Availability of Sex, Gender Identity, and Sexual Orientation Data: An Electronic Medical Record Review of a Catholic Healthcare System from 2012-2023

Whitney Linsenmeyer, PhD, RD, LD, Katie Heiden-Rootes, PhD, LMFT, Michelle R. Dalton, PhD, LPC, and Timothy Chrusciel, MPH

Abstract

Purpose: The study purpose was to describe the availability of sex, gender identity, and sexual orientation (SOGI) data in a large, Catholic health system.

Methods: A retrospective chart review on the Sisters of St. Mary (SSM) Health database was conducted from January 1, 2012, to March 27, 2024. The availability of SOGI data and number of sexual and gender minority patients was reported.

Results: Among the 5,759,869 records, data on sex was available for the majority of the population (99.9 percent); data on gender identity and sexual orientation were reported for smaller proportions (7.4 percent and 4.5 percent, respectively). Sex and gender were reported among 7.4 percent of the population. A total of 4,567 gender minority and 14,644 sexual minority patients were seen.

Conclusion: Though SOGI data were largely unavailable in the SSM Health database, the system has the capacity to separately record sex, gender, and sexual orientation, with a range of response options to capture gender and sexual orientation diversity.

Keywords: sex, gender identity, sexual orientation, demographic data

Introduction

Sex, sexual orientation, and gender identity (SOGI) are essential, demographic patient data.[1] The American Medical Association (AMA) defines *sex* or *sex assigned at birth* based on a subjective evaluation of external anatomic structure(s) and its comparison to various sex categories, whereas *gender identity* describes how people conceptualize themselves as gendered beings, including one's innate and personal experience of gender. *Sexual orientation* describes an inherent or immutable enduring emotional, romantic or sexual attraction to others.[2] Definitions of these key terms may vary among countries, cultures, and time periods.[3]

Recommendations for SOGI Data Collection

SOGI data are often conflated or omitted in clinical, research, and administrative settings. This practice undermines the accuracy and validity of patient data and resulting datasets for research purposes.[2] Accurate data collection is essential for all patients, but especially sexual and gender minority patients (SGM) who may otherwise be excluded. Sexual minority patients are those who identify as gay, lesbian, bisexual, or pansexual or who are attracted to or have sexual contact with people of the same gender; gender minority patients are those whose gender identity (man, woman, other) or expression (masculine, feminine, other) differs from their sex assigned at birth (male, female).[4] They may identify as transgender, gender queer, non-binary, or something other than their sex assigned a birth.

Leading organizations such as the AMA, the National Academies of Sciences, Engineering and Medicine (NASEM), and the Biden-Harris Administration have underpinned the importance of accurately distinguishing between these terms to ensure precision in data collection and reporting.[1,2,5] National and international surveys have utilized a two-step approach to separate query sex and gender, such as the censuses in Canada, England and Wales, New Zealand, and Scotland, among others.[1] In the United States, the NASEM published recommended language for the two-step approach with a separate question to query intersex status (Figure 1). A breadth of gender identity response options were recommended to capture gender diversity (i.e. transgender, two-spirit), as well as options to enter free text, "don't know," or "prefer not to answer."[1]

Religiously Affiliated Institutions

These recommendations present a unique question for religiously affiliated institutions where entities have asserted conflicting comments surrounding the healthcare of SGM patients. For example, the United States Conference of Catholic Bishops issued a mandate against gender-affirming medical interventions for gender minority patients at Catholic hospitals, which contradicts standards of care from international and national health organizations.[3,6,7]

Regardless of whether gender-affirming medicine occurs in Catholic healthcare settings, whether these institutions are collecting and reporting SOGI data on their patient population has yet to be explored. SGM patients may still utilize religiously affiliated medical care facilities for routine, urgent, and emergent needs, like any other patient, though perhaps with greater frequency given known health disparities in cancer, chronic illnesses, infectious disease and mental health.[8]

Study Purpose and Aims

The purpose of this study was to describe the reporting of SOGI patient data in a large, Catholic health system. The objectives were to: describe availability of sex, gender identity, and sexual orientation data; and report the number of SGM patients captured in the health system.

Methods

Study Design

This was a retrospective chart review of the Sisters of St. Mary (SSM) Health System patient population. This method was selected for its appropriateness of fit with the study purpose and aims focused on data availability.[9] The checklist by the Professional Society for Health Outcomes and Research (ISPOR) Task Force on Retrospective Databases was used to ensure methodological quality.[10]

SSM Health Database

SSM Health is a Catholic, non-profit integrated care network operating in Missouri, Illinois, Oklahoma, and Wisconsin in the United States. The network includes 12,800 providers, 23 hospitals, 300 physician offices, outpatients, and virtual are services, and 13 post-accurate facilities.[11] The SSM Health electronic health record (EHR) database, EPIC Clarity, is a subset of the SSM Health patient data that is available for research purposes and stored in Microsoft SQL.[12] The database contains records on over 11 million patients.

The database was queried for sex, gender identity, and sexual orientation data among all patients ages 12 or older during years January 1, 2012, to March 27, 2024. Age 12 was selected as a cut-off as it is often used as a research standard for adolescence and often coincides with puberty[13] and gender identity development.[14] Records were included if the patient had at least one encounter in the system within the study period. Patients were excluded if they did not have an encounter during the study period or if they were younger than 12 years old at the time of the pull.

Sex was reported as female, male, other, unknown, or null. Gender identity was reported as female, male, transgender female, transgender male, gender nonconforming, genderqueer, other, choose not to disclose, or null. Sexual orientation was reported as straight, gay, lesbian, lesbian or gay, bisexual, multiple listed, don’t know, something else, choose not to disclose, or null. Null indicated the data was unavailable, or not reported in the system.

Data Analysis

Data were analyzed using descriptive statistics. A very low prevalence was reported as <1.0.

Results

A total of 5,759,869 records were included in the analysis; 5,165,056 records were excluded due to lack of encounters in the system during the study period. Figure 2 depicts the availability of sex, gender identity, and sexual orientation data. Data on sex was available in the majority of the population (99.9 percent), while data on gender identity was available for a small proportion (7.4 percent). Both sex and gender data were reported among 7.4 percent of the population. Data on sexual orientation was reported in a smaller proportion of the population (4.5 percent).

The sex, gender identity, and sexual orientation of the population is depicted in Table 1. Regarding sex, patients were mostly female (53.7 percent) or male (46.2 percent), with small proportions of patient data that was reported as unknown, "other," or unavailable (<1.0 percent). Although gender identity data was unavailable for most patients (92.6 percent), a total of 4,567 patients, or 0.08 percent of the population identified as a gender minority (transgender female or transgender male, genderqueer, gender nonconforming, or "other.") Regarding sexual orientation, although data was unavailable for most patients (95.4 percent), a total of 14,644 patients, or 0.25 percent of the population identified as a sexual minority (gay, lesbian, bisexual, or "something else.")

Discussion

Gender identity and sexual orientation data were largely unavailable from the SSM Health database. This reflects larger trends in national surveys where SOGI data are largely omitted, or where sex and gender are conflated and limited to a male-female binary.[1,5,15] Promisingly, the database includes the fields of sex, gender identity, and sexual orientation, with a range of response options to capture gender and sexual orientation diversity. This suggests that the gap in data availability is occurring at the provider level where SOGI data is not routinely collected, rather than the database level that limits its availability. Provider education or recommendations from the network may improve demographic data collection practices. Future studies may track the frequency of SOGI data collection over time, the impact of provider education and network recommendations on data collection practices, the accuracy of SOGI data collection and reporting, and the number of sexual and gender minority patients cared for by SSM Health.

SSM Health cared for over 14,000 sexual minority patients and over 4,000 gender minority patients from January 1, 2012, to March 27, 2024, though these estimates are likely low. It is also probable that SOGI data were not collected in certain patient encounters where the data seemed irrelevant to the nature of care provided. For example, data on sexual orientation may not have been collected in an emergency room encounter for a fractured bone. The NASEM recommends collecting only necessary data to meet a defined purpose. Thus, a proportion of unavailable data may have reflected its lack of relevance to the encounter.

Routine SOGI data collection in Catholic healthcare settings will ensure that, at minimum, SGM patients are accurately counted. Administrators of Catholic health institutions must consider how they will adhere to standards of care for their SGM patients, while also responding to conflicting guidance from the United States Conference of Catholic Bishops.

Strengths, Limitations, and Future Research

Strengths of this study were the large study population across multiple states and the long observation period. One limitation was the potential discrepancy between clinical intake procedures, which likely vary by site and provider, and how data is reported in the SSM Health database; for example, providers may be routinely collecting SOGI data, but the language of the questions may differ from that of the database. A final limitation was the generalizability of the study findings to the Midwestern United States or similar health networks with a Catholic affiliation. Future research may explore the SOGI data collection practices in other regions of the United States and at institutions with various religious affiliations.

Lastly, future research is needed to explore how data collection and reporting practices respond to the government's charge to improve SOGI data in the United States.[6] Given that approaches may vary by agency, studies can report the data collection and reporting practices adopted.

Conclusion

Though SOGI data were largely unavailable in the SSM Health database, the system has the capacity to separately enter sex, gender, and sexual orientation, with a range of response options to capture gender and sexual orientation diversity. Provider education and recommendations from the network are needed to ensure SOGI data are treated as routine, essential demographic data.

Authors' Contributions: Chrusciel facilitated the data pull and synthesis. Linsenmeyer drafted a first draft of the article. All authors reviewed the article, provided comments, and agreed on the final draft.

Acknowledgements: This work is supported by work from the Saint Louis University School of Medicine and resources from the Advanced HEAlth Data (AHEAD) Research Institute. Irene Ryan of the AHEAD Institute completed the data pull and synthesis.

Statement on Conflicts of Interest: The authors have no conflicts of interest to report.

Authors:

Whitney Linsenmeyer, PhD, RD, LD (she/her) is an assistant professor of nutrition at Saint Louis University and co-founder of the Transgender Health Collaborative at SLU. Her research centers on gender-affirming nutrition care for the transgender population.

Katie Heiden-Rootes, PhD, LMFT (she/her) is an assistant vice president in the division of diversity & innovative community engagement at Saint Louis University and an associate professor of medical family therapy. She co-founded the Transgender Health Collaborative at SLU and centers her scholarship on the mental health and well-being of queer youth and their families.

Michelle R. Dalton, PhD, LPC (they/them) is an assistant professor of medical family therapy at Saint Louis University and a member of the Transgender Health Collaborative at SLU. Their research areas focus on gender identity and racial minority stress and transgender health.

Timothy Chrusciel, MPH (he/him) is a biostatistician with Saint Louis University's Advanced HEAlth Data (AHEAD) Research Institute. He has over 15 years of experience in a wide range of statistical methodologies.

References

1. National Academies of Sciences, Engineering, and Medicine. Measuring sex, gender identity, and sexual orientation. National Academies Press, Washington, DC (2022).
2. American Medical Association. Advancing health equity: A guide to language, narrative and concepts. Published 2021. Accessed July 17, 2023. https://www.ama-assn.org/system/files/ama-aamc-equity-guide.pdf
3. Coleman E, Radix AE, Bouman WP, et al. Standards of Care for the Health of Transgender and Gender Diverse People, Version 8. *Int J Transgend Health*. 2022;23(Suppl 1):S1-S259.
4. Center for Disease Control and Prevention, Department of Adolescent and School health. Health considerations for LGBTQ youth: Terminology. Reviewed December 23, 2022. Accessed July 17, 2023. https://www.cdc.gov/healthyyouth/terminology/sexual-and-gender-identity-terms.htm
5. National Science and Technology Council, Subcommittee on Sexual Orientation, Gender Identity, and Variations in Sex Characteristics (SOGI) Data, Subcommittee on Equitable Data. Federal Evidence Agenda on LGBTQI+ Equity. Washington, DC (2023).
6. United States Conference of Catholic Bishops, Committee on Doctrine. Doctrinal note on the moral limits to technological manipulation of the human body. Published March 20, 2023. Accessed July 18, 2023. https://www.usccb.org/resources/Doctrinal%20Note%202023-03-20.pdf
7. Hembree WC, Cohen-Kettenis PT, Gooren L, et al. Endocrine Treatment of Gender-Dysphoric/Gender-Incongruent Persons: An Endocrine Society Clinical Practice Guideline [published correction appears in *J Clin Endocrinol Metab*. 2018 Feb 1;103(2):699] [published correction appears in *J Clin Endocrinol Metab*. 2018 Jul 1;103(7):2758-2759]. *J Clin Endocrinol Metab*. 2017;102(11):3869-3903.
8. Center for Disease Control and Prevention. Lesbian, gay, bisexual and transgender health. Reviewed November 3, 2022. Accessed July 20, 2023.
9. Vassar M, Holzmann M. The retrospective chart review: important methodological considerations. *J Educ Eval Health Prof.* 2013;10:12.
10. SSM Health. Our heritage of healing. Publication date unknown. Accessed July 17, 2023. https://www.ssmhealth.com/resources/about/mission-vision-values/our-heritage
11. Motheral B, Brooks J, Clark MA, et al. A checklist for retrospective database studies--report of the ISPOR Task Force on Retrospective Databases. *Value Health*. 2003;6(2):90-97
12. Microsoft SQL [Computer Software]. Version 18.12.1. Washington: Microsoft; 2022.
13. Rosenfield RL, Lipton RB, Drum ML. Thelarche, pubarche, and menarche attainment in children with normal and elevated body mass index [published correction appears in Pediatrics. 2009 Apr;123(4):1255]. *Pediatrics*. 2009;123(1):84-88.

14. Steensma TD, Kreukels BP, de Vries AL, Cohen-Kettenis PT. Gender identity development in adolescence. *Horm Behav*. 2013;64(2):288-297.
15. Heiden-Rootes KM, Salas J, Scherrer JF, Schneider FD, Smith CW. Comparison of Medical Diagnoses among Same-Sex and Opposite-Sex-Partnered Patients. *J Am Board Fam Med*. 2016;29(6):688-693.

Table 1. Sex, Gender Identity, and Sexual Orientation of Patient Population.

	N (%)
Sex	
Female	3,095,729 (53.7%)
Male	2,658,409 (46.2%)
Unknown	2,084 (<1.0)
Other	6 (<1.0)
Data Unavailable	411 (<1.0)
Gender Identity	
Female	259,980 (4.5%)
Male	160,437 (2.8%)
Transgender Female	903 (<1.0)
Transgender male	1,421 (<1.0)
Genderqueer	583 (<1.0)
Gender Nonconforming	1,660 (<1.0)
Choose Not to Disclose	1,160 (<1.0)
Other	418 (<1.0)
Data Unavailable	5,330,077 (92.5%)
Sexual Orientation	
Straight	232,148 (4.0%)
Lesbian	2,916 (<1.0)

Gay	3,508 (<1.0)
Lesbian or Gay	33 (<1.0)
Bisexual	7,322 (<1.0)
Multiple Listed	898 (<1.0)
Something Else	1,633 (<1.0)
Don’t Know	2,244 (<1.0)
Choose Not to Disclose	9,656 (<1.0)
Data Unavailable	5,496,281 (95.4%)

The Impact of Professional Ethics Case-based Learning on the Ethical Sensitivity of Health Information Technology Students

Shahla Damanabi, PhD, Mozhgan Behshid, PhD, Zahra Moradi, Msc, Leila Ghaderi-Nansa, PhD

Introduction

Moral sensitivity is one criterion for competent professional ethics. This sensitivity can be reinforced by specific educational practices. The purpose of this study was to investigate the impact of professional ethics-based education on the ethical sensitivity of health information technology students.

Method

This quasi-experimental pre-post study was conducted in 2022 with 49 students. A researcher-created questionnaire based on Lutzen was used for data collection. Data were analyzed using descriptive statistics and paired t-tests.

Findings

Students' moral sensitivity score was 7.4 ± 0.7 before and 7.6 ± 0.8 after, a significant increase in post scores (P=0.031). The moral sensitivity score of students who had not previously received professional ethics training was statistically significantly increased by case-based learning.

Results

The professional ethics-based teaching method was effective in increasing the moral sensitivity of health information technology students, so it is recommended to use this method of teaching medical ethics courses.

Keywords

Moral sensitivity, moral strength, moral responsibility, case-based learning, health information technology.

Introduction:

Ethics is one of the most important, common, and challenging issues across all academic disciplines, especially in the medical sciences. Rapid advances in technology in the field of health, increases in public information, changing patterns of disease, and differences in the type and volume of health service requests have created new ethical issues[1, 2].

Often, professional staff do not know the solutions to some ethical conflicts or are unaware of ethical decisions. Today, ethical decision-making is essential to professional work[3]. Ethical decision-making is the decision-making process of identifying issues through analysis according to ethical criteria and deciding whether or not to do it[4-6]. In other words, ethical decision-making is a debate about good or bad, and the conflict between acting and not acting on one's values[7]. Identifying ethical conflicts is very important in the decision-making process and staff need to be able to recognize ethical situations and evaluate the situation quickly and accurately to make decisions that are ethically beneficial to patients[8-11]. For example, hospitals are responsible for responding to legal requests for patient information disclosure by the principle of confidentiality of patient data[12]. Because of this, deciding whether to disclose sensitive patient information, such as information about sexually transmitted diseases, is an ethical situation that challenges staff to make decisions. Making ethical decisions not only requires moral knowledge but also moral sensitivity. Ethical sensitivity means the ability to identify an ethical issue, understand the ethical consequences of the decision, and how one's actions affect others[13].

Lutzen defines moral sensitivity as one's "awareness of a sense of responsibility, moral burden, and moral ability." In other words, moral sensitivity includes cognition and awareness that a person's decision or action may affect others' interests, welfare, or expectations, and may conflict with one or more ethical standards[14]. Thus, people with moral sensitivity are better equipped to resolve ethical conflicts in complex situations[15].

One way to increase ethical sensitivity is through professional ethics training[16-17]. Teaching professional ethics and institutionalizing these ethical principles among employees is a principal concern among health authorities. Increasing students' ability to practice professional ethics as future employees of the health system is also a concern for education authorities[18-20]. If these professional principles and beliefs are not institutionalized during a student's educational career, it may reduce the student's moral sensitivity and, consequently, make it more difficult for them to practice ethical decision-making in the future[19-21].

Some texts provide specific training methods for increasing ethical sensitivity, such as group discussion[22], problem-solving, case-based or scenario-based methods[7,23,24], or workshops[25] More objective training (e.g., examples, case studies, and the use of teaching aids) affect moral sensitivity positively[26]. Student-centered teaching methods are also more useful than traditional teaching methods to help develop critical thinking skills, problem-solving, and decision-making[27]. One strategy that enhances active learning and prepares students for future careers and service in a real-world environment is case-based teaching[7, 23, 28]. Case-based teaching is a combination of traditional (lecture) and problem-solving teaching[29]. In this method, scenarios are used, real or realistic, which require problem-solving and decision-making. In this way, the teacher acts as a facilitator, guiding the students toward the goals, and students learn to think critically to solve the problem and make decisions in the new situation. This teaching method helps students organize knowledge, identify gaps in the field, communicate with students, develop problem-solving, and decision-making skills, and has the added benefit of increasing motivation for learning[30]. Today, the use of case-based methods in medical sciences pedagogy, especially in basic sciences, is expanding in most universities around the world. Therefore, given the benefits of this teaching method, this study investigated the impact of case-based ethics training on the ethical sensitivity of health information technology students.

Method

The present study was a quasi-experimental pre-post study with health information technology students at the Tabriz University of Medical Sciences. The sample consisted of 49 students in the fifth through eighth semesters who were enrolled in the study. The instrument used in this study was a researcher-created questionnaire on moral sensitivity designed based on the Lutzen questionnaire in two parts: demographic information and moral sensitivity dimensions. The first part of the questionnaire included information on age, sex, semester, work experience, duration of work experience, and prior study of professional ethics and the second part of the questionnaire included dimensions of ethical orientation (three items), moral strength (two items) and moral responsibility (five items).

The first dimension of moral inclination or moral burden means the "negative" aspect of morality and resembles the experience of moral stress[14], which is something that must be done ethically. The second dimension is about moral power, which means having the courage to act, being able to reason, and having flexibility and endurance[14]. The third dimension is a moral responsibility, reflecting the ethical commitment to work according to laws, regulations, and insights[14].

In this questionnaire, each question was scored on a five-point Likert scale, with a score of always =1, often = 0.75, in some cases = 0.5, rarely = 0.25, and never = 0. The total score range of the questionnaire was 0 to10, with the range of moral tendency score of 3, moral strength of 2, and moral responsibility of 5. Students' moral sensitivity to decision-making based on the total scores of the questionnaire was divided into four categories: very low (scores 0 to 2.5), low (scores 2.75 to 5.75), moderate (scores 5.25 to 7.5), and high (scores 10 to 7.75). The questionnaire was reviewed by a team of nine members (including health information management specialists, medical informatics from the Tabriz School of Management and Information Science, and members of the Medical Ethics Working Group of the Tabriz University of Medical Sciences). To test reliability, the retesting method was used at a distance of 10 days. (Cronbach's alpha= 0.73, CVI= 0.87, CVR= 0.96).

This study was carried out after approval by the Ethics Committee of the Tabriz University of Medical Sciences and control of ethical criteria in the study including confidentiality of information and informed consent of the subjects in the study. All questionnaire information was distributed in a professional ethics training session and completed by students both pre- and post. The data were entered into SPSS 23 software and a paired t-test was used to analyze the data at a significance level of 0.05.

Findings

Of the 49 students in the study, 40 (81.6 percent) were female and 9 (18.4 percent) were male. The mean age of the study samples was 21.5 and the mean work experience was 11 months. In total, 32 (65.3 percent) students had work experience and 17 (34.7 percent) did not. Additionally, 33 (67.3 percent) students had no prior professional ethics training, and 16 (32.7 percent) students had prior professional ethics training.

Table 1 shows average scores of students' moral sensitivity before and after ethics training. According to the results of the paired t-test, the ethical sensitivity score was 7.4 ± 0.7 before education and 7.6 ± 0.8 after education this increase was statistically significant (*P = 0.031*).

In all, 27 (55.1 percent) of students showed moderate moral sensitivity and 22 (44.9 percent) showed high moral sensitivity before ethics training took place, after training, 24 (49 percent) students showed moderate moral sensitivity and 25 (51 percent) students showed high moral sensitivity. In terms of ethical susceptibility, responsibility and moral power subscales scores before and after ethics training were statistically significant (PV< 0.05).

Table 1: Comparison of mean and standard deviation of moral sensitivity score and subscale scores in students before and after ethics training

Moral Sensitivity	Moral Burden Score	Moral Strength Score	Moral Responsibility Score	Moral Sensitivity Score

Before	2.1 ± 0.5	1.1± 0.3	4.4 ± 0.4	7.4 ± 0.7
After	2.1 ± 0.5	1.3 ± 0.3	4.7 ± 0.3	7.6 ± 0.8
Pv	0.51	0.001	0.000	0.031

There was a statistically significant difference between the mean scores of responsibility subscale ethical sensitivity of students without work experience at the time of education (*P =0.000*); however, the mean scores of students with work experience were not statistically significant (*P=0.35*).

As can be seen in Table 2, the moral sensitivity scores of sixth semester students before and after ethics training were statistically significant. There was also a statistically significant difference between the scores of students who did not have work experience while studying. Having work experience and passing a course in ethics seemed to have no significant effect on students' post-training sensitivity scores.

Table2: Comparison of mean and standard deviation of students' moral sensitivity scores before and after ethics training

		Moral Burden Score	Moral Strength Score	Moral Responsibility Score	Moral Sensitivity Score
work experience	Before	2.07 ± 0.5	0.78 ± 0.3	4.4 ± 0.5	7.2 ± 0.7
	After	1.96 ± 0.5	0.79 ± 0.4	4.6 ± 0.4	7.3 ± 0.8
	Pv	0.30	0.74	0.010	0.35
No work experience	Before	2.3 ± 0.4	0.86 ± 0.3	4.4 ± 0.3	7.7 ± 0.5
	After	2.4 ± 0.4	0.89 ± 0.4	4.8 ± 0.2	8.1 ± 0.6
	Pv	0.48	0.75	0.000	0.008
Passing the ethics unit	Before	2.2 ± 0.6	0.64 ± 0.3	4.4 ± 0.5	7.3 ± 0.9
	After	2 ± 0.6	0.54 ± 0.4	4.7 ± 0.4	7.3 ± 0.9
	Pv	0.18	0.13	0.032	0.75
No Passing the ethics unit	Before	2.1 ± 0.4	0.89 ± 0.3	4.4 ± 0.3	7.4 ± 0.6
	After	2.1 ± 0.5	0.96 ± 0.4	4.6 ± 0.3	7.8 ± 0.6
	Pv	1	0.19	0.000	0.017
5th semester students	Before	2.47 ± 0.6	1.38 ± 0.3	4.63 ± 0.5	7.72 ± 0.8
	After	2.31 ± 0.5	1.47 ± 0.3	4.9 ± 0.2	7.77 ± 0.7
	Pv	0.31	0.22	0.16	0.88
6th semester students	Before	2.1 ± 0.3	0.94 ± 0.3	4.4 ± 0.3	7.5 ± 0.4
	After	2.3 ± 0.4	1.05 ± 0.4	4.7 ± 0.3	8.1± 0.6
	Pv	0.16	0.32	0.001	0.001
7th semester students	Before	2 ± 0.6	1.22 ± 0.3	4.5 ± 0.3	7.2 ± 0.7
	After	1.8 ± 0.6	1.59 ± 0.3	4.7 ± 0.2	7 ± 0.7
	Pv	0.13	0.038	0.082	0.055
8th semester students	Before	2.03± 0.4	1.01± 0.23	4.1± 0.4	7. 2 ± 0.8
	After	2 ± 0.6	1.21± 0.28	4.4 ± 0.5	7.6 ± 0.9
	Pv	0.84	0.002	0.028	0.11

Discussion

The results of this study showed that case-based ethics training increased the score of ethical sensitivity of health information technology students. In this regard, the results of various studies have emphasized the positive impact of education on students' moral sensitivity and identifying ethical dilemmas[17, 31-33]. Some studies have addressed the impact of different educational practices on moral sensitivity[7]. For example, some point out that the more objective a training is by exemplifying and using teaching aids, the more moral sensitivity is affected[30]. Another study examines nursing students' experience with exposure to the first case of ethical decision-making in the clinical setting. Film screenings, creating situations similar to those that practitioners and nurses face, and role-playing, far beyond mere theory training, can be effective in teaching students about ethical decision-making[27]. Gaul quotes Bostani as the reason for inappropriate teaching of ethics and states that teaching ethics is not comprehensive enough that students get a good picture of the subject of ethical decision-making and reasoning[34].

Since students will be future employees, it is imperative that they are sensitive to the ethical issues of their profession and that they acquire the necessary skills and competencies before entering the workplace. Education is one of the most important and powerful ways of helping students acquire these skills and enhance their moral sensitivity[31]. Likewise, one of the most important principles in education is the use of appropriate teaching methods.

Accordingly, various studies have referred to a variety of ethics training methods such as simulated environments, formal lectures, group discussions, and so on[7, 20, 23, 24, 32]. One teaching method for ethics that has been emphasized in most of the literature is case-based teaching[3, 28, 32-34]. Case-based teaching is a student-centered approach to teaching that engages students as learners through active learning in small, collaborative groups to solve problems that uses ethically conflicting learning[35]. Case-based learning (CBL) is a teaching approach that engages students as learners through active learning in small, collaborative groups to solve problems that resemble real-world examples. The professor selects problematic situations and asks students to discuss the situation and examine the ethical conflicts of the case and the outcome of the various responses. Discussion of ethical conflicts introduces students to important ethical questions both professionally and socially. Working with cases of ethical conflict is a useful way to understand moral theory. These cases help students identify ethical situations and apply ethics and reasoning enhancing the ethical judgment of students[36, 37].

One reason for using this method is that students have the opportunity to discuss, debate, and present different opinions. As these cases and scenarios reflect real-world situations, students realistically exchange ideas, process and enjoy course content in a real situation[37, 38]. Numerous studies have shown that case-based teaching improves independent learning skills[32], critical thinking, decision-making skills, communication skills, problem-solving skills, the ability to identify relevant issues, and the ability to objectively judge and motivate learning.

The results of the current study showed that ethics education was more effective in those who had not previously studied ethics than in students who had prior exposure. In this respect, Myyry[16] states that technical and professional knowledge has nothing to do with moral sensitivity. To improve sensitivity to ethical issues and to increase awareness and judgment, ethics education must be included in curricula. A study conducted in Korea by Lutzen also found that moral sensitivity was influenced by several factors, including culture, religion, education, age, sex, experience, and education, and varied from person to person[14].

The greatest effect of ethics training in this study was seen in the subscales of moral responsibility and moral power. The results of the present study showed that students with no work experience while studying at school were more likely to have a higher level of moral sensitivity than students with work experience. However, other studies have shown a significant relationship between work experience and students' moral sensitivity scores[38]. In comparing ethical sensitivity scores between sixth and eighth-semester students, there was a statistically significant positive difference in the sensitivity score of the sixth-semester students and their power and moral responsibility subgroups compared to eighth-semester students.

This finding also underscores the impact of students' work experience on students' ethical sensitivity, which may be one reason for students' use of departments that incorporate more routine and repetitive health information management processes, where they perceive no need for complicated and case-based decisions. Therefore, students are in most cases not confronted with ethical conflicts and challenging situations in these settings.

One of the study limitations was the fact that the single-group pre-test and post-test design with no control group. Consequently, there was no conclusion on the causal relationship. Another limitation was the small sample size of the study, although all eligible students were included in it.

Conclusion

The findings of this study indicate that ethics education through case-based learning is an effective strategy for improving students' ethical sensitivity. This teaching method is also useful for actively engaging students in learning and facilitating the learning process. Thus, this method is recommended for ethics education to health information technology students.

References

1. Horton K, Tschudin V, Forget A. The value of nursing: a literature review. Nursing ethics. 2007;14(6):716-40.
2. Borhani fa AK, m;Fazayel,m. Compare moral reasoning ability of nurses and nursing students in Kerman University of Medical Science dealing with ethical issues. Ethics and history of medicine 2010;3(4):71-81.
3. Hasanpour M, Hosseini M, Falahi M. The Effect of Nursing Ethics Training on Ethical Sensitivity in Nurses' Decision-making in Kerman Hospitals. Medical Ethics & History. 2011;4(5).
4. Fitzgerald L, Van Hooft S. A Socratic dialogue on the question ‘What is love in nursing? Nursing Ethics. 2000;7(6):481-91.
5. Borhani F, Alhani F, Mohammadi I, Abbaszadeh A. Professional nursing ethics: its development and challenges. Iranian Journal of Medical Ethics and History of Medicine. 2009;2(3):27-38.
6. GORW R. Ethics in Information Technology2010.
7. Namadi F, Hemmatimaslakpak M, Moradi y, Gasemzade n. The Impact of Professional Ethics Case - Based learning on the Ethical Sensitivity of Nurse Students: A clinical trial. Orumieh Nursing-Midwifery School. 2019.

8. Zubairu U. The Impact of University Education on the Moral Development of Accounting Students. International Online Journal of Education and Teaching (IOJET). 2016;3(2):142-60.
9. Laurinda B. Ethical challenges in the Management of Health Information2006.
10. Khoobi M, Ahmadi F. Maintaining Moral Sensitivity as an Inevitable Necessity in the Nursing Profession. J Caring Sci, 2023, 12(4), x-x.
11. Gerrits EM, Assen LS, Noordegraaf-Eelens L, Bredenoord AL & van Mil MHW. Moral imagination as an instrument for ethics education for biomedical researchers. International Journal of Ethics Education. 2023; 8 (2):275-289
12. AHIMA Code of Ethics, 2019, https://bok.ahima.org/topics/industry-resources/code-of-ethics/
13. Martinov-Bennie N, Mladenovic R. Investigation of the impact of an ethical framework and an integrated ethics education on accounting students' ethical sensitivity and judgment. Journal of Business Ethics. 2015;127(1):189-203.
14. Lützén K, Dahlqvist V, Eriksson S, Norberg A. Developing the concept of moral sensitivity in health care practice. Nursing ethics. 2006;13(2):187-96.
15. Reynolds SJ. Moral awareness and ethical predispositions: investigating the role of individual differences in the recognition of moral issues. Journal of Applied Psychology. 2006;91(1):233.
16. Myyry L. Components of morality: A professional ethics perspective on moral motivation, moral sensitivity, moral reasoning and related constructs among university students. 2003.
17. Nolan PW, Markert D. Ethical reasoning observed: a longitudinal study of nursing students. Nursing Ethics. 2002;9(3):243-58.
18. Grace PJ. Nursing ethics and professional responsibility in advanced practice: Jones & Bartlett Learning; 2017.
19. Woods M. Nursing ethics education: are we really delivering the good (s)? Nursing Ethics. 2005;12(1):5-18.
20. Park M, Kjervik D, Crandell J, Oermann MH. The relationship of ethics education to moral sensitivity and moral reasoning skills of nursing students. Nursing ethics. 2012;19(4):568-80.
21. Stern DT. Practicing what we preach? An analysis of the curriculum of values in medical education. The American journal of medicine. 1998;104(6):569-75.
22. Caramporian A, Mohammadi N, Imany B, Dashtiy S. Evaluating the Consciousness of Professional Ethics Based on Class Based Training in Emergency Medical Students. http://umshaacir/psj. 2014;11(2).
23. Baker DF. When moral awareness isn't enough: Teaching our students to recognize social influence. Journal of Management Education. 2014;38(4):511-32.
24. Walker M. Evaluating the Intervention of an Ethics Class in Students' Ethical Decision-making. 2011.
25. Borhani F, pourama A, Abbaszade A. The effect of case study and simulation teaching method on nursing students' drug calculation skills. Development of Medical Education. 2015;7(16):42-9.
26. Hoseini M, Ebadi M, Farsi Z. The Effect of Moral Motivation Training on Moral Sensitivity in the Nurses of Military Hospitals. Military Caring Sciences Journal. 2018;4(4):249-57.
27. Hosseini M, Ebadi M, Farsi Z. The effect of ethical motivation program on ethical sensitivity of military hospital nurses. Military Care Sciences. 2018;4(4):257-49.

28. Naimi L, Alizadeh M, Shariati M. Case-based learning: Concepts, models, effectiveness, and challenges. Journal of Medical Development and Education. 2017;11(3):201-9.
29. Williams B. Case -based learning—a review of the literature: is there scope for this educational paradigm in prehospital education? Emergency Medicine Journal. 2005;22(8):577-81.
30. Giacalone D. Enhancing Student Learning with Case-Based Teaching and Audience Response Systems in an Interdisciplinary Food Science Course. Higher Learning Research Communications. 2016;6(3):n3.
31. Falakdami A, Takasi P, Effective interventions for improvement of moral sensitivity among nursing students: A systematic review. Journal of Nursing Reports in Clinical Practice. 2023;1:2.
32. Kaya A, BOZ İ. Effects of the Case-Based Learning Approach on the Ethical Sensitivity of Nursing Students: An Experimental Study. Turkiye Klinikleri Journal of Medical Ethics-Law & History / Türkiye klinikleri tıp Etiği, Hukuku ve Tarihi Dergisi, 2023;31(1): 60
33. Zia T, Sabeghi H, Mahmoudirad G. Problem-based learning versus reflective practice on nursing students' moral sensitivity. BMC Nursing. 2023;22:215.
34. Bostani S. Professional Ethics Promotion Solutions in Nursing Education System. Development Strategies in Medical Education. 2016;2(2):23-13.
35. Donkin, R., Yule, H. & Fyfe, T. Online case-based learning in medical education: a scoping review. BMC Med Educ 2023; 23, 564.
36. Miñano R, Uruburu Á, Moreno-Romero A, Pérez-López D. Strategies for teaching professional ethics to IT engineering degree students and evaluating the result. Science and engineering ethics. 2017;23(1):263-86.
37. Li J, Fu S. A systematic approach to engineering ethics education. Science and engineering ethics. 2012;18(2):339-49.
38. Lowry D. An investigation of student moral awareness and associated factors in two cohorts of an undergraduate business degree in a British university: Implications for business ethics curriculum design. Journal of Business Ethics. 2003;48(1):7-19.

Shahla Damanabi, PhD (damanabi46@gmail.com) is an associate professor of health information management in the School of Management and Medical Informatics, at Tabriz University of Medical Sciences, Tabriz, Iran.

Mozhgan Behshid, PhD (Mojganbehshid@hotmail.com) is an assistant professor in the Medical Education Research Center, Health Management and Safety Promotion Research Institute, Nursing & Midwifery Faculty, Tabriz University of Medical Sciences.

Zahra Moradi (moradiz2020@gmail.com) is a Master's student in the Department of Health Information Technology, School of Management and Medical Informatics, at Tabriz University of Medical Sciences.

Leila Ghaderi- Nansa, PhD (leila.gadery@gmail.com) (corresponding author) is an assistant professor of health information management in the School of Management and Medical Informatics, at Tabriz University of Medical Sciences.

Leveraging an Innovation Model to Facilitate ICD-11 Implementation

Kathy L. Giannangelo, MA, RHIA, CCS, FAHIMA, Michael B. Pine, MD, MBA, Christopher P. Tompkins, PhD

Abstract

The 11th Revision of the International Classification of Diseases (ICD-11) with its informatics-based infrastructure has transformed an antiquated classification system into a suite of 21st century computer applications. This manuscript proposes an innovation model to facilitate the implementation of ICD-11 by the US. The model introduces ICD-11 Comprehensive Clinical Linearization, Evolution and Response, or C-CLEAR, a fully coded comprehensive clinical linearization and syntactical rules for combining these codes. These enhancements can be incorporated into electronic coding tools that enable clinical reporters to transmit complex clinical concepts expressed in detailed natural clinical language by means of standardized clusters of ICD-11 stem and extension codes. The model can support rich clinical data captures such as condition acuity and severity, as well as pharmacological treatments. This approach shows promise to accelerate ICD-11 implementation with minimal disruption and maximal net benefits but will require vetting, testing and input from expert stakeholders.

Key words: ICD-11, ICD-10, ICD-10-CM, International Classification of Diseases, ontology, morbidity, interoperability, health information exchange, episode of care, value-based healthcare

Introduction

In 2007, the World Health Organization (WHO) began a revision and restructuring of the International Statistical Classification of Diseases and Related Health Problems, 10th revision (ICD-10) to transform this classification system into a flexible clinical and research friendly structure aligned with advances in information technology. The 11th revision was endorsed by the World Health Assembly at its 72nd meeting in 2019 for implementation beginning in January 2022.[1]

With the release of ICD-11 and its associated architecture, the National Committee on Vital and Health Statistics (NCVHS) began its stakeholder engagement around adoption of ICD-11 for morbidity in the US. Their work in 2019 and 2022 resulted in a set of recommendations to the Department of Health and Human Services (DHHS), [2-4] including that DHHS should conduct research to evaluate the impact of different approaches to implementing ICD-11.[3] Listed as the most important action for this recommendation was an assessment to determine whether ICD-11 can fully support morbidity data collection without the development of a US clinical modification (CM), and if not, which areas might require a CM version (or US-specific extension codes).[3] These recommendations provided a framework from which to create an innovation model to streamline ICD-11 implementation in the US.

In response to the NCVHS recommendations, this manuscript proposes an innovation model to facilitate a seamless transition to ICD-11 by the US. The model introduces ICD-11 Comprehensive Clinical Linearization, Evolution and Response (C-CLEAR), a fully coded comprehensive clinical linearization along with syntactical rules for combining these codes that can translate detailed natural clinical language into standardized coded patient records that exploit the significant advantages of ICD-11.

Background

It was not until 2015, or 25 years after it was endorsed by the WHO, that the US implemented ICD-10-CM for morbidity data collection. While ICD-10-CM was seen as an improvement over the 9th Clinical Modification of ICD,[5,6] a system used since 1979, the implementation was considered highly disruptive and time-consuming, and added significant financial and administrative burdens on physicians and other healthcare providers.[7-10]

More than four years have passed since the 11th revision was endorsed by the World Health Assembly.[1] WHO also has ceased updates to ICD-10.[3] As of February 2023, 64 Member States are in different stages of ICD-11 implementation.[11] Compared to ICD-10, Harrison et al.[12] noted in their review that, "ICD-11 is a different and more powerful health information system, based on formal ontology, designed to be implemented in modern information technology infrastructures, and flexible enough for future modification and use with other classifications and terminologies."

New in ICD-11, all clinical concepts are included in the Foundation, which is a medical knowledge base organized in a poly-hierarchy (in which an entity can descend from more than one branch or parent) that identifies relationships or connections among the entities.[13] Foundation entities of interest are extracted based on use case to form a subset (called a linearization) from the Foundation in the form of a single hierarchy of entities and a corresponding code set. A linearization is the means by which ICD-11 would be accessed by most users, and in the US it ideally would provide backward compatibility to ICD-10-CM.[14] Should each specialty move forward with its own linearization, this would tend to reinforce silos rather than promote integrated information systems and would not readily support use cases that require access to comprehensive and precise information across several or all clinical domains.

The WHO has provided a linearization or tabular list of codes, that is, ICD-11 for Mortality and Morbidity Statistics (MMS), as a potential system for countries to implement or transition in cases where an ICD-10 modification exists. Developed countries that have been using customized modifications of ICD-10 (e.g., ICD-10-CM in the US, -AM in Australia and parts of Asia, and -CA in Canada) are finding that MMS has gaps.[15] The application of ICD-11 codes differs substantially from the ICD-10 coding process. ICD-10-CM consists of tens of thousands of precoordinated codes from which a coder selects the best match to suit the situation. Although MMS has many fewer individual stem codes than ICD-10-CM, a coder can express a clinical concept of interest by using a single stem code if that is sufficient or can form a post-coordinated cluster of stem and extension codes that are combined to capture a more complex concept. Code clusters can represent millions of different clinical scenarios, far surpassing the extent of any library of precoordinated codes. Theoretically, ICD-11 can deliver code expressions for most or all such clinical concepts in clinical modifications of ICD-10. Hindrances include the gaps in MMS, general unfamiliarity among stakeholders with "post-coordination" (clustering and adjoining codes to describe a clinical scenario), and the absence of a sanctioned syntax to provide robust discipline and consistency in the formation of post-coordinated code clusters. A further hindrance is widespread recollection of the transition from ICD-9 to ICD-10 with its fanfare and promised benefits, which failed to materialize in the minds of many observers.

Recognizing the NCVHS recommendations and ICD-11's potential, the authors created a prototype innovation model with a volunteer group of professionals representing medicine, informatics, healthcare data, computer technology, analytics, performance evaluation, economics, and payment. This group developed the innovation model described in this manuscript as a novel approach that could facilitate implementation of ICD-11 by taking full advantage of ICD-11's informatics-based infrastructure and architecture and thereby streamlining transition.

Methods

Comprehensive Code Set

C-CLEAR is an expansion of MMS in which every ICD Entity available in the ICD-11 Foundation is assigned a code. The Foundation is a multidimensional collection of all WHO-Family of Classifications (WHO-FIC) entities.[16] An ICD entity represents a concept, such as a disease, disorder, sign or symptom, or extension code and is assigned a unique Uniform Resource Identifier (URI).[17]

To retain MMS as a common basis for C-CLEAR, all MMS blocks and codes were retained. Each Foundation URI in Chapters 1 through 25 that was included in an aggregated "other specified" code ending in the letter Y was assigned a sequential code in its appropriate series. These new C-CLEAR codes were demarcated with a terminal subscript "underscore CCL" ($\__{\text{CCL}}$). For example, ICD-10-CM has a specific code representing the ICD-11 index term "Hyperplasia, maxillary." In MMS, this clinical entity does not have its own code and is lumped into a Y (other specified) code.

ICD-11 MMS
DA0E.0 Major anomalies of jaw size
    DA0E.00 Micrognathia
    DA0E.0Y Other specified major anomalies of jaw size

whereas C-CLEAR has a specific code for maxillary hyperplasia by expanding on MMS and thus providing a one-to-one map back to ICD-10-CM.

ICD-11 C-CLEAR
DA0E.03$__{\text{CCL}}$ Hyperplasia maxilla
Foundation URI: http://id.who.int/icd/entity/1336634664

Other specialty linearizations of ICD-11 also have placed new specialty-specific codes based on Foundation URIs in the series nested under the appropriate clinical concept in MMS. The advantage of C-CLEAR is that it has already incorporated all such potential codes, making it optimal and user-friendly for all coders regardless of specialty or clinical perspective. Offering

C-CLEAR codes for each entity addresses the limitations of MMS where "other unspecified" codes mark the points where details are truncated. If all such C-CLEAR modifications were eliminated, the result would be MMS.

Composite Linearization

An academic concern might be that providing access to all clinical concepts does not conform to a restriction that is expected for linearizations, which is a single hierarchy that permits each child entity in the linearization to have only one parent. In other words, only one hierarchical set of relationships can be viewed at a time and all other valid branches or pathways that connect concepts in the Foundation are ignored. For example, MMS classifies salmonella pneumonia as a type of infectious or parasitic disease, while in the Foundation, it is both a type of infectious disease and a type of pneumonia. Similarly, in MMS, amebic abscess of the liver is classified as an infectious or parasitic disease but not as a disease of the liver although both are included in the Foundation's poly-hierarchy of clinical concepts.

To augment MMS and overcome this limitation, a composite linearization was created. All the hierarchical relationships residing in the Foundation, and unique C-CLEAR codes for all concepts, are made accessible to users according to their needs. The codes and relationships available in MMS are set as default values. However, alternative parent-child pairs and logical pathways are available whenever those are easier or clearer representations of the patient's situation from the perspective or specialty of the clinician describing the patient.

Going back to our previous examples, an infectious disease specialist might focus on treatment options for amebic abscess or salmonella pneumonia as well as other manifestations of those bacteria in the patient (e.g., other organs or body systems). Meanwhile, the hepatologist and pulmonologist might address the respective body systems and organs holistically, including the abscess or the pneumonia, with secondary mention of the underlying external causes being addressed by the infectious disease specialist.

Clinical Language Syntax

Another component of the C-CLEAR innovation model is its clinical language syntax. MMS imposes the rules of a statistical classification on ICD-11's richly expressive ontology. In contrast, C-CLEAR's syntax is designed to enable clinical reporters to indicate the intended ancestry of each stem code used in a cluster, starting with a primary stem code that best captures the clinical reporter's condition of interest. It then enables the clinical reporter to diverge from the reference MMS taxonomy by introducing a method of designating where and how a stem code's ancestry deviates from MMS's taxonomy. This feature of C-CLEAR enables a clinical reporter to communicate his or her clinical message utilizing the linguistic power of the entire ICD-11 ontology.

Another feature enables C-CLEAR to create uniquely ordered clusters of codes for complex concepts, mimicking the one-to-one relationship of a clinical concept to a single pre-coordinated code. C-CLEAR and its syntax follow ICD-11 MMS conventions including the use of stem codes as clinical concepts and extension codes as modifiers. Each unique C-CLEAR code can be mapped to a single ICD-11 Foundation URI.

Results

Over the past year, the authors and their collaborators have made progress with the conceptual logic and an instantiation of the proposed innovation model. This includes the creation of C-CLEAR and its syntax.

We also have created clinical scenarios demonstrating the capability of C-CLEAR and its syntax to provide clinically credible representations of the detailed evolution of patients' health status. These are intended to capture the clinical justification or appropriateness of medical interventions, document important changes in patients' health in response to medical care, and provide representations superior to anything to date based on either ICD-9-CM or ICD-10-CM.

Table 1 illustrates a simple medical scenario that compares the descriptive power of ICD-9-CM, ICD-10-CM, and C-CLEAR. The table depicts the six stages of the clinical progression of a female patient who first presents with aortic valve insufficiency due to aortic dilation and eventually is referred for a cardiac valve operation. The issue addressed here is how clearly the patient scenario is captured by each of the disease classification systems.

1. **ICD-9-CM and ICD-10-CM.** From ICD-9-CM, one can conclude this is a patient with aortic valve disease and thoracic aortic dilation. In ICD-10-CM, we know more specifically her condition was nonrheumatic aortic insufficiency (with the incremental details italicized in the table). Later, the patient had hyperpotassemia due to adverse effects from an antihypertensive agent, described in ICD-10-CM as hyperkalemia due to an ACE inhibitor. This complication apparently resolved. By stage four of this vignette, the patient had developed congestive heart failure (CHF). It is unclear why an aortic valve replacement was ultimately indicated and why it was recommended in stage six rather than in stage four or stage five, all of which appear identical in the coded data.

2. **ICD-11 C-CLEAR.** From C-CLEAR one learns that this patient had chronic mild aortic valve insufficiency due to thoracic aortic dilation that progressed from mild to moderate, after which it was treated with valsartan. This treatment resulted in hyperkalemia, so her medication was changed to benazepril. Her hyperkalemia resolved, but she developed chronic New York Heart Association (NYHA) Class II CHF as a complication of her aortic valve insufficiency. Treatment with furosemide resulted in lessening of her heart failure to NYHA Class I. However, her underlying aortic valve insufficiency progressed from moderate to severe, with a marked worsening of her chronic heart failure to NYHA Class III. The indications and timing for an aortic valve replacement are now made clear.

Precoordinated ICD-9-CM and ICD-10-CM codes, while lacking in important clinical detail, are easily interpretable representations. Because each accessible concept is represented by a single code, the challenge for coders is to identify which of a plethora of codes comes closest to the concept a clinical reporter wishes to convey. In contrast, C-CLEAR codes and syntax permit clinical reporters to convey nuanced clinical information in single, unique, systematically organized clusters of codes. However, these computer-friendly codes and clusters are not readily decipherable by a general clinical audience.

Fortunately, clinicians can generate and decipher C-CLEAR clusters with only rudimentary knowledge of the coding system itself. To make this possible, coding tools based on those created by the WHO to support ICD-11 MMS are being created. These enhanced coding tools will be able to translate 'natural clinical language' into C-CLEAR coded clusters. These clusters can support sophisticated analyses of the evolution of the health of individuals and populations and of the appropriateness, quality, and cost-effectiveness of diagnostic and therapeutic interventions and the care provided by healthcare practitioners and organizations. These tools also will be capable of transforming C-CLEAR clusters back into natural clinical language to allow clinicians to determine how well their clinical information has been captured, to revise their original input as needed to improve C-CLEAR coding, and to generate documentation consistent with coded data submitted for reporting, evaluation, and reimbursement.

Discussion

Accurate diagnoses are an essential element in providing appropriate and timely care to patients. However, as illustrated in Table 1, a wide range of clinical states can exist within single diagnostic categories. Curing diseases and reducing patient burden from diagnosed conditions are both essential elements of high-quality medical care. Similarly, patient-centered quality measurement, analytics, and fair payment of healthcare providers all require knowledge about each patient's diagnoses, general health status, and related functional and socioeconomic factors as addressed in ICD-11, along with detailed information about the progression and regression of individual diagnosed conditions.

Furthermore, risk-adjustment based solely on diagnosed conditions without clinical details regarding the severity and complexity of these conditions can be anemic at best, or even misleading when systematic biases are present. This is a fatal flaw in the current data used for quality comparisons, performance evaluations, and alternative payment models.

Finally, the rationale and appropriateness for medical treatment and management as embodied in clinical guidelines require details available only in patient records. The adoption of ICD-11 could upgrade standard claims databases from catalogs of diagnoses, procedures, and costs to genuine clinical and research tools with new applications to monitor, improve, and pay for healthcare.[18] Moreover, melding the EHR data and standardized claims data could eliminate current administrative redundancies.

The US is investing heavily in developing and supporting information technology innovation models. Several government agencies have established programs and provided funding for projects to accelerate the next generation of interoperable health information technology. For example, the Centers for Medicare & Medicaid Services (CMS) established the CMS Innovation Center to support development and testing of innovative healthcare payment and service delivery models.[19] In 2020, the Centers for Disease Control and Prevention (CDC) launched the data modernization initiative intended to modernize core data and surveillance infrastructure across the federal and state public health landscape.[20] The Office of the National Coordinator for Health

Information Technology (ONC) Leading Edge Acceleration Projects (LEAP) in Health Information Technology (IT) provides funding for projects that support the adoption of health IT and the promotion of nationwide health information exchange (HIE). ONC recently issued a Special Emphasis Notice stating that it "is critical that the field of health care innovate and leverage the latest technological advancements and breakthroughs far quicker than it currently does to optimize real-time solutions, especially in areas which are ripe for acceleration."[21] Our proposed innovation model would complement and enrich these initiatives.

The next logical step is to pilot test the innovation model. The NCVHS August 3rd meeting discussed the need to pilot test ICD-11 options prior to implementation.[22] In addition, WHO has indicated interest in pilot testing.[23] We welcome the opportunity to vet our approach among clinical and classification experts and to test C-CLEAR's codes, architecture, syntax, and associated electronic coding tools for their stated purposes.

It is hoped this system would:

1. facilitate efficient user-friendly coding that completely and accurately captures the clinical information clinical reporters wish to convey,
2. produce output that is intuitively obvious to clinicians,
3. support the generation of coded ICD-11 data directly from EHRs including free text, and
4. enable analysts to manipulate these coded data to support important use cases.

C-CLEAR and its syntax will also serve as a framework for upgrading the Episode Grouper for Medicare (EGM),[24] which was developed in response to a provision in the Affordable Care Act[25] that directed CMS to create a public sector episode grouper with standard definitions of clinical conditions and procedures to support analyses, reimbursement, and other applications by all healthcare stakeholders. Work is currently underway to create a standard set of clinically nuanced episodes of care that will take full advantage of enhanced ICD-11 data capabilities to support the exchange, interpretation, and application of information among healthcare providers and other stakeholders. This upgraded episode grouper could facilitate a wide variety of extremely useful applications and replace some ineffective, inefficient analytic and operational applications that appear to be creating as many or more problems than they were designed to solve.

Conclusion

Unlike the transition to ICD-10, the adoption and implementation of ICD-11 would represent a major advance in medical informatics. The transformation of a collection of words into an architecture and syntax, a global language of sorts, enhances available information beyond diagnosed conditions to include how clinical progression within diagnoses and progressive interactions among diagnoses in different states affect a patient's overall health status. ICD-11 also comes with a suite of 21st century computer applications that can be enhanced to support easy adoption, meaningful data sharing, and improved patient care.

When ICD-11's informatics-based infrastructure is utilized to its fullest extent within the proposed innovation model, there is great potential to support clinically useful evaluations of the evolution of the health status of individual patients and populations and the contribution of alternative healthcare services to health and well-being. For example, C-CLEAR and its syntactical rules for combining these codes could conceivably become a universally applicable translator of parochial terms into a language that retains the important clinical details required to accurately monitor each patient's clinical pathway. In addition, interaction with EHRs to locate and code clinical details, add key information to claims, increase interoperability, and invigorate many applications such as quality reporting and value-based payments may be possible. Using the model in such a manner could transform an inefficient healthcare payment and disjointed service delivery system into a cost-effective, coordinated, patient-centered healthcare ecosystem. And finally, the proposed approach shows promise of a faster and smoother transition to ICD-11, reduced administrative burden, seamless electronic healthcare information exchange, increased interoperability of electronic health information, and facilitation of a wide range of applications to foster the evaluation and improvement of the quality and cost-effectiveness of healthcare.

And finally, a next step is for the US and other countries should be to rigorously test ICD-11 linearizations for their ability to meet the many demands these countries have for accurate information in clinical care, clinical research, and secondary data use cases related to public health and policy. Specifically, the US needs to develop and embrace an approach that will justify an expensive Federal mandate to adopt ICD-11 via legislation or regulation. This research program should pilot test the implementation process by integrating ICD-11 into realistic health information technology environments and informing the industry with guidance and lessons learned on "how to" adopt ICD-11.

Furthermore, resulting data sets should be used to address the question of "why" adopt ICD-11. For example, with ICD-11, researchers could simulate potential net benefits to be expected in important use cases such as accurately and reliably measuring efficiency and quality of care. While the US should not rush to adopt ICD-11 merely because its developers wish it would, it should not remain stuck on the aging ICD-10-CM if it can do better, nor should it postpone meaningful reforms to accommodate stakeholders that are prospering despite its inability to achieve a sustainable, high-value healthcare system.

Acknowledgements
The authors thank the following individuals for their contribution during manuscript preparation: Charles Hobson, Denise Love, Greg Wozniak, and John Martin

References

1. "The 72nd World Health Assembly resolution for ICD-11 adoption | WHO." May 27, 2019. https://www.who.int/publications/m/item/eleventh-revision-of-the-international-classification-of-diseases-adoption-wha72

2. "Subcommittee on Standards – ICD-11 Evaluation Expert Roundtable Meeting." August 6-7, 2019. National Committee on Vital and Health Statistics. https://ncvhs.hhs.gov/meetings/subcommittee-on-standards-icd-11-evaluation-expert-roundtable-meeting/

3. "Recommendation to HHS Secretary on Preparing for Adoption of ICD-11." November 25, 2019. National Committee on Vital and Health Statistics. https://ncvhs.hhs.gov/wp-

content/uploads/2019/12/Recommendation-Letter-Preparing-for-Adoption-of-ICD-11-as-a-Mandated-US-Health-Data-Standard-final.pdf

4. “Updated recommendations for immediate action on ICD-11 to HHS Secretary.” September 10, 2021. National Committee on Vital and Health Statistics. https://ncvhs.hhs.gov/wp-content/uploads/2021/09/NCVHS-ICD-11-recommendations-for-HHS-Sept-10-2021-Final-508.pdf

5. “Transitioning to ICD-10 | CMS.” February 25, 2015. CMS. https://www.cms.gov/newsroom/fact-sheets/transitioning-icd-10

6. “ICD-10-CM - International Classification of Diseases, ICD-10-CM/PCS) Transition | CDC.” November 6, 2015. CDC. https://www.cdc.gov/nchs/icd/icd10cm_pcs_background.htm

7. Libicki, Martin and Irene Brahmakulam. March 2004. “The costs and benefits of moving to the ICD-10 code sets.” Rand Science and Technology. https://www.rand.org/content/dam/rand/pubs/technical_reports/2004/RAND_TR132.pdf

8. Deloitte. 2010. “ICD-10 implementation for health care providers: The business imperative for compliance. https://www.healthit.gov/sites/default/files/facas/us_lshc_icd-10implementationforhealthcareproviders_0810.pdf

9. Fiegl, Charles. “Organized medicine urges CMS to halt ICD-10 switch.” January 7, 2013. American Medical News. Accessed April 18, 2023. https://amednews.com/article/20130107/government/130109986/7

10. Nachimson Advisors, LLC. February 12, 2014. “The cost of implementing ICD-10 for physician practices – updating the 2008 Nachimson Advisors study.” https://docs.house.gov/meetings/IF/IF14/20150211/102940/HHRG-114-IF14-Wstate-TerryW-20150211-SD001.pdf

11. World Health Organization “ICD-11 2023 release is here.” February 14, 2023. . https://www.who.int/news/item/14-02-2023-icd-11-2023-release-is-here

12. Harrison, James E., Stefanie Weber, Robert Jakob, and Christopher G. Chute. 2021. “ICD-11: an International Classification of Diseases for the twenty-first century.” *BMC Medical Informatics and Decision Making*, 21 (S6). https://doi.org/10.1186/s12911-021-01534-6

13. Tu, Samson W. Olivier Bodenreider, Can Çelik, Christopher G. Chute, Sam Heard, Robert Jakob, Guoquian Jiang , Sukil Kim , Eric Miller , Mark M. Musen , Jun Nakaya, Jon Patrick, Alan Rector, Guillermo Reynoso, Jean Marie Rodrigues, Harold Solbrig, Kent A Spackman, Tania Tudorache, Stefanie Weber, and Tevfik Bedirhan Üstün. 2015. “A content model for the ICD-11 revision.” https://getinthepicture.org/resource/content-model-icd-11-revision

14. Chute, Christopher G. and Can Çelik, 2022. "Overview of ICD-11 architecture and structure." *BMC Medical Informatics and Decision Making*, 21 (Suppl 6), 378. https://doi.org/10.1186/s12911-021-01539-1

15. Fung, Kin W., Julia Zu, Shannon McConnel-Lamptey, Donna Pickett, and Olivier Bodenreider. 2023. "A practical strategy to use the ICD-11 for morbidity coding in the United States without a clinical modification." *Journal of the American Medical Informatics Association*. July. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10531107/

16. WHO-FIC Content Model Reference Guide, World Health Organization (WHO) 2021 https://icd.who.int/browse11. Licensed under the Creative Commons Attribution-NoDerivatives 3.0 IGO licence (CC BY-ND 3.0 IGO)

17. Ibid.

18. Fenton, Susan H., Kathy L. Giannangelo, and Mary H. Stanfill. 2021. "Preliminary study of patient safety and quality use cases for ICD-11 MMS." *Journal of the American Medical Informatics Association*. September. https://doi.org/10.1093/jamia/ocab163

19. "CMS Innovation Center Homepage | CMS Innovation Center." 2015. Cms.gov. 2015. https://innovation.cms.gov

20. "Data Modernization Initiative Basics | CDC." 2023. www.cdc.gov. January 9, 2023. https://www.cdc.gov/surveillance/data-modernization/basics/index.html

21. "Leading Edge Acceleration Projects (LEAP) in Health Information Technology (Health IT) Notice of Funding Opportunity (NOFO) | HealthIT.gov." April 10, 2023. www.healthit.gov. https://www.healthit.gov/topic/onc-funding-opportunities/leading-edge-acceleration-projects-leap-health-information

22. "Workgroup on Timely and Strategic Action to Inform ICD-11 Policy – ICD-11 Expert Roundtable Meeting." August 3, 2023. National Committee on Vital and Health Statistics. https://ncvhs.hhs.gov/wp-content/uploads/2023/10/2023-August-3_NCVHS_ICD-11-Expert-Roundtable-Meeting-Summary-FINAL-508.pdf

23. WHO-FIC ICD-11 Implementation or Transition Guide, World Health Organization (WHO) 2019 https://icd.who.int/en/docs/ICD-11%20Implementation%20or%20Transition%20Guide_v105.pdf. Licenses under the Creative Commons Attribution (CC BY-NC-SA 3.0 IGO)

24. Tompkins, Christopher P. 2016. Episode Grouper for Medicare (EGM): Design Report. https://i11forum.org/EGM_Final_Report.pdf

25. *The Patient Protection and Affordable Care Act*. 2010. §§3003, 3007. https://www.govinfo.gov/content/pkg/PLAW-111publ148/pdf/PLAW-111publ148.pdf

Kathy L. Giannangelo, MA, RHIA, CCS, FAHIMA*, (kathy.giannangelo@gmail.com) is a health information and informatics professional. As president of Kathy Giannangelo Consulting, LLC, her work includes clinical terminology feasibility assessments, adoption methodology appraisals, and implementation support.*

Michael B. Pine, MD, MBA*, is a board-certified internist and cardiologist. He is the founder and managing director of MJP Healthcare Innovations, a research and development firm focused on measuring and improving healthcare quality and cost-effectiveness.*

Christopher P. Tompkins, PhD*, is associate research professor at Brandeis University. He led the development of several key innovations including the Medicare shared savings payment model, the hospital value-based payment model, the Episode Grouper for Medicare, and the ACS-Brandeis advanced alternative payment model (with the American College of Surgeons).*

Table 1. How the Clinical Progression of Aortic Valve Insufficiency Due to Thoracic Aortic Dilation would be represented in ICD-9-CM, ICD-10 CM, and ICD-11 C-CLEAR (text only)

Stage	ICD-9-CM	ICD-10-CM	ICD-11 C-CLEAR
1	Aortic valve disorders Thoracic aortic ectasia	*Nonrheumatic* aortic *insufficiency* Thoracic aortic ectasia	*Chronic mild* aortic valve insufficiency *due to* thoracic aortic dilation
2	Aortic valve disorders Thoracic aortic ectasia	*Nonrheumatic* aortic *insufficiency* Thoracic aortic ectasia	*Chronic moderate* aortic valve insufficiency *due to* thoracic aortic dilation
3	Aortic valve disorders Thoracic aortic ectasia	*Nonrheumatic* aortic *insufficiency* Thoracic aortic ectasia	Initial encounter for hyperkalemia caused by drugs, medicaments, or biological substances (i.e., *valsartan*) associated with injury or harm in therapeutic use in context of correct administration or dosage for *chronic moderate* aortic valve insufficiency *due to* thoracic aortic dilation
	Coronary vasodilators causing adverse effects in therapeutic use	Adverse effect of *angiotensin converting enzyme inhibitors*, *initial encounter*	
	Hyperpotassemia	Hyperkalemia	
4	Congestive heart failure	Congestive heart failure, unspecified	*Initial encounter* for *acute NYHA Class II* congestive heart failure; *secondary to chronic moderate* aortic valve insufficiency *due to* thoracic aortic dilation, *treated with - benazepril*
	Aortic valve disorders Thoracic aortic ectasia	*Nonrheumatic* aortic *insufficiency* Thoracic aortic ectasia	
5	Congestive heart failure	Congestive heart failure, unspecified	*Subsequent encounter* for *chronic NYHA Class I* congestive heart failure, *prescribed - furosemide*; *secondary to chronic moderate* aortic valve insufficiency *due to* thoracic aortic dilation, *treated with - benazepril*
	Aortic valve disorders Thoracic aortic ectasia	*Nonrheumatic* aortic *insufficiency* Thoracic aortic ectasia	
6	Congestive heart failure Aortic valve disorders Thoracic aortic ectasia	Congestive heart failure, unspecified *Nonrheumatic* aortic *insufficiency* Thoracic aortic ectasia	*Subsequent encounter* for *chronic NYHA Class III* congestive heart failure, *prescribed - furosemide*; *secondary to severe chronic* aortic valve insufficiency *due to* thoracic aortic dilation, *treated with - benazepril*

TABLE 2. How the Clinical Progression of Aortic Valve Insufficiency Due to Thoracic Aortic Dilation would be represented in ICD-9-CM, ICD-10 CM, and ICD-11 C-CLEAR (codes and titles only)

Stage	ICD-9-CM	ICD-10-CM	ICD-11 C-CLEAR
1	424.1 - aortic valve disorders 447.71 - thoracic aortic ectasia	I35.1 - nonrheumatic aortic insufficiency I77.810 - thoracic aortic ectasia	BB71.1$_{\_CCL}$ - aortic valve insufficiency due to aortic dilation XT8W - chronic XS5W - mild
2	424.1 - aortic valve disorders 447.71 - thoracic aortic ectasia	I35.1 - nonrheumatic aortic insufficiency I77.810 - thoracic aortic ectasia	BB71.1$_{\_CCL}$ - aortic valve insufficiency due to aortic dilation XT8W - chronic XS0T - moderate
3	424.1 - aortic valve disorders 447.71 - thoracic aortic ectasia	I35.1 - nonrheumatic aortic insufficiency I77.810 - thoracic aortic ectasia	5C76 - hyperkalemia XY18 - initial encounter PL00 - drugs, medicaments or biological substances associated with injury or harm in therapeutic use XM29M2 - valsartan PL13.2 - drug-related injury or harm in the context of correct administration or dosage, as mode of injury or harm BB71.1$_{\_CCL}$ - aortic valve insufficiency due to aortic dilation XT8W - chronic XS0T – moderate
	E942.4 - coronary vasodilators causing adverse effects in therapeutic use	T46.4X5A - adverse effect of angiotensin converting enzyme inhibitors, initial encounter	
	276.7 - hyperpotassemia	E87.5 - hyperkalemia	
4	428.0 - congestive heart failure	I50.9 - congestive heart failure, unspecified	BD10.0$_{\_CCL}$ - acute congestive heart failure XY18 - initial encounter XS6B - NYHA Class II - slight limitation of physical activity BB71.1$_{\_CCL}$ - aortic valve insufficiency due to aortic dilation XT8W - chronic XS0T – moderate XM0HG1 - benazepril
	424.1 - aortic valve disorders 447.71 - thoracic aortic ectasia	I35.1 - nonrheumatic aortic insufficiency I77.810 - thoracic aortic ectasia	
5	428.0 - congestive heart failure	I50.9 - congestive heart failure, unspecified	BD10.1$_{\_CCL}$ - chronic congestive heart failure XY8S - subsequent encounter XS3A - NYHA Class I - no limitation of physical activity XM8UE3 - furosemide BB71.1$_{\_CCL}$ - aortic valve insufficiency due to aortic dilation XT8W - chronic XS0T – moderate XM0HG1 - benazepril
	424.1 - aortic valve disorders 447.71 - thoracic aortic ectasia	I35.1 - nonrheumatic aortic insufficiency I77.810 - thoracic aortic ectasia	
6	428.0 - congestive heart failure	I50.9 - congestive heart failure, unspecified	BD10.1$_{\_CCL}$ - chronic congestive heart failure XY8S - subsequent encounter XS9T - NYHA Class III - marked limitation of physical activity XM8UE3 - furosemide BB71.1$_{\_CCL}$ - aortic valve insufficiency due to aortic dilation XT8W – chronic XS25 – severe XM0HG1 - benazepril
	424.1 - aortic valve disorders 447.71 - thoracic aortic ectasia	I35.1 - nonrheumatic aortic insufficiency I77.810 - thoracic aortic ectasia	

Perspectives on Big Data and Big Data Analytics in Healthcare

Egondu R. Onyejekwe PhD, Dasantila Sherifi, PhD, MBA, RHIA, and Hung Ching, PhD, DABR

Abstract

Big data (BD) is of high interest for research and practice purposes because it has the potential to provide insights into the population served and healthcare practices. Much progress has been made in collecting BD and creating tools for big data analytics (BDA). However, healthcare organizations continue to experience challenges associated with BD characteristics and BDA tools. Utilization of BD impacts current decision-making, planning, and future use of artificial intelligence (AI) tools, which are trained on BD. This qualitative study focused on better understanding the reality of BD and BDA management and usage by healthcare organizations. Six structured interviews were conducted with individuals who work with healthcare BD and BDA. Findings confirmed the known challenges associated with BD/BDA and added rich insights into the structural, operational and utilization aspects, as well as future directions. Such perspectives are valuable for education and improvements in BD/BDA management and development.

Keywords: big data, big data analytics, health records, digital data, population health, artificial intelligence

Introduction

The implementation of electronic health records (EHRs) and widespread information systems and applications for providers, consumers, and other parties have led to tremendous growth of electronic health data. The current sources of data include mostly textual content, which can be structured, semi-structured or unstructured. They also include videos, audios, and images that constitute multimedia. They can come from a variety of platforms such as machine-to-machine communications, social media sites, sensor networks, cyber-physical systems, and Internet of Things (IoT).[1] These platforms begin to define big data (BD) because they make us think about size, volume, complexity, and heterogeneity of the data emanating every second from a variety of devices.

BD arrived sooner than the development of appropriate and efficient analytical methods for its analysis. In addition to the structured data, BD includes massive volumes of heterogeneous data in unstructured text, audio, video, and other formats, and so is not amenable to the inferences of statistical methods that are used for analyzing numerical structured data. Unstructured BD requires new tools for predictive analytics. In addition, there is a need for computationally efficient algorithms to handle the heterogeneity, noise, and massive size of structured BD. These are ways to dispel and/or avoid potential spurious correlations.

Artificial intelligence (AI) and data analytics are top technology priorities as they capitalize on sustainability through data analytics and adaptive AI.[2] For over a decade, Mayer-Schönberger and Cukier encouraged datafication of BD, where essentially, virtually anything is transformed into useful data (insights) by documenting, measuring, and capturing digitally.[3] Van Dijck asserted that the future of BD and big data analytics (BDA) will lie with machines, where data will be generated, shared, and communicated among data networks.[4] After a decade of progress, much of the structured and unstructured data stored in EHRs can be analyzed with the use of natural language processing (NLP) and machine language processing (MLP) algorithms, which can unlock the value of the text and galvanize the extraction of the hidden insights and connectors.[1] Transforming unstructured text into real patient insights holds great potential for improving health outcomes. Use of AI and BDA for clinical and non-clinical applications in healthcare has great potential, however, the majority of healthcare organizations have yet to reach the full benefits of their BD. This highlights the need to better understand the status quo of how big data is being handled and analyzed by healthcare organizations. What are some of the ways big data is being used and what are some of the challenges faced by healthcare organizations when it comes to working with big data? A deeper dive into how organizations use big data, how much they invest in big data technologies, and what challenges they experience creates an opportunity to identify and share some best practices, as well as identify potential gaps. Where the findings are translated into real patient insights and where such knowledge fosters better health outcomes, there may be opportunities for positive change in terms of improving population health, addressing health inequalities, improving operations, and reducing healthcare costs.

Background and Significance

Big Data

BD refers to data sets that are so large or complex with high volume, high velocity, and high variety that they cannot be processed by traditional data processing software in a reasonable amount of time, thus, requiring advanced techniques and technologies for management and analytics.[5,6,7,8] BD can be described by characteristics such as volume, variety, velocity, variability, veracity, and value.

BD is inherently defined by big *volume*.[9] The quantity of generated and stored data is usually reported in multiple terabytes and petabytes – where a terabyte stores enough data to fit on 1500 CDs or 220 DVDs. A terabyte of data would store approximately 16 million Facebook photographs. The volume of data in healthcare continues to grow because information is increasingly gathered not only systematically in systems used by hospitals, pharmacies, laboratories, insurance, research institutions, or genetic databases, but also by numerous information sensing IoT devices used by providers, patients, and other parties. The size of the data is believed to account for its value as well as its potential insight. Volume-related challenges are related to storage and data management technologies.

The type and nature or the structural heterogeneity of the data describes its *variety*.[9] Structured data, mostly tabular data, found in spreadsheets and relational databases constitute about 20 percent of healthcare data.[10] Unstructured data includes mostly text, images, audios, and videos. Semi-structured data may or may not conform to strict standards and include textual language for Web data exchange, called Extensible Markup Language (XML), that deploys user-defined data tags to make them machine readable. BD variety becomes even more complex given the diverse sources and formats, requiring that data from those sources be connected, matched, cleansed, and transformed.

At the heart of big data is *velocity,* which measures the rate of data generation and the speed at which the data is analyzed and acted upon to meet the demands and challenges that lie in the path of growth and development of organizations.[9] Smart phones, digital sensors, and other devices, using mobile apps produce enormous and useful information about customers (or patients) that include geospatial location, demographics, buying and viewing patterns, and even physical activity or other health indicators tracked by mobile apps. These types of data can be analyzed in real time to harness real-time intelligence.

Another dimension of BD is *variability,* which implies the inconsistency or variation in the data flow (whereas velocity shows periodic peaks and troughs).[9] Variability can hamper processes that manage BD.

Veracity reflects the "truthfulness" of data and was added as BD characteristics by IBM, given their specialization in removing and replacing BD errors.[11] Addressing the imprecision and uncertainty becomes relevant for BD because of the inherent unreliability in certain data sources. The quality of captured data may vary tremendously, thus affecting the accurate analysis and results.

Lastly, BD is generally associated with *value*, which means that when large volumes of BD are analyzed, it is possible to extract high value from them.[8] The original form of data has low value, but the information identified through its analysis can make a difference in its value. For that to happen, data should be relevant and of high integrity.

Big Data Analytics

BDA involves the analysis of BD. It is during this process that the value of big data for decision support and business intelligence is realized. Given BD characteristics, BDA cannot be derived by simple statistical analysis.[12,13] In fact, use of advanced BDA tools and extremely efficient, scalable, and flexible technologies are necessary to efficiently manage and analyze the substantial amounts and variety of data.[1,14] Technologies such as NoSQL Databases, BigQuery, MapReduce, Hadoop, WibiData, and Skytree have been in use for more than a decade.[15] AI tools such as Microsoft Power BI, Microsoft Azure Machine Learning QlikView, RapidMiner, Google Cloud AutoML, or IBM Watson Analytics are offering greater value in BDA. For example, Microsoft Power BI was successfully used to detect specific antenatal data for babies small for gestational age (SGA) and monitor them through a dashboard, thus allowing clinicians to intervene and plan delivery as necessary.[16]

BD management entails both the processes and the associated technologies that allow for the acquisition, storage, and retrieval of data, which can be done in three stages: acquisition/recording; extraction, cleaning, and annotation; and integration, aggregation, and representation.[17,18] Analytics involves the techniques applied in analyzing and acquiring intelligence from BD and can be completed in two stages: modeling and analysis; and interpretation. It becomes imperative that processing and management should be efficient enough to expose new knowledge in a timely manner, which is crucial for capitalizing on emerging opportunities, in providing a competitive edge, as well as rich business intelligence used to differentiate the organization, increase visibility, flexibility, and responsiveness to environmental changes.[19,20,21,22,23] The allure in healthcare BDA is the ability to examine and apply the patterns that emerge from various and vast amounts of healthcare data to predict trends in population health and ways to improve it, while limiting costs. BDA benefits are already visible in reduced administrative costs, improved clinical decision support, better care coordination, reduced fraud and abuse; as well as improved patient wellness.[24] Adoption of mHealth, eHealth and wearable technologies will push the increase in BD volume. Increased integration of such data with EHRs, imaging, patient generated data, or sensor data create even greater opportunities to leverage BD in healthcare.

Much of the BD and BDA research demonstrates success in use of BD and BDA tools such in monitoring SGA babies, response to COVID in Taiwan, or use of BD in mental health care.[16,25,26] One study also highlights issues with big data privacy [27] (Golbus, W Nicholson & Brahmajee 2020.)[27] Other studies help in understanding BD and BDA concepts through reviews, analyses, and summaries.[19,28,29] In our study, we focused on the healthcare organizational structure regarding big data, the approach in integrating big data into operations, issues and challenges experienced, and the vision for BDA. Our research question was “How are healthcare organizations handling BD and BDA?” Better understanding of this reality serves not only to share best practices or challenges but also to inform decisions on resource allocation and opportunities for education of professionals to work with BD and BDA.

Methodology

The purpose of this study was to gain greater understanding on how BD and BDA are handled within healthcare organizations. To gain such perspective, the study evaluated experiences of professionals with healthcare BD and BDA. For this applied research, we followed the case study method, a qualitative research design.[30] Case studies help explore an activity or process in depth and allow for detailed data collection through interviews of one or more individuals.[31,32] The research was approved by the Institutional Review Board at Walden University.

The sampling strategy was purposeful and convenient. The research team focused on identifying individuals from various settings who worked with BD and BDA. Based on professional connections and LinkedIn profiles, we reached out to nine individuals in such roles (not all at once); over time, only six of them were available to participate in the study. We conducted six structured interviews with individuals whose main work was managing and/or analyzing healthcare big data. The interviews were completed virtually via Zoom and lasted between 45 and 60 minutes each. The principal investigator conducted structured interviews by following the pre-established interview protocol, which included an introduction to the study and researchers, verbal agreement to participate in the study, and questions in order, as presented below. Probes were also used at times to elaborate on some of the answers with further details and/or examples. The other two researchers were present during all interviews, recorded, and took notes. All interviewees were asked the following 11 standard open-ended questions:

1. Can you please describe your role and how your organization’s big data team is structured for data collection and data analytics?
2. What investments has your organization made to drive or support big data analytics?
3. Can you briefly describe the *types of questions* your organization answers by using big data analytics?
4. Can you briefly describe the *types of decisions* that are based on big data?
5. What is your organization’s approach for integrating data analytics into operations?
6. Sometimes a game changing opportunity arises, but the opportunity does not get vetted with evidence from the big data. Have you seen this happen in your organization? If so, can you give an example?
7. How does your organization use big data to support population health?
8. Now I’d like to focus on challenges in using big data. What are some of the frequent problems that big data analysts in your organization encounter?
9. What are some solutions or approaches you have employed to overcome those challenges?
10. Now, let’s talk about non-healthcare organizations that use healthcare big data.
    a. What are your thoughts on how device manufacturers, pharma, and insurance companies benefit from healthcare big data?
    b. What are your thoughts on how data companies such as Google, Amazon, and Microsoft benefit from healthcare big data?

11. Finally, let's talk about the future.
    a. What are your thoughts on how your organization will use big data in the future?
    b. Are there any new tools or resources your organization plans to use to improve the usage of big data and the experience with big data analytics?
    c. Given sufficient resources, what is your vision for an effective and efficient data analytics program in your organization?

After each interview, researchers discussed the main points that came out during the interviews. After the sixth interview, it was determined that the saturation point was reached, and no further outreach was made for additional interviews.[33]

The transcribed interviews were analyzed by using a summative content analysis approach. The summative approach focuses on identifying the essential aspect of the text and has been used successfully in analyzing interviews from healthcare professionals to examine complex text from diverse sources, including innovation in services or technology, which is similar to our research.[34] This approach is also accommodating to differences (as opposed to only similarities), which is important in our study, given the diverse roles of interviewees and their experiences with BD and BDA.

Responses were coded based on the topics addressed through questions. Codes were aggregated into concept maps to group related codes into themes and show relations. While the use of standardized open-ended questions facilitated the data organization and analysis, some portions of answers that were provided under a certain question were moved to areas where they fit the topics better. For example, responses to questions 1 through 6 were categorized into: interviewee roles; organizational structure for BD and BDA; purpose of using BD and BDA; and dynamics/processes of using BD and BDA. The rest of the themes such as use of BD for population health, BD/BDA challenges, approaches in addressing such challenges, use of BD by non-healthcare organizations, and future directions were consistent with the questions asked. Another important note is that due to the diversity of the interviewees and organizations they represented, response analysis are mostly broken down by the type of organization.

Responses were coded by two researchers independently and discussed. No discrepancies were found, and 100 percent consensus was reached among the research team. All researchers engaged in recording, transcribing, discussing the text, identifying themes, key points, counting and comparisons of keywords and/or content, as well as the interpretation of the underlying context. Results of the surveys are organized and presented below.

Results

Six interviews were conducted with seven professionals who work with big data in different capacities and settings. To clarify the context of the results, where necessary, responses from interviewees that represented care provider organizations are discussed first, and responses from the quality management and the data platform representatives are summarized right after. Following are the findings from those interviews.

Interviewee **Roles**

Interviewee roles included the manager of healthcare data analytics at a large healthcare system in Pennsylvania, the chief research information officer at a university hospital in Ohio, the director of analytics and performance measurement along with a team member from a national quality organization in Virginia, a consultant and program manager at a private not-for-profit healthcare system in New Mexico, the senior director of engineering application at a large global data platform company in California, and the director of a data analytics consulting company in Missouri.

Organizational Structure for BD and BDA

Interviewees were asked about the formal organizational structure dedicated to working with BD, and they indicated that there is either a dedicated team/function, or department (such as a data analytics department) that is focused on working with health data. These teams were composed of business analysts, developers, data architects, engineers, clinicians, and occasionally health information specialists, and the size varied from a few to about 100 (the larger numbers correspond to larger health systems and the global data platform company). Additionally, staffing is done with internal employees and consultants. Consolidation of prior data analytics teams into one large function was mentioned by three of the interviewees. Despite the use of external resources, BD work is led and driven internally.

The way these teams function varies significantly, depending on the type and size of organization, as well as resources available. Two interviewees indicated that much of the BD work is conditioned by EPIC, the EHR used in the facility. In those cases, EPIC data and claims data are brought together into a common data governance platform. Physical servers are used, but cloud-based infrastructure is expanding.

How Are BD and BDA Used by Organizations – Purpose

Four interviewees shared that healthcare systems use BD and BDA to respond to regulatory requirements from the federal government, payers, or audit needs, as well as to fulfill executive and business unit requests. Requests mostly follow the industry trends and benchmarking, and a desire to stay ahead of the curve. One of the interviewees went into greater detail that BD and BDA are used to support optimal operations, shared saving, commercial contracts, Medicare shared savings, risk optimization, cost and utilization, as well as quality measures. Another interviewee shared that the organization uses BD and BDA to explore better ways of bundling services so that the facility does not lose money and possibly makes a profit to compensate for communities and services that are harder to pay for. A third interviewee shared that BD and BDA are used for predictive analytics around readmissions or to address questions pertaining to the health of communities around.

How Are BD and BDA Used by Organizations – Dynamics/Processes

Interviews revealed that the way BD/BDA are used varies from one organization to another. The care provider organizations that use EPIC had more in common. They capitalize on the templates and predictive models pushed by EPIC, given they run daily, and provide users with opportunities to act on the findings. Even when templates or models are not fully understood, there is trust in the vendor who provides the idea and tool. Often, such tools are integrated without a clear plan on how the information will be used, as in the case of a model that predicts the risk of a patient dying in the next year. Yet, three interviewees shared that some units have plans, or some have ideas about what they want but have no tool to develop it. Generally, the business side drives the types of analyses by telling IT what's needed. IT explains what's possible with the data and tools available. Results of BDA are used as a basis for operational and senior-level decisions, justification of investments, public health, care management, patient outreach, education, vendors, and for potential restructuring of the organization.

The interview with the individuals at the national quality organization showed a different process. Given that they are an organization that creates measures, ideas for quality measures are prioritized, and once decided, a technical expert panel defines the specifications for that measure. Then, the company uses the BD and BDA to apply specifications and test the measure for reliability and validity. For example, an opioid measure is tested, and then adjusted by removing certain populations, such as hospice or cancer patients. Measures are sometimes imposed by the Centers for Medicare and Medicaid Services (CMS), as well as driven by the National Quality Forum. Measures are often risk-adjusted for age, sickness, living location, race, ethnicity, and low-income status for Medicare. BD and BDA are also used to interpret clinical guidance with the data available. Lastly, they are used to maintain measurements; as clinical guidelines or literature review change, measures are re-tested.

The other distinct organization, the data platform company uses BD and BDA to assess how well the client company is using the data. They are able to trace and identify user-errors (as per regulations pertaining to data hosting services), identify faulty software, and use BDA to decide on how to prevent similar errors in the future. Such insight helps build better technologies to manage an organization's data and test software as needed. Additionally, the company uses BD to understand product features, identify whether the product is working as it should, and proactively check quality of operations in the cloud platform and SAS platform.

How Is BD and BDA Used to Support Population Health?

When asked about how the organizations use BD and BDA to support population health initiatives, responses pertaining to care provider organizations had three areas in common: claims analytics; risk optimization; and quality measures. Claims data are heavily analyzed to identify opportunities for reducing costs and clinical variation, comparing utilization indicators, with peers, improving utilization and efficiency, as well as informing and supporting value-based contracts. One of the interviewees shared that geospatial analytics is also used to identify heat map areas in terms of cost-utilization for primary care facilities. Discussion on risk optimization was focused on better documentation of the level of risk, rather than BDA. Quality measures pertaining to the internal patient population are collected and reported. Additionally, there are efforts to understand the populations outside internal data sources. Depending on the request, the organization may include state or national level data that is publicly available. Two interviewees have community partnerships to address issues like health equity and social determinants of health. One organization uses the internal data available to make broad assumptions about the population (although access to the clinical data of that larger population may be limited or not available). Another organization is actively engaged with tribal leaders for outreach to minority communities and better population health management. The latter organization also performs spatial analysis and uses a geographic information system (GIS) and Microsoft platform, QlikView. Additionally, one interviewee shared progress in customizing a wellness program and diabetes predictive model for employees.

The interviewees from the national quality organization shared that they support population health tangentially by creating measures that drive incentives in the marketplace, which then drive health plans to manage population health and intervene as necessary. GIS or mapping algorithms are not used currently, but a machine learning algorithm would help identify the highest risk patients, or those most likely to be impacted.

The data platform company is mostly engaged in data collection exercises to understand peoples' behaviors and trends in relation to data. For example, spatial analysis is used to monitor air quality during California fires and decisions can be made accordingly. There is potential to build use cases software that help healthcare organizations not only monitor health data but also recognize patterns. Additionally, it was pointed out that there is potential to capitalize on data derived by sensors and IoT devices for better management of population health.

Challenges Pertaining to BD and BDA

When asked about the challenges observed in relation to BD and BDA, interviewees identified various aspects that are grouped into four categories: leadership; data literacy; system integration; and data characteristics. Challenges related to data characteristic are organized by volume, variety, velocity, veracity, value, and integrity.

Leadership-Related Challenges

All interviewees shared that organizational leadership is focusing on BD and BDA and dedicated teams (large or small) are in place. However, aside from the data platform company, others have yet to establish clear strategies, alignment of strategy with BD and BDA, and pathways for optimal BD use and collaboration within the various units and external parties. One interviewee said that there is lack of ownership of all required data sources to perform desired analytics, as well as lack of foundational infrastructure to support business needs. Another interviewee pointed out the leadership vacuum in certain areas. For example, in a university hospital, there are three important parties: clinicians; researchers; and administration. Clinicians are data generators, while researchers are data consumers. The administration follows the legal requirements: Family Educational Rights and Privacy Act controls teaching data, Health Insurance Portability and Accountability Act controls patient data, and Institutional Review Boards control research data. Tensions exist over the data management, and trusted relationships need to be developed among the three parties.

Data Literacy Challenges

All interviewees addressed that there is a need to improve data literacy across business operations. There are misinterpretations of graphs, and often, decisions are based on assumptions. There is a gap in translating business needs into what is possible to do with existing

BD/BDA or how it could be possible. One interviewee mentioned that BD and BDA "is not something you can learn in a book. It is understanding what the data is telling you."

The data platform company shared that most users don’t use proper search terms, cannot do data analysis, or build a dashboard by using the SAS platform because they do not know the language to engage the platform. However, they are working on training users, as well as making the platform easier to use.

System Integration-Related Challenges

Five interviewees brought up that information is siloed. Integrating hospital clinical data with billing data, or claims data, or data from various practices is a challenge. People working with data question practices around patient duplication across the system or even proper physician identification in the multiple databases, given lack of proper integration. There are questions on how to index the data. As per one interviewee, “System standards exist but EHRs are customizable. For example, heparin control could be recorded in four different EHR locations in different organizations depending on the system customization. In the absence of guardrails, interoperability means relatively little; theoretically possible but it's pragmatically difficult because of choice.” There is concern that vendor competition and the market system in the US add to the challenge of integration.

Data Characteristics-Related Challenges

Five interviewees shared that there are challenges associated with handling large amounts of data. Not all organizations are provided with the equipment needed to analyze such volume. As per interviewee, “Using SAS in our computers or Optum landsite on a remote desktop can be limiting. We don’t use Hadoop or anything like that, where the processing resources are distributed across multiple machines.”

In terms of data variety, interviewees from the healthcare organizations and consulting companies indicated that the unstructured data was not being used for BDA, yet. The data platform company, on the other hand, indexes unstructured data and makes it structured by following certain schemas. After that, data cannot be changed. There is a risk of corrupting the data, so it is important to understand the data well prior to indexing it.

From the perspective of the data platform company, data velocity presents a challenge. One interviewee says that “based on budgets, there are limits on the daily ingestion rate, and we need to make the incoming data fit into those limits. Data is bursty.” Data flow and need for data varies throughout the day. So, from an operational perspective, decisions need to be made on how much of the system is needed throughout the day for a particular client. At the same time, the infrastructure must be provisioned to handle peak times, and adapted for scaling up and down.

Another data challenge aspect that was brought up by two interviewees was incorrect matching of data elements from old legacy systems with new systems. This process is not always accurate, and as per one interviewee, “variability of denominators should be questioned.” There are no tools or sufficient resources to assure data integrity for such issues, and most rely on manual reviews by users and analysts. This brings up data veracity challenges.

One interviewee shared that “a potential problem exists when buying a clinical or administrative dataset or billing dataset, like a market scan. Such data is used to determine that the most cost-effective treatment for an individual who has a heart attack, in general, is to run a drug-eluting stent. However, our population-level studies are subject to our specific population. Given the difference in population heterogeneity, how are we sure that the general treatment works for our group?” This data selection bias results in solutions or recommendations that do not work for certain populations, which diminishes its value.

When asked about overall data integrity, all interviewees addressed security. All comply with HIPAA regulations, but data security is a big challenge and a barrier that still prevents organizations from trusting cloud services. Data security was also mentioned as a constraint to patient matching by one interviewee, who said, “Despite having Medicare and Medicaid datasets, we can match some variables at a level of 30 percent but not the rest because of privacy.” Furthermore, on data security, the interviewee from the data platform company explained that their software can monitor whether a healthcare system is being hacked. Certifications guide them to features that are used and pushed as well as who can access the data. The amount and type of data coming in is a moving target and users need to understand where the data is going. If it goes beyond firewalls, it is inherently vulnerable. Questions about ownership are also discussed as part of agreements: “At what point do we own the data and at what point does the customer own the data? When does the exchange happen? That moment needs to be heavily secured.”

Another issue related to data integrity was data definition, discussed by four interviewees. There are inconsistent data structures and lack of standardization (mostly due to system customization, as pointed out on the example above). For example, there is no clear definition of a hospital admission in the system. There is also the absence of meta data standards, as well as lack of data dictionaries. As per one of the interviewees, “one data set has over 10,000 tables. How do you navigate 10,000 tables? How do you find the variables you're looking for? It's there, and analysts have to go digging around to find them.” Clear definitions would also help with query inclusion and exclusion criteria.

Data completeness challenges (another aspect of data integrity) were also identified by four interviewees. As per one interviewee, “health equity relies on race, ethnicity, language data, and that data is not captured well.” Claims data also does not tell the whole picture. As per another interviewee, “claims data is not perfect, as it does not include encounters paid in cash.” These comments relate to data capture, availability, and comprehensiveness of a data governance program needed to properly address various healthcare initiatives, such as population health and health equity.

How Are BD and BDA Challenges Being Addressed?

As per all interviewees, organizations have recognized the BD and BDA challenges, and have discussed the work in progress to address them. All plan to add data analyst positions, and some plan to restructure data teams under one leadership. One organization plans to create an analyst class position acting as middleman between the “research or operations question” and the data.

Such a position would help with the better understanding of the business needs and the data needed to support those needs, as well as simplify and create abstractions in the data that could be analyzed.

Five interviewees brought up the need for robust data governance programs, documenting the true sources of data, monitoring data movement, potentially bringing some data in house or potentially using third-party payers to augment data, and using an agile methodology regarding the Master Patient Index project. Four interviewees were evaluating current tools and options offered by the EHR, working to improve the matching of data elements, and bringing solutions to the warehouse or the data layer. This would address existing problems of the visualization layer, which is currently suffering because of the siloed data. From a process perspective, one interviewee said, “more frequent touch points should be added with the management to clarify technical aspects and how to tailor them to meet business needs.” Organizations are also supporting data standardization and integration projects. As per the interviewee from the consulting company, consultants are also helping organizations with data governance and integration issues.

Thoughts on Non-Healthcare Organizations Using BD and BDA

All interviewees were asked about what they think about non-healthcare organizations that use healthcare big data, how device manufacturers, pharma, and insurance companies benefit from healthcare big data, and how data companies such as Google, Amazon, and Microsoft benefit from healthcare big data. Responses were interesting, as they showed a variety of perspectives from the interviewees.

The main emerging theme was an overall positive view of tech companies, including Google, Apple, Amazon, and Microsoft. They are viewed as building value in general and for healthcare. It was pointed out that healthcare can learn from such industries, or even the manufacturing industry, when it comes to BD and BDA. Additionally, industries trying to enter the healthcare space should be collaborating better with the healthcare providers. The interviewee from the data platform company mentioned that sometimes tech companies engage in “fun” activities that produce incidental findings that provide great insight into certain healthcare behaviors. Such capabilities are not tapped but could be with greater collaboration.

The second theme was the need for non-healthcare organizations to be more responsible in working with health data. All interviewees mentioned that there is a risk of incorrect interpretation or misinterpretation of health data. While in the business world that may affect sales of a product, in healthcare it affects patient safety and “it puts patients at risk.” There is a belief (brought up by three interviewees) that when non-healthcare organizations use machine learning to support certain programs, the primary goals are financial, which makes them less trusted in their motives for creating patient programs. One interviewee said that non-healthcare organization tend to hoard data, and not share it to make more money.

One of the interviewees brought up the idea of distinguishing between clinical and health data. For example, clinical data is generated by clinicians, while health data is generated by medical devices such as Fitbit watches or wearable devices used for continuous glucose monitoring or other medical issues. This same interviewee discussed the economic perspective by saying, “you could not have built these devices if not for the healthcare and technological infrastructure that exists. They don’t pay for data collection, there is no cost of input, and they extract value without contributing to the overall healthcare system. This is a distortion of market factors.” It was believed that players such as these add value, but they are working in an environment without principles. “They should be building value for the public good. As they increase their financial earnings, they have greater potential to make their ideas more tangible. I would like to see some of these successful companies sponsor people who do not have access to money, power, and resources; so, they can also bring their ideas to life and possibly change the healthcare system for the better. That additional level of connectedness that really separates; that creates social disparities should be addressed.”

Future Directions in Relation to BD and BDA

The five interviewees from healthcare organizations indicated that the vision is to invest in BD and BDA infrastructure and technology and expand use of BD while aligning better with strategy. Ideas for investment included: (1) integration of external data sources with internal analytical capabilities; (2) creation of a data lake with potential to coordinate data integration across multiple organizations and allow for data abstraction; (3) development of better approaches for software acquisition, with the goal continuing to use what’s available in the current data warehouse, despite the new software; (4) formation of teams that specialize in analytics and decision support across the various business units that would guide operational leadership in making informed decisions with data; (5) support for greater patient access to their own data; (6) increased interoperability through data standards; (7) cloud-based expansion; and (8) better data standards for population health needs.

Ideas for expanding BD uses included leveraging data better to monitor population trends, support population health initiatives, and reusing the data for multiple types of research. Coordinating the brain power of different and more experts is very important, and “work should be done on signing agreements with other healthcare organizations (including optometry and dental services) to share identifiable data in a HIPAA compliant manner, with proper security layers to protect the data from misuse.” Additionally, there was a suggestion to consider a centralized versus a federated approach, where every clinical area does their own analytical work. With the right tools and knowledge, as well as proper data governance, people can engage in self-service discovery more effectively. Patient engagement was also identified as an important future direction, from the perspective of allowing data sharing for research, as well finding ways to engage patients when not in clinical care. Furthermore, one interviewee suggested shifting focus from administrative-facing to public-facing conversations, meaning greater engagement with patients and the public regarding health data.

Four interviewees indicated that their organizations are discussing the use of AI but did not have details for specific AI tools. One of the healthcare organizations is using machine learning to predict the likelihood that a patient may develop sepsis. The data platform company already uses machine learning heavily during data acquisition. They follow user requests such as identify and acquire only the data that matches a particular model and discards the rest. The company plans to continue to evaluate the amount of data being collected and the data filters. Filters intend to clean the noise but may remove valuable data in the process, so finetuning of the algorithms will be ongoing. Balance between too much or not enough data will continue to be monitored, while acknowledging that certain data need to be kept for compliance reasons, as evidence. Additionally, the vision is to elevate the search process and make it easier for the user to run queries and use the data.

Discussion

Our findings add rich context and details in understanding how healthcare organizations are managing BD and BDA. While formal structures for BD and BDA exist in many organizations, there is a variety of approaches on how that is done, and organizations are seeking improvement in that area. The variety of structures, tools, and practices in managing BD and BDA can generate questions for further research that could dive deeper into specific operational aspects. As found in literature review, it is important to develop evaluation tools, comparative studies, best practices, or structural models that will help providers or other users of healthcare BD and BDA to make more informed decisions when they are purchasing BD sets and BDA technologies, hiring staff, or structuring/restructuring their data analytics functions.[20,35,36] Decision-makers should be able to evaluate such investment proposals based on well-established criteria and resources (and not solely those served by the vendors). This also highlights opportunities for building trust and improving collaboration with non-healthcare organizations whose strength is working efficiently with BD and BDA.[37]

Challenges associated with BD and BDA are also consistent with those in the literature review.[1,9,11,38,39] Data definition, accuracy, and completeness become even more important when data is aggregated and analyzed in larger scales. Not only does data quality affect current decisions about patient populations, but it also affects future decisions, given that machines learn based on those inaccurate, incomplete, poorly defined data pools. This has multiple implications. First, healthcare organizations need to invest resources in improving the integrity of their existing data pools (and this should go beyond patient matching). Second, everyone who touches and uses health data should be educated and/or trained in managing health data with integrity during all stages, including acquisition, extraction, cleaning, integration, and aggregation. Third, any tools that are used for automated data collection should be carefully evaluated at the beginning and monitored throughout to assure the data integrity is intact.

Health data literacy came up as an important aspect of BD/BDA challenges. Lack of skilled staff was pointed out early on and continues to come up in research as the need for upskilling of the workforce is ongoing.[40,41,42,43] There is an opportunity for health information professionals to contribute in the process of elevating health data literacy among healthcare professionals, as well as non-healthcare professionals who work with health data. This will also contribute positively in trust building and collaboration efforts mentioned above.

Closely related to health data literacy is another important finding, the potential to improve data governance programs. Data governance is one of the most important domains for health information professionals. As organizations use their home-grown BD or acquire BD from other organizations, it is imperative to accompany such activities with data dictionaries and relevant terminologies. Other data governance aspects are data lifecycle, data architecture, metadata, data quality, and security, all of which present opportunities for health information professionals' leadership and sharing of expertise.[44]

Findings show great efforts in addressing population health issues and health equity, along with the need for more complete and accurate social determinants of health data. AHIMA has already recognized this opportunity and is leading the way in creating relevant standards and identifying specific training needs for healthcare workers.[45]

In addition to the significance for further research or potential work areas for health information professionals, findings from this study may be useful for educational purposes. Current literature and textbooks used in health information, healthcare management, and healthcare services provide general overviews, discuss the importance and potential of BD and BDA, as well as specific tools used successfully in particular settings.[1, 37, 46] Knowledge about BD and BDA operational and technical aspects is usually part of information technology programs and is currently lacking in the space of health information and other healthcare studies. As per findings, IT professionals are working very closely with health information professionals, clinicians, and administration, but they currently experience barriers in communication, *i.e.*, they do not fully understand each other's language. It is imperative to bridge this gap and improve data literacy among all clinical and non-clinical participants in healthcare who touch health data or make decisions related to BD/BDA. Details of this study contribute to such a body of knowledge.

Limitations

This study focused on better understanding of BD and BDA operations and practices in healthcare. The open-ended structured interview protocol enabled collection of rich answers filled with details and examples. It also allowed for greater comparability of responses and getting a complete data set for each question or subtopic.[30] The questions asked required mostly facts and objective information that reflected the interviewee's knowledge and experiences, which are recent (since BD and BDA are a recent reality in healthcare). This is a strength as findings rely less on perceptions or subjective data. Findings from this study may also serve as stimulation for new research pertaining to BD and BDA. As with most qualitative studies, findings are not highly generalizable; however, there are opportunities for similar research by expanding the interviewee pool and the types of organizations they represent. While details shared are mostly related to the interviewee's recent roles with BD and BDA, it is possible that there may have been errors of memory and/or judgment. Certain details may not have come up, which creates a less than full picture of BD and BDA reality. Additionally, the authors recognize the fast-paced technological environment and growth of AI tools between 2023 and 2024 (which is after the interviews were conducted). Such progress has yet to be realized in all settings of the healthcare industry, and the findings from the study are still relevant as pointed out in the discussions section.

Conclusions

This study provides greater insight into how BD and BDA are being managed and used in various healthcare organizations as well as by vendors servicing healthcare providers. Given the very complex and diverse healthcare landscape in the US, our attempt was not to obtain a full picture of such reality but to better recognizing some of the BD/BDA realities. Such knowledge, details, and examples help all who work with health data to better understand their role and potential contribution to the management and use of BD. They also contribute to greater effectiveness and efficiency in processing and using BD and BDA meaningfully, in today's digital healthcare environment. Lastly, some of the findings validate the work and role of health information professionals when it comes to BD and BDA in healthcare.

References

1. Shilo, S., Rossman, H. and Segal, E. "Axes of a revolution: challenges and promises of big data in healthcare." *Nat Med* 26, 29–38 (2020). https://doi.org/10.1038/s41591-019-0727-5.
2. Gartner. "Gartner Identifies the Top 10 Strategic Technology Trends for 2023." (October 2022). Accessed September 2023. https://www.gartner.com/en/newsroom/press-releases/2022-10-17-gartner-identifies-the-top-10-strategic-technology-trends-for-2023.
3. Mayer-Schönberger, Viktor and Cukier, Kenneth. *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt. (2013). Accessed October 2022. https://psycnet.apa.org/record/2013-17650-000.
4. Van Dijck. "Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology." *Surveillance and Society* 12 No. 2 (2014). https://doi.org/10.24908/ss.v12i2.4776.
5. Iyamu, Tiko. "Advancing Big Data Analytics for Healthcare Service Delivery." London; Routledge, 2023.
6. Sivarajah, Uthaysankar, Kamal, Muhhamad Mustafa, Irani, Zahir and Weerakkody, Vishanth. "Critical analysis of Big Data challenges and analytical methods." *Journal of Business Research* 70, no. C (2017): 263-286. https://doi.org/10.1016/j.jbusres.2016.08.001.
7. Reichman OJ, Jones Matthew B, and Schildhauer Mark P. "Challenges and opportunities of open data in ecology." *Science* 331, no. 6018 (February 2011):703-5 https://pubmed.ncbi.nlm.nih.gov/21311007/.
8. Segaran, Toby & Hammerbacher, Jeff. *Beautiful data: the stories behind elegant data solutions.* O'Reilly, 2009.
9. Schintler, Laurie A. (Laurie Anne), and Connie L. McNeely, eds. 2022. Encyclopedia of Big Data. 1st ed. Cham: Springer International Publishing, 2022. https://doi.org/10.1007/978-3-319-32010-6.
10. Kong, Hyoun-Joong. "Managing Unstructured Big Data in Healthcare System." *Healthcare Informatics Research* 25, no.1 (January 2019): 1-2. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6372467/.
11. Mønsted T. "Achieving veracity: A study of the development and use of an information system for data analysis in preventive healthcare." *Health Informatics Journal* 25(3) (2019):491-499. https://pubmed.ncbi.nlm.nih.gov/30198372/.
12. Liu, Yong-Chuan, Atefeh Farzindar, and Mingbo Gong. "Transforming Healthcare with Big Data and AI." Charlotte, North Carolina: Information Age Publishing, Inc., 2020.

13. Sandhu, R and Sood, SK. “Scheduling of big data applications on distributed cloud based on QoS parameters.” *Cluster Computing*, 18 no. 2 (December 2014) doi:10.1007/s10586-014-0416-6.
14. Zhang, Feng, Min Liu, Feng Gui, Weiming Shen, Abdallah Shami, and Yunlong Ma. "A distributed frequent itemset mining algorithm using Spark for Big Data analytics." *Cluster Computing* 18, no. 4 (2015): 1493-1501.
15. Yi, Xiaomeng, Fangming Liu, Jiangchuan Liu, and Hai Jin. "Building a network highway for big data: architecture and challenges." *IEEE* 28, no. 4 (2014): 5-13.
16. Hugh, Oliver, and Jason Gardosi. “Use of Microsoft Power BI to Display Pregnancy Related Performance Statistics within NHS Trusts.” *International Journal of Population Data Science* 8 (2) (2023). https://doi.org/10.23889/ijpds.v8i2.2342.
17. Tiko, Iyamu. *Advancing Big Data Analytics for Healthcare Service Delivery,* 1st ed. Routledge. New York, 2023.
18. Jagadish, Hosagrahar V, Johannes Gehrke, Alexandros Labrinidis, Yannis Papakonstantinou, Jignesh M. Patel, Raghu Ramakrishnan, and Cyrus Shahabi. "Big data and its technical challenges." *Communications of the ACM* 57, no. 7 (2014): 86-94.
19. Dash, Sabyasachi, Sushil Kumar Shakyawar, Mohit Sharma, and Sandeep Kaushik. “Big Data in Healthcare: Management, Analysis and Future Prospects.” *Journal of Big Data* 6 (1) (2019): 1–25. https://doi.org/10.1186/s40537-019-0217-0.
20. Senthilkumar, S. A., Bharatendara K. Rai, Amruta A. Meshram, Angappa Gunasekaran, and S. Chandrakumarmangalam. "Big data in healthcare management: a review of literature." *American Journal of Theoretical and Applied Business* 4, no. 2 (2018): 57-69.
21. Murdoch, Travis B, and Detsky, Allan S. “The inevitable application of big data to health care.” *JAMA*, 309, no. 13 (2013): 1351–1352. https://doi.org/10.1001/jama.2013.393.
22. Zeng, Jing, and Glaister, Keith W. “Value creation from big data: Looking inside the black box.” *Strategic Organization* 16, no. 2 (2018): 105–140. https://doi.org/10.1177/1476127017697510.
23. Chen, Jinchuan, Yueguo Chen, Xiaoyong Du, Cuiping Li, Jiaheng Lu, Suyun Zhao, and Xuan Zhou. "Big data challenge: a data management perspective." *Frontiers of Computer Science* 7, no. 2 (2013): 157-164.
24. Agrawal, R. and Prabakaran, S. “Big data in digital healthcare: lessons learnt and recommendations for general practice.” *Heredity* 124 (2020): 525–534. https://doi.org/10.1038/s41437-020-0303-2.
25. Wang, C. Jason, Chun Y Ng, and Robert H Brook. 2020. “Response to COVID-19 in Taiwan: Big Data Analytics, New Technology, and Proactive Testing.” *The Journal of the American Medical Association* 323 (14): 1341–42. https://doi.org/10.1001/jama.2020.3151.
26. Simon, Gregory E. 2019. “Big Data from Health Records in Mental Health Care: Hardly Clairvoyant But Already Useful.” *JAMA Psychiatry* (Chicago, Ill.) 76 (4): 349–50. https://doi.org/10.1001/jamapsychiatry.2018.4510.
27. Golbus, Jessica R, W Nicholson Price, and Brahmajee K Nallamothu. “Privacy Gaps for Digital Cardiology Data: Big Problems with Big Data.” *Circulation* (New York,

N.Y.) 141 (8) (2020): 613–15. https://doi.org/10.1161/CIRCULATIONAHA.119.044966.

28. Frakt, AB, and Pizer, SD. “The promise and perils of big data in healthcare.” *The American Journal of Managed Care* 22, no. 2 (2016): 98–99.
29. Sivarajah, U., Kamal, MM, Irani, Z., and Weerakkody, V. “Critical analysis of Big Data challenges and analytical methods.” *Journal of Business Research*. 70 (2017); 263-286.
30. Patton, Michael Quinn. *Qualitative Research and Evaluation Methods*, 3rd ed. Sage Publications, Inc. 2002, p. 339-427.
31. Creswell, John W. *Research Design*, 3rd ed. Sage Publications, Inc. 2009, p. 173-200.
32. Ritchie, Jane and Lewis, Jane. *Qualitative Research Practice: A Guide for Social Science Students and Researchers*, 1st ed. Sage Publications, Inc. 2003, p. 173-200.
33. Ishak, NM and Bakar, AYA. “Qualitative data management and analysis using NVivo: An approach used to examine leadership qualities among student leaders.” *Education Research Journal*. Vol 2.(3) (March 2012): 94-103.
34. Rapport, Frances. “Summative Analysis: A Qualitative Method for Social Science and Health Research.” *International Journal of Qualitative Methods*. (September 2010). https://doi.org/10.1177/160940691000900303.
35. Riaz Ahmed, Sumayya Shaheen, and Simon P. Philbin. “The role of big data analytics and decision-making in achieving project success.” *Journal of Engineering and Technology Management*, 65, (July–September 2022): 101697. https://doi.org/10.1016/j.jengtecman.2022.101697.
36. Dobre, Ciprian and Xhafa, Fatos. “Intelligent services for Big Data science.” *Future Generation Computer Science*. 137 (July 2014): 267-281. https://www.sciencedirect.com/science/article/abs/pii/S0167739X13001593.
37. Batko, K., & Ślęzak, A. “The use of Big Data Analytics in healthcare.” *Journal of big data*, 9(1), (2022): 3. https://doi.org/10.1186/s40537-021-00553-4.
38. Gandomi, Amir, and Murtaza Haider. "Beyond the hype: Big data concepts, methods, and analytics." International Journal of Information Management 35, no. 2 (2015): 137-144.
39. Frost and Sullivan White Paper. “Drowning in Big Data? Reducing Information Technology Complexities and Costs for Healthcare Organizations.” Accessed October 3, 2023. https://www.academia.edu/6563567/A_Frost_and_Sullivan_White_Paper_Drowning_in_Big_Data_Reducing_Information_Technology_Complexities_and_Costs_For_Healthcare_Organizations_CONTENTS.
40. Kim, Gang-Hoon, Trimi, Silvana, and Chung, Ji-Hyong. “Big-Data Applications in the Government Sector.” *Communications of the ACM* 57 (2014): 78-85. https://dl.acm.org/doi/10.1145/2500873.
41. NORC at the University of Chicago & AHIMA. “Health Information Workforce: Survey Results on Workforce Challenges and the Role of Emerging Technologies.” October 2023. https://www.norc.org/research/projects/workforce-challenges-technology-adoption-health-information-professionals.html.
42. Sørensen K. “From Project-Based Health Literacy Data and Measurement to an Integrated System of Analytics and Insights: Enhancing Data-Driven Value Creation

in Health-Literate Organizations." *Int J Environ Res Public Health*. 2022 Oct 14;19(20):13210. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9603602/.

43. Lubasch JS, Voigt-Barbarowicz M, Lippke S, et al. "Improving professional health literacy in hospitals: study protocol of a participatory codesign and implementation study." *BMJ Open* 11 (2021): e045835. https://pubmed.ncbi.nlm.nih.gov/34400444/.
44. Oachs P & Watters AL. *Health Information Concepts, Principles and Practice*, 6th ed. AHIMA Press, 2020.
45. NORC at the University of Chicago & AHIMA. "Social Determinants of Health Data: Survey Results on the Collection, Integration, and Use." February 2023. https://www.norc.org/content/dam/norc-org/pdf2023/AHIMA-Workforce-Survey-Report-Final-2023.pdf.
46. Chan, Chien-Lung, and Chi-Chang Chang. "Big Data, Decision Models, and Public Health" *International Journal of Environmental Research and Public Health* 19(14) (2022): 8543. https://doi.org/10.3390/ijerph19148543.

Author Bios

Egondu R. Onyejekwe, PhD, MSc., MA., MA., MS. Engr., is a director of Serve, Educate, Elevate (S.E.E.) program. She currently works for the National Council of Negro Women Inc., Columbus, Ohio Section (NCNWCSO). Prior to this, Onyejekwe was a faculty at Walden University.

Dasantila Sherifi, PhD, MBA, RHIA, is an assistant professor and program director for Health Information Management at Rutgers University. She received her Doctor in Philosophy degree in Health Services with specialization in Public Health Policy from Walden University and her Master's in Business Administration from Southern Illinois University.

Hung Ching is a senior medical physicist at Memorial Kettering Cancer Center. He received his Doctor of Philosophy degree in public health from Walden University and his master's and bachelor's degrees in physics from the State University of New York at Stony Brook. He is licensed to practice medical physics in New York and New Jersey. He is board-certified by the American Board of Radiology in the field of diagnostic medical physics.

A Process of User-Centered Design to Create a Social Determinants of Health Data Platform

By Danequa Forrest, Jeremy Pyne, Laura McKieran, Cristina E. Martinez

Summary: A model of user-centered design, including community input, is tested to increase the accessibility and use of health and social determinants of health data.

Abstract

Bexar Data Dive, an online data platform, was created to increase accessibility and use of health and social determinants of health data, such as education, economic barriers to healthcare, and hospitalization rates, to decrease racial/ethnic health disparities throughout Bexar County. A model of user-centered design helped us incorporate community input into the platform. We conducted four interviews and five focus groups to gather information on how people use data — specifically beginner and intermediate-level data users from various educational, governmental, and nonprofit organizations. Then, we launched a community survey to assess specific data needs. Lastly, once the alpha version of Bexar Data Dive was ready, we conducted user testing sessions to measure usability, identify bugs, and gather final feedback before launch. Our findings included many recommendations for incorporating user-centered design in health data management. Participants wanted a health data tool that was easy to use, had the indicators they commonly need, and would provide visualizations for presentations, grants, and other projects.

Keywords: user-centered design (UCD) methodology, health data, platform, SDOH, social determinants of health (SDOH), qualitative

Introduction

Bexar Data Dive is a health data platform created by Community Information Now (CINow), a data nonprofit in Bexar County (San Antonio), Texas. The user-centered design (UCD) process helped CINow build a platform that allows users to access the same integrated social determinants of health (SDOH) and health data in multiple ways, depending on how detailed they would like the data. Our goal was for beginner and intermediate data users to use our platform and visualize their communities, create maps, charts, and graphs, and get quick statistics for their projects and work — be it grant writing, social work, community outreach, or other community-centered projects.

Background

Through our work with educational institutions, government agencies, health organizations, community workers, and other nonprofits, CINow has heard the need for more accessible and easy-to-use health and SDOH data. We have also felt this need ourselves, particularly with regard to having data disaggregated by multiple demographics, like age, race/ethnicity, and sex, and nested geographies. Many of the organizations we provide data to regularly check the SDOH conditions of the communities they work in, guide them in disbursing assistance and aid, and assist in completing both small and large projects that benefit those communities. We set out to create Bexar Data Dive to help local people and organizations in Bexar County have an easier way to analyze, visualize, and disseminate data to continue helping their communities decrease racial/ethnic health barriers. We employed a UCD approach to incorporate community feedback in an iterative and ongoing process through every step of creating Bexar Data Dive.

UCD is a design process of continual improvement that takes into consideration the needs of those who would be using the product or service. The UCD process CINow used to create the data platform Bexar Data Dive involved qualitative methods to gather input from data users, to take into account what they would need from a data platform. Qualitative analysis in UCD provides detailed information on how to best serve the target audience and has been used and advocated for within many types of health research.[1] Focus groups were used to gather rich qualitative data on how data users would prefer for a data platform to function and what content it would contain. While CINow conducted focus groups and interviews through Zoom, research has found virtual focus groups and interviews generate the same amount of unique ideas as in-person.[2] The UCD process has been shown to be beneficial in creating technology-enabled services,[3] and this process is novel in applying the process to building a health data platform.

Methods and Materials

CINow developed and deployed a UCD process to understand requirements and features desired by beginner and intermediate-level data users, who were defined as people in grassroots organizations, churches, small nonprofits, small government agencies, and students. The approach and instruments were designed with technical assistance from the UCD-experienced Data Driven Detroit, one of CINow's peer organizations in the National Neighborhood Indicators Partnership (NNIP). The UCD process included interviews, focus groups, qualitative analysis, user stories for the web developer, a community survey, wireframes and web development, and user testing, all within the span of a year. IRB approval was not needed, as the information we were gathering was solely to inform the data platform's development. All materials used during the UCD process can be found in Appendices A-C, and are also described below.

Interviews and Focus Groups

To create the interview questions, we researched other community surveys on data needs, such as those put out by three other NNIP partners: SAVI (a program at the Polis Center at Indiana University-Purdue University Indianapolis), DataHaven (a data nonprofit in Connecticut), and MORPC (Mid-Ohio Regional Planning Commission). Some question topics included participants' organizations, roles, commonly used datasets, problems they were trying to solve using data, preferred features in a data tool, and preferred ways of receiving data training. You can view the interview and focus group guides in Appendix A.

Recruitment for the interviews and focus groups involved assistance from our partners on the project, as well as our prior connections to data users and students at the UTHealth Houston School of Public Health in San Antonio, Texas. Our project partners included The Health Collaborative, C3 Health Information Exchange, COSA Metro Health, and COSA Information Technology Services Department.

We conducted four interviews (two beginner-level data users and two intermediate-level data users). The interviews helped us refine the questions for the focus groups, which we held in January 2022, and the two advanced-user interviews, which occurred in early February 2022.

The focus groups had similar questions to the interviews, with a few minor edits for clarity. We held five focus groups of three to four people each (mixed beginner and intermediate-level data users). All interviews and focus groups were held through Zoom, and CINow used a function on Zoom to save video, audio, and transcriptions of the participants. The interview and focus group guides in Appendix A also show the project information we provided to participants about how their information would be used and stored responsibly.

Qualitative Analysis and User Stories

Qualitative analysis of the interviews and focus groups was performed by synthesizing participants' answers to the questions. We began with open coding of general themes that emerged, then axial coding to reveal over-arching and sub-categories from the themes, and lastly, selective coding to gather the final themes that represented most of the participants' thoughts. Word was the only program used, and it was more helpful to write up results in "list" format with bullet points, explanations, and quotes, rather than a traditional qualitative narrative. This made it easier to translate themes into user stories for the web developer.

From the thematic analysis, we created user stories, which are a common part in UCD, to translate the participants' desires into a standardized format for the web developer. The user stories do not include how the functionality would be accomplished. That part of the process is saved for the web developer, who has training in implementing end-users' wants into an intuitive design. With these user stories, the contracted web developer, The Johnson Center for Philanthropy at Grand Valley State University, was able to draft the wireframes of Bexar Data Dive.

Community Survey

Next, we created a community survey in Qualtrics to supplement the information we learned from the interviews and focus groups. The answer choices on the community survey were populated from the themes in the qualitative analysis. This allowed us to get input from more people and narrow down what was most important to data users. For example, during the focus groups and interviews, we asked people about the datasets and indicators they used. And when we emailed the community survey, we provided lists of those same indicators and datasets and asked additional people which ones they would like to access. While the community survey was created to be more closed ended than the focus groups and interviews, we still allowed people a way to write in their answers if they had more say. We may release more community surveys in the future to gauge data users' reviews of Bexar Data Dive. The questions from the community survey are in Appendix B.

User Testing

CINow conducted user testing in Fall 2022, drawing participants from those who joined us for interviews and focus groups previously. This was conducted through Zoom and involved a short demo of the alpha testing version of Bexar Data Dive, engaging users in a group discussion about how they would find data on the platform and gathering input on how to improve functionality before the beta launch of Bexar Data Dive in October 2022. We held three sessions of three to five people each. The user testing guide can be found in Appendix C.

Findings

Interviews, Focus Groups, and Community Survey Results

We expected most of our users to be beginner and intermediate level data users. Based on the interviews and focus groups, many of the users work in the public sector — connected to nonprofits, universities, or public organizations — and are generally acquainted with data, though they do not consider themselves experts. Their main focuses are accessibility and ease-of-use, as their greatest obstacles include data being too difficult to find or too complicated to use.

Difficulties Encountered with Data

The most common difficulties people had with data were related to two obstacles:

- Finding Data — It was difficult for people to find data for a few reasons. For some, it required too much time to find data because they had to comb through so much of it from multiple data sources. Then, by the time they found what they needed, it would be incomplete or not disaggregated by race/ethnicity, age, other demographics and not at the geographic level they needed.
- Using Data — Once people found the data they needed, the second difficulty they had was using the data. This could be because they did not understand the data and would like more training on data literacy. Or, for other participants, this was because they had trouble formatting the data properly (for one user, they had difficulty formatting longitudinal data).

Data Training

The desire for data training aligned with the difficulties encountered with data. Participants specifically wanted training on how to:

- Find data and get community level information;
- Understand and create visuals/infographics, such as bar charts, pie graphs, and tables in multiple programs (including Excel and Google Slides);
- Learn data literacy and data best practices (such as when should databases be updated with data, how to run queries and capture multiple demographics, and how to understand technical data language); and
- Present data.

Tool Design

Participants had great ideas about their dream data tool. They described specific features they would like to experience:

> "*If we were able to run our own queries and also have it translate into visualizations that would be really helpful and amazing."*
> Participant 1
>
> *"Something that can be more user-friendly and flexible*." –
> Participant 2. They go on to say how they know people who would like to be able

> to manipulate the data in the platform, but also download it and manipulate it themselves.

There were also users who had a very descriptive idea of how a tool could be useful to them:

> "*For me, the perfect tool would be like if we were looking at a map of San Antonio and then it goes to the street level. It would be so cool if you could take that map and put whatever it is that you want to look at. You click on it and put a person there, and it pops up with all the information. On the side, there are little tabs or little checkboxes, and you could say 'I'm looking for poverty level, race/ethnicity'… And then I could do layers and say this is what their income looks like. And then, you could print it out, and in that print-out it's an infographic. That would be great. Like mind-blown right there."*
> Participant 4
>
> Participant 3 agreed and added on to this by saying it would be cool for students to be able to use an "*explore your neighborhood, explore your city*" type of feature – especially because they have found that students love dropping the little Google Maps guy down to "walk" around their neighborhoods.

Additionally, several participants mentioned wanting a data tool that:

- Is accessible to different types of users (multiple languages, color-blind mode)
- Provides examples of how the data could be used, or how other people had used the data before
- Is similar to other familiar data platforms without being redundant
- Has training on how to use the tool. Most people preferred live training/live virtual training, video tutorials, or written manuals in that order
- Has good documentation on the data (field names, etc.)
- Has the capability to see indicators interact on maps

A very useful recommendation we received from an advanced user (Participant 5) was to focus on a few datasets that were easy to maintain and update, and a few indicators that people usually have difficulty obtaining. This way, the tool is manageable for a small staff while also not being redundant to other tools that offer common indicators.

As for visualizations and outputs, participants wanted a diverse array of options. Some included trend lines (for different numbers of years), charts, bar graphs, pie charts, and heat maps to show density — all disaggregated by demographics, such as race and age, where possible. It was of particular interest for these visualizations to be presented in fact sheets, data sheets, and infographics. Participants wanted a way to easily print out information and disperse it to clients, co-workers, peers, or themselves for reference.

Datasets and Indicators

Some of the most common datasets discussed were:

- Census Bureau's American Community Survey (ACS)
- Centers for Disease Control and Prevention (CDC)

- County Health Rankings
- City of San Antonio (COSA) data
- San Antonio Metro Health data

The most prevalent indicators needed by participants were:

- Demographics (age, race, sex, income, education, marital status)
- Food insecurity
- Transportation
- Health insurance rates
- Vaccination rates (COVID-19 and other vaccines)
- COVID-19 rates
- Housing
- Business ownership

The community survey launched shortly after the interviews and focus groups, and it received 31 responses. Primarily, it helped us decide the order we would launch indicators, geographies, and functionalities, based on how participants ranked them. Some items were prioritized for phase one of the site's beta launch, and some items were delayed for a future iteration. High priority items included Spanish translation of the site, demographic indicators, SDOH indicators, census data, and the ability to compare by certain demographics (such as race and age).

User Stories Results

From the qualitative thematic analysis of interviews and focus groups, we created user stories for the web developer. The three parts of a standard user story are:

1) Who wants it?

2) What do they want?

3) Why do they want it?

Below are the most prevalent user stories we derived from the thematic analysis:

As an *entry-level user* **I want** *the data I retrieve turned into maps and charts* **so that** *I don't have to visualize the data myself.*

As an *advanced user* **I want** *to be able to download the data and metadata* **so that** *I can do my own calculations.*

As a *community health worker* **I want** *current COVID-19 vaccination rates by neighborhood* **so that** *I know where to target our outreach.*

As a *data user* **I want** *the tool to be in Spanish* **so that** *I can access the data in the language I'm most comfortable speaking.*

As a *grant writer* **I want** *quick and easy access to data and statistics* **so that** *I don't have to spend much time looking for the numbers I need.*

As a *data user* **I want** *to be able to click checkboxes of multiple indicators on a map* **so that** *I can view multiple indicators (like race/ethnicity and income) at once.*

As an *instructor who uses data* **I want** *an "explore your neighborhood" feature* **so that** *students can engage with data on a personal level.*

The web developer then took the user stories and planned out how to implement the data users' wants in an intuitive way. Some functionalities included were the ability to download the data in csv format, trend charts, an interactive map to select geographies, and a page to quickly explore the data in fact sheet form. Some user stories did not translate well to the wireframes process for various reasons. For example, "As a data user, I want to be able to click checkboxes of multiple indicators on a map so that I can view multiple indicators (like race/ethnicity and income) at once" was not incorporated into the wireframes because implementing this type of functionality is difficult and complicated for users to interpret on a map.

User Testing Results

Lastly, user testing provided CINow with feedback on the functionality and ease-of-use of the tool. Participants helped us find bugs in the site, make the functionalities more visible and intuitive, and affirm the usefulness of the platform. Some of the changes implemented as a result of user testing included: Adding an on/off toggle for viewing labels, moving the filter box to be more viewable, and having the selected indicator be more prominent while applying filters.

Discussion and Conclusion

Bexar Data Dive, a health data platform, was created and launched in 2022 through a UCD process which included interviews, focus groups, qualitative analysis, a community survey, user stories, web development, and user testing. This process was valuable in building a data tool that addressed certain needs of users, including availability of data, access to meaningful geographies, and the ability to quickly pull local stats for grant work, policy writing, health assessments, and other data needs.

While we succeeded in gathering community input on how to build a useful data platform, our greatest obstacle was recruitment, particularly of beginner-level data users. Additionally, not all user stories were able to be translated into the tool, due to limitations of the project. Future projects with similar goals would benefit from establishing fresh connections in the community beforehand so that they can easily contact their target audience.

UCD is a beneficial model for creating health data platforms, particularly the way it was used to create Bexar Data Dive. The iterative process of engaging data users and gathering their feedback in a systematic way resulted in a fully realized health data platform. While UCD is typically implemented by user experience (UX) designers and product managers, this methods research is novel in showing it can be beneficial in creating online tools for public health data users. Additionally, our methods have produced specific materials (Appendices A-C) which can be referenced and replicated by others who manage public health data to incorporate UCD in gathering community input. Future expectations include seeing a decrease in local health disparities, as public health data users have easier accessibility of health data to use in grant writing, policy recommendations, patient care, and other concerted efforts to improve community health conditions. We recommend establishing your audience and building connection early in the process, so that outreach for focus groups, interviews, and user testing is more targeted, personable, and beneficial to the research and data users. This will also help with engagement after the data tool is complete, to ask about their experiences since its launch and evaluate efficacy.

Resources

1. McIlvennan, Colleen K, Megan A Morris, Timothy C Guetterman, Daniel D Matlock, and Leslie Curry. 2019. "Qualitative Methodology in Cardiovascular Outcomes Research: A Contemporary Look." *Circulation: Cardiovascular Quality and Outcomes* 12 (9): e005828, https://doi.org/10.1161/CIRCOUTCOMES.119.005828

2. Richard, Brendan, Stephen A Sivo, Marissa Orlowski, Robert C Ford, Jamie Murphy, David N Boote, and Eleanor L Witta. 2021. "Qualitative Research via Focus Groups: Will Going Online Affect the Diversity of Your Findings?" *Cornell Hospitality Quarterly* 62 (1): 32–45, https://doi.org/10.1177/1938965520967769

3. Graham, Andrea K, Jennifer E Wildes, Madhu Reddy, Sean A Munson, C Barr Taylor, and David C Mohr. 2019. "User-centered Design for Technology-enabled Services for Eating Disorders." *International Journal of Eating Disorders* 52 (10): 1095–1107, https://doi.org/10.1002/eat.23130

Appendix A

Interview and Focus Group Guides

CINow Data Use Interview (Zoom)

Location: _Zoom_________________

Date: ___________________________

Facilitator: ______________________________________

Note taker(s)/Recorder(s): ____

Participant: ____

[Make other facilitator a co-host]

[To get the transcribed closed captions: 1) Make sure on Zoom.us>settings>meeting that the options "closed captions" and "save captions" are enabled. 2) During the meeting, enable "live transcription" and disable "live captions" so that you don't see the subtitles at the bottom of the screen. 3) Make sure you record to cloud. Once the meeting is over, the transcribed captions will be sent to the Zoom cloud.]

Introduction:

Hello, and thank you for speaking with us today. I am _____________________________. I will be facilitating today's interview. This is __________________________ (*facilitator introduces themselves*). They will be taking notes.

Before we begin, if we have any connection issues, we can reconnect through email. (*provide email information if they don't already have it*). Also, if you would like turn on or off subtitle settings, you can do so at the bottom of the screen by clicking Live Transcript.

To start I would like to tell you about CINow and the Office of Minority Health project. CINow is a nonprofit organization housed at UT Health Science Center Houston School of Public Health in San Antonio. We help the community by opening up access to information that helps assess community conditions, we help people better understand data and how to use it, and help to define results of that information.

The OMH project is to create an accessible online tool that will strengthen local efforts to reduce health disparities through use of local data. This will allow the San Antonio community to have greater capacity to use information to make changes in our community to address health disparities among racial and ethnic minority populations.

The purpose of this interview is to examine if and how you use data. We specifically would like to know how data can help you in your work and what we can do to make data more accessible to you. Findings from this interview will help build a tool that people, like you, can use to find data, and if you would like to track our progress with making this tool, you can view our monthly updates on CINow.info or email us.

Today's interview will last about thirty minutes to an hour. During the conversation, we want to get your reaction to some questions about your experience with data. We're here to listen and learn.

Do you have any questions about the information shared so far?

If it's okay with you, we will start recording the session so that we have the audio transcript. This will only be used internally.

[Start recording session, make sure to click "Record to Cloud" to get the transcription]

Note: Add question to Zoom in the chat as they're being asked.

Okay, let's begin.

1. What community or organization do you represent?

2. What is your role in the community or in your organization?

3. What problem(s)/issue(s) are you, your community, or your organization trying to solve?
    a. What kind of questions do you try to answer?

4. How do you use data in your daily work?

5. What tools do you use to capture, understand, or analyze data?

6. What datasets do you commonly use? *Can also ask: What sources do you usually get your data from?*

7. What difficulties do you encounter when looking for or using data?

8. What are your data goals, that is, what information do you want from the data or wish you could get from data?
    a. What data do you hope to have for current or future work, like your dream dataset?

9. If you needed a data tool/platform, what would you imagine an ideal data tool to look like to be useful for you, your role in your organization, and your organization?
    a. What indicators does your ideal data tool have?
    b. What output would you like to get from this tool, e.g., trend line, bar charts, tables, maps, etc.?
    c. What technical support do you think you might need for the data tool you just described?

10. Are there any data skills you would like to improve?
    a. What kind of skills would you like develop or improve?

11. What format works best for you to learn a new tool?
    a. Videos, in-person trainings, manuals?

12. Lastly, would you be available for questions or user testing in January, February, April, or July of next year?
    a. How much advance notice would you like before a session?

That was the last question. Thank you very much for your input today. Are there any last comments that anyone would like to make?

[Stop recording. The audio transcript should get sent to your Zoom Cloud Recordings online]

CINow Data Use Focus Group Guide (Zoom)

Location: Zoom_________________________________

Date: ___________________________

Facilitator: ______________________________________

Note taker(s)/Recorder(s): ________________________

Participant: _______________

[Make other facilitator a co-host]

[To get the transcribed closed captions: 1) Make sure on Zoom.us>settings>meeting that the options "closed captions" and "save captions" are enabled. 2) During the meeting, enable "auto live transcription" and disable "live captions" so that you don't see the subtitles at the bottom of the screen. 3) Make sure you record to cloud. Once the meeting is over, the transcribed captions will be sent to the Zoom cloud.]

Introduction:

Hello, and thank you for speaking with us today. I am _____________________________. I will be facilitating today's interview. This is ___________________________ (*facilitator introduces themselves)*. They will be taking notes.

Before we begin, if we have any connection issues, we can reconnect through email. (*Provide email information if they don't already have it*). Also, if you would like turn on or off subtitle settings, you can do so at the bottom of the screen by clicking Live Transcript.

I would like to review the Zoom etiquette for this focus group. Conducting focus groups virtually does occur in many fields of practice. However, it does present unique challenges that in person focus groups do not have. Let's review some Zoom best practices:

Zoom Guidelines

- It is preferable for you to be on camera.
- You should not be driving while on Zoom. If in a car, the car should be stopped and safely parked.

- If you are not speaking, please make sure to keep your microphone muted so that it does not interfere with everyone's audio and does not create additional noise distractions during the focus group.
- Please, no screen shots or photos taken during the focus group.
- If choosing to use the Chat feature, be mindful that these messages can be saved automatically and shared.
- If something unexpected happens, turn off your camera and remain muted until you can resolve it.

To start I would like to tell you about CINow and the Office of Minority Health project. CINow is a nonprofit organization housed at UT Health Science Center Houston School of Public Health in San Antonio. We help the community by opening up access to information that helps assess community conditions, we help people better understand data and how to use it, and help to define results of that information.

The OMH project is to create an accessible online tool that will strengthen local efforts to reduce health disparities through use of local data. This will allow the San Antonio community to have greater capacity to use information to make changes in our community to address health disparities among racial and ethnic minority populations.

The purpose of this focus group is to examine if and how you use data. We specifically would like to know how data can help you in your work and what we can do to make data more accessible to you. Findings from this interview will help build a tool that people, like you, can use to find data, and if you would like to track our progress with making this tool, you can view our monthly updates on CINow.info or email us.

Today's focus group will last about an hour. During the conversation, we want to get your reaction to some questions about your experience with data. We're here to listen and learn.

Some general guidelines for today's session are:

1) Have a "kitchen table" conversation. You can bounce ideas off of one another, and I will try to make sure I hear from everyone for each question.
2) There are no "right answers." Draw on your own experiences, views and beliefs — you do not need to be an expert.
3) It's okay to have the same or similar answers to different questions, or even if you would like to skip a question altogether.
4) If at any point you forget what the question was, you can ask me to repeat it, or check the chat where my colleague is placing the questions.
5) Have Fun!!

Do you have any questions about the information shared so far?

If it's okay with you, we will start recording the session so that we have the audio transcript. This will only be used internally.

[Start recording session, make sure to click "Record to Cloud" to get the transcription]

Note: Add question to Zoom in the chat as they're being asked.

Okay, let's begin.

13. Ask Individually: What community or organization do you represent?
    a. What is your role in the community or in your organization?

14. Ask Group: What problem(s)/issue(s) are you, your community, or your organization trying to solve?
    a. What kind of questions do you try to answer?
    b. How do you usually find information?
    c. Do you have a preferred way of finding this information?
    d. What information would be useful to help work towards solutions and goals?

15. Ask Individually: How do you use data in your daily work?
    a. What tools do you use to capture, understand, or analyze data?

16. Ask Group: What datasets do you commonly use? *Can also ask: What sources do you usually get your data from?*

   a. What levels of geography would be helpful to you in your work? (Example: zip codes, census tracts, neighborhood level, county, state, etc)
      i. Do you have trouble when using zip code level data?

17. Ask Individually: What difficulties do you encounter when looking for or using data?

18. Ask Group: What are your data goals, that is, what information do you want from the data or wish you could get from data?
   a. What data do you hope to have for current or future work, like your dream dataset?

19. Ask Group: If you needed a data tool/platform, what would you imagine an ideal data tool to look like to be useful for you, your role in your organization, and your organization?
   a. **Optional Question**:What indicators does your ideal data tool have?
   b. What output would you like to get from this tool, e.g., trend line, bar charts, tables, maps, etc.?
   c. What technical support do you think you might need for the data tool you just described?

20. Ask Individually: Are there any data skills you would like to improve?
   a. What kind of skills would you like develop or improve?

21. Ask Group: What format works best for you to learn a new tool?
   a. Videos, in-person trainings, manuals, live virtual trainings?

22. Ask Individually: Lastly, would you be available for user testing in July?
    a. How much advance notice would you like before a session?

That was the last question. Thank you very much for your input today.

Ask Group: Are there any last comments that anyone would like to make?

[Stop recording. The audio transcript should get sent to your Zoom Cloud Recordings online]

Appendix B

Community Survey (Administered Through Qualtrics)

Contact

1. Name (write-in option)
2. Email (write-in option)
3. What organization do you work for, and what is your role? (write-in option)

Demographics

4. Race/Ethnicity (Select all that apply):
    a. American Indian or Alaska Native
    b. Asian
    c. Black or African American
    d. Hispanic or Latino
    e. Native Hawaiian or Other Pacific Islander
    f. Two or More Races
    g. White
    h. Other
5. Age
    a. Under 18
    b. 18-24
    c. 25-34
    d. 35-44
    e. 45-54
    f. 55-64
    g. 65+

Familiarity with Data

6. Did you know about Community Information Now before we contacted you?
    a. Yes
    b. No
    c. Other (please specify)
7. Which, if any, of CINow's data tools have you used?
    a. None
    b. Viz-a-lyzer.
    c. Somos Neighbors.
    d. ACS Sidekick.
    e. Data Explorer.
    f. COVID-19 Scatterplot.

8. Choose the response you most closely identify with:
   a. I do not use data or know how to find it.
   b. I'm aware of data and use data occasionally.
   c. I use data regularly and am aware of a few data sources.
   I am very familiar with data, how to find it, how to analyze it, and how to interpret it.

Data Tool Preferences. The following questions are about what you would prefer from a data tool.

9. What indicators/variables would you like to see in a data tool? (select all that apply)
   a. Basic Demographics (race/ethnicity, age, sex, population counts, etc.)
   b. Income
   c. Food Insecurity/Food Deserts
   d. Transportation
   e. Housing/Food/Social Assistance
   f. Crime Rates
   g. Voter Registration and Turnout
   h. Births/Prenatal Care/Birthweight
   i. Educational Attainment
   j. Child Abuse Prevalence
   k. Sexual Assault Prevalence
   l. Uninsured Rates
   m. Poverty Rates
   n. Qualified for Chip/Medicaid
   o. Cancer Incidence Rates
   p. COVID-19 Rates
   q. COVID-19 Vaccination Rates
   r. Disease Prevalence
   s. Other (please specify)
10. What data sources would you like included in the tool? (select all that apply)
   a. Census/American Community Survey
   b. U.S. Bureau of Labor Statistics
   c. Feeding Texas
   d. Local Area Unemployment Statistics (LAUS)
   e. Community Health Needs Assessment (CHNA)
   f. Behavioral Risk Factor Surveillance System (BRFSS)
   g. Texas Health and Human Services
   h. CDC Wonder
   i. Every Texan
   j. San Antonio Metro Health
   k. County Health Rankings
11. What geographies would you like to see in a data tool? (select all that apply)
   a. National
   b. State
   c. County

d. City
e. MSA
f. Zip Codes
g. Census Tracts
h. Neighborhood
i. City Council Districts
j. County Precincts
k. School Districts
l. Other (Please specify)

12. What visualizations would you like to see in a data tool? (select all that apply)
   a. Maps
   b. Data Tables
   c. Trend Line Graphs
   d. Comparison Bar Charts (Geography)
   e. Comparison Bar Charts (Race/Ethnicity)
   f. Comparison Bar Charts (Age)
   g. Other (please specify)
13. What sort of training would you like for this data tool? (select all that apply)
   a. None
   b. Live Workshops
   c. Virtual Workshops
   d. Video Tutorials
   e. Written Manuals
   f. Other (please specify)
14. What do you use data for? (select all that apply)
   a. Grant Writing
   b. Program Planning
   c. Budgeting
   d. Other (please specify):
15. On a scale of 1 to 5, where 1 is “not at all” and 5 is “very much”, how much do you trust the data you have access to?
   a. 1
   b. 2
   c. 3
   d. 4
   e. 5
16. How do you usually get updates about data?
   a. Email/Newsletters
   b. Social Media
   c. Print
   d. Other (please specify):
17. Is there anything else you would like to add about what you would like to have in a data tool? (write-in option)

Thank you for taking our survey! For data resources or to keep up with our progress creating this data tool, visit us at cinow.info. You can sign up for our newsletter here.

Appendix C

User Testing Guide

Note: PowerPoint presentation has been converted to text below.

Community Information Now: Bexar Data Dive User Testing

Introduction

CINow is a nonprofit housed in the School of Public Health. Our mission is improved lives through democratized data. CINow provides data, tools, analysis, and training to inform decisions and improve Texas communities. We find, collect, link and analyze, and visually display the data that our neighbors need to improve neighborhood and regional conditions. We serve Bexar County and 11 surrounding counties in South-Central Texas.

The Office of Minority Health awarded us a grant to design and implement a data platform which will give users access to health and SDOH data. The long-term goal is to reduce health disparities by providing organizations, nonprofits, government entities, and businesses with an easy way to access health information and use it to inform decisions and policy. We have also created a new geographic measure, which we are calling Statistically Small Areas, to help visualize indicators and health disparities across meaningful regions across Bexar County.

We began this project almost a year ago. Early in the process, we conducted interviews and focus groups to get input from people about what they would want in a data platform. This allowed us to create user stories based on people's roles and needs. An example of one of our user stories was "**As a** *community health worker* **I want** *current COVID-19 vaccination rates by neighborhood* **so that** *I know where to target our outreach."* This was something we heard from the interviews and focus groups, along with the desire for maps and charts that can change colors, access to multiple indicators as once (such as insurance coverage rates by race/ethnicity), and an easy-to-use interface. We implemented as many of them as we could, within our restraints, and we will be uploading more data to it soon.

The general purpose of Bexar Data Dive is to offer data users access to multiple indicators and data sources in one place. We have designed it to be useful to beginner, intermediate, and advanced data users – and we hope that we can improve the health outcomes of Bexar County, especially those in marginalized populations.

User Testing

Expectations for the session

- We want your honest feedback (don't worry, you won't hurt our feelings)
- Currently, our focus is on the functionality of the platform, not the design or data

Demo

(Provide brief demo of Bexar Data Dive)

Explore Bexar Data Dive

Take 5 minutes to explore the platform on your own: https://dive.cinow.info

(Use this as a chance to note any questions they have as they explore)

Activity 1

https://dive.cinow.info

Please navigate to the My Community tool

Try to answer the following questions on your own:

- In SSA 34, what percentage of the population is between 5-17 years of age?
- How does that compare to Bexar County as a whole?

(Answers: In SSA 34, what percentage of the population is between 5-17 years of age? 19.3%

How does that compare to Bexar County as a whole? 18.5%)

Activity 1 Continued

We need a volunteer to share their screen

Please navigate to the My Community tool

Please walk us through how you tried to answer the questions by narrating your experiences and expectations.

- In SSA 34, what percentage of the population is between 5-17 years of age?
- How does that compare to Bexar County as a whole?

Everyone can chime in with feedback

(Answers: In SSA 34, what percentage of the population is between 5-17 years of age? 19.3%

How does that compare to Bexar County as a whole? 18.5%

Take notes on the following:

What did the user tester do or try to do?

Ex: (Capture method: Searched "hiking" or display pathway Clicked "Activities > Outdoor > Hiking")

What were they expecting?

Ex: (Have testers narrate their experience and ask them what they were expecting)

What actually happened? Did it meet expectations?

Ex: (Take notes about what actually happened "Clicked on Start button but nothing loaded" or "She searched for "hiking 60601" and relevant results showed up immediately")

Did the tester successfully complete the task? (first time, second?)

Ex: (Tester first tried to click "Hiking," but couldn't find results. Tried using search second time… and thought the results were correct. Never saw the proper result)

How could their experience be improved?

Ex: (Capture ideas and suggestions - it can be process oriented or design… Anything!))

Activity 2

https://dive.cinow.info

Please navigate to the Explore Data tool
Try to answer the following questions on your own:

- Choose the indicator "18 – 34"
  - What percentage of Hispanic or Latino females are ages 18 to 34 in zip code 78244?
  - How has that percentage changed between 2015 and 2020?
  - How does this compare to Asian females?

(Answers: What percentage of Hispanic or Latino females are ages 18 to 34 in zip code 78244? 27.6%
How has that changed between 2015 and 2020? In 2015, it was 23.8%, so it has increased. Hint: Navigate to Trend Chart, or toggle the year filter.
How does this compare to Asian females? It's 15.9% for Asian females. Hint: Navigate to Comparison Chart and compare by race -> Asian)

Activity 2 Continued

We need another volunteer to share their screen
Please navigate to the Explore Data tool
Please walk us through how you tried to answer the questions by narrating your experiences and expectations.

- Choose the indicator “18 – 34”
  - What percentage of Hispanic or Latino females are ages 18 to 34 in zip code 78244?
  - How has that percentage changed between 2015 and 2020?
  - How does this compare to Asian females?

Everyone can chime in with feedback

(Answers: What percentage of Hispanic or Latino females are ages 18 to 34 in zip code 78244? 27.6%
How has that changed between 2015 and 2020? In 2015, it was 23.8%, so it has increased. Hint: Navigate to Trend Chart, or toggle the year filter.
How does this compare to Asian females? It’s 15.9% for Asian females. Hint: Navigate to Comparison Chart and compare by race -> Asian
Take notes on the following:

What did the user tester do or try to do?
Ex: (Capture method: Searched “hiking” or display pathway Clicked “Activities > Outdoor > Hiking”)

What were they expecting?
Ex: (Have testers narrate their experience and ask them what they were expecting)

What actually happened? Did it meet expectations?
Ex: (Take notes about what actually happened “Clicked on Start button but nothing loaded” or “She searched for “hiking 60601” and relevant results showed up immediately”)

Did the tester successfully complete the task? (first time, second?)
Ex: (Tester first tried to click “Hiking,” but couldn’t find results. Tried using search second time… and thought the results were correct. Never saw the proper result)

How could their experience be improved?
Ex: (Capture ideas and suggestions - it can be process oriented or design… Anything!))

Activity 3

https://dive.cinow.info

Please navigate to the Tables & Downloads tool
Try to answer the following questions on your own:

- Choose the indicator “Under 5”
  - In zip code 78109, what percentage of Black or African American males are under 5 years of age, according to 2020 data?
- How would you download this table if you wanted to?

        - What's the data source for this indicator?

(Answers: In zip code 78109, what percentage of Black or African American males are under 5 years of age, according to 2020 data? 3.6%
How would you download this table if you wanted to? With the download button in the top right
What's the data source for this indicator? ACS, 2020)

Activity 3 Continued

We need another volunteer to share their screen
Please navigate to the Tables & Downloads tool
Please walk us through how you tried to answer the questions by narrating your experiences and expectations.

- Choose the indicator "Under 5"
    - In zip code 78109, what percentage of Black or African American males are under 5 years of age, according to 2020 data?
- How would you download this table if you wanted to?
- What's the data source for this indicator?

Everyone can chime in with feedback
(Answers: In zip code 78109, what percentage of Black or African American males are under 5 years of age, according to 2020 data? 3.6%
How would you download this table if you wanted to? With the download button in the top right
What's the data source for this indicator? ACS, 2020
Take notes on the following:

What did the user tester do or try to do?
Ex: (Capture method: Searched "hiking" or display pathway Clicked "Activities > Outdoor > Hiking")

What were they expecting?
Ex: (Have testers narrate their experience and ask them what they were expecting)

What actually happened? Did it meet expectations?
Ex: (Take notes about what actually happened "Clicked on Start button but nothing loaded" or "She searched for "hiking 60601" and relevant results showed up immediately")

Did the tester successfully complete the task? (first time, second?)
Ex: (Tester first tried to click "Hiking," but couldn't find results. Tried using search second time… and thought the results were correct. Never saw the proper result)

How could their experience be improved?
Ex: (Capture ideas and suggestions - it can be process oriented or design… Anything!))

Feedback

What were some things you liked about using the platform? (if anything)

What were some things that could use improvement? (if anything)
What were some things that made it very difficult to use the platform? (if anything)
Did everything work the way you expected it to?
(This portion runs similarly to a focus group)

Author Biographies

Danequa Forrest, PhD, (danequa.forrest@uth.tmc.edu) is a research coordinator with Community Information Now, a local data intermediary in San Antonio, Texas. With a PhD in sociology and double minors in applied statistics and African and African-American studies, she uses her mixed methods skillset to analyze and interpret qualitative and quantitative data. Most of her work involves contextualizing data through the lens of health and racial equity.

Jeremy Pyne, MPA, is a project manager with Community Information Now with a background in public administration and Geographic Information Systems. He has two decades of experience managing and supporting community-based research projects including local indicator reports, online dashboards and interactive mapping tools with the focus to support equitable community change.

Laura McKieran, DrPH, is executive director of Community Information Now, a nonprofit local data intermediary, and an Associate Professor of Management, Policy and Community Health at the UTHealth Houston School of Public Health in San Antonio. She has over 25 years of experience connecting data and community via informatics, evaluation and outcome measurement, community assessment, and data-informed planning.

Cristina E. Martinez, PhD(c), MPH, is a research coordinator at Community Information Now with a background in public health and demography. She specializes in applied demographic methods and advanced statistical analysis.

Unlocking Patient Portals: Health Information Professionals Navigating Challenges and Shaping the Future

Jennifer L. Peterson, PhD, RHIA, CTR, and Shannon H. Houser, PhD, MPH, RHIA, FAHIMA

Abstract

Due to recent regulations and the COVID-19 pandemic, patient portals have increased in use and importance as a tool for both patients and providers. While patient portals have many benefits, the recent increase in use has resulted in additional complexities in managing these portals. Health information (HI) professionals are ideally suited to manage these tools. While past efforts may have focused on increasing portal use, current efforts must include ensuring patient access, data quality, portal policies and procedures, and more. This study was designed to explore the experiences and perspectives of a group of HI directors and patient portal managers who are deeply involved in portal use and management. The findings of this study are used to assess the patient portal management role that HI professionals currently play and could play in the future, develop guidelines for best practices, and determine educational needs for both higher and professional education.

Key Words: patient portals, data management, interoperability, patient engagement, 21st Century Cures Act, patient-provider communication, patient access

Introduction

Ten years ago, the use of patient portals was rare; most patients and physicians were not communicating electronically. However, with the advent of the Health Information Technology for Economic and Clinical Health (HITECH) Act, Meaningful Use, and the Merit-based Incentive Payment System (MIPS), the availability and use of patient portals has increased dramatically in recent years.[1,2,3] These platforms do more than just store health records — they play an instrumental role in enhancing patients' participation in their healthcare and providing an efficient means to relay crucial health information to various medical stakeholders. Portals provide easy communication between patients and providers and allow patients access to important health information that they can use themselves or share with other providers. These multi-faceted tools provide not only patient health information but also a variety of tools to help patients better manage their health.

A decade ago, patient portals were in their early stages, with the basic abilities to share information and allow communication between healthcare providers and patients. HI professionals' focus on patient portals was minimal: to increase usage and help patients enroll in the portal. However, transformative policies like the HITECH Act, Meaningful Use, and MIPS prompted the widespread adoption of patient portals. By 2020, as HealthIT.gov reveals, almost 40 percent of Americans had interacted with a patient portal, marking a significant growth of 13 percent since 2014.[3] A recent ONC publication noted that this use increased even further during the COVID-19 pandemic as "the share of individuals nationwide who were offered and accessed their online medical records or patient portals more than doubled between 2014 and 2022".[4] This trend is not confined to only the tech-inclined younger population. Data from the University of

Michigan's National Poll on Healthy Aging demonstrates that nearly 78 percent of those aged between 50 and 80 have engaged with at least one patient portal.[5]

The COVID-19 pandemic further underscored the relevance and usefulness of these portals.[3] Faced with the rising demand for virtual medical consultations and communication, patient portals became essential tools, bridging the physical divide, facilitating uninterrupted care, and providing digital touchpoints between patients and healthcare providers. This enabled patients to connect safely with their healthcare teams, access diagnostic results, request prescription updates, manage appointments, and participate in telehealth sessions seamlessly.

The 21st Century Cures Act marked a significant shift in health information regulations, particularly related to health information technology (IT) and patient access to electronic health records (EHRs).[6] It emphasized the patient's right to access their electronic health data. Consequently, patient portals became essential platforms, enabling patients to engage with their health records, manage appointments, communicate with healthcare providers, and review billing details. The Cures Act encouraged health providers and IT developers to create more user-friendly systems, facilitating easy patient interactions with their information. Furthermore, the push for more complete and timely electronic access led numerous healthcare institutions to refine or broaden their patient portal services.

While the adoption of patient portals has brought numerous benefits, recent trends and initiatives have intensified the complexities of managing these digital platforms. A particularly important challenge is the increase in patient-provider messaging. Secure communications between patients and providers escalated from 48 percent in 2017 to almost 60 percent in 2020.[3] The integration of secure messaging into patient portals is lauded as a breakthrough in healthcare communication, meeting contemporary patient demands for rapid, transparent access to their healthcare teams. This heightened emphasis on digital patient-provider communication is evident in MIPS, particularly where "providing patients electronic access to their health information" is a key objective. [7]

However, the increased messaging frequency within these portals has presented its own set of challenges. Healthcare providers report that they are inundated with messages, leading to extended work hours, uncompensated time, and, in some cases, burnout. The volume of these messages has, in some instances, resulted in “something closer to a clinical encounter”.[8] This escalation prompted some leading health care systems to start charging patients for patient portal messages or e-visits. There are a number of considerations to billing for portal messaging. These include the potential inability of an EHR to create a billable encounter, the fact that the billing is only based on time spent, and the lack of coding guidelines.[8] In addition, if e-visit messaging is billed, the physician or nurse practitioner must be the respondent; nurses’ or assistants’ responses cannot be billed.[9]

While billing may seem like a practical solution to compensate for a clinician's time, its introduction has broader implications. A study completed at UCSF Health following the implementation of billing for e-visits showed that “a reduction in patient portal messaging (both threads and individual messages) was observed that may be attributable to awareness of the possibility of being billed.”[10] Instituting such charges might discourage patients from using these platforms, which, in turn, could diminish the overall efficacy of patient portals. The long-term impact on patient satisfaction, retention, and health outcomes remains a topic for further exploration.

Clearly, the use and importance of portals has increased dramatically in recent years. Due to these changes, HI professionals are positioned to play an increasingly pivotal role. While their past contributions were largely geared towards driving patient portal adoption, today, there is a need for supporting patient access to portals, ensuring data accuracy, establishing strong portal management protocols, meeting regulatory requirements, overseeing billing procedures, and gauging patient feedback.

It is not clear that HI professionals are consistently in management and oversight roles for patient portals. The literature in this area is sparse, with little guidance as to the role HI professionals play or should play in the oversight of portals. This study was developed to address this gap and determine the current role that HI professionals play in portal management, as well as to initiate the development of best HI practices for patient portals. This study utilized a case study design aimed to probe deeper into the current trends in patient portal and messaging utilization, highlight associated challenges, and draw attention to the increasingly significant role of HI professionals. Moreover, it was designed to evaluate the need for educational programs, continuing education, and policy development to guide the future of patient portal engagement and communication.

Methods

Research Design and Participants

This research adopted a qualitative case study methodology to explore the experiences and perspectives of its participants. A total of 11 participants were recruited, all of whom held positions in their respective facilities either as HI directors or patient portal managers. These professionals originated from two specific states: Illinois and Alabama.

This qualitative case study was not designed to provide widely generalizable data. However, it was designed to provide initial insight for the following questions:

1. What is the current role that HI professionals play in patient portal management?
2. What are the current trends and challenges in patient portal management?
3. What is the need for education for HI professionals and others involved in patient portal management?
4. What is the need for policy development for patient portal management?

Participants were selected through a convenience sampling method due to the practical benefits of easy availability and accessibility. As HI professionals in the field, the researchers used state professional association lists, to which they had easy access, to identify potential participants, ensuring they were active professionals in the field who were working with portals. By selecting participants from this list, the study aimed to gather in-depth insights, challenges, and best practices related to patient portal management from those with firsthand experience. While the majority of participants were HI professionals, the results are felt to be pertinent to any professionals serving as patient portal managers.

Ethical Considerations

Prior to the collection of any data, the research study and the interview questionnaire were submitted to the Illinois State University Institutional Review Board (IRB) for review. The IRB found the study to be exempt from IRB review.

Recruitment and Data Collection

The participants were interviewed to gain insights into:

1. Current patient portal usage and policies;
2. Existing and future portal message billing practices;
3. Current challenges associated with patient portals; and
4. Needs concerning education, training, and policy development related to patient portals.

The interview process was based on a structured questionnaire comprised of 22 predefined questions, with the majority being open-ended. In order to address the above issues, the researchers developed survey questions designed to elicit information regarding the respondent's facility, the respondent's role in patient portal management, the facility's current patient portal management policies and procedures, and the facility's current messaging billing practices. In addition, the researchers included questions regarding respondent's challenges, the need for policies and procedures, and other issues the respondent felt were pertinent. The questionnaire was pilot tested by four HI professionals; three in Illinois and one in Alabama. Following the pilot test, no changes were recommended or required. The questionnaire can be found in Appendix A. The participant interviews were conducted through Zoom meetings or via telephone calls.

Data Analysis

Qualitative data were obtained through in-depth interviews. This data was thoroughly analyzed on two levels: individual case studies to understand unique scenarios and perspectives, and aggregate analysis to identify common themes and patterns across participants. The data was analyzed using the constant comparative method. The interviews were transcribed and data were then coded to identify themes and categories. Data were further analyzed to determine an in-depth understanding of the relationships between themes and categories and to summarize the data and over-arching themes. This was done both within and between individual interview data in order to analyze both individual case studies and aggregate information.

Limitations

The results may not be generalizable based on the small sample size of 11 participants. The convenience sampling methodology and the fact that participants were from only two states could also reduce the generalizability of the findings.

Results

Participant Overview

A total of 11 participants were interviewed to gain insights into their involvement and experiences with patient portals and their associated management. The demographic breakdown of the participants showed a majority holding positions as HI directors or managers. Other roles included office manager, patient portal representative, and patient liaison. This aligned with the target group as the goal was to interview individuals in these roles. All but one of the participants worked in a hospital or health system; one worked in an outpatient private office.

Involvement with Patient Portals

In examining their roles in relation to patient portals, it was found that six of the participants were deeply involved, either having direct responsibilities associated with patient portals or overseeing front line patient portal employees. Four served in a patient portal management support role. One participant stated that their role was "TBD (to be determined)" as specific patient portal management duties had not yet been clarified. All but one participant collaborated with their IT department or outside IT provider in the use and development of patient portals. All participants' facilities' patient portals were either offered through their EHR or through a system contracted with their EHR. Participants' facilities in Illinois were most likely to use Epic, whereas participants' facilities in Alabama were more likely to use Cerner. Other EHRs used were OncoEMR, Paragon, Allscripts, and Meditech.

EHR Systems and Patient Portal Platforms

All patient portals were provided either directly through the facilities' EHRs or through a system contracted with the EHR. When analyzing the EHR systems used, regional preferences were evident.

Data and Services Available on Patient Portals

Facilities provided a variety of data in their patient portals as can be seen below. In addition, there were a variety of patient services provided in portals. These can be seen in Table 1.

Table 1: Types of Data and Services Available within Patient Portals

Data Type	**Services**
Appointments – past and future	Access to Health Reference Library
Clinical data	Make/Change/Cancel Appointments
Correspondence	Payment submission
Demographic information	Prescription refill requests
Financial information	Preventative care reminders
Immunizations	Secure messaging with provider
Links to other facility's data	
Medications	
Patient history	
Test results	

Some participants stated their portals included “everything except nursing,” “a complete record (that) can be downloaded in one site,” or “all data except ‘sensitive’ data.” It should be noted that some participants stated that more data is now included in their patient portals following the implementation of the Cures Act; in some cases, patient portal data now includes nursing notes. It was also noted that not all portals included billing information and billing/payment functionality. Some facilities utilized a billing platform separate from the patient portal.

All facilities had at least some data automatically and immediately pushed to the patient portal from the EHR. There was some variability in types of data that may have a delayed release in being made available to patients. These delayed times were tied to requirements for providers to sign off or approve data before it was released to the portal. There were a variety of policies in regard to this. While most facilities did not require provider sign-off or release of data, three of the respondents stated that there was some requirement in place for provider review and release. However, in two of these three cases, data was automatically pushed to the portal after a set period of time, which ranged from 36 to 72 hours. “Sensitive data” was frequently not released until the provider signed off on the release. Two participants noted that they have experienced some issues with the speed at which results are pushed to the patient portal. They both stated that their facilities had experiences in which ER patients received their lab or radiology results through the portal, and then, knowing the results, left the ER before the provider was able to talk with them about their results. These facilities have contemplated slowing the pushing of results to the portal in the ER setting.

Document Handling

When asked about scanned documents and whether these are pushed to the patient portal, the results were mixed. Four of the respondents’ facilities pushed scanned records into the patient portal. Three of these four facilities pushed the scanned documents to the patient portal automatically and immediately. Each of these three facilities had experienced issues with scanned documents being scanned into the wrong patient chart and released to the wrong patient portal. None of these facilities had experienced HIPAA violation because of the scanning error as the errors were caught prior to patient viewing. One site, however, noted that they have implemented a new system with AI which has decreased the error rate. All respondents were aware of the potential for erroneously scanned records to become a HIPAA violation and data breach.

Communication

Participants were next asked about patient-provider messaging. When asked which staff members answer patient portal messages, four stated that nurses screen and/or answer messages. Six sites stated that centralized system staff or patient portal staff/patient liaisons screen and/or answer messages. One site stated that doctors screen and answer all of their own messages. Most sites had nursing or other staff work the patient message inbox and push messages that required a physician response to the providers. This information is pertinent as related to the recent trend of billing for patient messaging. None of the participants’ facilities were billing for portal messaging at the time of the interviews.

Challenges in Portal Management

Participants provided a wide variety of responses regarding challenges of patient portal use and management. After review and analysis, it was found that these fell into three main categories. These categories and challenges can be seen in Table 2.

Table 2: Types of Patient Portal Management Challenges

Timing Issues	**Patient**	**Other**
Other hospital in town releases information faster	Need for more education for patients (both use and content); patients interpret data wrong or Google information	Patients complain about lack of results from other facilities (i.e., MyChart at one facility should have all MyChart info from all facilities)
ER patients get results then leave before seeing provider	Need to create emails for patient access (i.e., no email, couple shares email)	Patients upset about content (i.e., "patient refuses treatment")
Patients try to enroll before they receive invite	Complaints about parent limited access for ages 12-17	Issues with copy/paste in notes (i.e., diagnosis from 2015)
Staff remembering to reply to messages in timely manner	Inappropriate parent/proxy access	Dependent on patient to maintain privacy/security
Providers/staff slow in answering messages	Patients' technical challenges take time and attention	Maintaining current provider list for messaging
	Patients email with problems but don't include ID info	Ensuring no information blocking
	Numerous password/login reset requests	Low patient portal usage
	Patients get text that results are ready and go to physical facility for copy of results	Increase in amendment requests
	Managing proxy requests	Scanned documents going into the wrong portal – HIPAA violation

Policies and Procedures for Patient Portal Use and Management

Participants stated that their facilities had a variety of policies and procedures regarding patient portal use and management. The policies and procedures that were in place among the participants' facilities varied based on setting, inpatient vs outpatient, as well as by facility. The majority of the facilities relied on the policies and procedures provided by the EHR provider, a few had facility specific basic documents, one had very in-depth policies and procedures, and two were developing further documentation. Policies and procedures fell into two main categories: basic use, and management and regulatory issues. Some of the basic use policies and procedures covered enrollment, time frame for information release, use at bedside, parent/child access regulations, responsibilities by department or position, proxy use and proxy authorization forms, and documentation included in the portal. Management and regulatory policies and procedures covered guidelines for specific management of different issues, revocation of patient access due to misuse, code of ethics for portal use, inappropriate language in messages, and proxy removal. While most sites had policies and procedures for basic use, very few sites had more in-depth policies and procedures for managing the portal and ensuring regulatory compliance.

Patient Engagement and Relationship in Portal Use

Finally, participants were asked to add any other important feedback or information regarding patient portal use and management. These end comments seemed to focus more on patient engagement and relationship issues. These comments are seen in Table 3.

Table 3: Additional Comments Regarding Patient Portal Use or Management

Providers like portals as they decrease patient calls.
Portals are good for billing/payment.
Press-Ganey surveys could be added to portals.
Portals result in a culture change since patients have more information.
Portals should be user friendly for patients and providers.
Working with patients on portals is a new role requiring a positive attitude and desire to serve the customer.

Discussion

It is clear that HI professionals who participated in this study are currently taking on many roles in the management of patient portals. Participants are intricately involved in ensuring quality patient information in the portal, ensuring privacy and security of portal information, ensuring patient accessibility, and improving integration with other health information systems. It is noted that billing for patient messaging was not being done in any of the participant's facilities at the time of the study, therefore, they were not involved in this aspect of patient portal management.

The participants in this study mentioned a variety of challenges that they are experiencing with patient portal use and management. In many instances, a challenge faced by one professional may have been addressed by another. The researchers, therefore, compiled a list of best practices based on participant comments and expertise, as well as documentation in the literature. These can be found in Table 4.

Table 4: Patient Portal Management Best Practices

Adjust timing of results to avoid problems with patient misunderstanding or cancelling of interactions with providers (i.e., release results after provider review, especially in the ER setting, allow time for providers to discuss results with patient for certain tests/circumstances).
Engage providers in review of timing of results to enable the best balance for providers and patients.
Ensure compliance with 21st Century Cures Act – if information is not included in the patient portal ensure that patients understand it is still available to them through other means.
Provide education to nurses regarding the fact that patients may see their documentation.

Develop policies and procedures for not only basic use of patient portals but more complex management of patient portals and potential issues that may arise.
Provide ongoing patient education on patient portal use and health literacy.
Advocate for patients to help meet patient needs and ease of use.
Train staff in consumer facing skills and equip them with the skills to aid patients in portal use.
Watch for provider burn-out and increases in provider uncompensated time due to patient messaging.
Prepare for potential billing for complex patient messaging encounters. If billing is initiated ensure knowledge of regulations and develop policies and procedures.
If billing is initiated, design and implement data analyses to evaluate the effect of billing for patient messaging to ensure appropriate patient usage and to monitor for changes in quality of care.

This list clearly points to skills that fall within the HI professionals' domain. HI professionals have a unique understanding of data management, provider engagement, and patient facing aspects of patient portal management. These professionals are well placed to manage these areas and use these guidelines to improve patient portal efficiency and effectiveness for both patients and providers. In addition, HI professionals are ideally suited to use their knowledge and be involved in discussions about patient portal design weaknesses and to work with patient portal vendors to improve design.

With the recent growth in the use of patient portals, HI professionals are being asked to take on more duties related to patient portal management. Both HI students and professionals can benefit from education in the management of patient portals. The AHIMA Council for Excellence in Education™ 2018 Health Information Management Baccalaureate Degree Curriculum Guidance (2022 edition)[11] includes multiple suggestions for integrating education regarding patient portal management throughout the curriculum competencies. This study of current HI professional involvement with patient portal management reinforces the need for new graduates with patient portal management knowledge and skills. In addition, the rapid growth of patient portal use may have left HI professionals with little time to master portal management skills, therefore prompting the need for continuing education offerings in portal management and best practices.

Limitations

While this study was limited to a small sample size with participants from only two states, the data gathered is valuable in that it provides insight into current patient portal management and associated challenges. Further study in this area could provide additional insight into patient portal management best practices. For example, this study found that one facility is using AI to decrease scanning errors. Further research into such new technologies and processes can assist others in providing high quality patient portal data. As patient portal usage continues to increase and healthcare systems expand their patient portal offerings, additional research will be needed to provide further insight into best practices and sharing of successes.

Conclusion

This in-depth study was designed to evaluate current uses of patient portals and messaging, current challenges surrounding patient portal management, and responses to these challenges, including planned billing for messaging. The results of this study provided insight into these issues as well as additional information on the role that HI professionals currently are playing and could play in the future, guidelines for best practices, and educational needs at the higher education and professional levels.

The recent increase in patient portal use and the need for new skills in managing these portals has placed new responsibilities on HI professionals. As portal use continues to increase and portals become more sophisticated, HI professionals will be called upon to take on even more new roles. Ongoing review and study of these roles and the associated managerial needs will allow HI professionals to grow with portal use and development and lead them to further improve patient portals which will result in a more positive portal experience and improved patient outcomes.

Support

There was no financial support or grant for this study. The authors report there are no conflicts of interest.

References:

1. Office of the National Coordinator for Health Information Technology. “A Decade of Data Examined: Patient Access to Electronic Health Information.” Washington, DC, December 2023. Available online at https://www.healthit.gov/buzz-blog/a-decade-of-data-examined/a-decade-of-data-examined-patient-access-to-electronic-health-information
2. Turner, Kea, Young-Rock, Hong, Sandhya, Yadav, et. al. “Patient Portal Utilization: Before and After Stage 2 Electronic Health Record Meaningful Use.” *Journal of the American Medical Informatics Association.* 26, no. 10 (2019): para 11.
3. Office of the National Coordinator for Health Information Technology. “Individuals’ Access and Use of Patient Portals and Smartphone Health Apps, 2020.” Washington, DC, September 2021. Available online at https://www.healthit.gov/data/data-briefs/individuals-access-and-use-patient-portals-and-smartphone-health-apps-2020
4. Office of the National Coordinator for Health Information Technology. “Individuals’ Access and Use of Patient Portals and Smartphone Health Apps, 2022.” Washington, DC, October 2023: para 3. Available online at https://www.healthit.gov/data/data-briefs/individuals-access-and-use-patient-portals-and-smartphone-health-apps-2022
5. Murez, Cara. “More Older Americans Use Online 'Patient Portals' to Access Care.” *HealthDay*. May 24, 2023. Available online at https://consumer.healthday.com/health-care-tech-2660468555.html
6. Public Law 114-255. December 13, 2016. Available online at https://www.govinfo.gov/content/pkg/PLAW-114publ255/pdf/PLAW-114publ255.pdf
7. Federal Register. “Quality Payment Program.” Washington, DC. November 23, 2018. Available online at https://www.federalregister.gov/documents/2018/11/23/2018-

24170/medicare-program-revisions-to-payment-policies-under-the-physician-fee-schedule-and-other-revisions#p-3196

8. Dowling, Robert A, “Charging for Portal Use: Here is What Urologists Should Know.” *Urology Times*. January 16, 2023: para 7. Available online at https://www.urologytimes.com/view/charging-for-portal-use-here-is-what-urologists-should-know
9. CMS.gov Newsroom. “Medicare Telemedicine Health Care Provider Fact Sheet.” Washington, DC. March 17, 2020. Available online at https://www.cms.gov/newsroom/fact-sheets/medicare-telemedicine-health-care-provider-fact-sheet
10. Holmgren, A. Jay, Byron, Maria E, Grouse, Carrie K, et al. “Association Between Billing Patient Portal Messages as e-Visits and Patient Messaging Volume.” *Journal of the American Medical Association*. 329, no. 4 (2023): para. 6.
11. AHIMA Council for Excellence in Education, “2018 Health Information Management Baccalaureate Degree Curriculum Guidance (2022 edition).” Chicago, IL.

Appendix A

Interview Questionnaire:

1. What is your position?
2. What type of facility to you work in?
3. What is your role/level of involvement in patient portals?
4. Do you work with IT in the use/development of patient portals/data?
5. What EHR do you use?
6. What data is included in your patient portal?
7. How is that data pulled into the portal from your EHR?
8. How do results of tests, exams, etc. appear in your portal?
9. How quickly do results, exam summaries, etc. appear in your portal?
10. Does a practitioner have to approve the release of the above information to the portal?
11. Do you scan results in to the EHR that then go into the patient portal?
12. How quickly do these scanned results appear in the patient portal?
13. Do you have any issues with scanned images going into the wrong patient chart, being released immediately, and resulting in HIPAA violations?
14. Who answers patient portal messages?
15. How do your physician’s approach portal messages?
16. Do you bill for portal messaging?
17. If so, how are these billed (specifics, which types of messages, coding, etc.)?
18. Are patients notified that they may be billed for messaging?
19. What has the patient response been to the potential for billing?
20. What are your/your practice’s challenges regarding patient portals?
21. What policies and procedures do you have regarding patient portal use/practice?
22. What other issues do you feel are important regarding patient portal use/patient response/practitioner response, etc.

Jennifer L. Peterson, **PhD, RHIA, CTR,** is an associate professor and program director for Health Informatics and Management in the department of health sciences at Illinois State University in Normal, IL.

Shannon H. Houser, **PhD, MPH, RHIA, FAHIMA,** is a professor at the University of Alabama at Birmingham's department of health services administration in Birmingham, AL.

Using Electronic Health Records Data to Identify Strong Performers in Healthcare Quality Improvement

By Adam Baus, PhD, MA, MPH, Andrea Calkins, MPH, Cecil Pollard, MA, Craig Robinson, MPH, Robin Seabury, MS, Marcus Thygeson, MD, MPH, Curt Lindberg, MHA, DMan, Andrya Durr, PhD

Summary: This case study presents an adaptable, straightforward framework for identifying positive deviance, or strong performers, within the healthcare setting and is intended for any primary care health system tracking quality measures and aiming to understand the performance of their providers, clinic sites, or organization.

Abstract

Assessing for positive deviance is one method of identifying individuals, teams, or organizations that perform substantially better than their peers. This approach has been used to support quality-of-care improvement processes in healthcare settings by identifying healthcare team members who perform comparatively well within a given environment and sharing their opinions, actions, and practices with others. This case study presents an adaptable, straightforward framework for identifying positive deviance, or strong performers, within the healthcare setting and is intended for any primary care health system tracking quality measures and aiming to understand the performance of their providers, clinic sites, or organization. Moreover, this protocol does not require the use of more time-consuming methods, such as interviews, and is instead based on repurposing data already being documented in the electronic health record.

Keywords: Electronic Health Record Data, Positive Deviance, Primary Care, Health System, Healthcare Provider, Quality Improvement

Introduction

Positive deviance is a strengths-based approach to identifying individuals within an organization who excel or perform substantially better than their peers, and then understanding the reasons for their better than average performance.[1–4] This field of study has been used as a support to the quality of care improvement process in healthcare settings by identifying healthcare team members who perform comparatively well within a given environment, understanding the reasons for their high performance including their attitudes and actions, promoting those positive practices to others within the organization, and determining ways in which positive changes can be sustained.[5–9]

Background

Several methods exist to identify strong performers, as presented in methodological reviews of peer reviewed publications and/or staff and physician interviews. Regardless of the methods used, all assessments are performed with the goal of identifying individuals or organizations that are performing above average and that serve as good role models for strategies to improve patient care. Finding ways in which the information gathering process can be supported in primary care is important, given the time constraints of busy clinicians and administrators.

This case study examines whether the identification of strong performers can be supported in a rural safety-net health system by retrospective analysis of national standardized quality metrics housed in an electronic health record (EHR). This study took place in Cabin Creek Health Systems, a West Virginia-based federally qualified health center with the mission to promote the health and well-being of all people in their community, especially the most vulnerable, through healthcare guided by science, compassion, and respect, and to contribute to the education of skilled and caring health professionals. The goal of this study was to examine the feasibility and utility of using existing data to identify strong performers among healthcare providers and generate a standardized data analysis methodology that is applicable to any EHR capable of providing quality metrics by provider. For context, this study contributed to the evaluation of a self-measured blood pressure monitoring initiative implemented at Cabin Creek Health Systems. The aim was to gain insights into healthcare providers who excel not only in blood pressure quality measures but also in other potential quality indicators. The concept of positive deviance was central to efforts in reducing hypertension via understanding quality measures in addition to exploring patient/provider dynamics, delivery of care, and lifestyle coaching. Concurrently, this study supports the recent charge to help standardize positive deviance methodologies across research settings.[10]

Methods

Data Collection

The concept of "positive deviance" can be used to identify providers who perform above average, or exceptionally well, based on a list of pre-determined quality measures that are tracked within an EHR. For this project, we utilized athenahealth. In this example, a .csv file was collected from athenahealth containing selected quality measures for all available providers. Quality measures stem from Health Resources and Services Administration Uniform Data System clinical measures reporting and other national standard adult preventative metrics, including:

- Blood pressure control among patients with diabetes (systolic <140, diastolic, <90)
- Breast cancer screening
- Cervical cancer screening
- Colorectal cancer screening
- Comprehensive diabetic foot exam
- Controlling high blood pressure
- Counseling medication adherence for patients on a statin
- Diabetes: HbA1C poor control (>9 percent)
- HIV (Human Immunodeficiency Virus) screening
- Ischemic vascular disease (IVD): Use of aspirin or another antiplatelet
- Lipid monitoring for patients with atherosclerotic cardiovascular disease (ASCVD)
- Preventative care and screening: body mass index (BMI) screening and follow-up plan
- Preventative care and screening: screening for depression and follow-up plan
- Statin therapy for the prevention and treatment of cardiovascular disease
- Tobacco use: screening and cessation intervention

Data Cleaning and Quality Checks

Data cleaning and quality checks may be necessary. For example, the data files used in this project included key pieces of data, such as primary care provider and measure name, but also included a variety of unnecessary fields that were omitted as they were not pertinent to identifying strong performers. In this example, the data file was connected to Tableau, and satisfied rates for performance measures by provider were distributed along box and whisker plots to determine quartiles for each quality measure. Any software with the ability to create box and whisker plots visualizing data quartiles could be used in place of Tableau. Physicians with little or no data across quality measures were removed, including providers who were no longer with the health system or providers who did not engage in care delivery related to the quality measures of interest. Data cleaning led to the removal of one provider with no available data. In Tableau, the exclusion of a provider removes them from all quality measures, but the data cleaning process is adaptable and subjective depending on the methods and/or software used.

Data Organization and Visualization

Based on the box and whisker plots for each quality measure created in Tableau, a "Top Quartile Score" column was added to the data file scoring providers who appeared within the upper quartile with a score of "1" and providers outside of the upper quartile with a score of "0" for each quality measure. Each provider could get a maximum score of equal to the number of quality measures, in this case 15, or a minimum score of 0. This file, and the quartile scores within, was used to color code each provider according to the number of quality measures they were marked within the "top quartile," or as a top performer. In the case of this study, strong performers were identified as those providers scoring in the top quartile in 12 or more (80 percent) of the 15 quality measures examined.

Results

Figure 1 represents output from Tableau scoring providers by their top quartile score. Each dot within the box and whisker plot represents a single provider and is colored according to the number of quality measures for which that provider placed within the top quartile. Darker green indicates higher scores for quality measures and becomes more yellow with fewer top quartile satisfied rates. Provider 16 consistently achieved above-average rankings on 12 of 15 quality measures, demonstrating a strong performance compared to other providers within the same health system.

Discussion

A broad range of publications exist surrounding and evaluating the concept of positive deviance, but three primary steps for identifying strong performers exist: review of methodology; performer identification; and the determination of commonalities. A scoping review of 1,140 studies determined that most studies used objective measures of health or survey-based responses to identify positive deviants and focused primarily on identifying positive deviants within individual patient outcomes.[10] At the organizational level, a study demonstrated that you can identify the best and worst performing primary healthcare centers by utilizing semi-structured and in-depth interviews with managerial and clinical staff from each of the primary healthcare centers.[11]

Additionally, several studies utilize systematic review, empirical studies, and interviews to apply the same concept of "performer identification" at the provider and staff level.[4,12,13] Lastly, a review of literature is often used to determine commonalities, including things like strategies, procedures, and routines.[14,15]

As outlined above, the three primary methods used to identify strong performers include the review of methodology, performer identification, and the determination of commonalities, often using scoping literature review, surveys, or informant interviews. This manuscript describes an alternative method of identifying strong performers by repurposing quality outcome measures housed within EHR data. By re-purposing these data, one can identify individuals who are performing exceptionally well without implementing surveys or interviews. Of note, once the positive deviants have been identified, surveys and informant interviews are often the next best step to determine why each individual is performing at the level they are at.

A few limitations exist for this case study. Specifically, the study assesses providers within a single health system composed of six clinic sites, rather than comparing between organizations. Because of this, the data collected and repurposed all come from the EHR, athenahealth. Each EHR collects and stores data in different ways, with athenahealth specifically housing and labeling the quality outcomes measures utilized by many federal agencies for quality improvement reporting. Due to these limitations, we present this work as a generalizable method of determining positive deviance at the individual levels.

Conclusion

This case study represents a pragmatic, easy to implement, methodology aimed at any primary care health system tracking quality measures across a variety of providers, and aiming to understand the performance of their individuals, clinic sites, or organization. This protocol does not require the use of more time-consuming methods, such as surveys or interviews, and is instead based on repurposing data from quality measures likely already being documented in the EHR.

References

1. O'Malley, R., O'Connor, P., Madden, C, & Lydon, S. A Systematic Review of the Use of Positive Deviance Approaches in Primary Care; 2022. *Family Practice,* 39(3), 493-503.
2. Goff, S., Mazor, K., Priya, A., Moran, M., Pekow, P., & Lindenauer, P. Organizational Characteristics Associated with High Performance on Quality Measures in Pediatric Primary Care: A Positive Deviance Study; 2021. *Healthcare Management Review,* 46(3), 196-205. https://doi.org/10.1097/HMR.0000000000000247.
3. Foster, B. A., Seeley, K., Davis, M., & Boone-Heinonen, J. Positive deviance in health and medical research on individual level outcomes - a review of methodology; 2022. *Annals of Epidemiology,* 69, 48–56. https://doi.org/10.1016/J.ANNEPIDEM.2021.12.001.
4. Toscos, T., Carpenter, M., Flanagan, M., Kunjan, K., & Doebbeling, B. N. Identifying Successful Practices to Overcome Access to Care Challenges in Community Health Centers: A "Positive Deviance" Approach; 2018. *Health Services Research and Managerial Epidemiology,* 5, 233339281774340. https://doi.org/10.1177/2333392817743406.

5. Baxter, R., Taylor, N., Kellar, I., & Lawton, R. What methods are used to apply positive deviance within healthcare organisations? A systematic review; 2016. *BMJ Quality & Safety*, 25(3), 190–201. https://doi.org/10.1136/bmjqs-2015-004386.

6. Rose, A. J., & McCullough, M. B. A Practical Guide to Using the Positive Deviance Method in Health Services Research; 2017. *Health Services Research*, 52(3), 1207–1222. https://doi.org/10.1111/1475-6773.12524.

7. Lindberg, C. M., Lindberg, C. C., D'Agata, E. M. C., Esposito, B., & Downham, G. Advancing Antimicrobial Stewardship in Outpatient Dialysis Centers Using the Positive Deviance Process; 2019. *Nephrology Nursing Journal: Journal of the American Nephrology Nurses' Association*, 46(5), 511–518.

8. D'Agata, E. M. C., Lindberg, C. C., Lindberg, C. M., Downham, G., Esposito, B., Shemin, D., & Rosen, S. The positive effects of an antimicrobial stewardship program targeting outpatient hemodialysis facilities; 2018. *Infection Control & Hospital Epidemiology*, 39(12), 1400–1405. https://doi.org/10.1017/ice.2018.237.

9. Lindberg, C., Norstrand, P., Munger, M., DeMarsico, C., & Buscell, P. (n.d.). Letting Go, Gaining Control: Positive Deviance and MRSA Prevention. Available at: https://static1.squarespace.com/static/5a1eeb26fe54ef288246a688/t/5df46b424b14fd7c686a67b2/1576299330698/Lindberg+-+Letting+Go+Gaining+Control+-+PD+and+MRSA+Prevention+-+Clinical+Leader+12-09+FINAL.pdf.

10. Foster, B. A., Seeley, K., Davis, M., & Boone-Heinonen, J. Positive deviance in health and medical research on individual level outcomes – a review of methodology; 2022. *Annals of Epidemiology*, 69, 48–56. https://doi.org/10.1016/j.annepidem.2021.12.001.

11. Lewis, T. P., Aryal, A., Mehata, S., Thapa, A., Yousafzai, A. K., & Kruk, M. E. Best and worst performing health facilities: A positive deviance analysis of perceived drivers of primary care performance in Nepal; 2022. *Social Science & Medicine* (1982), 309. https://doi.org/10.1016/J.SOCSCIMED.2022.115251.

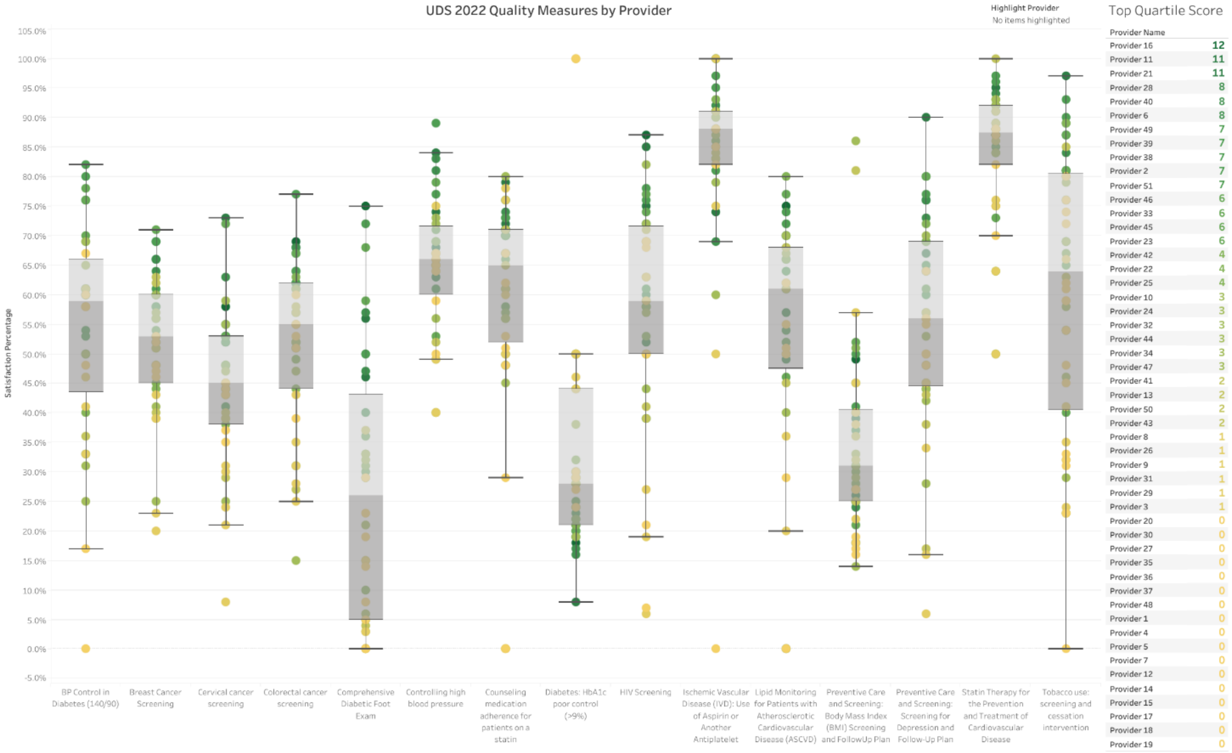
12. Cohen, R., Gesser-Edelsburg, A., Singhal, A., Benenson, S., & Moses, A. E. Translating a theory-based positive deviance approach into an applied tool: Mitigating barriers among health professionals (HPs) regarding infection prevention and control (IPC) guidelines; 2022. *PloS One*, 17(6). https://doi.org/10.1371/JOURNAL.PONE.0269124.

13. Ellenbogen, M. I., Wiegand, A. A., Austin, J. M., Schoenborn, N. L., Kodavarti, N., & Segal, J. B. Reducing Overuse by Healthcare Systems: A Positive Deviance Analysis; 2023. *Journal of General Internal Medicine*. https://doi.org/10.1007/S11606-023-08060-3.

14. Singh, S., Mazor, K. M., & Fisher, K. A. Positive deviance approaches to improving vaccination coverage rates within healthcare systems: a systematic review; 2019. *Journal of Comparative Effectiveness Research*, 8(13), 1055–1065. https://doi.org/10.2217/CER-2019-0056.

15. de Kok, E., Weggelaar-Jansen, A. M., Schoonhoven, L., & Lalleman, P. A scoping review of rebel nurse leadership: Descriptions, competences and stimulating/hindering factors; 2021. *Journal of Clinical Nursing*, 30(17–18), 2563–2583. https://doi.org/10.1111/JOCN.15765.

Figure 1. Quartile scoring by primary care provider


UDS 2022 Quality Measures by Provider
Highlight Provider
No items highlighted
Satisfaction Percentage
105.0%
100.0%
95.0%
90.0%
85.0%
80.0%
75.0%
70.0%
65.0%
60.0%
55.0%
50.0%
45.0%
40.0%
35.0%
30.0%
25.0%
20.0%
15.0%
10.0%
5.0%
0.0%
-5.0%
BP Control in Diabetes (140/90)
Breast Cancer Screening
Cervical cancer screening
Colorectal cancer screening
Comprehensive Diabetic Foot Exam
Controlling high blood pressure
Counseling medication adherence for patients on a statin
Diabetes: HbA1c poor control (>9%)
HIV Screening
Ischemic Vascular Disease (IVD): Use of Aspirin or Another Antiplatelet
Lipid Monitoring for Patients with Atherosclerotic Cardiovascular Disease (ASCVD)
Preventive Care and Screening: Body Mass Index (BMI) Screening and FollowUp Plan
Preventive Care and Screening: Screening for Depression and Follow-Up Plan
Statin Therapy for the Prevention and Treatment of Cardiovascular Disease
Tobacco use: screening and cessation intervention

Top Quartile Score

Provider Name	
Provider 16	12
Provider 11	11
Provider 21	11
Provider 28	8
Provider 40	8
Provider 6	8
Provider 49	7
Provider 39	7
Provider 38	7
Provider 2	7
Provider 51	7
Provider 46	6
Provider 33	6
Provider 45	6
Provider 23	6
Provider 42	4
Provider 22	4
Provider 25	4
Provider 10	3
Provider 24	3
Provider 32	3
Provider 44	3
Provider 34	3
Provider 47	3
Provider 41	2
Provider 13	2
Provider 50	2
Provider 43	2
Provider 8	1
Provider 26	1
Provider 9	1
Provider 31	1
Provider 29	1
Provider 3	1
Provider 20	0
Provider 30	0
Provider 27	0
Provider 35	0
Provider 36	0
Provider 37	0
Provider 48	0
Provider 1	0
Provider 4	0
Provider 5	0
Provider 7	0
Provider 12	0
Provider 14	0
Provider 15	0
Provider 17	0
Provider 18	0
Provider 19	0

Figure 1 (also available online) displays quartile scoring by primary provider. Each dot within the box and whisker plot represents a single provider and is colored according to the number of quality measures for which that provider placed within the top quartile. Darker green indicates higher scores for quality measures and becomes more yellow with fewer top quartile satisfied rates.

Adam Baus, PhD, MA, MPH, (abaus@hsc.wvu.edu) is a research assistant professor in the Department of Social and Behavioral Sciences at the West Virginia University School of Public Health and the director of the West Virginia University School of Public Health, Office of Health Services Research.

Andrea Calkins, MPH, is a program coordinator senior with the West Virginia University School of Public Health, Office of Health Services Research.

Cecil Pollard, MA, is the assistant director of the West Virginia University School of Public Health, Office of Health Services Research.

Craig Robinson, MPH, is the executive director of Cabin Creek Health Systems.

Robin Seabury, MS, is the former hypertension care manager of Cabin Creek Health Systems.

Jessica McColley, DO, is the chief medical officer of Cabin Creek Health Systems.

Marcus Thygeson, MD, MPH, is the executive director of Adaptive Health.

Curt Lindberg, MHA, DMan, is the principal and senior consultant of Partners in Complexity.

Andrya Durr, PhD, is a research specialist with the West Virginia University School of Public Health, Office of Health Services Research.

The Validation Of COVID-19 Information In The Pharmacoepidemiological Research Database of Spain's Public Health System Data by Vaccination Status

Oliver Astasio, MD, PhD, Belén Castillo-Cano, MSc, Beatriz Sánchez Delgado, MSc, PharmD, Fabio Riefolo, PhD, Rosa Gini, PhD Elisa Martín-Merino, PhD, PharmD

Purpose

To validate COVID-19 information records in The Pharmacoepidemiological Research Database for Public Health System (BIFAP) of Spain.

Methods

The recorded COVID-19 cases in primary care or positive test registries (gold-standard) were identified among vaccinated patients against COVID-19 infection and their matched unvaccinated controls, between December 2020 and October 2021. The sensitivity, specificity, positive (PPV) and negative (NPV) predictive values were estimated for primary care records.

Results

Among 21,702 patients with positive tests and 20,866 with recorded COVID-19 diagnoses, the sensitivity, specificity, PPV and NPV were, respectively, 79.98 percent, 99.95 percent, 80.24 percent, and 99.94 percent among vaccinated, and 78.67 percent, 99.96 percent, 84.51 percent and 99.94 percent among controls.

Conclusions

Primary care COVID-19 diagnosis recorded in BIFAP showed that sensitivity was similar and PPV was slightly lower among vaccinated than unvaccinated controls. Among the elderly, COVID-19 diagnosis was less recorded. These findings permit the design of informed algorithms for performing COVID-19-related studies.

Keywords: COVID-19, primary care, validation, predictive values; misclassification, measurement errors, electronic health records, vaccination status

Key Points

1. Data on SARS-CoV-2 tests, vaccination and primary care (PC) consultations were rapidly unified in one of the most populated Spanish healthcare databases (BIFAP) with the purpose to study the effectiveness and safety of COVID-19 vaccines.
2. COVID-19 diagnoses in PC showed high sensitivity to detect true infections (i.e., positive tests) that was lower among ≥70 years old than younger patients, probably influenced by the different healthcare settings.
3. PPV for COVID-19 diagnoses in PC was high and more predictive among unvaccinated and oldest people, probably due to be at-high risk of complications.
4. Specificity of COVID-19 diagnoses was very high.
5. This validation helps understand under- or over- estimations of associated vaccine effectiveness and develop informed algorithms to detect true COVID-19 outcomes in future studies.

Summary

Does the Spanish-collected primary care data about patients suffering from COVID-19 reflect the real pandemic situation in Spain? Patients' healthcare records are, in an anonymized form, used for different research purposes. COVID-19 data has been widely used to study pandemic and vaccination campaign effects, guiding authorities' decisions in this regard. Validating whether the recorded COVID-19 diagnoses reliably reflect true positive laboratory tests is fundamental to trust the performed research outcomes. Herein, we demonstrated that COVID-19 diagnoses in the Spanish public primary care records are truly associated with infection-positive tests, especially for patients >70 years old, and that most of the patients with positive tests also have a diagnosis of infection in primary care. Thus, the Spanish data on COVID-19 is a valid research tool.

Introduction

The SARS-CoV-2 pandemic triggered the need to rapidly share patient-level data across different healthcare institutions, giving them vital importance to promptly monitor pandemic-setting evolution, as well as conditionally approve COVID-19 vaccines' safety and effectiveness, in different world countries through real-world-data evidence.

In Spain, several efforts have been invested among public healthcare institutions to merge patients' information through the creation of common data models (CDM) in order to facilitate and guarantee timely pharmacoepidemiology research related to COVID-19 matters. To this extent, a clear example of the work performed in Spain is given by the Spanish Pharmacoepidemiological Research Database for Public Health System (Base de datos para la Investigación Farmacoepidemiológica en el Ámbito Público or BIFAP) database, a single integrated electronic health record (EHR) system, able to link and merge patient information from several Spanish regional data sources with different settings[1].

A Spanish royal decree regulates the epidemiological surveillance network by making mandatory the case reporting of specific diseases to national authorities. COVID-19 was a mandatory notifiable disease during the pandemic. Since 2020, primary care EHRs directly gathered by BIFAP have been merged in a CDM with SARS-CoV-2 positive laboratory tests, and hospital and intensive care unit (ICU) admissions of external healthcare institutions.

The pandemic data unification allowed the execution of different COVID-19 vaccination studies and the production of significant real-world data evidence during the last years[2,3]. Thus, the EHR CDM creation has been crucial for studying and understanding COVID-19-related matters on the population, undoubtedly supporting important urgent national authorities' decisions about public health measures[4,5,6]

COVID-19 information linked from different data sources may not always overlap and must be evaluated for identification of true cases for research. The data regarding COVID-19 diagnosis in some sources[3] have a positive predictive value (PPV) between 81percent and 94 percent of the true cases depending on the calendar period, whereas there was a sensitivity of 94.4 percent among all episodes. The implication of this could be substantial. For instance, if PPV were different between vaccinated- and unvaccinated-compared groups, the estimations of vaccines effectiveness would be confounded.

While significant advantages have been achieved by using the CDM strategy in terms of promptly available outcomes with large population sizes, further validation studies to quantify the risk of data bias due to case misclassification in the performed pharmacoepidemiology studies are needed[7]. Research using primary care (PC) databases required practical definitions based on the information recorded to identify COVID-19 and, more in general, defining validation parameters would be a useful tool for correctly designing future studies. In the current study, we aimed to estimate and describe the validation parameters of the collected SARS-CoV-2 disease information among vaccinated patients and their unvaccinated controls in BIFAP.

Methods

Data sources and COVID-19 information

Patients' data from the Spanish public National Health System (SNS) data sources were linked and unified in BIFAP[1]

- Data about COVID-19 diagnosis, birth year, sex, and COVID-19 vaccination of around 13.7 million patients (7.4 million of them aged ≥18 years) were obtained from the public PC source for four geographical regions (Aragón, Asturias, Castilla y León, and Murcia). The recorded episodes of COVID-19 diagnosis were identified through SNOMED (Systematized Nomenclature of Medicine) codes, as reported in Table 1. SNOMED codes were mapped to COVID-19 diagnosis codes that were introduced in 2020 into the International Classification of Primary Care ICPC-2[8] and the International Classification of Diseases ICD-9[9] used in PC settings.

- Positive test due to COVID-19 infections were tracked from a COVID-19 registry linked to PC data on the date of the testing result. Infections might be confirmed through positive PCR, antigens, or any other confirmatory criteria established by clinical protocols whose definition is out of the scope of the current study. Herein, COVID-19 positive tests were the gold standard.

BIFAP has been previously validated for research in pharmacoepidemiology, including the estimations of the precision of both, clinical outcomes[10,11] and vaccination records[12]. BIFAP is fully funded by the Spanish Agency on Medicines and Medical Devices (AEMPS), belonging to the public Department of Health, and is maintained with the collaboration of the participant Spanish regions.

The study protocol was approved by the BIFAP Scientific Committee (Reference Number 02_2021).

Study Design and COVID-19 Case Ascertainment

A validation study of COVID-19-related data identified in two study cohorts (3.805.279 COVID-19 vaccinated and unvaccinated control individuals) was performed as designed in the study protocol[12]. In summary, individuals of any age were included when vaccinated against COVID-19 (time0) during the study period, from December, 27 2020 to October, 31 2021. The corresponding unvaccinated controls were matched 1:1 based on the date of the first vaccination of the vaccinated pair, birth year, sex, and region. All the study participants were free of prior SARS-CoV-2 infection. Follow-up was until the end of the study period (October, 31 2021) or until diagnosis of COVID-19.

In the study cohorts, the COVID-19 outcomes described above were identified during the study period (i.e., between time0 and the latest available data, death date, or study end date).

Statistical Analysis

Using as gold standard the COVID-19 positive laboratory tests (main analysis), we estimated the sensitivity, specificity, positive (PPV), and negative (NPV) predictive values as well as the accuracy of the diagnosis date recorded by the PC physicians in the patients' clinical histories.

Parameters were estimated by vaccination status (i.e., vaccinated or control), age band (<70 or ≥70 years old), and sex (female or male). The results of the study were calculated using STATA v.16.1.

Results

Out of 3.80 million pairs of vaccinated and controls study participants (mean age: 53.4 years), 21,702 had a positive test and 20,866 had a recorded COVID-19 episode (18,926 [90.7 percent percent] of them were recorded using two COVID-19 diagnosis codes, see Table 1).

Table 2 shows the validation parameters of tracked COVID-19 cases stratified by vaccination status and age. Considering COVID-19 diagnosis codes, sensitivity was similar among vaccinated (79.8 percent) and unvaccinated (78.7 percent) patients or among women (79.2 percent) and men (79.2 percent). However, differences appeared amongst age groups, i.e. sensitivity ranged from 82.1 percent to 79.6 percent for subjects aged <70 years old and from 71.2 percent to 72.9 percent for older patients (≥70 years old) among vaccinated and unvaccinated controls, respectively. PPV was lower among vaccinated (80.2 percent) than unvaccinated (84.5 percent) subjects and also lower among <70 years old (79.3 percent, vaccinated-84.0 percent, unvaccinated) than ≥70 years old (84.7 percent, vaccinated-88.0 percent, unvaccinated) individuals. Specificity was ≥99.94 percent over all groups.

When recorded codes for suspected COVID-19 or contact with COVID-19 cases were included in the analyses, PPV decreased to 44.0 percent among vaccinated and to 57.6 percent among unvaccinated, while the other predictive values remained similar to their exclusion results (data not shown in tables).

Regarding the accuracy of the COVID-19 diagnosis records, COVID-19 of true positive cases were recorded within five days (in a median value of zero days) from the confirmatory positive laboratory test.

Conclusions

During the fourth and fifth SARS-CoV-2 epidemiological waves with incidences ranging between 21 (October 2021) and 800 (August 2021) cases per 100,000 inhabitants in Spain in 14 days as reported by the public institutions[13] , the recorded COVID-19 diagnoses in BIFAP PC EHRs showed high sensitivity in detecting confirmed SARS-CoV-2 infections and very high specificity to track non-cases of the disease, both among vaccinated and their unvaccinated control group. The estimated predictive values suggested certain differential misclassification of the COVID-19 records and timing of infection when identified based on SNOMED codes in BIFAP or with laboratory positive tests. Quantifying such misclassification permits to understand potential under- or over-estimations in the associated absolute (i.e. incidences; considering that up-to 30 percent of cases could be missed if only primary care diagnosis are collected) and relative risks (at least in unvaccinated vs vaccinated individuals, considering that confirmation seems slightly different among them) of COVID-19 episodes.

On the other hand, we do not recommend the inclusion of codes for suspected SARS-CoV-2 infection or contact with the virus in the definitions of COVID-19 outcomes. In fact, while sensitivity values remained similar, those records' inclusion strongly decreased the PPV, especially among vaccinated individuals, increasing the probability to include misdiagnosed cases of SARS-CoV-2 infections. This misclassification may be due to frequent PC physician consultations of those individuals or other unknown reasons.

The validation parameter of COVID-19 cases in PC and its accuracy, herein provided, can be potentially used as a supportive design tool for outcome definitions in other studies. For example, in studies interested only in PC consultations, when a decision should be taken over including only COVID-19 events linked to positive test results (to increase the PPV), or whether using COVID-19 diagnoses regardless of any associated positive laboratory test. This latter case may not include up to one-third (from 17.9 percent to 28.8 percent among vaccinated and unvaccinated) of individuals with COVID-19, especially for the elderly group (≥70 years old). Alternatively, for studies interested in all infection regardless of the setting, whether using both types of records i.e., people with a positive test and/or a clinical diagnosis (given challenges in accessing testing and/or primary care during the pandemic) or only positive laboratory tests.

Concerning age, PC records' sensitivity for the detection of COVID-19 cases was lower among the oldest patients (≥70 years old), especially those vaccinated, while PPV was higher in this group compared to <70 years old participants. The identified differences in sensitivity across the different ages may be due to the tendency of ≥70 years old patients of seeking medical attendance directly at the hospital. Another point that should be taken into account is related to patients living in nursing homes. They receive in-house medical attention directly from the nursing homes' experts, thus, may not visit their PC physician to communicate the COVID-19 infection. Nursing homes' cases of COVID-19 are not systematically collected by the BIFAP data source. Other cofactors that may justify the sensitivity differences in identifying COVID-19 cases between the two age categories above/below 70 years old are, among others, the higher number of elders experiencing the infection during long stays in the hospital for other reasons or when receiving special care directly at their own home and may also die of COVID-19. These cases might not be correctly tracked by the BIFAP data sources and could explain the higher numbers of losses when compared to the <70 years old population.

Differently, our results suggest that if the COVID-19 diagnosis is recorded in the PC clinical registries, the PPV of those aged ≥70 years old is 5 percent and 14 percent, among vaccinated and unvaccinated, respectively, more accurate than the younger group. This variation could be led to different reasons such as more frequent testing of COVID-19 cases due to more clear infection symptoms in the eldest population. We also observed that the accuracy of the infection diagnosis date in BIFAP was also high since almost all COVID-19 positive laboratory test have been recorded within five days in PC registries. This is of fundamental importance when time-window analyses are needed to evaluate if and when taking preventative measures and decisions, such as promoting large vaccination campaigns for specific age categories.

Finally, comparing our study with an already-published work on COVID-19 diagnosis validation carried out in the national medical product safety surveillance program funded by the Food and Drug Administration (FDA) in 2020, we can highlight comparable results. The study[3] showed that the PPV of COVID-19 diagnoses codes across all participating data sources was between 81.2 percent and94.1 percent (variability depends on the considered time period), values almost close to our PPVs of 80.2 percent and 84.5 percent among vaccinated and unvaccinated, respectively, whereas the sensitivity was reported to 94.4 percent, which is a higher value than our estimations of ≈79 percent in both vaccinated and unvaccinated groups. The differences in sensitivity among the two works can be the result of our chosen study cohorts (which, in our case, have been selected according to the characteristics of the vaccinated patients and may not represent the entire BIFAP population), diverse healthcare settings (population-based versus claim data sources), or diverse healthcare systems, age, socioeconomic status or geographical areas of the covered populations, healthcare data recording habits, or virus epidemiology. Thus, the parameters observed in our study may mainly be used to interpret studies performed in the same data source and period and may not be generalisable to other contexts or settings.

Some limitations must be acknowledged.

Race, ethnicity and other demographic characteristics potentially associated with unequal burden of COVID-19 were not available to assess any differential parameters among them.

In the BIFAP data source, the tracked COVID-19 diagnoses in PC records have high validation parameters with a low misclassification of their timing. Both COVID-19 vaccination status and old age of the patients influenced the recordings of infection diagnoses and the accuracy of their timing. Thus, the PPV in PC should be a parameter to be taken into account in COVID-19 research studies. These findings reinforce the reliability of using the linked healthcare registries to BIFAP clinical histories as a source of data for performing observational studies on SARS-CoV-2 infection.

Electronic healthcare databases share common challenges, including the accurate identification of healthcare outcomes of interest for observational studies. Considering the evolving fundamental role of real-world data and healthcare databases, the validation process, to what this study contributes, is crucial for assuring the quality and accuracy of the produced evidence in pharmacoepidemiology studies.

Acknowledgements

This study is based on data from the “Pharmacoepidemiological Research Database for Public Heatlh System” (BIFAP) in Spain.

BIFAP is a public program for independent research financed by the Spanish Agency of Medicines and Medical Devices (AEMPS). The results, discussion and conclusions of this work are only of the authors and do not represent in any way the position of the AEMPS on this subject. The authors would like to acknowledge the excellent collaboration of the primary care physicians (general practitioners/paediatricians), and patients taking part the primary care records as well as the support from the regional health administrations providing BIFAP data.

Conflict of Interest

Authors declare they do not have conflict of interest in the publication of this article.

Ethics Statement

The study protocol was approved by the Ethical Committee Comité de ética de la investigación con medicamentos regional de la Comunidad de Madrid (CEIm-R) with the reference Number BIFAP_02_2021.

Authors

Oliver Astasio, MD, PhD, is a physician in clinical pharmacology and PhD in Biomedical Investigation at Complutense University in Madrid. At the time of the study, Astasio was affiliated with the clinical pharmacology department at the Hospital Clínico San Carlos’ Health Research Institute in Madrid and external expert at the Spanish Agency of Medicines and Medical Devices. At this time, he is medical advisor in Novartis pharmaceutical for haematology diseases.

Belén Castillo-Cano, **MSc,** is working in the pharmacoepidemiology and pharmacovigilance division at the Spanish Agency of Medicines and Medical Devices in Madrid. She is collaborating as a junior biostatistician on different projects with the BIFAP team. At this time, she is studying for a PhD in technology in the department of computer science, applied mathematics and statistics at Girona University. She has a Bachelor's degree of mathematics at the University of Almería and a Master's degree of statistics at the University of Granada.

Beatriz Sanchez-Delgado, mailto:is a pharmaco-epidemiologist in BIFAP at the Spanish Agency of Medicines and Medical Devices in Spain. She has a bachelor of pharmacy from the University of Salamanca, Spain and a Masters degree on both

International Public Health (Queen Margaret University in Edimburgh, UK) and pharmacoepidemiology and pharmacovigilance (Alcala University in Madrid, Spain)

Fabio Riefolo, **PhD**, worked on the development of cholinergic nervous system drugs at the University of Milan (Italy) and Wuerzburg (Germany) during his Master's thesis in pharmaceutical chemistry & technology. He obtained a Ph.D. in medicinal chemistry at the Institute for Bioengineering of Catalonia in Barcelona (Spain), working on cardiovascular diseases and neurological disorders and was also a researcher for the Biomedical Research Networking Centre in Bioengineering, Biomaterials, and Nanomedicine (CIBER-BBN). From a post-doc position at IBEC, he moved to Teamit (Barcelona, Spain) as scientific study manager and regulatory science advisor, expanding his knowledge of medicines development and their regulatory roadmap, from preclinical to clinical regulation, post-marketing authorization studies based on real-world-evidence (participation in several HMA-EMA-registered studies), medical device regulation, and working in various healthcare-related public European proposals.

Rosa Gini, is a data scientist focused on secondary use of EHRs for pharmacoepidemiology, epidemiology and health services research. Her specific interest is in developing culture and tools for accurate, reliable, transparent, and fast generation of evidence to support health policy making at a national, European, and international level. In ARS Toscana, the Regional Agency for Public Health of Tuscany, she is the head of the pharmacoepidemiology unit and conducts methodological studies, providing expertise for studies using real-world evidence on the use and safety of medicines and vaccines on an international distributed networks of databases.

Elisa Martín-Merino, PhD, PharmD (emartinm@aemps.es), is a senior pharmacoepidemiologist at the Spanish Agency of Medicines and Medical Devices in Madrid, Spain. She earned her PhD in preventive medicine and public health, with her doctoral research focusing on assessing the risk of acute coronary syndrome associated with the use of non-steroidal anti-inflammatory drugs in a field study. Martín-Merino has actively contributed to pharmacoepidemiological research studies aimed at evaluating potential adverse reactions to medications used by individuals in real-world settings- outside the controlled context of clinical trials. Additionally, she is interested in studying the precision of electronic health records for research on medication use and its effects.

References

1. “BIFAP Base de Datos Para La Investigación Farmacoepidemiológica En El Ámbito Público.”. Accessed January 28, 2021. http://bifap.aemps.es/.

2. Brown CA, Londhe AA, He F, Cheng A, Ma J, Zhang J, Brooks CG, et al. 2022. “Development and Validation of Algorithms to Identify COVID-19 Patients Using a US Electronic Health Records Database: A Retrospective Cohort Study,” no. May: 699–709.

3. Kluberg SA, Hou L, Dutcher SK, Billings M, Kit B, Toh S, Dublin S,, et al. 2022. “Validation of Diagnosis Codes to Identify Hospitalized COVID-19 Patients in Health Care Claims Data.” Pharmacoepidemiology and Drug Safety 31 (4): 476–80. https://doi.org/10.1002/pds.5401.

4. Bots SH, Riera-Arnau J, Belitser SV, Messina D, Aragón M, Alsina E, Douglas IJ, et al. 2022. “Myocarditis and Pericarditis Associated with SARS-CoV-2 Vaccines: A Population-Based Descriptive Cohort and a Nested Self-Controlled Risk Interval Study Using Electronic Health Care Data from Four European Countries.” Frontiers in Pharmacology 13: 1038043.

5. Willame C, Dodd C, Durán CE, Elbers R, Gini R, Bartolini C, Paoletti O, et al. 2023. “Background Rates of 41 Adverse Events of Special Interest for COVID-19 Vaccines in 10 European Healthcare Databases - an ACCESS Cohort Study.” Vaccine 41 (1): 251–62. https://doi.org/10.1016/j.vaccine.2022.11.031.

6. Riefolo F, Castillo-Cano B, Martín-Pérez M, Messina D, Elbers R, Brink-Kwakkel D, Villalobos F, et al. 2023. “Effectiveness of Homologous/Heterologous Booster COVID-19 Vaccination Schedules against Severe Illness in General Population and Clinical Subgroups in Three European Countries.” Vaccine 41 (47): 7007–18. https://doi.org/10.1016/j.vaccine.2023.10.011.

7. Seeger, JD, Jonsson, M, Layton, JB and Clarke, TC. 2022. “Considerations of Misclassification and Confounding on COVID-19 Vaccines Effectiveness Studies - A Vaccine SIG Endorsed Symposium.” In International Conference of Pharmacoepidemiology. Pharmacoepidemiology. 2022. Vol. 67.

8. Oxford University Press. 1998. “ICPC-2. International Classification of Primary Care.” Second Edi.

9. “World Health Organization. WHO IRIS: International Classification of Diseases: [9th] Ninth Revision, Basic Tabulation List with Alphabetic Index.” 1978. 1978. http://www.who.int/iris/handle/10665/39473.

10. Martín-Merino E, Martín-Pérez M, Castillo-Cano B, Montero-Corominas D. 2020. “The Recording and Prevalence of Inflammatory Bowel Disease in Girls’ Primary Care

Medical Spanish Records." Pharmacoepidemiology and Drug Safety 29 (11): 1440–49. https://doi.org/10.1002/pds.5107.

11. Maciá-Martínez MA, Gil M, Huerta C, Martín-Merino E, Álvarez A, Bryant V, Montero D; BIFAP Team. 2020. "Base de Datos Para La Investigación Farmacoepidemiológica En Atención Primaria (BIFAP): A Data Resource for Pharmacoepidemiology in Spain." Pharmacoepidemiology and Drug Safety 29 (10): 1236–45. https://doi.org/10.1002/pds.5006.

12. Martin-Merino, E, Seco-Meseguer, E, Castillo-Cano, B, Limia-Sanchez, A, Olmedo-Lucerón, C, Monge-Corella, S, and Larrauri, A. 2021. "Real-World Effectiveness of Different COVID-19 Vaccines in Spain: A Cohort Study Based on Public Electronic Health Records (BIFAP)." Spain. EU PAS Register (study EUPAS42668)

13. "Instituto de Salud Carlos III. Informes COVID-19. Informe No 103. Situación de COVID-19 En España a 3 de Noviembre de 2021.". Accessed April 6, 2023. https://www.isciii.es/QueHacemos/Servicios/VigilanciaSaludPublicaRENAVE/EnfermedadesTransmisibles/Documents/INFORMES/Informes%20COVID-19/INFORMES%20COVID-19%202021/Informe%20nº%20103%20Situación%20de%20COVID-19%20en%20España%20a%203%20de%20noviembre%20de%202021.pdf

Table 1. SNOMED description of COVID-19 diagnosis mapped to available ICPC/ICD-9 codes in primary care clinical histories and frequency of true positives found against SARS-CoV-2 lab positive test.

SNOMED description	**SNOMED codes**	**Frequency**	**Percentage**
Coronavirus infection (disorder)	186747009	10,249	49.12
Disease caused by severe acute respiratory syndrome coronavirus 2 (disorder)	840539006	8,677	41.58
Diagnosis of COVID-19 infection confirmed by laboratory testing (disorder)	63681000122103	1,740	8.34
Pneumonia caused by Human coronavirus (disorder)	713084008	107	0.51
Pneumonia caused by severe acute respiratory syndrome coronavirus 2 (disorder)	882784691000119100013084008	62	0.30
Disease caused by Coronaviridae (disorder)	27619001	20	0.10
Polymerase chain reaction positive for severe acute respiratory syndrome coronavirus 2 (finding)	62531000122108	7	0.03
Asymptomatic severe acute respiratory syndrome coronavirus 2 infection (finding)	189486241000119100	1	0.00
Procedure for action related to case of disease due to SARS-CoV-2 (procedure)	64121000122109	1	0.00
Testing positive for IgG against SARS-CoV-2 (finding)	64671000122103	1	0.00

Outcome: case of COVID-19 still under follow-up (finding)	63511000122107	1	0.00
Positive result of rapid test for detection of IgM and IgG antibodies against SARS-CoV-2 in blood (finding)	63621000122102	0	-
Detection of severe acute respiratory syndrome coronavirus 2 (observable entity)	871562009	0	-
SARS-CoV-2 antigen testing positive (finding)	64731000122108	0	-
Secondary triage for severity level in patient with disease due to SARS-CoV-2 (procedure)	64031000122106	0	-
Diagnosis of COVID-19 infection confirmed by laboratory testing (disorder)	63681000122103	0	-
Detection of severe acute respiratory syndrome coronavirus 2 antigen (observable entity)	871553007	0	-
Positive serologic study for COVID-19 (finding)	62951000122108	0	-
Total		20,866	100.00

Table 2. Validation parameters of COVID-19 Codes recorded in primary care clinical histories using as gold-standard SARS-CoV-2 lab positive test.

	N. Positive Covid test (gold-standard)	**N. Covid Recorded in PC**	**N. in both sources (True positive)**	**N. recorded in PC without +test (% False positives)**	**N. Positive test without PC record**	**Sensitivity of PC records**	**Specificity of PC records**	**PPV of PC records**	**NPV of PC records**	**Missing in PC overall positive test (%)**
Vaccinated	10,439	10,381	8,330	2,051 (19.76%)	2,109	79.80	99.95	80.24	99.94	20.20
<70	8,248	8,540	6,771	1,769 (20.71%)	1,477	82.09	99.94	79.29	99.95	17.91
≥70	2,191	1,841	1,559	282 (15.32%)	632	71.15	99.97	84.68	99.93	28.85
Unvaccinated	11,263	10,485	8,861	1,624 (15.49%)	2,402	78.67	99.96	84.51	99.94	21.33
<70	9,657	9,156	7,691	1,465 (16.00%)	1,966	79.64	99.95	84.00	99.93	20.36
≥70	1,606	1,329	1,170	159 (11.96%)	436	72.85	99.98	88.04	99.95	27.15

Perspectives in Health Information Management

Winter/Spring 2024 Introduction

Welcome to the Winter / Spring 2024 issue of *Perspectives in Health Information Management* (PHIM), the peer-reviewed journal of AHIMA. I am honored to serve as the editor for the publication. In addition, a new editorial board will be meeting soon with the goal of continuing to build this important peer-reviewed journal to increase its relevance and importance for the health information industry. More details will be forthcoming. We encourage authors to consider PHIM for their HI-related research.

In this Winter/Spring 2024 issue we are pleased to present a variety of manuscripts for you, addressing vital topics such as recording sex gender identity, ICD-11 innovation, professional ethics, patient portals, utilizing EHR data for quality improvement, the creation of a SDOH platform, and the validation of COVID-19 data in a research database. The breadth of these manuscripts highlights the diversity found in health information professional practice. We hope you enjoy them.

- A Process of User-Centered Design to Create a Social Determinants of Health Data Platform
- Availability of Sex, Gender Identity, and Sexual Orientation Data: An Electronic Medical Record Review of a Catholic Healthcare System from 2012-2023
- The Impact of Professional Ethics Case-based Learning on the Ethical Sensitivity of Health Information Technology Students
- Leveraging an Innovation Model to Facilitate ICD-11 Implementation
- Perspectives on Big Data and Big Data Analytics in Healthcare
- Unlocking Patient Portals: Health Information Professionals Navigating Challenges and Shaping the Future
- Using Electronic Health Records Data to Identify Strong Performers in Healthcare Quality Improvement
- The Validation Of COVID-19 Information In The Pharmacoepidemiological Research Database of Spain's Public Health System Data by Vaccination Status

Susan H. Fenton, PhD, RHIA, ACHIP, FAMIA
Dr. Doris L. Ross Professor
Vice Dean for Education
UT System Distinguished Teaching Professor
Editor, *Perspectives in Health Information Management*