The current issue and full text archive of this journal is available on Emerald Insight at:
www.emeraldinsight.com/2050-3806.htm

Evaluating physicians' serendipitous knowledge discovery in online discovery systems

A new approach

Mark E. Hopkins
INTEGRIS Health, Oklahoma City, Oklahoma, USA, and
Oksana L. Zavalina
Department of Information Science, University of North Texas, Denton, Texas, USA

Received 15 February 2019
Revised 31 May 2019
4 August 2019
13 August 2019
15 August 2019
19 August 2019
Accepted 28 August 2019

Abstract

Purpose – A new approach to investigate serendipitous knowledge discovery (SKD) of health information is developed and tested to evaluate the information flow-serendipitous knowledge discovery (IF-SKD) model. The purpose of this paper is to determine the degree to which IF-SKD reflects physicians' information behaviour in a clinical setting and explore how the information system, Spark, designed to support physicians' SKD, meets its goals.

Design/methodology/approach – The proposed pre-experimental study design employs an adapted version of the McCay-Peet's (2013) and McCay-Peet *et al.*'s (2015) serendipitous digital environment (SDE) questionnaire research tool to address the complexity associated with defining the way in which SKD is understood and applied in system design. To test the IF-SKD model, the new data analysis approach combining confirmatory factor analysis, data imputation and Monte Carlo simulations was developed.

Findings – The piloting of the proposed novel analysis approach demonstrated that small sample information behaviour survey data can be meaningfully examined using a confirmatory factor analysis technique.

Research limitations/implications – This method allows to improve the reliability in measuring SKD and the generalisability of findings.

Originality/value – This paper makes an original contribution to developing and refining methods and tools of research into information-system-supported serendipitous discovery of information by health providers.

Keywords Health professionals, Information behaviour, Serendipity, Confirmatory factor analysis, Monte Carlo simulations, Questionnaire design

Paper type Research paper

1. Introduction

Due to the growth and complexity of the biomedical literature, as well as the increasingly specialised nature of medicine, there is a need for advanced systems that can quickly present information and assist physicians to discover new knowledge through serendipity. Over the years, the idea of serendipitous knowledge discovery (SKD) – chance, or accidental discovery of new knowledge – has been studied using a variety of methods. This paper presents the data collection and data analysis approaches developed and tested by Dr Mark E. Hopkins at the University of North Texas in 2017–2018 in the dissertation research which sought to address the gaps in the literature and some of the important limitations of research tools and methods used for assessing information seeking behaviour of physicians, and more specifically serendipitous information discovery with the help of specialised

The authors would like to thank the experts who greatly contributed to developing and refining this methodological approach and to implementation of the study in which it was tested: Drs Richard Herrington, T. Elizabeth Workman and LeRoy Southmayd.

Aslib Journal of Information Management
Vol. 71 No. 6, 2019
pp. 755-772

2050-3806
DOI 10.1108/AJIM-02-2019-0045

information systems in a clinical care setting. In this research, SKD has been defined as "the chance, or accidental discovery of new knowledge, where its encountering happens without the expressed or known information of interest at the time of initial searching/browsing" (Hopkins, 2018, p. 7). While an anticipated outcome of this research is a better understanding of the complexity associated with defining and measuring how SKD is operationalised and applied in research, the primary aim is to evaluate whether the Spark information system (discussed in detail next) contributes to physicians' SKD and if the information flow-serendipitous knowledge discovery (IF-SKD) model recently proposed by Workman and colleagues (2014) is a good representation of this type of information behaviour.

2. Literature review

While the concept of serendipity has been present in the literature since the 1960s (e.g. Bernier, 1960), its targeted study has shown enormous growth in the literature in the past 20 years (Erdelez *et al.*, 2016). According to Agarwal (2015, "literature review"), SKD is a logical extension of Wilson's (1999) nested model of information behaviour. Information behaviour refers to the information seeking, information needs, and gaps encountered in information acquisition. Information behaviour models seek to explain how humans interact with information, whether that be in their daily life or through online systems (Case and Given, 2016). There are numerous types of information behaviour models, and several of them focus on the idea of SKD, such as information encountering and accidental information discovery (e.g. Erdelez, 1997; McCay-Peet and Toms, 2010).

Because of serendipity's elusive and unpredictable nature, SKD is challenging to understand within existing information behaviour models (Foster and Ford, 2003). Multiple factors such as age, education, task, personality, information need and prior knowledge influence SKD (Burkell *et al.*, 2012; Heinström, 2006; Spink, 2004). Yet, despite these fundamental complexities, the study of SKD is vital in today's information world. It is particularly important for the health information domain: for example, in a recent review that looked at the opportunities to utilise existing scientific knowledge to assist with the identification of new drugs to treat diseases, and the costs often associated with these endeavours, Prasad *et al.* (2016) noted that serendipity was, and remains, an integral factor in many major drug discoveries.

Physicians' information behaviour research is quite rich. Gorman (1995) identified five pieces of "information used" – data that play a role in physician information behaviour: patient data, population statistics, medical knowledge, logistical information and social influences. He also categorised physicians' information needs as recognised, pursued, satisfied or unrecognised (this unrecognised information need relates to SKD). Studies demonstrate that physicians look for information such as treatment modalities, procedures, equipment and medication (Case and Given, 2016). Capturing how physicians find this type of information and use it is challenging (Chen *et al.*, 2006). For example, in addition to busy and complex routines, physicians were found to face the information overload which occurs when "information received becomes more of a hindrance rather than a help when the information is potentially useful" (Bawden *et al.*, 1999).

The format, presentation, access and modes of using information have changed greatly over the past 30 years, with a strong move towards utilising electronic resources to answer clinical questions. However, studies consistently demonstrate that physicians prefer colleagues and textbooks as sources of information over electronic information resources, and note challenges related to usability and ease of access to the latter (e.g. studies reviewed by Younger, 2010). The narrative nature of physician questions makes it difficult to express their information need as a query that is understandable by information systems. This narrative nature and complexity may be part of the reason why physicians rely on human sources of information (Gorman, 1995, 1999; Clarke *et al.*, 2013).

Prior research has focussed on maximising the breadth of content available to physicians and studying how that content (or system presenting it) was used, and whether access to the information impacted their clinical decision making. For example, several studies examined the use of clinical alerting mechanisms in electronic medical records (EMRs) designed to provide safety precautions for activities such as drug administration, when known potentially harmful issues (e.g. drug interactions) exist. Cimino *et al.* (2002) explored physicians' questions within the EMR workflow to ascertain the situational factors likely to resolve unmet information needs through the implementation of solutions, and Currie *et al.* (2003) collected and categorised the types of unmet information needs of physicians. These studies, however, are all predicated on a known (or anticipated) user information need. Accounting for the unformed and unknown needs (e.g. what Taylor (1968) called visceral need and Wilson and Koepp (1968) referred to as dormant need) of physicians, modelling those needs, and designing new tools and systems for them, is an area needing further exploration.

As the information landscape, its systems and resources continue to grow, there is an increased need to study SKD as a type of information behaviour. Yet, within the field of information science, there are few models specifically focussed on SKD, and within the context of the clinical setting, they are almost non-existent. General models of SKD that have been developed are also relatively new in their application to system design. One major reason for this is the difficulty in measuring the central concept, serendipity. Studies by Erdelez (2004), Björneborn (2008) and McCay-Peet and Toms (2010) illustrate the need to develop and understand the quantitative tools that assist in measuring serendipity and how to relate those to system design. Additional challenges exist in the administration, acquisition and collection of information from physicians engaged in patient care. This is due in part to physicians' routines, which are complex and busy. There is a need to explore the theories and models that can explain, and moreover reinforce, the conceptual framework of SKD. Research in the environments that users, in this case physicians, engage in as part of their normal information behaviour is critical to capturing real world variables that can influence models within the field.

Workman and colleagues (2014) developed an IF-SKD model of information behaviour. The model outlines the stages in the iterative process that ultimately results in knowledge discovery: initial information engagement, visual representation of retrieved information, conceptual short-term memory evaluation, and iterative clarifications or refinements of that searching. Four components derived by the developers of this model from the information science literature underpin the IF-SKD: SKD is an iterative process; SKD often involves change or clarification of information interests the user had initially, which may involve integrating new topics; SKD is grounded in the user's prior knowledge; and information organisation and presentation have fundamental roles in SKD.

Essential to both physicians' information behaviour and the idea of SKD is an understanding of existing information resources and content that comprise the biomedical literature, including the rich taxonomies, metadata and controlled vocabularies contributing to it (e.g. those integrated in the Unified Medical Language System). Within the biomedical information space, there are numerous information resources. The US National Library of Medicine (NLM) has been central to the creation and growth of these online databases and resources (e.g. MedlinePlus, PubMed, ClClinicalTrials.gov, TOXNET) which offer unique and powerful access to information. Years of careful and meaningful curation of underlying data have, in large part, made this possible. However, for many resources, there is the inherent assumption that the user has a goal, or a known information need.

These rich information resources, and their underlying metadata, provide the ideal springboard from which to build new systems that can promote SKD. Through improved system design, the meaningful identification of semantic relationships, and the use of information visualisation, new tools can assist in modelling common for information behaviour

in general and for health information seeking in particular non-linear, iterative information seeking processes. Tools supporting this iterative process in which a user relies on search results to refine information need, reformulate query and continue searching not only improve outcomes, but also reduce "the cognitive demands of information organization" by ultimately increasing the chance for SKD (Workman *et al.*, 2014). New systems should be built to support SKD within the clinical setting. The task of future researchers is to better understand how the design of these systems should be examined. This will allow us to determine how system design reflects the discipline's understanding of SKD as a type of information behaviour. In turn, this helps address another major challenge, which is the growth and specialisation of biomedical information.

Only recently have tools been designed to support SKD for situations where a goal (or information need) is not present, or potentially unknown by the user. In the USA, these efforts are exemplified by the Semantic MEDLINE project of the NLM Lister Hill Center for Biocommunication focussed on identifying and visualising semantic relationships in the biomedical literature to support knowledge discovery. This project led to the development of a new online information discovery system, Spark, a system design that incorporates conceptual short-term memory acknowledges users have limited short-term memory space to make associations between concepts. The aim of Spark application promotes SKD by assisting users in maximising the use potential of their conceptual short-term memory by allowing them to iteratively search for, engage, clarify and evaluate information presented from the biomedical literature. Spark is designed based on the IF-SKD model (Workman *et al.*, 2014) and, particularly, the four core major components of SKD.

The Spark application supports an iterative step process as shown in this brief YouTube demo recorded by the first author of this paper for the purposes of this study (www.youtube.com/watch?v=TpShpHCL3_o). Through an initial search, or topic of interest, the user can refine and visually explore semantic relationships found within the biomedical literature. Core features that make up the Spark application include: Work Space, Graph Presentation and Retrieval Affordance Mechanisms. Work Space is the layout of Spark, in particular, the major left and right pane sections that permit information organisation geared to support the conceptual short-term memory process. This includes the radial connected graph in the left pane and the saved connections of interest in the right pane. Graph Presentation is the structure and visual layout of the results from an information search (the use of colours and lines, as well as graph type). Retrieval Affordance Mechanisms in Spark allow users to adjust the visual presentation of semantic relationships. These mechanisms include frequency of occurrence in the literature (all, common or rare), concept type (disorder, drugs, genes, etc.) and relation type by relation or concept: therapy (e.g. therapy and drugs or chemical), diagnosis, and comorbidity.

2.1 Context: purpose, research questions and results of the broader study

At the present time, there is no understanding of Spark's efficacy to address the goal of its development – supporting SKD. A thorough analysis is needed of Spark's ability, within an actual clinical setting, to promote SKD. Because the purposeful, direct and intentional study of SKD within the information science literature is relatively early in its development, the furthering of new models to explain this behaviour, coupled alongside research tools, is imperative. By studying the IF-SKD model and by analysing Spark, the study methodology presented in this paper provides a better understanding of the use of Spark in promoting SKD within the clinical context.

This research offers an opportunity to narrow, within the information science literature, the gap of how the concept of SKD can be operationalised to study information systems (in this case, the Spark system). Recently developed research instruments have helped capture the concept of SKD in relation to information behaviour models. This work extends understanding of how these new research instruments reflect the operationalised meaning

of SKD and in particular how novel analysis approaches can support the evaluation of these research instruments where the complexity of capturing large sets of user responses is challenging (which is often the case in studying information behaviour of health providers).

While this paper focusses on the methodology, it is important to include here the research questions of the broader study that this methodological approach was developed to address and the answers this analysis provided. These research questions and answers include:

R1. Does Spark successfully contribute to physicians' SKD?

It was determined through frequency analysis that Spark information discovery system does successfully contribute to physicians' SKD:

R2. Does the IF-SKD model reflect physician SKD information behaviour in the clinical setting?

Using confirmatory factor analysis, it was demonstrated that the IF-SKD model was able to reflect physicians' SKD on several fit statistics, however, not on all. Further research is warranted to better understand the relationship between this model and this type of information behaviour.

3. Methods

Our study developed a combination of a questionnaire research instrument and statistical approaches suitable for small sample survey data analysis to explore whether Spark contributes to physicians' SKD and to what degree the IF-SKD model reflects physicians' SKD in a clinical context by capturing and analysing physician feedback. Our study proposed a mapping of the constructs of IF-SKD model to the McCay-Peet's (2013) and McCay-Peet *et al.* (2015) serendipitous digital environment (SDE) questionnaire and Perception of Serendipity Scale (SDE) and adapted the SDE with adjustments based on these mappings for data collection and data analysis. This study evaluated the IF-SKD model using confirmatory factor analysis – "a type of structural equation modelling that deals specifically with measurement models measuring the relationship of factors (concepts) and items (questions or variables)" (Brown, 2015, p. 1). Confirmatory factor analysis procedures used in this study are detailed in Section 3.3.1 of this paper.

It is important to note that the analysis techniques discussed are focussed on describing a novel approach at analysing small sample survey data to arrive at confidence in a broader extrapolation of the findings in the context of studying an information behaviour model with the focus on SKD. It is noteworthy to point out that while much can be learned from these findings, how serendipity is understood and operationalised in different information behaviour models, and measured through surveys (or other methods), will be an ongoing and evolving area of study, that hopefully this study will support.

Studies by Erdelez (2004), Björneborn (2008) and McCay-Peet and Toms (2010) contributed to the development of the research tool employed in this project. We mapped the IF-SKD model's core components to that of the McCay-Peet (2013) and McCay-Peet *et al.* (2015) SDE questionnaire and Perception of Serendipity Scale (SDE) questionnaire. While some recent research, such as Sun *et al.*'s (2011, "research methods and activities") quick diary technique and Jiang *et al.*'s (2018) diary process using critical incident technique, may be a path towards a middle ground between quantitative and qualitative methods, these approaches are not effective for consistent, ongoing, organisational independent data collection, particularly in a clinical setting. Makri and Blandford's (2012) literature review and qualitative analysis reveals that the event, or trigger, for serendipity and the outcome often overlap, which creates a challenge for measuring serendipity, while Dantonio *et al.* (2012) note that serendipity is non-reproducible in a controlled setting. This sentiment reinforces the need to evaluate tools such as the questionnaire employed for this project, despite any limitations it may pose.

This evaluation helps to better understand what aspects of serendipity measurement can withstand cross-organisation use and assist in paving the generalised role serendipity plays in today's information-rich world. The development of the research instrument used here seeks to address the complexity associated with defining the way in which serendipity is understood and applied. This would improve the reliability in measuring serendipity and make findings more generalisable across different settings. It should also help continue to further how the concept of serendipity is understood and operationalised in research methods.

In this pre-experimental design study, feedback on the research instrument was collected using review by a small interdisciplinary group of experts who contributed to the study as dissertation committee members: developers of the biomedical SKD tools, health providers, information scientists and statistics researchers. As there is no known established quantitative approach for measuring SKD in a clinical setting, the following method was developed: a single treatment sample group was provided a video introduction on the use of the Spark, and then asked to complete the research instrument. This method is preferred due to the nature of a clinical setting: physicians' extremely busy work schedule, as well as the challenges associated with the time constraints and accessibility of participants.

Results of the study that developed and tested the new methodological approach reported in this paper are beyond the scope of this methodology-focussed special issue and are reported elsewhere (Hopkins, 2018).

3.1 Data collection process

This research employed participant self-selection as a means of identifying participants for inclusion. Study participants included health providers: physicians, with Doctor of Medicine (MD) or Doctor of Osteopathic Medicine (DO) credentials, working for the INTEGRIS Health system in the state of Oklahoma in the USA. INTEGRIS Health operates numerous hospitals, standalone primary and specialty clinics throughout Oklahoma, as well as specialty facilities, such as Jim Thorpe Rehabilitation, Lakeside Women's Hospital and the INTEGRIS Cancer Institute. In total, 235 physicians had an opportunity to participate, with 23 ultimately responding, representing a 9.78 per cent response rate. The low response rate observed in this study is quite common for studies of physicians' information behaviour (e.g. Cunningham *et al.*, 2015) and illustrates the need for novel approaches to data collection and data analysis that allow to make meaningful conclusions from small samples. The Institutional Review Board of the University of North Texas reviewed and approved the data collection process and research instruments.

The setting for the study, described as the clinical setting, is inclusive of the locations and of the workflows used by the health providers participating in the study and could include a physician's office, the patient's room, the physician's home, the physicians' lounge(s) or other settings. Because workflow surrounding the acquisition of information can differ among participants, the goal was not to assume where an SKD event should occur, but rather understand how physicians' information behaviour in using Spark correlated to the clinical care setting.

An introduction to Spark was provided to participants using a brief, yet meaningful, summary video of Spark being used to explore a medical question. The invitation to participate in the study was distributed to physicians through e-mail and word of mouth. Additionally, we relied on help from the gatekeeper, the Medical Director, Inpatient Informatics for INTEGRIS system, who helped communicate with physicians regarding the opportunity to participate in this research. The questionnaire was administered online using Qualtrics, with the link emailed to participants.

3.2 Research instrument

The research instrument used in this study is a variation of the McCay-Peet (2013) and McCay-Peet *et al.* (2015) 37-item SDE questionnaire and 4-item Perception of Serendipity Scale, available at https://dalspace.library.dal.ca/handle/10222/42727. The SDE questionnaire represents a consolidated pluralistic approach to capturing the presence of serendipity and measuring serendipity by accounting for its varying definitions. In the process of developing the SDE instrument, McCay-Peet conducted two forms of content validity testing on the SDE questionnaire to evaluate the questions, their meaning and wording, and the appropriateness of their facet (or broader factor grouping) assignments. First, a review of the questions and the underlying meaning behind them was performed by experts in the field (McCay-Peet, 2013). In addition to this, McCay-Peet utilised an online survey that asked participants to rate how well an item matched the definition provided of its facet, where the relationship of item-to-facet differed between surveys. McCay-Peet's (2013) analysis of variance to evaluate online survey responses considered items with the highest mean rating and items that had a significantly higher mean rating ($p < 0.05$) on their posited facet (p. 98). This approach provided a mechanism to evaluate how well the proposed item-to-facet relationships could potentially work as a model of information systems' serendipitous characteristics.

The study reported in this paper employed the research instrument presented in the same manner as McCay-Peet's (2013). This research also took a confirmatory factor analysis approach to analyse proposed item-to-facet relationships in consideration of the IF-SKD model to understand how well the models (and questions based on them) represented the information behaviour being studied. It also provided an opportunity to consider, separate from this primary confirmatory factor analysis model fit analysis, the conceptual space of the item questions in the survey and how they relate to how systems' serendipitous characteristics match to broader facets, or components, identified in the research literature Please see the larger study (Hopkins, 2018) for details.

In our study, the IF-SKD model was incorporated in the data collection instrument based on the SDE questionnaire by mapping IF-SKD components to SDE question groupings and individual questions (Tables I–III). As the IF-SKD model is derived from the literature on serendipity and information behaviour, the questionnaire does not reflect the question grouping laid out by the IF-SKD model. An important reason to bring together into a single research instrument the IF-SKD model and the SDE questionnaire is to help reflect serendipity as a process, which is central to the IF-SKD model. This helps, during data analysis, broaden the consideration for any variables that might correlate to the refinement

McCay-Peet *et al.*'s (2015) concepts	Workman *et al.* (2014) IF-SKD model proposed mappings
Enables exploration	Iterative process
	Change/clarification/integration
	Information organisation and presentation have fundamental role
Trigger-rich	Grounded in prior knowledge
Enables connections	Iterative process
	Change/clarification/integration
	Grounded in prior knowledge
Highlights triggers	Grounded in prior knowledge
	Information organisation and presentation have fundamental role
Leads to the unexpected	Iterative process
	Change/clarification/integration
	Grounded in prior knowledge
	Information organisation and presentation have fundamental role

Table I.
IF-SKD concept mappings to SDE questionnaire

SDE questions	SDE question mapping to IF-SKD component
Enables exploration: a user's assessment of the degree to which a digital environment supports exploration and examination of its information, ideas or resources (A)	[E6]: 2 [E7]: 2 [E8]: 2 [E9]: 2
1. E1: it is easy to explore [the digital environment]'s content	
2. E2: [The digital environment] supports exploration	
3. E3: it is easy to wander around in [the digital environment]	
4. E6: there are many ways to explore information in [the digital environment]	
5. E7: [The digital environment] invites examination of its content	
6. E8: [The digital environment] is an instrument for discovery	
7. E9: [The digital environment] is a tool for exploration	
Trigger-rich: a user's assessment of the degree to which a digital environment contains a variety of information, ideas or resources that is interesting and useful to the user (B)	[T5]: 3 [T6]: 3
8. T1: the content contained in [the digital environment] is diverse	
9. T2: [The digital environment] is rich with interesting ideas	
10. T3: the digital environment] offers a wide variety of information	
11. T4: there is a depth of information in [the digital environment]	
12. T5: [The digital environment] is full of information useful to me	
13. T6: I find information of value to me in [the digital environment]	
14. T7: [The digital environment] is a treasure trove of information	
Enables connections: a user's assessment of the degree to which a digital environment makes relationships or connections between information, ideas or resources apparent (C)	[C1]: 1 [C2]: 1
15. C1: [The digital environment] enables me to make connections between ideas	[C3]: 4
16. C2: associations between ideas become obvious in [the digital environment]	[C4]: 4
17. C3: I can see connections between topics in [the digital environment]	[C6]: 3
18. C4: it is easy to see links between information in [the digital environment]	[C8]: 4
19. C6: I make useful connections in [the digital environment]	[C9]: 3
20. C8: the features of [the digital environment] help me see connections between its content	
21. C9: I come to understand relationships between ideas in [the digital environment]	
Highlights triggers: a user's assessment of the degree to which a digital environment brings interesting and useful information, ideas or resources to the user's attention (D)	[H2]: 4 [H3]: 4
22. H1: I am directed towards valuable information in [the digital environment]	[H4]: 4
23 H2: [The digital environment] has features that ensure that my attention is drawn to useful information	[H5]: 4 [H7]: 4
24. H3: information that interests me is highlighted in [the digital environment]	[H9]: 4
25. H4: the way that [the digital environment] presents content captures my attention	[H10]: 4
26. H5: I am alerted to information in [the digital environment] that helps me	
27. H7: I notice content I would not normally pay attention to in [the digital environment]	
28. H8: [The digital environment] has features that draw my attention to information	
29. H9: I am pointed towards content in [the digital environment]	
30. H10: [The digital environment] has features that alert me to information	
Leads to the unexpected: a user's assessment of the degree to which a digital environment provides opportunities for unexpected interactions with information, ideas or resources (E)	[U1]: 1 [U2]: 2 [U3]: 3
31. U1: I bump into unexpected content in [the digital environment]	[U6]: 1
32. U2: I encounter the unexpected in [the digital environment]	[U7]: 1
33. U3: I am surprised by what I find in [the digital environment]	
34. U4: I come across topics by chance in [the digital environment]	
35. U5: [The digital environment] exposes me to unfamiliar information	
36. U6: my interactions in [the digital environment] are unexpectedly valuable	
37. U7: I stumble upon information in [the digital environment]	

Table II.
IF-SKD individual question concept mappings

SDE instrument questions grouped by IF-SKD model components	Resulting questions in the SDE instrument adaptation for this study
SKD is an iterative process (1) C1: [The digital environment] enables me to make connections between ideas C2: associations between ideas become obvious in [the digital environment] U1: I bump into unexpected content in [the digital environment] U6: my interactions in [the digital environment] are unexpectedly valuable U7: I stumble upon information in [the digital environment]	SKD is an iterative process (1) C1: Spark enables me to make connections between ideas U1: I bump into unexpected content in Spark U6: my interactions in Spark are unexpectedly valuable U7: I stumble upon information in Spark
SKD often involves change or clarification of initial information interests, which may involve integrating new topics (2) E6: there are many ways to explore information in [the digital environment] E7: [The digital environment] invites examination of its content E8: [The digital environment] is an instrument for discovery E9: [The digital environment] is a tool for exploration U2: I encounter the unexpected in [the digital environment]	SKD often involves change or clarification of initial information interests, which may involve integrating new topics (2) E6: there are many ways to explore information in Spark E7: Spark invites examination of its content E8: Spark is an instrument for discovery E9: Spark is a tool for exploration U2: I encounter the unexpected in Spark
SKD is grounded in the user's prior knowledge (3) T5: [The digital environment] is full of information useful to me T6: I find information of value to me in [the digital environment] C6: I make useful connections in [the digital environment] C9: I come to understand relationships between ideas in [the digital environment] U3: I am surprised by what I find in [the digital environment]	SKD is grounded in the user's prior knowledge (3) T5: Spark is full of information useful to me T6: I find information of value to me in Spark C9: I come to understand relationships between ideas in Spark U3: I am surprised by what I find in Spark
Information organisation and presentation have fundamental roles (4) C3: I can see connections between topics in [the digital environment] C4: it is easy to see links between information in [the digital environment] C8: the features of [the digital environment] help me see connections between its content H3: information that interests me is highlighted in [the digital environment] H4: the way that [the digital environment] presents content captures my attention H7: I notice content I would not normally pay attention to in [the digital environment] H2: Spark has features that ensure that my attention is drawn to useful information H9: I am pointed towards content in [the digital environment] H5: I am alerted to information in [the digital environment] that helps me H10 [The digital environment] has features that alert me to information	Information organisation and presentation have fundamental roles (4) C3: I can see connections between topics in Spark C4: it is easy to see links between information in Spark H3: information that interests me is highlighted in Spark H4: the way that Spark presents content captures my attention H7: I notice content I would not normally pay attention to in Spark H2: Spark has features that ensure that my attention is drawn to useful information H9: I am pointed towards content in Spark H5: I am alerted to information in Spark that helps me

Table III.
Questions grouped by proposed IF-SKD mappings and resulting 21-item questionnaire

and understanding of the core meaning of serendipity as used throughout the questionnaire. This also assists in better understanding what characteristics influence the concept of serendipity in the clinical setting.

Table I reflects high-level conceptual mappings of the IF-SKD model to each of the SDE questionnaire groupings from "enables exploration" to "leads to unexpected". For example, all four major components of IF-SKD model exhibit semantic similarity with "leads to the unexpected" concept of SDE but only two of them overlap in meaning with the "highlights triggers" SDE concept.

Table II presents the specific SDE questionnaire items mapped to the IF-SKD model.

The following key is used for the IF-SKD model specified in the right column of the table:

(1) iterative process;

(2) change/clarification/integration;

(3) grounded in prior knowledge; and

(4) information organisation and presentation have fundamental role.

Table III shows original questions on the McCay-Peet's SDE Questionnaire according to the IF-SKD groupings proposed by this study in the left-hand column. In the right-hand column of this table, the final 21-item SDE questionnaire used in the study is shown. As part of the expert review, to help overcome the challenge of physicians' information overload, four questions were removed, leaving 21 final questions that were used in the study and could be completed by participants within 15 min.

3.3 Data analysis

Components of our data analysis approach were carefully designed to support two goals: exploring the information seeking behaviour of physicians using the Spark system developed to support SKD in a clinical context, and evaluation of the efficiency of the proposed way to use the IF-SKD model as part of the research instrument. In addition to descriptive statistics and frequency analysis, our study employed confirmatory factor analysis for both the SDE questionnaire's original groupings and the proposed mappings to the IF-SKD model. Data were analysed using RStudio and various R packages, along with SPSS Statistics. This technique allowed us to explore earlier studied and explored theories, as well as identified relationships within the literature, from an *a priori* perspective. This makes possible the examination of latent constructs to determine appropriateness of fit with respect to the IF-SKD model. Results from the questionnaire used in this study support the evaluation of SDE-to-IF-SKD mappings, but also determine how well the questionnaire captures the meaning and significance of serendipity and the aspects of it that contribute to system design. In addition to answering the core research questions of the study for which the new methodological approach reported in this paper was developed, one of the goals of this analysis was to determine in what ways the questionnaire could be improved in the future. The analysis of the IF-SKD mappings may help present valuable insights into refinements to better capture the meaning of serendipity, as well as improve its utility within the clinical setting.

Below is an overview of our data analysis approach as well as discussion on sample size and methods utilised to enhance the existing data to support the data analysis. An overview of the software packages and processes used to conduct the confirmatory factor analysis are presented. For each model, the same confirmatory factor analysis process, fit statistics and output were analysed to support individual and between model comparisons.

3.3.1 Confirmatory factor analysis overview and the SDE questionnaire. Confirmatory factor analysis of the SDE questionnaire included evaluating how well the three different

models represent the data collected from physicians as part of this study. The confirmatory factor analysis approach is especially useful when the overall study of a topic has a strong conceptual underpinning and initial efforts to measure it are in the early development stages. As Brown (2015, p. 1) has stated, confirmatory factor analysis "is almost always used during the process of scale development to examine the latent structure of a test instrument (e.g. a questionnaire)". Work by McCay-Peet (2013) in evaluating a serendipity data collection instrument using exploratory factor analysis was a precursor to the use of a confirmatory factor analysis in this study. McCay-Peet's (2013) work pointed towards a likely four-factor model, though a five-factor model was proposed. In effect, this approach allows for the evaluation of the second research hypothesis, of whether the IF-SKD model reflects physicians' SKD in a clinical setting.

This section focusses on the presentation of the models analysed using confirmatory factor analysis and delves into each model's fit statistics to help evaluate them. Moreover, these findings are evaluated in consideration of the SDE questionnaire to assess how well the questions capture aspects of serendipity among respondents and how effectively the instrument performed.

The overall process performed to support the data analysis is captured in Figure 1 and explained next.

Due to the fact that some of the participant survey responses contained missing data, additional steps were necessary to allow for an effective set of confirmatory factor analyses. This required that an estimated population be generated following data imputation. Data imputation involves an estimation of the raw data set to approximate what values should be selected to replace missing data.

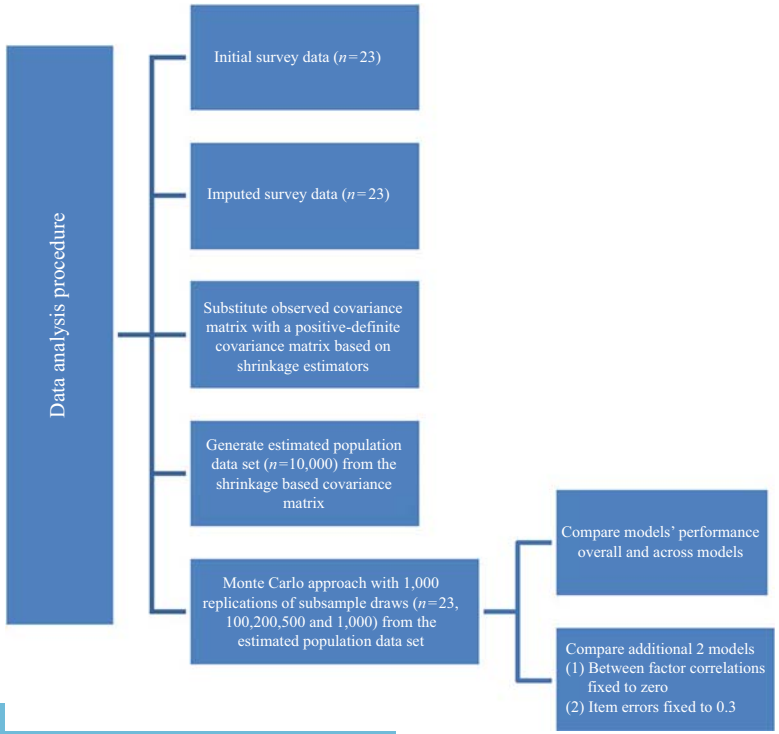

Figure 1.
Data analysis approach

Below we outline the tools, strategy and mathematical approaches used to arrive at a final data set that could be studied with the proposed models, it is important to first point out that all these steps were not undertaken simply to get the data in a functionally usable state. Rather, the literature has supported the use of this approach in producing viable data for this type of analysis. Specifically, Krinsky and Robb (1986) demonstrated that the use of Monte Carlo simulations to describe the mean and variances of a random sample was as effective as other methods at representing reliable standard errors. The *missForest* R package was chosen for the data imputation task because it handles "categorical data including complex interactions and nonlinear relations" (Stekhoven and Buhlmann, 2012). The bootstrap approach was not selected because it is not recommended with sample sizes that are less than 200 (Nevitt and Hancock, 2001). The data imputation resulted in a new data set with statistics for the SDE questionnaire.

The following steps were used specifically for the data set related to the SDE questionnaire. In addition to data imputation using *missForest*, the R package *corpcor* was used to assist with the MASS package in creating the estimation population data by ensuring that the covariance matrix used as input with the MASS function was positive definite. This process ensured that subsequent samples drawn from that population would generate a positive definite covariance matrix by *lavaan* when performing the confirmatory factor analysis. More specifically, *corpcor* performs the following steps:

(1) Each random variable's empirical variance is calculated and shrunken towards the mean.

(2) The shrinkage intensity is then computed using the following formula by Opgen-Rhein and Strimmer (2007):

$$\lambda^*_{var}\left(\sum_{k=1}^{p}\mathrm{Var}(s_{kk})\right)/\sum_{k=1}^{p}(s_{kk}-\mathrm{median}(s))^2.$$

In the formula, the median refers to the median of the empirical variances.

(3) The covariance matrix shrinkage is calculated towards the identity matrix using the following formula by Schäfer and Strimmer (2005):

$$\lambda^* = \sum_{k\neq l}\mathrm{Var}(r_{kl})/\sum_{k\neq l}r^2_{kl}$$

It is not possible to always have an ideal sample from which to run a set of statistics. Regularisation is intended to minimise the variance in the small imputed data set so that the implied covariance matrix produced is still representative of the underlying data, and capable of being analysed in a confirmatory factor analysis framework. The concept of regularisation within the literature has taken different forms and matured over time to account for different types of data, such as normal vs non-normal. Ridge regression is one of the ways regularisation has been employed. For example, "in the case of severe multicollinearity in a regression model, without imposing a bit of bias on the regression coefficient estimates via ridge regression, it would be impossible to obtain estimates of these coefficients" (Mooney and Duval, 1993, p. 44). Another way to envision regularisation is as a process whereby additional new information is introduced in an effort to address an ill-posed question (Neumaier, 1998).

Evaluating the least impactful approach to regularising data to support the goals of our research, within realistic bounds of interpretation, is the goal of regularisation. For our research, due to the low sample size, data imputation, along with covariance shrinkage, was used to obtain an estimated population ($n = 10{,}000$) from which Monte Carlo simulation of

subsamples were drawn and then fitted to each model to support fit statistic comparisons and to understand the changes of sample size occurring on each of the models. Tofighi and MacKinnon (2016) noted that while there are different approaches to performing summary analysis in structural equation modelling, "the Monte Carlo method produces more accurate results especially for smaller sample sizes" (p. 194).

The rationale underlying the use of the Monte Carlo method in this study is to generate many Monte Carlo replications (e.g. 1,000 replications) of subsample size draws of $n = 23$, 100, 200, 500 and 1,000 from the estimated populations. This allows evaluation of confirmatory factor analysis models, and fit statistics, as the sample size increases. Moreover, this method allows valid estimates of standard errors for factor loadings and factor correlations for the original small sample size of $n = 23$. The diagonally weighted least squares (DWLS) parameter estimation method is used in combination with the Monte Carlo simulations to estimate the confirmatory factor analysis models. Essentially, this research utilises, in order to deal with the small sample size problem, Monte Carlo-based DWLS parameter estimation, utilising shrinkage estimators for the observed covariance matrix, referred to here as MC-SDWLS.

Marcoulides and Saunders (2006) discuss the use of Monte Carlo analysis in two different ways: proactive and reactive. The former, while identified as preferable and more likely to produce valid confidence intervals, is not always easy to conduct since the information about the entire population may not be known. Instead, this study used Monte Carlo in a reactive way to allows for comparison of the fit indices of the models with more confidence. By conducting a thousand iteration runs of varying sample sizes, researchers can see at what level of sample size one begins to assess meaningful information about the sampling error. Marcoulides and Saunders (2006) proposed doing this through the use of a *t*-test to assess the significance of one statistic between models.

Using a simulated population model to further validate the MC-SDWLS method utilised, a Monte Carlo simulation was performed with the McCay-Peet model as the known true model that generates the population data. Observed shrunken covariance matrix (with $n = 23$) was generated from the McCay-Peet population model. The simulation was accomplished using the function *simulateData* within the R package *lavaan*. Specifically, the shrunken observed covariance matrix, based on the imputed data set of $n = 23$, was used in conjunction with the unconstrained McCay-Peet model as the true population model. Thus, a 10,000-record data set was created to represent the population under the McCay-Peet model.

Using these population data, the MC-SDWLS simulation was performed to estimate winsorised mean point estimates, winsorised mean fit statistics and standard errors for these point estimates and fit statistics, for both the McCay-Peet model and the IF-SKD model. A two-sample *t*-test was performed between the McCay-Peet and the IF-SKD model, using the mean *Fmins*, across Monte Carlo replications, and standard errors obtained from these Monte Carlo replications (using 1,000 Monte Carlo replications). As expected, the *t*-test statistically significantly favoured the McCay-Peet model, which was actually the true generating model, when compared with the IF-SKD model.

The *Fmin* is the objective function that is minimised during optimisation of the *lavaan* confirmatory factor analysis model. When the data are drawn from a multivariate normal distribution, minimising the *Fmin* (the difference between the observed and implied covariance matrix) also minimises the so-called Kullback–Liebler divergence. Wang and Jo (2013, p. 409) explained that the Kullback–Liebler divergence "can be viewed as a measure of the information loss in the fitted model relative to that in the reference model". This fact motivates the use of the *Fmin* statistic as a way of discriminating the relative differences in goodness of fit between each respective model assumed generating population model. Consequently, a *t*-test of statistical significance between the *Fmin* of two different models can determine which model is better at approximating the respective reference models.

In summary, in the situation of the Monte Carlo simulation with a known population structure, having the statistical test on the *Fmin* objective function values favour the McCay-Peet model fit over the IF-SKD fit, when the true generating model was the McCay-Peet mode. This provides some confidence that the MC-SDWLS methodology developed as part of this study can work to select a best approximating model, in a relative sense (as opposed to an absolute goodness of fit).

3.3.2 Confirmatory factor analysis technique for all models. In this section, the technique, measurements and evaluation criteria used for each model are presented along with justification for these approaches as outlined in the literature, based upon the research instrument and the stated goals of the research. Before discussing the fit statistics and interpretation guidance criteria for this study, the estimation method used to conduct the confirmatory factor analysis must be addressed. There is an array of different estimation methods for conducting a confirmatory factor analysis. Maximum Likelihood (ML) is a common and effective method focussed on the analysis of continuous data, influenced by low sample sizes (e.g. Brown, 2015). Other estimation methods include: generalised least squares; weighted least squares; DWLS, sometimes also referred to by the acronym WLSMV; unweighted least squares (ULS); and variants, including robust ML and ML with different standard error reporting. ML, which is considered one of the better confirmatory factor analysis estimator methods, suffers from small sample sizes (Brown, 2015).

Of all the estimation approaches available, the DWLS was chosen. Li (2016) utilised a Monte Carlo approach to evaluate DWLS, ULS and Robust ML under a variety of different ordinal data conditions and distributional shapes. Li (2016, p. 369) showed that DWLS performed best, especially in accounting for the factor loading and in producing "more accurate inter-factor correlation estimates". Using a diagonally weighted matrix, as opposed to an inverse matrix, in computing fit statistics, DWLS allows for easier comparison for small sample sizes and handles well with non-normal data (Rhemtulla *et al.*, 2012). Marsh and Grayson (1995) summarised the decision to choose an approach well, stating that "a general approach is to establish that the model is identified, that the iterative estimation procedure converges, that all parameter estimates are within the range of permissible values, and that the standard errors of the parameter estimates have reasonable size" (p. 198). Selecting DWLS and evaluating the models relative to each other, while also looking at the corrected fit indices, allows for rich analysis and comparison on a variety of different fronts, which is a goal for this type of analysis.

To improve the understanding of the significance of these estimates for each specific model and also between the models, the confidence intervals, point estimates, standardised point estimates and percentiles (2.5 and 97.5 per cent) are calculated. The calculations are performed across all the samples and averaged to provide information about the 1,000 simulations for each model. Tofighi and MacKinnon (2016) found the Monte Carlo approach to evaluating results was an effective way to draw on the law of large numbers to evaluate these statistics, further finding that the Monte Carlo approach was as effective as bootstrapping or alternative methods. The reason the percentiles are evaluated is to determine if the distribution of the data is non-normal, which helps provide a better conservative indication of the upper and lower bounds of likely values for any specific model. This helps demonstrate what type of fit is represented by the numbers that are 2.5 and 97.5 per cent underneath the distribution curve.

A summary of the fit statistics, compiled by work from Schreiber *et al.* (2006), is available at www.tandfonline.com/doi/abs/10.3200/JOER.99.6.323-338. This was used to guide the interpretation of the results from the study. In addition, the fit statistics are grouped into categories, type of fit statistic, highlighting their value in interpreting the findings in this study, as well as areas where they are impacted by limitations of the study. While there are specific cut-offs listed, the approach taken in this analysis is to evaluate each model against the other, in addition to looking at its overall score on certain indices. This allowed for the

evaluation of the second null hypothesis, asking whether the IF-SKD model reflects physicians' SKD in a clinical setting, while considering its score in comparison to other proposed models. This approach also allowed for a more generalisable interpretation that support calls for future research.

4. Discussion and implications

There are multiple novel aspects in the research approach reported in this paper. This research presented the second application of a relatively new research instrument, the Perception of Serendipity and SDE (McCay-Peet, 2013; McCay-Peet *et al.*, 2015) questionnaire. We developed the research tool based on the SDE and studies by Erdelez (2004) and Björneborn (2008). Our research was the first to assess the SDE instrument using confirmatory factor analysis. In addition, the study of multiple confirmatory factor analysis models has helped provide broader context regarding the application of the indicators to the proposed models' structures.

Another contribution is the novel data analysis approach, with the focus on utilising small sample size to conduct confirmatory factor analysis. This approach allows us to derive statistically meaningful results from a small sample of questionnaire responses and thus helps address the common problem in physicians' information behaviour research: the extremely busy work schedules and information overload resulting in low participation rate in information behaviour studies. The methods undertaken to successfully analyse these data presented meaningful statistical metrics to compare one model to another, which offers insight into how future analysis can be conducted when small samples are encountered. This is especially useful in the study of serendipity and the application of a research instrument such as the SDE questionnaire which is relatively lengthy, depending on the audience to which it is posed. Moreover, the use of the Omega coefficient to test generalisability and reliability of the SDE questionnaire on a four and five-factor model is an important finding in this study, particularly given the small sample size.

There are several limitations to the method proposed in this paper that researchers designing a study of health providers' information behaviour need to keep in mind. Potentially unknown environmental factors such as interruptions due to patient care, participants' technology familiarity, and the study duration, could influence results (Bawden *et al.*, 1999). When conducting analyses within a context that considers system design aspects and underlying assumptions governing the model, other salient influencing variables could be missed. While an enhanced understanding of how to operationalise the concept of serendipity, and better measure it, were the products of our study, the concept of serendipity itself remains challenging to convey and measure in practice (Foster and Ford, 2003; Dantonio *et al.*, 2012; Makri and Blandford, 2012). Through analysis of the research methods and instruments used, including their ability to successfully measure SKD, improvements to future research could be possible.

Our research potentially has broader applications beyond developing methodological approaches for small sample survey data analysis discussed above. For example, relevant findings from this study and other studies utilising the proposed method could later be incorporated into the development of new research tools and avenues for future research (e.g. confirmatory factor analysis studies with nurses and physician's assistants to evaluate proposed models; multi-site comparative studies of perception of physicians' serendipitous information discovery support in different clinical settings – immediate care, family medicine and specialty care). Additionally, the application of the IF-SKD model to system design is significant and an area that warrants future research and discussion. As a reflective model of SKD, the model can serve as a springboard for the future development of various information systems, in the medical field and in other domains, and future studies can further test the reliability of the model in being able to support SKD.

Overall, this research demonstrates two principal outcomes. First, it allows for the assessment of Spark, a new online information resource designed to engage users and

promote SKD and the associated IF-SKD information behaviour model (Workman *et al.*, 2014). Second, this study presents a novel approach to data analysis, with the aim of improving the field's overall understanding of how SKD is operationalised and to demonstrate effective ways that can support an evaluation of research instruments where access to data is limited or difficult to acquire due to the complexity of this type of information behaviour. In line with the special issue's focus on novel methodology, this paper maintains focus on the second principal outcome of the research project.

References

Agarwal, N.K. (2015), "Towards a definition of serendipity in information behavior", *Information Research*, Vol. 20 No. 2, p. 675, available at: www.informationr.net/ir/20-3/paper675.html (accessed 14 August 2019).

Bawden, D., Holtham, C. and Courtney, N. (1999), "Perspectives on information overload", *Aslib Proceedings*, Vol. 51 No. 8, pp. 249-255.

Bernier, C.L. (1960), "Correlative indexes VI: serendipity, suggestiveness, and display", *Journal of Documentation*, Vol. 11 No. 4, pp. 277-287.

Björneborn, L. (2008), "Serendipity dimensions and users' information behaviour in the physical library interface", *Information Research*, Vol. 13 No. 4, p. 370, available at: www.informationr.net/ir/13-4/paper370.html (accessed 14 August 2019).

Brown, T.A. (2015), *Confirmatory Factor Analysis for Applied Research*, 2nd ed., The Guilford Press, New York, NY, 462pp.

Burkell, J., Quan-Haase, A. and Rubin, V.L. (2012), "Promoting serendipity online: recommendations for tool design", *Proceedings of the 2012 iConference, ACM, Toronto*, pp. 525-526.

Case, D.O. and Given, L.M. (2016), *Looking for Information: A Survey of Research on Information Seeking, Needs, and Behavior*, 4th ed., Emerald, Bingley.

Chen, E.S., Bakken, S., Currie, L.M., Patel, V.L. and Cimino, J.J. (2006), "An automated approach to studying health resource and infobutton use", *Studies in Health Technology and Informatics*, Vol. 122 No. 2, pp. 273-278.

Cimino, J.J., Li, J., Bakken, S. and Patel, V.L. (2002), "Theoretical, empirical and practical approaches to resolving the unmet information needs of clinical information system users", *AMIA Annual Symposium Proceedings, American Medical Informatics Association, Bethesda, MD*, pp. 170-174.

Clarke, M.A., Belden, J.L., Koopman, R.J., Steege, L.M., Moore, J.L., Canfield, S.M. and Kim, M.S. (2013), "Information needs and information-seeking behaviour analysis of primary care physicians and nurses: a literature review", *Health Information and Libraries Journal*, Vol. 30 No. 3, pp. 178-190.

Cunningham, C.T., Quan, H., Hammelgarn, B., Noseworthy, T., Beck, C.A., Bixon, E., Samuel, S., Ghali, W.A., Sykes, L.L. and Jette, N. (2015), "Exploring physician specialist response rates to web-based surveys", *BMC Medical Research Methodology*, Vol. 15 No. 1, pp. 1-8, available at: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-015-0016-z (accessed 14 August 2019).

Currie, L.M., Graham, M., Allen, M., Bakken, S., Patel, V. and Cimino, J.J. (2003), "Clinical information needs in context: an observational study of clinicians while using a clinical information system", *AMIA Annual Symposium Proceedings, American Medical Informatics Association, Bethesda, MD*, pp. 190-194.

Dantonio, L., Makri, S. and Blandford, A. (2012), "Coming across academic social media content serendipitously", *Proceedings of the American Society for Information Science and Technology*, Vol. 49 No. 1, pp. 1-10.

Erdelez, S. (1997), "Information encountering: a conceptual framework for accidental information discovery", *Proceedings of an International Conference on Information Seeking in Context, Taylor Graham, Tampere*, pp. 412-421, available at: https://dl.acm.org/citation.cfm?id=267217 (assessed 14 August 2019).

Erdelez, S. (2004), "Investigation of information encountering in the controlled research environment", *Information Processing & Management*, Vol. 40 No. 6, pp. 1013-1025.

Erdelez, S., Beheshti, J., Heinström, J., Toms, E., Makri, S., Agarwal, N.K. and Björneborn, L. (2016), "Research perspectives on serendipity and information encountering", *Proceedings of the 79th ASIS&T Annual Meeting: Creating Knowledge, Enhancing Lives through Information & Technology*, *Copenhagen*, *14-18 October*, available at: https://asistdl.pericles-prod.literatumonline.com/doi/full/10.1002/pra2.2016.14505301011 (accessed 14 August 2019).

Foster, A. and Ford, N. (2003), "Serendipity and information seeking: an empirical study", *Journal of Documentation*, Vol. 59 No. 3, pp. 321-340.

Gorman, P. (1999), "Information seeking of primary care physicians: conceptual models and empirical studies", in Wilson, T.D. and Allen, D.K. (Eds), *Information Behaviour: Proceedings of the Second International Conference on Research in Information Needs, Seeking and Use in Different Contexts*, Taylor Graham, Sheffield, pp. 226-240.

Gorman, P.N. (1995), "Information needs of physicians", *Journal of the American Society for Information Science*, Vol. 46 No. 10, pp. 729-736.

Heinström, J. (2006), "Psychological factors behind incidental information acquisition", *Library & Information Science Research*, Vol. 28 No. 4, pp. 579-594.

Hopkins, M. (2018), "A study of physicians' serendipitous knowledge discovery: an evaluation of Spark and the IF-SKD model in a clinical setting", PhD thesis, University of North Texas, available at: https://digital.library.unt.edu/ark:/67531/metadc1157586/ (accessed 14 August 2019).

Jiang, T., Zhang, C., Li, Z., Fan, C. and Yang, J. (2018), "Information encountering on social Q&A sites: a diary study of the process", *Proceedings of the 13th International Conference, iConference 2018, Sheffield, Springer*, *New York, NY*, pp. 476-486.

Krinsky, I. and Robb, A.L. (1986), "On approximating the statistical properties of elasticities", *The Review of Economics and Statistics*, Vol. 68 No. 4, pp. 715-719.

Li, C.H. (2016), "The performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables", *Psychological Methods*, Vol. 21 No. 3, pp. 369-387.

McCay-Peet, L. (2013), Investigating work-related serendipity, what influences it, and how it may be facilitated in digital environments" (Doctor of Philosophy), Dalhousie University, Halifax, available at: https://dalspace.library.dal.ca/handle/10222/42727 (accessed 14 August 2019).

McCay-Peet, L. and Toms, E.G. (2010), "The process of serendipity in knowledge work", *Proceedings of the Third Symposium on Information Interaction in Context, ACM*, *New Brunswick, NJ*, pp. 377-382, available at: https://dl.acm.org/citation.cfm?id=1840842&dl=ACM&coll=DL (accessed 14 August 2019).

McCay-Peet, L., Toms, E.G. and Kelloway, K.E. (2015), "Examination of relationships among serendipity, the environment, and individual differences", *Information Processing & Management*, Vol. 51 No. 4, pp. 391-412.

Makri, S. and Blandford, A. (2012), "Coming across information serendipitously", *Journal of Documentation*, Vol. 68 No. 5, pp. 684-724.

Marcoulides, G.A. and Saunders, C. (2006), "Editor's comments: PLS: a silver bullet?", *Management Information Systems Quarterly*, Vol. 30 No. 2, pp. iii-ix.

Marsh, H.W. and Grayson, D. (1995), "Latent variable models of multitrait-multimethod data", in Hoyle, R.H. (Ed.), *Structural Equation Modeling: Concepts, Issues, and applications*, Sage, Thousand Oaks, CA, pp. 177-198.

Mooney, C. and Duval, R. (1993), *Bootstrapping: A Nonparametric Approach to Statistical Inference*, Sage, Newbury Park, CA.

Neumaier, A. (1998), "Solving ill-conditioned and singular linear systems: a tutorial on regularization", *Society for Industrial and Applied Mathematics Review*, Vol. 40 No. 3, pp. 636-666.

Nevitt, J. and Hancock, G.R. (2001), "Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling", *Structural Equation Modeling: A Multidisciplinary Journal*, Vol. 8 No. 3, pp. 353-377.

Opgen-Rhein, R. and Strimmer, K. (2007), "Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach", *Statistical Applications in Genetics and Molecular Biology*, Vol. 6 No. 1, pp. 1-20, available at: https://pdfs.semanticscholar.org/8ead/8c271bd27d93797809d43e1e32848be8a304.pdf (accessed 14 August 2019).

Prasad, S., Gupta, S.C. and Agarwal, B.B. (2016), "Serendipity in cancer drug discovery: rational or coincidence?", *Trends in Pharmacological Sciences*, Vol. 37 No. 6, pp. 435-450.

Rhemtulla, M., Brosseau-Liard, P.É. and Savalei, V. (2012), "When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions", *Psychological Methods*, Vol. 17 No. 3, pp. 354-373.

Schäfer, J. and Strimmer, K. (2005), "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics", *Statistical Applications in Genetics and Molecular Biology*, Vol. 4 No. 1, pp. 1-30, available at: www.strimmerlab.org/publications/journals/shrinkcov2005.pdf (accessed 14 August 2019).

Schreiber, J.B., Nora, A., Stage, F.K., Barlow, E.A. and King, J. (2006), "Reporting structural equation modeling and confirmatory factor analysis results: a review", *Journal of Educational Research*, Vol. 99 No. 6, pp. 323-338.

Spink, A. (2004), "Multitasking information behavior and information task switching: an exploratory study", *Journal of Documentation*, Vol. 60 No. 4, pp. 336-351.

Stekhoven, D.J. and Buhlmann, P. (2012), "MissForest – non-parametric missing value imputation for mixed-type data", *Bioinformatics*, Vol. 28 No. 1, pp. 112-118.

Sun, X., Sharples, S. and Makri, S. (2011), "A user-centered mobile diary study approach to understanding serendipity in information research", *Information Research*, Vol. 16 No. 3, p. 492, available at: www.informationr.net/ir/16-3/paper492.html (accessed 14 August 2019).

Taylor, R.S. (1968), "Question negotiation and information seeking in libraries", *College & Research Libraries*, Vol. 29 No. 3, pp. 178-194.

Tofighi, D. and MacKinnon, D.P. (2016), "Monte Carlo confidence intervals for complex functions of indirect effects", *Structural Equation Modeling: A Multidisciplinary Journal*, Vol. 23 No. 2, pp. 194-205.

Wang, C.-P. and Jo, B. (2013), "Applications of a Kullback-Leibler divergence for comparing non-nested models", *Statistical Modelling*, Vol. 13 Nos 5-6, pp. 409-429.

Wilson, P. and Koepp, D.W. (1968), *Two Kinds of Power: An Essay on Bibliographical Control*, The University of California Press, Berkeley, CA.

Wilson, T.D. (1999), "Models in information behaviour research", *Journal of Documentation*, Vol. 55 No. 3, pp. 249-270.

Workman, T.E., Fiszman, M., Rindflesch, T.C. and Nahl, D. (2014), "Framing serendipitous information seeking behavior for facilitating literature-based discovery: a proposed model", *Journal of the Association for Information Science and Technology*, Vol. 65 No. 3, pp. 501-512.

Younger, P. (2010), "Internet-based information-seeking behaviour amongst doctors and nurses: a short review of the literature", *Health Information & Libraries Journal*, Vol. 27 No. 1, pp. 2-10.

Further reading

Box, G.E.P. and Draper, N.R. (1987), *Empirical Model-Building and Response Surfaces*, Wiley, New York, NY.

Corresponding author
Oksana L. Zavalina can be contacted at: oksana.zavalina@unt.edu