Decision-making under uncertainty is a difficult and unavoidable challenge in clinical contexts. Technologies such as probabilistic programming languages (PPLs) allow their users to explicitly model and reason with uncertainty. By taking a user-centric approach to the deployment of these technologies, we believe there is an opportunity to involve clinicians in the modelling process. In this paper, we present a field study of decisions taken to manage postoperative bleeding. From analysis of the findings, we outline three central themes that emerge and discuss implications for design, developing a set of evaluative design principles to assess a PPL-based tool in this context. These include visualising zones of optimal intervention, surfacing relative risk trade-offs between teams, and accessing specialist views within a holistic picture. These findings provide a structure for critically exploring PPL-based tools to support clinical reasoning under uncertainty.

1. INTRODUCTION

Decision-making under uncertainty is discussed widely in the clinical literature with calls to include it rigorously in medical education and training programs (Cooke and Lemay 2017) and arguments for greater explicit acknowledgement of uncertainty in clinical practice (Dunlop and Schwartzstein 2020). Some argue that a proclivity to seek certainty is embedded in medical culture and can lead to excessive testing, among other adverse consequences (Simpkin and Schwartzstein 2016).

One way of addressing this challenge of reasoning with uncertainty in clinical practice is by designing decision support tools (DSTs). Clinical DSTs are computational tools that are designed to augment decision-making in one of three main tasks: diagnosis, treatment or prognosis (Yang et al. 2016).

As a basis for a DST for decision-making under uncertainty, we propose that probabilistic programming languages (PPLs) are a promising way forward. PPLs provide a method of expressing and developing models to reason with uncertainty (Ghahramani 2015; Taka et al. 2020), enabling inclusion of expert knowledge, and stating explicitly the relations between variables (Taka et al. 2020). With these languages, users can construct a model of a given outcome of interest using probability distributions and run inference algorithms over their assumptions, combining them with observed data to make predictions (Martin 2018). PPLs have been applied in a number of fields including pharmaceuticals and public health (see for example: Presanis et al. (2019); Sakrejda and Novik (2019); Swiers (2019)). However, the main authors of programs in PPLs tend to be computer scientists and statisticians even when the tools are designed for application in other domains. This is due to the high
level of statistical knowledge that is a prerequisite of using these tools.

In order to critically assess the possibilities of PPLs as a tool to support medical decision-making under uncertainty, we will report on a field study of postoperative bleeding decisions. Severe blood loss after cardiac surgery is associated with adverse outcomes for patients, ranging from higher mortality to complications including stroke, heart attack, and acute kidney injury (Christensen et al. 2012; Karkouti et al. 2004; Ranucci et al. 2017). Many strategies have been proposed and applied to address this including preoperative risk scores, structured treatment protocol decision trees, and various pre- and post-operative interventions (Ranucci et al. 2017; von Heymann and Boer 2019; Petrou et al. 2016; Vuylsteke et al. 2011). Despite these advances, postoperative bleeding remains a critical issue in post cardiac surgical practice.

In addition to clinical relevance, postoperative bleeding decisions provide a wealth of material for studying decision-making under uncertainty. Though the target of reducing postoperative bleeding sounds straightforward, there are many areas of uncertainty beginning with the target itself, since the definition of excessive bleeding is not widely agreed (von Heymann and Boer 2019). Though it is straightforward to quantify the amount of blood in chest drains, it is not the bleeding itself but the harm that this causes to the patient that is the problem. Thus, the difficulty is in determining which values of blood loss relate to adverse outcomes, which requires an analysis of multiple variables influencing the individual patient’s ability to cope with loss of blood (von Heymann and Boer 2019). The definition of excessive bleeding is just the first of many areas of uncertainty throughout the process of managing postoperative bleeding.

The goal of this research is to critically assess whether PPLs would be a good fit for a decision support tool for reasoning with uncertainty in this context. Based on a detailed understanding from fieldwork, we present three key themes that emerge from the findings and draw from these evaluative design principles for assessing the success of PPL-based decision support tools in this context. Though the focus of this paper will be on the postoperative bleeding context, we believe the goal of developing end-user PPL-based tools is one that can be applied across many areas of medicine as well as outside it into other domains that require decision-making under uncertainty.

2. RELATED WORK

2.1. Clinical Decision Support Tools and Barriers to Adoption

Much has been done in research and development of clinical decision support tools since the first suggestion of using computers to aid in medical decision-making in 1959 (Ledley and Lusted 1959), (see for example (Kawamoto et al. 2005; Garg et al. 2005; Yang, Zimmerman, and Steinfeld Yang et al.)). However, many of these tools perform well in laboratory settings yet lack clinical relevance in application to healthcare environments (Yang et al. 2016; Yang, Zimmerman, and Steinfeld Yang et al.; Sendak et al. 2020) (Sendak et al. 2020). Several researchers argue that taking account of user-centred HCI considerations is a promising way forward to ensure that these tools are clinically impactful (Yang et al. 2016; Yang, Zimmerman, and Steinfeld Yang et al.; Sendak et al. 2020) (Sendak et al. 2020).

Some of these considerations relate to how the DST fits into clinical workflows and social processes of decision-making and action. Lack of recognition of social decision processes (Yang et al. 2019, 2016) can stem from conceptualising medical decision-making as an individual, cognitive activity rather than as a negotiated and social process of sensemaking (Berg 1992; Goodwin 2009). Issues for DSTs designed without sufficient recognition of clinical workflows include lack of mapping to the timelines of when clinicians need particular information (Khajouei and Jaspers 2010; Kawamoto et al. 2005), data entry burden (Shiffman et al. 1999), and provision of overly general recommendations (Shiffman et al. 1999; Kawamoto et al. 2005). Additionally, past work has demonstrated further aspects of workflows that need to be taken into account, such as at what point in the process clinicians are in front of a computer during decision-making (Yang et al. 2016) or how to maintain sterility while interacting with a DST (Johnson et al. 2011).

Another key consideration in the integration of decision support tools into medical contexts is the translational work necessary in making the outputs of DSTs useful (Morrison et al. 2016). Though clinical DSTs aim to make decision-making easier, they can often introduce further uncertainty into the reasoning and decision-making process due to a lack of clarity as to how outputs are arrived at or the reliability of their predictions (Cai et al. 2019; Hartswood et al. 2003). The work of translation is clearly emphasised by Sendak et al. (Sendak et al. 2020) (Sendak et al. 2020) in their research on designing a tool to predict the onset of sepsis. As part of the deployment of their system, they
trained a specialist team of nurses who were in charge of monitoring the tool and interpreting and communicating its outputs to others on the team.

2.2. Uncertainty in Clinical Decision Support Tools

One source of uncertainty in clinical contexts is a "lack of ground truth" (Cai et al. 2019; Sendak et al. 2020). This could be related, for example, to inter-rater variability in scoring of disease severity (Cai et al. 2019) or lack of agreed diagnostic definitions of particular conditions (Sendak et al. 2020). Another area of uncertainty reported was the difficulty of weighing different aspects of the decisions, such as which of two conflicting test results to trust or the importance of a test result or a patient report in influencing a course of action (Yang et al. 2016).

In field studies clinicians mentioned strategies for dealing with uncertainty or minimising it, including dynamic decision-making that factors in interventions that might change a prognosis (Yang et al. 2019) or asking colleagues (Yang et al. 2016; Cai et al. 2019; Hartswood et al. 2003). Cai et al. (2019) reported that clinicians solicited opinions from colleagues or external experts when uncertain. They further detailed how clinicians would ask colleagues with similarly liberal or conservative diagnostic tendencies for advice, in order to zero in on disagreements that were surprising and needed further investigation. This common strategy can result in a tendency of clinicians to try to equate DST outputs to the opinions of colleagues (Cai et al. 2019; Hartswood et al. 2003). This can be both useful as a shorthand way of understanding a system but also dangerous in that it may elide the ways in which the systems of reasoning between humans and technical models diverge.

A few strategies have been proposed in representing uncertainty in different contexts of DSTs including various visualisations (Klüber et al. 2020; Wang et al. 2008), though in some cases the uncertainty was misinterpreted by the clinician end-users (Wang et al. 2008). One strategy to minimise uncertainty about the workings of a DST was to design interactive tools for clinical users to explore the workings of the system (Cai et al. 2019). In one case, an "ambiguity aware" AI was used to flag contentious cases and a Wizard of Oz implementation generated explanations for these cases by human experts in order to explain the uncertainty (Schaeckermann et al. 2020). Medical experts reported that this was useful to track the reasoning of the system as compared to a percentage likelihood estimate.

2.3. End-User Approaches to Probabilistic Programming Languages

Probabilistic programming languages (PPLs) provide a framework for using code to build Bayesian statistical models and then running inference algorithms to fit those models to observed data (Martin 2018). Advances in both the design of the languages and the inference algorithms they use have made the frameworks for probabilistic modelling more powerful and accessible but there are still usability issues in understanding and interpreting these models to make them useful in decision-making by domain experts outside of statistics and computer science (Taka et al. 2020).

End-user approaches to PPLs are a relatively new area of research with just a few examples so far of systems designed to make PPLs more accessible to a wider range of users (Blackwell et al. 2019). Spreadsheets and databases have been an area of focus for some of this work. The most widespread end-user programming tool is the spreadsheet, where users can make use of tools such as formulas to serve their own purposes (Nardi 1993) and PPLs have been explored in this context (Geddes et al.) (Borghouts et al. 2019; Blackwell et al. 2019). In the context of databases, BayesDB (Mansinghka et al. 2015) is a general platform that uses program synthesis approaches that allow end-users to create queries for a database to build models using a SQL-like language.

Visualising the outputs of a Bayesian probabilistic model in a way that makes sense to clinicians, who are often trained in frequentist statistics (Harish et al. 2021) and may not be as comfortable with this approach will be an important challenge. In visualising the model and its outputs, this work will build on research by Gorinova et al. (2016) who built an interactive IDE for the probabilistic programming language Infer.NET and Taka et al. (2020) who created novel visualisations for the PyMC3 language to support decision-making across several application contexts.

We propose that PPLs present a unique opportunity for building clinical decision support tools that could incorporate uncertainty and allow clinicians to include their own assumptions and adjust and inspect the resultant models. In order to evaluate this claim, we take a field study approach which we describe in what follows, in order to gain a context-specific empirical foundation.

3. METHODOLOGY

The methodology of the field study consisted of three phases: 1. Preliminary interviews with key
The purpose of the preliminary interviews was problem scoping, to identify a set of decisions that involved uncertainty. The area of postoperative bleeding was identified by the lead clinician. This round of interviews involved a purposive sample chosen by the lead clinician, of five people across Royal Papworth Hospital, covering a broad range of roles involved in postoperative bleeding management. These included an anaesthetist (who is also the Clinical Director of Surgery, Transplant, and Anaesthetics), surgeon, haematologist, laboratory manager for blood transfusion, and Electronic Patient Record Configuration Developer (previously the Postoperative Blood Usage Analyst). More detail on interviewee roles and years of experience is attached in Appendix A. The format was semi-structured interviews (interview protocol attached in Appendix B), which were each approximately 60 minutes long. Topics of discussion included how decisions were made, which information was relevant, who was involved, which data was recorded and available, examples of difficult cases, and current decision-making tools and strategies. Member checking (Miles et al. 2019) was employed throughout to ensure accuracy.

The fieldwork in the next phase consisted of nine hours of observation of eight nurses in the ICU as well as handovers from surgical staff to ICU nurses and the transition from day to night nursing staff. Observations were also carried out in the blood transfusion laboratory, studying the process from delivery of blood products to the lab through to transfusion. Another piece of fieldwork in this phase was virtual attendance of a training for surgeons led by the haematologist on how to use and interpret a new blood testing system.

In the third phase, detailed follow up interviews were conducted with the anaesthetist, surgeon, and haematologist. These interviewees were chosen out of the larger group because they held key decision-making roles. An additional interview was conducted with a US-based anaesthetist who specialises in postoperative bleeding, to explore the questions outside of a UK context and gain an external perspective.

Throughout these phases, audio recordings were made of virtual interviews and observations in the blood transfusion lab (recording was prohibited in the ICU), which were then manually transcribed. Analysis was performed on written notes taken in observations as well as transcriptions of audio-recorded interviews. These notes were analysed first for the overall decision structure, then to identify statements that were made related to uncertainty, exploring whether and how a probabilistic model might usefully be designed to support them. A participatory design (Simonsen and Robertson 2012) approach was taken, in which hypotheses and preliminary findings from initial phases of research were used as probes for subsequent phases so that the findings and results could be challenged by clinical experts and further input could be incorporated.

4. OVERVIEW OF CLINICAL CONTEXT

From the outset, defining excessive bleeding is difficult because the amount of bleeding that will have detrimental impacts on a patient varies according to many patient specific factors. So while the amount of blood loss can be measured, the quantity is not always indicative of adverse consequences. Prior to cardiac surgery, the patients’ blood is altered in a number of ways, for example, by giving blood thinners so that the blood does not clot in the machine that takes over the function of the heart and lungs. These interventions, coupled with invasive surgery can lead to problems with postoperative bleeding. The type of surgery and how it went both have an impact. There are also many different components in the blood that need to be at the right levels to form a clot and stop the bleeding. Finally, there can be internal bleeding which manifests through other symptoms such as agitated mental state or low blood oxygen levels.

The task facing the clinical teams is therefore to figure out which components in the blood are at the wrong levels, and then decide whether the bleeding requires transfusion or surgical intervention. This work is further complicated by clinicians of multiple specialties working in teams, each with different knowledge and observing the patient with different granularity. And these clinicians are continuously handing over the decision-making between specialties and between day and night staff.

5. KEY THEMES FROM THE FIELD STUDY AND IMPLICATIONS FOR DESIGN OF PPL-BASED DECISION SUPPORT TOOLS

Three key themes emerged from analysis of the field study and several design principles follow from them which can be used to assess a PPL-based tool in this context. A summary of these can be found in Table 1.
5.1. Visualising Zones of Optimal Intervention

In postoperative bleeding decisions, there is a spectrum of action in which doing either too little or too much both carry risks for the patient and an area between these points in which an intervention is beneficial.

While running different blood tests can help to steer the course of treatment, there is a point where further testing is detrimental. In the words of the haematologist,

“Before you send for the blood test, you have to check if your patient is actually bleeding - there is an expected amount that you know they’ll recover from. Once you start testing, you’ll start to fix things. Surgeons often say to me, ‘Bleeding begets bleeding’. But I respond that ‘testing begets testing’... not all abnormalities require correction.”

Haematologist

He explained that a key part of clinical expertise is in knowing when to stop testing, which of the coagulation issues require intervention and which will resolve naturally.

An analogous case was revealed in discussions with the surgeon:

“[there is] some ooziness that you can’t specifically stop. When you close the chest there’s pressure and what you need is that sometimes - you can’t keep chasing every minor thing.”

Surgeon

So there is an optimal zone here too where enough sutures are added by the surgeons to resolve most of the bleeding but they stop at a reasonable point where pressure can be applied to resolve the remaining oozing.

There is an overarching tension in the decisions between taking action, even with imperfect information, or waiting for further information to dictate more clearly which actions to take. Waiting for too long with a severely bleeding patient would be detrimental, but each intervention also presents risks, so it is important to be as sure as possible within the time constraints.

Two quotes from the surgeon an anaesthetist illustrate this tension:

“[Waiting and seeing] is very important here – that is exactly what is usually done. By requesting blood tests, you’re sort of imposing a stop point of getting the blood results and in that allowing time to pass to the next hour to see what happens. Very valid to wait for an hour to see what happens next – that’s the reality of what happens. Often the outcome of a discussion even with the consultants.”

Surgeon

“In decision-making mode, you flip to process. A decision is better than no decision, even if the patient dies, you did the best in the moment.”

Anaesthetist

A design principle that follows from this theme is that visualising where these zones of action are could be critical in determining when and how to act. These zones would be developed through input of clinical assumptions in the model coupled with training on historical patient cases and outcomes. For the purpose of scenario analysis, visualising these zones might inform when a particular intervention should be discontinued in favour of another course of action. For example, a visualisation of an approaching threshold where transfusion might pose more harm than benefit could signal a greater likelihood of return to theatre.

Thus, one criterion along which DST designs for postoperative bleeding can be evaluated is through an assessment of how accurately a visualisation can capture the zones of optimal intervention for a given course and how changes to potential actions in the model show movement through this zone. And further, a successful design will capture the connections between optimal zones across different intervention options. A probabilistic programming based tool is a strong candidate for visualising these zones due to the posterior distributions that are its output which can show an action with a probabilistic likelihood of success. Probability distributions will also be useful for contextualising an intervention against historical outcomes for other patient cases contained in institutional data.

5.2. Surfacing Relative Risk Evaluations between Teams

Multiple specialists work in teams on postoperative bleeding management after cardiac surgery. This leads to uncertainty in the shifting hierarchies of decision-making and handovers of information and patient care between specialists.

The anaesthetist explained it this way:

“Someone will have final say but not always the same person - often a more senior person will decide, but if the same seniority then they will need an arbitrator who will often be the haematologist. If the haematologist disagrees or believes strongly that the patient doesn’t need it [a transfusion], he will probably have much more weight. . . However there is a risk in under transfusing by someone
In postoperative bleeding management, multiple experts observe the patient with different granularity. A tool could help clinicians who have an intricate involvement with the patient to see the wider trends and those who lack this intricate knowledge of the patient to see the more granular view.

Multiple clinicians mentioned the importance of contextualising individual decisions within the wider context of the patient’s care:

“The hierarchy is not a straight line. The more you move up, the more you have pattern recognition, dealing with uncertainty more quickly. . . There are multiple layers of decision making with each performing at a slightly different level even within same specialty. . . Need to contextualise the tool for my level of need - take away what I don’t want and keep what I need.”

Haematologist

This leads to a design principle for the user interface of a model of customisation of views such that different experts can focus on the variables of interest to them, but also zoom out to see how they connect with those that others are trying to optimise. Another design principle that would affect the model itself might be a hierarchical structure that could be customised and adjusted by clinicians as they discuss relative priorities of different factors in the patient’s care. Being able to concretely and tangibly have these discussions while modelling the decisions could be beneficial in resolving disagreements in care plans in a way that illuminates issues and incorporates opinions of multiple experts.

6. LIMITATIONS

This work is limited by the small sample size of clinicians interviewed as well as being a single-centre study (though this second limitation is in part...
Table 1: Summary of themes and design principles from analysis of field study findings.

| Theme | Design Principles |
|-------|-------------------|
| 1. Visualising zones of optimal intervention | 1.1 Emphasise the zones of optimal intervention for a given course of action. |
| | 1.2 Highlight how changes to actions in the model (scenario analysis) show movement through the zone of optimal intervention. |
| | 1.3 Capture the connections between optimal zones for multiple interventions on the overall outcome. |
| 2. Surfacing relative risk evaluations between teams | 2.1 Quantify uncertainty in the model in a way that allows risk trade-offs to be surfaced. Design for drilling down into uncertainty associated with specific actions to compare them. |
| | 2.2 Design the user interface of the tool for inputting assumptions and doing scenario analysis such that the process can help to manage and facilitate important discussions in patient care. |
| 3. Accessing specialist view within a holistic picture | 3.1 Provide several customised views of the model such that different experts can focus on the variables of interest to them. |
| | 3.2 Design a dashboard of an overall view of all decisions and the target outcome such that clinicians can zoom out to see how the factors they are focused on connect with those that others are trying to optimise. |
| | 3.3 Construct the model such that there is a hierarchical structure that can be re-arranged and reordered by clinicians as they discuss relative priorities of interventions and outcomes in the patient's care. |

7. CONCLUSION

Reasoning under uncertainty is a fundamental challenge in medical decision-making. Decision support tools based on probabilistic programming present a promising way forward but there is significant work needed to build them in a way that is useful and clinically relevant in practice. In service of this translational work, we have presented three key themes that emerged from a field study of postoperative bleeding decisions in the cardiothoracic ICU and drawn a set of evaluative design principles for assessing a PPL-based tool in this context. These include visualising zones of optimal intervention, surfacing relative risk trade-offs between teams, and accessing specialist views within a holistic picture. These principles give us a framework to further the design of PPL-based decision support for use by clinical experts in reasoning under uncertainty. By taking an end-user approach, new opportunities are possible for making the modelling process available to clinicians and allowing them to rigorously capture uncertainty in their decision-making.

8. FUTURE WORK

In this work, we have provided preliminary evidence to motivate the use of PPLs as a basis for decision support in postoperative bleeding decision making. In order to further develop and test these ideas, future work will look specifically at which technical approaches to modelling are most useful in this context. A prototype tool will be developed using the design principles proposed above as a guide to evaluate iterations. This will then be evaluated in a study with clinicians.

REFERENCES

Berg, M. (1992). The construction of medical disposals medical sociology and medical problem solving in clinical practice. *Sociol. Health Illn. 14*(2), 151–180.

Blackwell, A. F., T. Kohn, M. Erwig, A. G. B. Baydin, L. Church, J. Geddes, A. Gordon, M. Gorinova, B. Gram-Hansen, N. Lawrence, V. K. Mansinghka, B. Paige, T. Petriceck, D. Robinson, A. Sarkar, and O. Strickson (2019, August). Usability of probabilistic programming languages. In M. Marasoiu, L. Church, and L. Marshall (Eds.), *PPIG 2019 - 30th Annual Workshop*, pp. 53–68.

Borghouts, J., A. D. Gordon, A. Sarkar, and N. Toronto (2019). End-user probabilistic programming. In *Quantitative Evaluation of Systems*, Lecture notes in computer science, pp. 3–24. Cham: Springer International Publishing.

Cai, C. J., E. Reif, N. Hegde, J. Hipp, B. Kim, D. Smilkov, M. Wattenberg, F. Viegas, G. S. Corrado, M. C. Stumpe, and M. Terry (2019, May). Human-Centered tools for coping with imperfect algorithms during medical Decision-Making. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI ’19)*.

Cai, C. J., S. Winter, D. Steiner, L. Wilcox, and M. Terry (2019, November). “hello AI”: Uncovering the onboarding needs of medical practitioners for Human-AI collaborative Decision-Making. *Proc. ACM Hum.-Comput. Interact. 3(CSCW), 1–24*.

Christensen, M. C., F. Dziewior, A. Kempel, and C. von Heymann (2012, February). Increased chest tube drainage is independently associated with adverse outcome after cardiac surgery. *J. Cardiothorac. Vasc. Anesth. 26*(1), 46–51.
Cooke, S. and J.-F. Lemay (2017, June). Transforming medical assessment: Integrating uncertainty into the evaluation of clinical reasoning in medical education. *Acad. Med.* 92(6), 746–751.

Dunlop, M. and R. M. Schwartzstein (2020, September). Reducing diagnostic error in the intensive care unit. Engaging uncertainty when teaching clinical reasoning. *ATS Sch 1*(4), 364–371.

Garg, A. X., N. K. J. Adhikari, H. McDonald, M. P. Rosas-Arellano, P. J. Devereaux, J. Beyene, J. Sam, and R. B. Haynes (2005, March). Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *JAMA* 293(10), 1223–1238.

Geddes, J., O. Strickson, and T. Counsell. Nocell: Probabilistic programming for spreadsheet experts. https://www.turing.ac.uk/research/research_projects/nocell-probabilistic-programming-spreadsheet-experts. Accessed: 2021-9-7.

Ghahramani, Z. (2015, May). Probabilistic machine learning and artificial intelligence. *Nature* 521(7553), 452–459.

Goodwin, D. (2009, February). *Acting in Anaesthesia: Ethnographic Encounters with Patients, Practitioners and Medical Technologies*. Cambridge University Press.

Gorinova, M. I., A. Sarkar, A. F. Blackwell, and D. Syme (2016, May). A live, Multiple-Representation probabilistic programming environment for novices. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 2533–2537. New York, NY, USA: Association for Computing Machinery.

Harish, V., F. Morgado, A. D. Stern, and S. Das (2021, January). Artificial intelligence and clinical decision making: The new nature of medical uncertainty. *Acad. Med.* 96(1), 31–36.

Hartswood, M., R. Procter, M. Rouncefield, R. Slack, J. Soutter, and A. Voss (2003). 'repairing' the machine: A case study of the evaluation of Computer-Aided detection tools in breast screening. *ECSWC 2003: Proceedings of the Eighth European Conference on Computer Supported Cooperative Work*, 1–20.

Johnson, R., K. O’Hara, A. Sellen, C. Cousins, and A. Criminisi (2011, May). Exploring the potential for touchless interaction in image-guided interventional radiology. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI ’11, New York, NY, USA, pp. 3323–3332. Association for Computing Machinery.

Karkouti, K., D. N. Wijeyasurya, T. M. Yau, W. S. Beattie, E. Abdelsaam, S. A. McCluskey, M. Ghannam, E. Yeo, G. Djalani, and J. Karaki (2004, October). The independent association of massive blood loss with mortality in cardiac surgery. *Transfusion* 44(10), 1453–1462.

Kawamoto, K., C. A. Houlihan, E. A. Balas, and D. F. Lobach (2005, April). Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *BMJ* 330(7494), 765.

Khajouei, R. and M. W. M. Jaspers (2010). The impact of CPOE medication systems’ design aspects on usability, workflow and medication orders: a systematic review. *Methods Inf. Med.* 49(1), 3–19.

Klüber, S., F. Maas, D. Schrauder, G. Hermann, O. Happel, and T. Grundgeiger (2020, July). Experience matters: Design and evaluation of an anesthesia support tool guided by user experience theory. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, DIS ’20, New York, NY, USA, pp. 1523–1535. Association for Computing Machinery.

Ledley, R. S. and L. B. Lusted (1959, October). Probability, logic and medical diagnosis. *Science* 130(3380), 892–930.

Mansinghka, V., R. Tibbetts, J. Baxter, P. Shafto, and B. Eaves (2015, December). BayesDB: A probabilistic programming system for querying the probable implications of data.

Martin, O. (2018, December). *Bayesian Analysis with Python: Introduction to statistical modeling and probabilistic programming using PyMC3 and ArviZ, 2nd Edition*. Packt Publishing Ltd.

Miles, M. B., A. Michael Huberman, and J. Saldana (2019, January). *Qualitative Data Analysis: A Methods Sourcebook*. SAGE Publications.

Morrison, C., K. Huckvale, B. Corish, J. Dorn, P. Kontschieder, K. O’Hara, Team, ASSESS MS, A. Criminisi, and A. Sellen (2016, July). Assessing multiple sclerosis with kinect: Designing computer vision systems for Real-World use. *Human–Computer Interaction* 31(3-4), 191–226.

Nardi, B. A. (1993). *A Small Matter of Programming: Perspectives on End User Computing*. MIT Press.

Petrou, A., P. Tzimas, and S. Siminelakis (2016, July). Massive bleeding in cardiac surgery. definitions, predictors and challenges. *Hippokratia* 20(3), 179–186.
Presanis, A., C. Jackson, D. De Angelis, P. Kirwan, A. Brown, A. Miltz, R. Harris, C. Chau, S. Michelsen, H. Mohammed, K. Davison, S. Croxford, S. Desai, K. Lowndes, V. Delpech, and N. Gill (2019, August). Estimating the prevalence of HIV infection in England using Bayesian evidence synthesis. StanCon 2019.

Ranucci, M., E. Baryshnikova, V. Pistuddi, L. Mencanti, A. Frigiola, and Surgical and Clinical Outcome REsearch (SCORE) Group (2017, February). The effectiveness of 10 years of interventions to control postoperative bleeding in adult cardiac surgery. *Interact. Cardiovasc. Thorac. Surg.* 24(2), 196–202.

Sakrejda, K. and E. Novik (2019, August). A Decision-Theoretic journey from early clinical data to late stage efficacy using hierarchical joint models. StanCon 2019.

Schaekermann, M., G. Beaton, E. Sanoubari, A. Lim, K. Larson, and E. Law (2020, April). Ambiguity-aware AI assistants for medical data analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–14. New York, NY, USA: Association for Computing Machinery.

Sendak, M., M. C. Elish, M. Gao, J. Futoma, W. Ratliff, M. Nichols, A. Bedoya, S. Balu, and C. O’Brien (2020). “the human body is a black box” supporting clinical decision-making with deep learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 99–109.

Sendak, M. P., W. Ratliff, D. Sarro, E. Alderton, J. Futoma, M. Gao, M. Nichols, M. Revoir, F. Yashar, C. Miller, K. Kester, S. Sandhu, K. Corey, N. Brajer, C. Tan, A. Lin, T. Brown, S. Engelbosch, K. Anstrom, M. C. Elish, K. Heller, R. Donohoe, J. Theiling, E. Poon, S. Balu, A. Bedoya, and C. O’Brien (2020, July). Real-World integration of a sepsis deep learning technology into routine clinical care: Implementation study. *JMIR Med Inform* 8(7), e15182.

Shiffman, R. N., Y. Liaw, C. A. Brandt, and G. J. Corp (1999, March). Computer-based guideline implementation systems: a systematic review of functionality and effectiveness. *J. Am. Med. Inform. Assoc.* 6(2), 104–114.

Simonsen, J. and T. Robertson (2012). *Routledge international handbook of participatory design*. Routledge.

Simkin, A. L. and R. M. Schwartzstein (2016, November). Tolerating uncertainty - the next medical revolution? *N. Engl. J. Med.* 375(18), 1713–1715.

Swiers, R. (2019, August). Handling missing data, censored values and measurement error in machine learning models using multiple imputation for early stage drug discovery. StanCon 2019.

Taka, E., S. Stein, and J. H. Williamson (2020). Increasing interpretability of Bayesian probabilistic programming models through interactive representations. *Frontiers in Computer Science* 2, 52.

von Heymann, C. and C. Boer (Eds.) (2019). *Patient Blood Management in Cardiac Surgery*. Springer, Cham.

Vuylsteke, A., C. Pagel, C. Gerrard, B. Reddy, S. Nashef, P. Aldam, and M. Utley (2011, June). The papworth bleeding risk score: a stratification scheme for identifying cardiac surgery patients at risk of excessive early postoperative bleeding. *Eur. J. Cardiothorac. Surg.* 39(6), 924–930.

Wang, T. D., C. Plaisant, A. J. Quinn, R. Stanchak, S. Murphy, and B. Schneiderman (2008, April). Aligning temporal data by sentinel events: discovering patterns in electronic health records. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI ’08, New York, NY, USA, pp. 457–466. Association for Computing Machinery.

Yang, Q., A. Steinfeld, and J. Zimmerman (2019, May). Unremarkable AI: Fitting intelligent decision support into critical, clinical Decision-Making processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Number Paper 238 in CHI ’19, New York, NY, USA, pp. 1–11. Association for Computing Machinery.

Yang, Q., J. Zimmerman, and A. Steinfeld. Review of medical decision support tools: Emerging opportunity for interaction design. https://www.researchgate.net/profile/QianYang29/publication/282121061_Review_of_Medical_Decision_Support_Tools_Emerging_Opportunity_for_Interaction_Design/links/56035cf608ae08d4f1716009.pdf. Accessed: 2020-9-2.

Yang, Q., J. Zimmerman, A. Steinfeld, and J. F. Antaki (2016, May). Investigating the heart pump implant decision process: Opportunities for decision support tools to help. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI ’16, New York, NY, USA, pp. 4477–4488. Association for Computing Machinery.
A. CLINICAL INFORMANT ROLES AND EXPERIENCE

| Role                                    | Years of Experience |
|-----------------------------------------|---------------------|
| Clinical Director of Surgery, Transplant and Anaesthetics, Anaesthetist (quoted after this as "the anaesthetist") | 27                  |
| Surgeon                                 | 5                   |
| Haematologist                           | 13                  |
| Laboratory Manager for Blood Transfusion | 14                  |
| Electronic Patient Record (EPR) Configuration Developer (previously Postoperative Blood Usage Analyst) | 22                  |

Table 2: Roles and years of experience for each clinical informant interviewed for the study.

B. SEMI-STRUCTURED INTERVIEW PROTOCOL

Intro (2-3 min) - explaining context and goals of the study

The focus of this research is on machine learning based clinical decision support tools in the area of postoperative bleeding. It would be extremely helpful to learn more about the process from your expertise. I’m interested specifically in talking through concrete episodes you have experienced (adverse bleeding events and how these are managed) to get an understanding of if and where in this process a tool might be helpful and in what ways. I want to find out where in the process key decisions take place, where there might be a need for more or different type of information/analysis, on what kind of timescales these decisions are being made.

Decisions Processes and Uncertainty (60-75 min) - questions probing decision processes, workflow, difficulties, uncertainty, and any current decision support tools

(i) Can you take me through which points in the patient pathway you’re involved in?

(ii) What actions and recommendations do you take? E.g. ordering tests, giving coagulants (protamine, tranexamic acid, are there others?), transfusing blood products

(iii) How are the decisions to return to theatre currently made?

(iv) What are some of the top variables you would include in a decision to start with? (going to build out the model in smaller steps, might be hierarchical, this would inform some of those decisions)

(v) How long does it take to see a patient response after transfusion?

(vi) What information do you gather to make your decision? Who does it come from? Which IT systems do you use and when? How are lab results delivered (printed out or integrated in the systems)?

(vii) What are the dilemmas? (e.g. are there particular symptoms that might be consistent with multiple courses of action?)

(viii) Who do you consult with?

(ix) What are some difficult decisions you have experienced and what information would you have wanted to have in those cases?

(x) What parts of the decision would it help to have an early warning on? How much time would define “early” for this?

(xi) What kind of data exploration would you want to do if you could?

(xii) Are there any other decision support tools that you use?

(xiii) Where in the process do you have time for reflective decision-making?