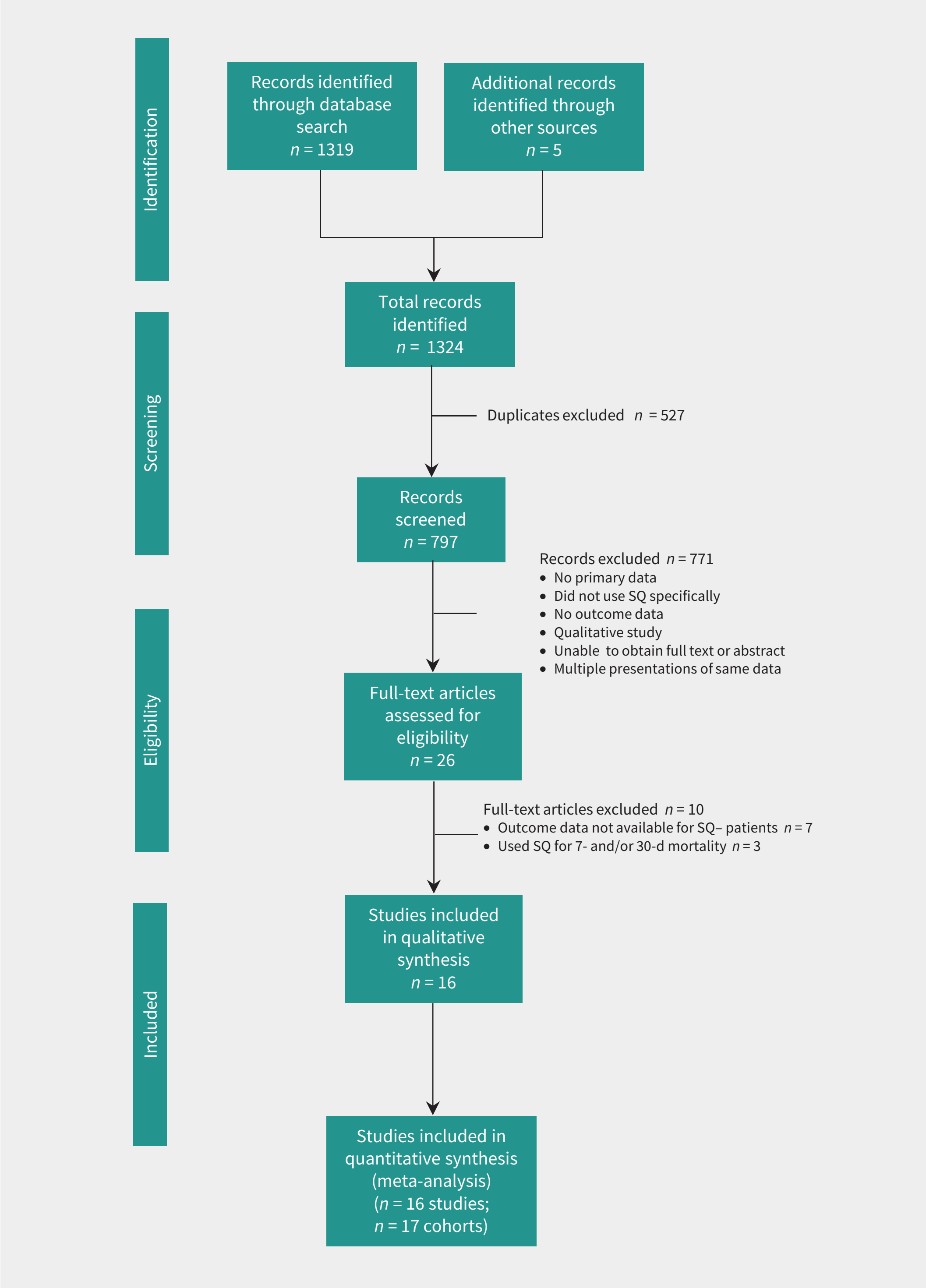
Identification
Screening
Eligibility
Included
Records identified through database search n = 1319
Additional records identified through other sources n = 5
Total records identified n = 1324
Duplicates excluded n = 527
Records screened n = 797
Records excluded n = 771
• No primary data
• Did not use SQ specifically
• No outcome data
• Qualitative study
• Unable to obtain full text or abstract
• Multiple presentations of same data
Full-text articles assessed for eligibility n = 26
Full-text articles excluded n = 10
• Outcome data not available for SQ– patients n = 7
• Used SQ for 7- and/or 30-d mortality n = 3
Studies included in qualitative synthesis n = 16
Studies included in quantitative synthesis (meta-analysis) (n = 16 studies; n = 17 cohorts)

Figure 1: Study selection for the systematic review. SQ = surprise question, SQ– = response to SQ is yes.

In most studies, only 1 clinician answered the SQ for each patient.[1,20,21,23–26,31,33,35] One study asked the SQ to a multidisciplinary team, which arrived at an answer by consensus.[27] In 1 study of patients receiving hemodialysis, every physician and nurse working with that patient was asked the SQ independently; to be consistent, we considered that the patient was SQ+ if either of the 2 physicians attending in their unit responded "no" to the SQ.[30] One study of patients receiving peritoneal dialysis[22] asked the SQ to 3 nephrologists for each patient and reported slight to fair[39] agreement (κ 0.34–0.41).

We assessed the methodologic quality of the included studies using the 6 domains of the Quality in Prognosis Studies tool[7] (Table 2) and assigned an overall risk of bias for each study according to the highest risk score in any domain. The rationale behind our judgment of risk is provided in Appendix 1. Only 2 studies were thought to have a low risk of bias in all domains; 10 had a moderate risk of bias in at least 1 domain, and 4 had a high risk of bias in at least 1 domain. The test of funnel plot asymmetry was not significant, suggesting no evidence of publication bias ($p = 0.3$).

The median incidence of death for the 17 cohorts was 15.1% (IQR 8.6%–20.5%). Prognostic properties of each included study are listed in Table 3. We contacted 6 authors to clarify or obtain additional data.[21,26,30,31,33,35] Using data from all 17 cohorts (Table 4), the meta-analytic estimates were as follows: of sensitivity 67.0% (95% confidence interval [CI] 55.7%–76.7%), specificity 80.2% (73.3%–85.6%), LR+ 3.4 (95% CI 2.8–4.1), LR– 0.41 (95% CI 0.32–0.54), PPV 37.1% (95% CI 30.2%–44.6%) and NPV 93.1% (95% CI 91.0%–94.8%). The sROC model showed moderate discriminatory ability (Figure 2), with an AUC of 0.81 (95% CI 0.78–0.86).

Prognostic properties of the SQ were better in studies involving patients with cancer than those involving patients with noncancer illness (Table 4); meta-analytic estimates of DOR, PPV and AUC were statistically significantly different between these patient populations. However, even in patients with cancer,[25,26,28,33,35] likelihood estimates would generate small changes in pretest probability,

Table 1 (part 1 of 2): Characteristics of studies included in the systematic review

Study, year	Centre	Diagnosis/ procedure/ practice	No. of eligible patients	Proportion of eligible patients enrolled, %	Proportion of enrolled patients with SQ response, %	Proportion of patients with "No" as SQ response, %	Incidence of death, % (95% CI)	No. of evaluators per patient; total no. of evaluators per study	Follow-up, mo
Barnes et al.,[21] 2008	16 general practice clinics	Congestive heart failure	1555	231/1555 (15)*	212/231 (92)	76/212 (36)	6.1 (3.4–10.0)	1 general practitioner; unclear	12
Moss et al.,[1] 2008	3 dialysis units looked after by 1 nephrology group	Hemodialysis	150 (consecutive)	147/150 (98)	147/147 (100)	34/147 (23)	15.0 (9.6–21.8)	1 nurse practitioner; 3	12
Cohen et al.,[20] 2010	5 dialysis units looked after by 1 nephrology group	Hemodialysis	1026 (consecutive)†	1026/1026 (100)	874/1026 (85)	127/874 (15)	Derivation, 6.0 (4.0–8.7); validation, 8.4 (6.0–11.5)	1 nephrologist; unclear	6
Moss et al.,[26] 2010	1 cancer centre	Cancer	853 (consecutive outpatients)	853/853 (100)	826/853 (97)	130/826 (16)	8.6 (6.8–10.7)	1 oncologist; 4	12
Da Silva Gane et al.,[30] 2013	3 dialysis units	Hemodialysis	344 (prevalent)	344/344 (100)	344/344 (100)	220/344 (64)	15.1 (11.5–19.3)	1 nephrologist; 6‡	12
Pang et al.,[22] 2013	1 dialysis unit	Peritoneal dialysis	367 (prevalent)	367/367 (100)	367/367 (100)	109/367 (30)	12.0 (8.8–15.8)	3 nephrologists;§ 3	12
Reilly et al.,[23] 2013	1 inpatient ward	Respiratory disease	Unclear	85 (randomly selected); unclear denominator	85/85 (100)	67/85 (79)	32.9 (23.1–44.0)	Respirologists (*n* unclear); unclear	12
Khan et al.,[24] 2014	1 medical–surgical ICU	Critically ill (mixed)	500 (consecutive)	500/500 (100)	500/500 (100)	238/500 (48)	36.0 (31.8–40.4)	Intensivists (*n* unclear); unclear	6
Moroni et al.,[25] 2014	42 general practitioners (unclear number of clinics)	Cancer	Unclear	231 (enrolled by participating physicians)¶	231/231 (100)	126/231 (55)	45.0 (38.5–51.7)	1 general practitioner; 42	12
Vick et al.,[28] 2015	Single centre	Cancer	4779	"All patients seen" by participating clinicians; unclear no. of nonparticipating physicians	4779/4779 (100)	732/4779 (15)	10.0 (9.2–10.9)	1 oncologist or nurse practitioner; 76	12

with a pooled LR+ of 4.2 (95% CI 2.9–6.0) and LR– of 0.41 (95% CI 0.32–0.53). In studies involving patients with noncancer illness,[1,20–23,27,30–32] likelihood ratios were less helpful (pooled LR+ 2.7 [95% CI 2.1–3.6] and LR– 0.53 [95% CI 0.46–0.61]). Heterogeneity was absent or low for DOR and LRs, but generally high for sensitivity, specificity, PPV and NPV for both cancer and noncancer studies.

Interpretation

In this systematic review, we identified 16 studies with 17 distinct cohorts that studied the SQ as a tool to predict death in patients with serious illness. Overall, pooled results suggest poor to modest accuracy of the SQ for predicting death at 12 months, with low sensitivity and positive predictive values for the studied populations. The pooled likelihood ratios are in the range of those that generate small changes from pretest to posttest probability.[40] Prognostic performance was worse for noncancer illness, missing more than one-third of those who died and more than two-thirds of positive results proved to be false. The NPV of the SQ was high, meaning that SQ– patients ("I would be surprised if this patient died") had a high probability of living. However, this finding was largely driven by the low prevalence of death in the included studies (e.g., if 12% of the population dies, then a coin flip would yield an NPV of 88%). One study also reported only a slight to fair interobserver reliability, and most included studies were felt to have at least a moderate risk of bias.

Our results may not be surprising to some, because physicians are inaccurate prognosticators,[2] and the SQ was not originally conceived as a prognostic tool but rather as a screening test for patients who might benefit from a palliative approach. The first reported use of the SQ, albeit without outcomes data, was among a primary care population in Tacoma, Washington.[41] Since then, the SQ has become widely used and promoted as a trigger for palliative or end-of-life interventions,[3,4] and has been incorporated as a screening test into the Gold Standards Framework in the United Kingdom[5] and the Necesidades Paliativas (NECPAL) program in Catalonia.[6]

The basic assumption of the SQ as a screening test is that patients in their final year of life may have unmet palliative care

Table 1 (part 2 of 2): Characteristics of studies included in the systematic review

Study, year	Centre	Diagnosis/ procedure/ practice	No. of eligible patients	Proportion of eligible patients enrolled, %	Proportion of enrolled patients with SQ response, %	Proportion of patients with "No" as SQ response, %	Incidence of death, % (95% CI)	No. of evaluators per patient; total no. of evaluators per study	Follow-up, mo
Gerlach et al.,[33] 2016	Single centre	Cancer	828 (consecutive outpatients)	828/828 (100)	828/828 (100)	146/828 (18)	17.0 (14.5–19.8)	1 oncologist or palliative care physician; 13	18
Amro et al.,[31] 2016	2 dialysis units looked after by 1 nephrology group	Hemodialysis	201 (prevalent)	201/201 (100)	201/201 (100)	50/201 (25)	19.4 (14.2–25.6)	1 nephrologist; 9	12
Lefkowits et al.,[35] 2016	Single centre	Cancer	263 (prevalent, receiving chemotherapy or radiotherapy)	263/263 (100)	263/263 (100)	87/263 (33)	20.5 (15.8–25.9)	1 gynecology–oncology physician; 7 physicians††	12
Carmen et al.,[32] 2016	1 dialysis unit	Hemodialysis	49 receiving hemodialysis for more than 3 mo in 2014; unclear if incident or prevalent	49/49 (100)	49/49 (100)	20/49 (41)	18.4 (8.8–32.0)	"Medical staff"; unclear	12
Lakin et al.,[34] 2016	Single centre	Primary care practice	Unclear	Patients screened for "high-risk care management program" over 18-month period, unclear number of eligible patients	1737/1737 (100)	114/1737 (7)	6.4 (5.3–7.7)	1 primary care physician; unclear	12

Note: CI = confidence interval, ICU = intensive care unit, SQ = surprise question, SQ+ = response to SQ is no, SQ– = response to SQ is yes.
*Interest in participation expressed by 748 patients; 587 of these were asked for demographic information and 542 returned the survey they were sent; 11/16 general practice clinics agreed to participate, for a total of 231 patients.
†Data from the derivation cohort (*n* = 512) were published. We obtained data for the validation cohort (*n* = 514) from the authors.
‡SQ+ ("I would not be surprised if this patient died in the next 12 mo") was defined as a positive response from one of the nephrologist evaluators.
§SQ+ ("I would not be surprised if this patient died in the next 12 mo") was defined as a positive response from any of the 3 evaluators.
¶Eight of 50 (16.0%) eligible general practitioners refused to enrol their patients.
**SQ+ and SQ– for each patient were defined by consensus.
††Data for physician responses are reported here for consistency with other studies.

needs. In general, a screening test should have high sensitivity; lower specificity is acceptable, if the cost of a false positive is minimal, or if each positive on the screening test is followed by a more specific confirmatory test. In the case of the SQ, the sensitivity is not high and, when applied to some populations, the number of positive results could be large: 7.7% of the population over age 65 years in the Osona region of Catalonia,[6] up to one-third of inpatients in 1 UK study[42] and 78% of patients admitted with advanced congestive heart failure in the abstract of 1 study in the United States.[43] The studies included in this meta-analysis had SQ+ rates that ranged from 7% to 79%. It is possible that the SQ is more accurate for identifying those with unmet palliative needs than those in the final year of life. But if being in the final year of life is a surrogate for unmet end-of-life care needs, including palliative care, then our results show that the SQ will simultaneously generate referrals of hospice and palliative care for patients who may not benefit from assessment and miss many patients who might benefit.

The high false-positive rate may not be surprising, because clinicians answering the SQ must consider whether death is possible rather than probable. However, several studies highlighted other concerns about the systematic use of the SQ. First, clinicians may only have moderate agreement when answering the SQ, with more experienced clinicians generally more accurate than those with less experience.[30,44] Second, although many of the single-centre studies showed a high response rate, some clinicians may be reluctant to adopt the SQ into routine practice;[21,25,45] 1 study reported that discomfort among staff led to the SQ being dropped as a trigger for a care bundle for patients in hospital.[46] This discomfort is an important problem for a test that depends on clinician participation. Third, the poorer sensitivity and positive likelihood ratio in patients with noncancer illness, such as those with end-stage organ failure and frailty,[47,48] suggests that the SQ may be particularly unhelpful in this population. Fourth, physicians may be reluctant to commit to a positive answer on the SQ even if they suspect that the patient is dying.[20,49] Thus, the SQ may have more success as a trigger for hospice and palliative care when used in selected settings, such as oncology, with a small number of experienced assessors.

It is possible that a false-positive SQ result would be beneficial, because a palliative approach or referral to hospice and palliative care may benefit patients who are not in the last year of life. The American Society of Clinical Oncology recommends “early” palliative care referral for patients with cancer,[50] but there is no consensus about the ideal timing of the referral. Two large randomized controlled trials (RCTs) showed a benefit to “early” referral to hospice and palliative care for patients with newly detected metastatic cancer,[51,52] but neither used the SQ as a trigger. A recent analysis of RCTs of hospice and palliative care found that authors used “no clear definitions” of hospice and palliative care patients, and that there was a “lack of consensus concerning the attributes of illnesses needing palliation.”[53] Although Advance Care Planning may be appropriate for patients at any stage of illness, some authors have highlighted the “Goldilocks phenomenon,” in which such planning may be ineffective or deleterious if implemented too early or too late.[3] Furthermore, patients with noncancer illness

Table 2: Risk of bias for studies included in the systematic review

Study, year	Domain 1: Study participation	Domain 2: Study attrition	Domain 3: Prognostic factor measurement	Domain 4: Outcome measurement	Domain 5: Study confounding	Domain 6: Statistical analysis and reporting	Overall risk of bias
Barnes et al.,[21] 2008	High	Low	Low	Low	Moderate	Low	High
Moss et al.,[1] 2008	Low	Low	Low	Low	Moderate	Low	Moderate
Cohen et al.,[20] 2010	High	Low	Low	Low	Moderate	Low	High
Moss et al.,[26] 2010	Moderate	Low	Low	Low	Low	Low	Moderate
Da Silva Gane et al.,[30] 2013	Low	Low	Low	Low	Moderate	Low	Moderate
Pang et al.,[22] 2013	Low	Moderate	Low	Low	Moderate	Low	Moderate
Reilly et al.,[23] 2013	Moderate	Low	Low	Low	Moderate	Low	Moderate
Khan et al.,[24] 2014	Low	Low	Low	Low	Moderate	Low	Moderate
Moroni et al.,[25] 2014	Moderate	Low	Low	Low	Low	Low	Moderate
Feyi et al.,[27] 2015	Low	Low	Low	Moderate	High	Low	High
Vick et al.,[28] 2015	Moderate	Moderate	Low	Low	Low	Low	Moderate
Gerlach et al.,[33] 2016	Low	Low	Low	Low	Low	Low	Low
Amro et al.,[31] 2016	Low	Low	Low	Low	Moderate	Low	Moderate
Lefkowits et al.,[35] 2016	Low	Low	Low	Low	Low	Low	Low
Carmen et al.,[32] 2016	Low	Low	Low	Low	Moderate	Low	Moderate
Lakin et al.,[34] 2016	High	Low	Low	Low	Low	Low	High

Note: Appendix 1 (available at www.cmaj.ca/lookup/suppl/doi:10.1503.cmaj.160775/-/DC1) provides the details of how judgments were made for the individual domains.

may benefit from earlier referral to hospice and palliative care, because they may have a longer period of high support needs.[54] But these patients can survive for years even with poor prognostic signs, suggesting that the burden on most hospice and palliative care services would be unsustainable if they were routinely referred early. Given the high false-positive rate, any hospice and palliative care service using the SQ as a referral trigger would need to screen out many referrals to avoid being overwhelmed. The broader issue of limited hospice and palliative care resources may be addressed by initiatives to educate and better support primary care providers in delivering palliative care without referral to hospice and palliative care services.

Ultimately, our findings do not indicate that the SQ is a "bad" or "good" test, but clinicians could benefit from better ways to identify patients with palliative health trajectories. One study reported improved prognostic accuracy by combining SQ with clinical predictors among a population of patients receiving hemodialysis.[20] Conversely, 1 study in our review reported that the SQ no longer predicted death when included in a multivariable model with the Clinical Frailty Scale.[32] The Gold Standards Framework[5] and NECPAL[6] both use the SQ serially with general and disease-specific indicators to identify those with hospice and palliative care needs, but the Supportive and Palliative Care Indicators Tool (SPICT)[15] uses these indicators without the SQ, which was removed. Further studies will be needed to determine whether the SQ combined with other clinical indicators improves the identification of patients with hospice and palliative care needs.

Table 3: Prognostic properties of studies included in the systematic review

Study, year	Sensitivity, % (95% CI)	Specificity, % (95% CI)	LR+ (95% CI)	LR– (95% CI)	PPV, % (95% CI)	NPV, % (95% CI)	DOR (95% CI)
Barnes et al.,[21] 2008	78.6 (49.2–95.3)	67.2 (60.2–73.7)	2.4 (1.7–3.4)	0.32 (0.12–0.87)	14.5 (7.5–24.4)	97.8 (93.7–99.5)	7.50 (2.50–22.54)
Moss et al.,[1] 2008	45.5 (24.4–67.8)	80.8 (72.8–87.3)	2.4 (1.3–4.2)	0.68 (0.46–1.00)	29.4 (15.1–47.5)	89.4 (82.2–94.4)	3.51 (1.58–7.78)
Cohen et al.,[20] 2010 (derivation cohort)	63.0 (42.4–80.6)	87.4 (83.8–90.4)	5.0 (3.4–7.3)	0.42 (0.26–0.69)	24.3 (14.8–36.0)	97.3 (95.2–98.7)	11.77 (5.85–23.67)
Cohen et al.,[20] 2010 (validation cohort)	47.2 (30.4–64.5)	89.8 (86.3–92.6)	4.6 (2.9–7.3)	0.59 (0.43–0.8)	29.8 (18.4–43.4)	94.9 (92.1–96.9)	7.85 (4.25–14.51)
Moss et al.,[26] 2010	74.6 (62.9–84.2)	89.8 (87.4–91.9)	7.3 (5.7–9.4)	0.28 (0.19–0.42)	40.8 (32.2–49.7)	97.4 (95.9–98.5)	25.93 (15.88–42.34)
Da Silva Gane et al.,[30] 2013	57.7 (43.2–71.3)	74.7 (69.3–79.5)	2.3 (1.7–3.1)	0.57 (0.41–0.78)	28.8 (20.4–38.6)	90.8 (86.5–94.2)	4.02 (2.41–6.70)
Pang et al.,[22] 2013	61.4 (45.5–75.6)	74.6 (69.5–79.3)	2.4 (1.8–3.3)	0.52 (0.35–0.76)	24.8 (17.0–34.0)	93.4 (89.7–96.1)	4.67 (2.69–8.10)
Reilly et al.,[23] 2013	100 (87.7–100)	31.6 (19.9–45.2)	1.4 (1.2–1.7)	0.05 (0.00–0.87)	41.8 (29.8–54.5)	100 (81.5–100)	26.70 (2.44–291.88)
Khan et al.,[24] 2014	82.2 (75.8–87.5)	71.9 (66.6–76.7)	2.9 (2.4–3.5)	0.25 (0.18–0.34)	62.2 (55.7–68.4)	87.8 (83.2–91.5)	11.82 (8.08–17.29)
Moroni et al.,[25] 2014	83.7 (75.1–90.2)	69.3 (60.5–77.2)	2.7 (2.1–3.6)	0.24 (0.15–0.37)	69.0 (60.2–77.0)	83.8 (75.3–90.3)	11.55 (6.74–19.79)
Feyi et al.,[27] 2015	66.7 (50.5–80.4)	77.9 (70–84.6)	3.0 (2.1–4.4)	0.43 (0.28–0.66)	48.3 (35.0–61.8)	88.3 (81.2–93.5)	7.07 (3.74–13.36)
Vick et al.,[28] 2015	57.9 (53.4–62.4)	89.4 (88.5–90.3)	5.5 (4.9–6.2)	0.47 (0.42–0.52)	38.0 (34.4–41.6)	95.0 (94.3–95.7)	11.66 (9.81–13.85)
Gerlach et al.,[33] 2016	48.2 (39.7–56.8)	88.6 (86.0–90.9)	4.3 (3.2–5.6)	0.58 (0.50,0.69)	46.6 (38.3–55.0)	89.3 (86.7–91.5)	7.27 (5.17–10.22)
Amro et al.,[31] 2016	56.4 (39.6–72.2)	82.7 (76.0–88.2)	3.3 (2.1–5.0)	0.53 (0.37–0.76)	44.0 (30.0–58.7)	88.7 (82.6–93.3)	6.19 (3.29–11.65)
Lefkowits et al.,[35] 2016	64.8 (50.6–77.3)	75.1 (68.7–80.8)	2.6 (1.9–3.5)	0.47 (0.32–0.68)	40.2 (29.9–51.3)	89.2 (83.7–93.4)	5.56 (3.25–9.52)
Carmen et al.,[32] 2016	77.8 (40.0–97.2)	67.5 (50.9–81.4)	2.4 (1.4–4.2)	0.33 (0.10–1.14)	35.0 (15.4–59.2)	93.1 (77.2–99.2)	7.27 (1.74–30.40)
Lakin et al.,[34] 2016	20.5 (13.5–29.2)	94.4 (93.2–95.5)	3.7 (2.4–5.6)	0.84 (0.77–0.93)	20.2 (13.2–28.7)	94.5 (93.3–95.6)	4.36 (2.85–6.65)

Note: CI = confidence interval, DOR = diagnostic odds ratio, LR+ = positive likelihood ratio, LR– = negative likelihood ratio, NPV = negative predictive value, PPV = positive predictive value.

Another study used administrative data for all nonpsychiatric admissions of adult patients to hospitals in Ontario and Alberta, and to the Brigham and Women's Hospital in Boston to derive a score with excellent discrimination (AUC 0.92) for predicting 1-year mortality among patients admitted to hospital for nonpsychiatric indications.[55,56] If the findings are reproducible, the score could be implemented to generate automatic prompts for clinicians to consider discussions of advanced care planning or referral to hospice and palliative care. However, without a consistent definition of a patient who is "palliative" or a consensus about what problems hospice and palliative care is meant to address, we do not have a gold standard to which we can compare any of these triggers.

Strengths and limitations

Strengths of this systematic review include its novelty; a rigorous and transparent search strategy involving multiple databases; independent duplicate review for study selection, data abstraction and quality assessment; efforts to obtain unpublished data from primary study authors; and robust statistical methods. Limitations include the predominant conduct of the studies, in single centres with a small number of evaluators, which limited generalizability and assessment of interrater reliability. Many studies had methodologic limitations that were primarily related to incomplete participation of eligible patients and lack of detail on whether a SQ+ response triggered a discussion with the patient regarding limita-

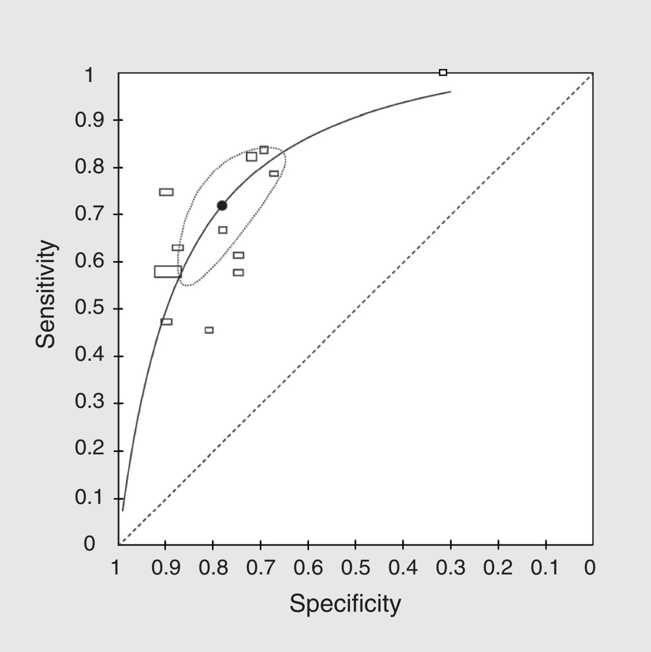

Figure 2: Summary receiver operating characteristic (sROC) curve. Each rectangle represents 1 study; the width of the rectangle is proportional to the standard error (SE) of the sensitivity, and the height is proportional to the SE of the specificity. The summary point is the meta-analytic sensitivity and specificity obtained from the hierarchical sROC model, and the dotted area is the 95% confidence interval region, based on the same model.

Table 4: Meta-analytic estimates of prognostic properties*

Parameter	Estimate for all patients (95% CI) *n* = 11 621	Heterogeneity (I^2), %	Estimate for patients with noncancer illness (95% CI) *n* = 2457	Heterogeneity (I^2), %	Estimate for patients with cancer (95% CI) *n* = 6927	Heterogeneity (I^2), %	*p* value†
DOR	8.21 (6.21–10.87)	0	5.94 (4.57–7.81)	0	10.69 (7.06–16.19)	28.0	0.02
LR+	3.4 (2.8–4.1)	0	2.7 (2.1–3.6)	0	4.2 (2.9–6.0)	17.4	0.06
LR–	0.41 (0.32–0.54)	0	0.53 (0.46–0.61)	0	0.41 (0.32–0.53)	39.0	0.09
Sensitivity, %	67.0 (55.7–76.7)	88.7	60.7 (52.6–68.1)	40.0	66.4 (54.1–76.8)	89.2	0.4
Specificity, %	80.2 (73.3–85.6)	96.5	75.9 (67.6–82.6)	92.7	84.3 (77.6–89.3)	95.0	0.08
PPV, %	37.1 (30.2–44.6)	89.5	31.3 (25.0–38.3)	67.8	46.8 (36.4–57.5)	90.1	0.02
NPV, %	93.1 (91.0–94.8)	86.4	93.2 (90.5–95.2)	68.6	92.4 (87.3–95.6)	94.3	0.7
AUC	0.81 (0.78–0.84)	NA	0.77 (0.74–0.8)	NA	0.83 (0.79–0.87)	NA	0.02

Note: AUC = area under the ROC curve, CI = confidence interval, DOR = diagnostic odds ratio, LR+ = positive likelihood ratio, LR– = negative likelihood ratio, NA = not applicable, NPV = negative predictive value, PPV = positive predictive value.
*Parameter estimates for all patients (16 studies, 17 cohorts) used hierarchical summary receiver operating characteristic (sROC) models. For cancer studies involving patients with cancer (5), studies involving patients with noncancer illness (9, 10 cohorts) and for all measures of heterogeneity (including for all studies combined), analyses used univariable models (see text for details). Heterogeneity measures are not available for AUC, which is derived from DOR.
†The *z* test was used for *p* values to compare studies involving patients with cancer with studies involving patients with noncancer illness.

tions of life-sustaining treatments. These limitations may have led to overestimates of the prognostic performance of the SQ. In addition, the meta-analytic estimates of predictive values will differ in populations with different risks of death.

Conclusion

In summary, the SQ is intended to be a simple and feasible screening test to identify patients with hospice and palliative care needs, but it performs poorly to modestly when used to predict death at 6 to 18 months, with poorer performance among patients with noncancer illness. Based on these findings, the SQ should not be used as a stand-alone prognostic tool, and we do not know whether it is more accurate for identifying patients with unmet palliative needs than it is for those in the final year of life. The high false-positive rate for SQ may be of concern if it used as a routine trigger for time-consuming, costly or poorly available assessments for hospice and palliative care. Developing accurate, reliable and automated means of identifying patients with hospice and palliative care needs in a variety of settings remains a high-priority area of research.

References

1. Moss AH, Ganjoo J, Sharma S, et al. Utility of the "surprise" question to identify dialysis patients with high mortality. *Clin J Am Soc Nephrol* 2008;3:1379-84.
2. Christakis NA, Lamont EB. Extent and determinants of error in doctors' prognoses in terminally ill patients: prospective cohort study. *BMJ* 2000;320:469-72.
3. Billings JA, Bernacki R. Strategic targeting of advance care planning interventions: the Goldilocks phenomenon. *JAMA Intern Med* 2014;174:620-4.
4. Spreeuwenberg C, Raats I, Teunissen S, et al. Development of a national care standard for palliative care in the Netherlands. *Palliat Med* 2014;28:634-5.
5. Kersun L, Gyi L, Morrison WE. Training in difficult conversations: a national survey of pediatric hematology-oncology and pediatric critical care physicians. *J Palliat Med* 2009;12:525-30.
6. Gómez-Batiste X, Martínez-Muñoz M, Blay C, et al. Identifying patients with chronic conditions in need of palliative care in the general population: development of the NECPAL tool and preliminary prevalence rates in Catalonia. *BMJ Support Palliat Care* 2013;3:300-8.
7. Hayden JA, Côté P, Bombardier C. Evaluation of the quality of prognosis studies in systematic reviews. *Ann Intern Med* 2006;144:427-37.
8. Hayden JA, van der Windt DA, Cartwright JL, et al. Assessing bias in studies of prognostic factors. *Ann Intern Med* 2013;158:280-6.
9. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med* 2001;20:2865-84.
10. Walter SD. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. *Stat Med* 2002;21:1237-56.
11. Macaskill P, Gatsonis CA, Deeks JJ, et al. Chapter 10: analysing and presenting results. In: Deeks JJ, Bossuyt PM, Gatsonis CA, editors. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version* 1.0. London (UK): The Cochrane Collaboration; 2010.
12. Higgins JP, Thompson SG, Deeks JJ, et al. Measuring inconsistency in meta-analyses. *BMJ* 2003;327:557-60.
13. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *J Clin Epidemiol* 2005;58:882-93.
14. Wachterman MW, Marcantonio ER, Davis RB, et al. Relationship between the prognostic expectations of seriously ill patients undergoing hemodialysis and their nephrologists. *JAMA Intern Med* 2013;173:1206-14.
15. Highet G, Crawford D, Murray SA, et al. Development and evaluation of the Supportive and Palliative Care Indicators Tool (SPICT): a mixed-methods study. *BMJ Support Palliat Care* 2014;4:285-90.
16. Quibell R, McAleer N. Improving end of life care for patients with COPD. An evaluation of a service innovation using the "would you be surprised?" question. *BMJ Support Palliat Care* 2014;4(Suppl 1):A92-3.
17. Johnson M, Nunn A, Hawkes T, et al. Planning for end-of-life care in heart failure: experience of two integrated cardiology–palliative care teams. *Br J Cardiol* 2012;19:71-5.
18. South G, Reddington O. End of life in COPD: There may be no surprises! *Eur Respir J* 2011;38(Suppl 55):1241.
19. Mitchell G, Zhang J, Burridge L, et al. Case conferences between general practitioners and specialist teams to plan end of life care of people with end-stage heart failure and lung disease: an exploratory pilot study. *BMC Palliat Care* 2014;13:24.
20. Cohen LM, Ruthazer R, Moss AH, et al. Predicting six-month mortality for patients who are on maintenance hemodialysis. *Clin J Am Soc Nephrol* 2010;5:72-9.
21. Barnes S, Gott M, Payne S, et al. Predicting mortality among a general practice-based sample of older people with heart failure. *Chronic Illn* 2008;4:5-12.
22. Pang W-F, Kwan BC-H, Chow K-M, et al. Predicting 12-month mortality for peritoneal dialysis patients using the "surprise" question. *Perit Dial Int* 2013;33:60-6.
23. Reilly L, Reilly K, Mc Closkey M, et al. Prognostic significance of the 'surprise question' in an respiratory inpatient population in a DGH. *Ir J Med Sci* 2013; 182:S484.
24. Khan S, Hadique S, Culp S, et al. Efficacy of the "surprise" question to predict 6-month mortality in ICU patients. *Crit Care Med* 2014;42(Suppl):A1457.
25. Moroni M, Zocchi D, Bolognesi D, et al. The 'surprise' question in advanced cancer patients: a prospective study among general practitioners. *Palliat Med* 2014;28:959-64.
26. Moss AH, Lunney JR, Culp S, et al. Prognostic significance of the "surprise" question in cancer patients. *J Palliat Med* 2010;13:837-40.
27. Feyi K, Klinger S, Pharro G, et al. Predicting palliative care needs and mortality in end stage renal disease: use of an at-risk register. *BMJ Support Palliat Care* 2015;5:19-25.
28. Vick J, Pertsch N, Hutchings M, et al. The utility of the "surprise question" in identifying patients with cancer most at risk of death: preliminary results [abstract]. American Society of Clinical Oncology Annual Meeting; October 2015; Boston.
29. Hamano J, Morita T, Inoue S, et al. Surprise questions for survival prediction in patients with advanced cancer: a multicenter prospective cohort study. *Oncologist* 2015;20:839-44.
30. Da Silva Gane M, Braun A, Stott D, et al. How robust is the 'surprise question' in predicting short-term mortality risk in haemodialysis patients? *Nephron Clin Pract* 2013;123:185-93.
31. Amro OW, Ramasamy M, Strom JA, et al. Nephrologist-facilitated advance care planning for hemodialysis patients: a quality improvement project. *Am J Kidney Dis* 2016;68:103-9.
32. Carmen JM, Santiago P, Elena D, et al. Frailty, surprise question and mortality in a hemodilaysis cohort question and mortality in a hemodialysis cohort. *Nephrol Dial Transplant* 2016;31:i553.
33. Gerlach C, Halbe L, Goebel S, et al. The role of the 'surprise'-question (SQ) in haemato-oncology: quantitative and qualitative analysis of a pilot project [abstract]. European Association for Palliative Care meeting; 2016 June 9–11; Dublin.
34. Lakin JR, Robinson MG, Bernacki RE, et al. Estimating 1-year mortality for high-risk primary care patients using the "surprise" question. *JAMA Intern Med* 2016;176:1863-5.
35. Lefkowits C, Chandler C, Sukumvanich P, et al. Validation of the "surprise question" in gynecologic oncology: comparing physicians, advanced practice providers, and nurses [abstract]. *Gynecol Oncol* 2016;141:128.
36. South G, Reddington O, Hatfield L, et al. End of life in COPD: there may be no surprises! *Eur Respir J* 2011;38:1241.
37. Strout TD, Haydar SA, Han PJK, et al. Utility of the modified "surprise question" for predicting inpatient mortality in emergency department patients. *Ann Emerg Med* 2015;66(Suppl):S81.
38. Strout TDS, Haydar SA, Eager E, et al. Identifying unmet palliative care needs in the ED: use of the 'surprise question' in patients with sepsis [abstract]. *Acad Emerg Med* 2016;23(Suppl 1):S196.
39. Byrt T. How good is that agreement? *Epidemiology* 1996;7:561.
40. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. *JAMA* 1994;271:703-7.
41. Pattison M, Romer AL. Improving care through the end of life: launching a primary care clinic-based program. *J Palliat Med* 2001;4:249-54.
42. Gardiner C, Gott M, Ingleton C, et al. Extent of palliative care need in the acute hospital setting: a survey of two acute hospitals in the UK. *Palliat Med* 2013;27:76-83.
43. Abell JM, Gafford E, Gustin J. CHF patients: Would an admission checklist increase identification of patients at high risk for unmet palliative care needs? [abstract]. *J Palliat Med* 2014;17:A21.

44. Gómez-Batiste X, Martínez-Muñoz M, Blay C, et al. Prevalence and characteristics of patients with advanced chronic conditions in need of palliative care in the general population: a cross-sectional study. *Palliat Med* 2014;28:302-11.
45. Elliott M, Nicholson C. A qualitative study exploring use of the surprise question in the care of older people: perceptions of general practitioners and challenges for practice. *BMJ Support Palliat Care* 2014 Aug. 28 [Epub ahead of print]. doi:10.1136/bmjspcare-2014-000679.
46. Carey I, Shouls S, Bristowe K, et al. Improving care for patients whose recovery is uncertain. The AMBER care bundle: design and implementation. *BMJ Support Palliat Care* 2015;5:12-8.
47. Pal LM, Manning L. Palliative care for frail older people. *Clin Med (Lond)* 2014; 14:292-5.
48. Boyd K, Kimbell B, Murray S, et al. A "good death" with irreversible liver disease: Talking with patients and families about deteriorating health and dying. *Clin Liv Dis (Hoboken)* 2015;6:15-8.
49. Thiagarajan R, Morris J, Harkins KJ. Can simple intuitive questions identify patients in the last year of life? — A pragmatic study comparing the "Paired surprise questions" with the "Single surprise question" *Age Ageing* 2012;41:61.
50. Smith TJ, Temin S, Alesi ER, et al. American Society of Clinical Oncology provisional clinical opinion: the integration of palliative care into standard oncology care. *J Clin Oncol* 2012;30:880-7.
51. Temel JS, Greer JA, Muzikansky A, et al. Early palliative care for patients with metastatic non-small-cell lung cancer. *N Engl J Med* 2010;363:733-42.
52. Zimmermann C, Swami N, Krzyzanowska M, et al. Early palliative care for patients with advanced cancer: a cluster-randomised controlled trial. *Lancet* 2014;383:1721-30.
53. Van Mechelen W, Aertgeerts B, De Ceulaer K, et al. Defining the palliative care patient: a systematic review. *Palliat Med* 2013;27:197-208.
54. Lunney JR, Lynn J, Hogan C. Profiles of older medicare decedents. *J Am Geriatr Soc* 2002;50:1108-12.
55. van Walraven C. The Hospital-patient One-year Mortality Risk score accurately predicted long-term death risk in hospitalized patients. *J Clin Epidemiol* 2014;67:1025-34.
56. van Walraven C, McAlister FA, Bakal JA, et al. External validation of the Hospital-patient One-year Mortality Risk (HOMR) model for predicting death within 1 year after hospital admission. *CMAJ* 2015;187:725-33.

Competing interests: None declared.

This article has been peer reviewed.

Accepted: Jan. 16, 2017

Affiliations: Divisions of Respirology/Critical Care and Palliative Care, University Health Network; and Temmy Latner Centre for Palliative Care (Downar), Sinai Health System; Temmy Latner Centre for Palliative Care (Goldman), Sinai Health System; Department of Critical Care Medicine (Pinto), Sunnybrook Health Sciences Centre; Library and Information Services (Englesakis), University Health Network, Toronto General Hospital; Department of Critical Care Medicine (Adhikari) and Sunnybrook Research Institute, Sunnybrook Health Sciences Centre; Interdepartmental Division of Critical Care (Adhikari), University of Toronto, Toronto, Ont.

Contributors: James Downar, Russell Goldman and Neill Adhikari conceived and designed the study. Ruxandra Pinto performed the statistical analyses. All of the authors were involved in data collection and analysis, drafted the manuscript and revised it critically for important intellectual content, approved the final version to be published and agreed to be accountable for all aspects of the work.

Funding: This study was supported by the Toronto General/Toronto Western Hospital Foundation. The funders had no role in the design of the study, interpretation of the results or writing of this manuscript.

Acknowledgements: This research was supported by funding from the Toronto General/Toronto Western Hospital Foundation. James Downar is supported by a Phoenix Fellowship from Associated Medical Services, Inc.

Correspondence to: James Downar, james.downar@utoronto.ca